Techniques for Speech Information Hiding and Its Applications

Shengbei, WANG

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

Techniques for Speech Information Hiding and Its Applications

Shengbei, WANG

Supervisor: Masashi UNOKI

School of Information Science
Japan Advanced Institute of Science and Technology

September, 2015

Abstract

The development in digital technologies has facilitated speech signal to be reduplicated and edited at high fidelity. Although many applications benefit from these developments, new social issues related to malicious attacks and unauthorized tampering to speech have arisen. For example, by using advanced speech analysis/synthesis tools, ordinary people are capable to produce high naturalness of tampered speech without leaving perceptual clues. Since these tools enable the speech to be tampered in a much easier and credible way, it is becoming difficult to confirm the originality of speech signal. As an important information carrier, the originality of speech signals should be strictly confirmed. To avoid the unauthorized tampering as well as the negative influence that they may cause, it is necessary to conduct relevant research about speech protection and tampering detection to protect speech signal.

Information hiding technique which can hide or embed digital data such as copyright notice or serial number in the original speech signal has been considered as an effective solution for the above issues. The embedded digital data is generally referred as watermarks, and this kind of information hiding methods is specified as watermarking methods. To be effective, watermarking methods should satisfy several requirements: (1) inaudibility to human auditory system, (2) blindness for watermark extraction, (3) robustness against allowable speech processing and common attacks, and (4) fragility against tampering. The first three requirements are required for general watermarking methods, and the last one is an additional requirement when watermarking methods are used for tampering detection. However, it is proven to be difficult for watermarking methods to satisfy all these requirements simultaneously. Our research aim is to solve the problem of unauthorized tampering with information hiding and watermarking methods. The first target is to realize a general watermarking method that can satisfy all the first three requirements. After that, this watermarking method will be applied to other applications, such as tampering detection by exploring the fragility, and hybrid watermarking.

Since human auditory system is usually not sensitive to tiny changes of speech parameters, watermarks are possible to be inaudibly embedded by subtly modifying speech parameters. According to the source-filter model, the linear prediction (LP) coefficients can provide accurate estimation of formants. The line spectral frequencies (LSFs), as substitute parameters of LP coefficients, can not only represent the formants but also have several excellent properties: (i) they are less sensitive to noise and (ii) the influences caused by the deviation of LSFs can be limited to the local spectral, thus the distortion introduced by LSFs deviation in both spectral and sound quality can be minimized. In addition, since LSFs are universal features in different speech codecs, if watermarks are embedded into LSFs, they are possible to survive from the encoding/decoding process. Therefore, embedding watermarks into LSFs also enables the watermarking method to be robust against difficult speech codecs.

Since LSFs can directly represent the formants, the modifications to LSFs made by watermark embedding can be physically considered as make tuning to the formants of speech signal. Therefore, our main concept for watermarking is formant tuning. Based on this concept, we propose two watermarking schemes. One is watermarking based on quantizing LSFs with quantization index modulation (QIM) (LSFs-QIM based watermarking). In this method, different watermarks are embedded into the LSFs of speech signal with different quantization steps. In the watermarking extraction process, watermarks are blindly extracted by re-quantizing the LSFs obtained from the watermarked signal with the same quantization steps. However, it is found that, since the QIM based modifications to LSFs are quite unintentional, the original formant structure of speech signal is easily disrupted, which will degrade the sound quality of speech signal. Moreover, the performance of this method is characterized by the quantization step, i.e., small quantization step is benefit for good sound quality of watermarked signal but strong robustness cannot be obtained, and vice versa. Therefore, it is difficult for this method to get a trade-off between inaudibility and robustness.

As to overcome these drawbacks, the original formant structure of speech signal should be considered for better performance. As we have found, in the field of speech synthesis, formant which is a crucial acoustic feature for speech perception, can be enhanced to improve the quality and intelligibility of speech when the speech is impaired by environmental noise or other reasons. Since formant can be enhanced to improve the speech quality, and such modifications do not cause perceptual distortion to the original speech, watermarking based on formant enhancement is possible to be inaudible to human auditory system. Based on this concept, we propose another watermarking scheme, i.e., watermarking based on formant enhancement (formant-enhancement based watermarking). In this method, different watermarks are embedded by enhancing different formants: ``0" is embedded by enhancing the sharpest formant and ``1" is embedded by enhancing the second sharpest formants, after which different bandwidth relationships between the sharpest and the second sharpest formants are established. These different bandwidth relationships can be used to blindly extract watermarks in the extraction process.

We evaluate the proposed two watermarking methods with respect to inaudibility and robustness (both methods are blind). For the LSFs-QIM based watermarking, the performance of inaudibility and robustness are evaluated with different quantization steps. The results from inaudibility evaluation reveal that the proposed method can satisfy inaudibility when quantization steps are small. The results from robustness evaluation suggest that the proposed method has good bit detection rate for normal extraction and some of general speech processing. However, the weak robustness of this method against speech codecs, down-sampling, and low-bit quantization has greatly restricted its effectiveness. For the formant enhancement based watermarking, evaluations are carried out for both this method and other watermarking methods to make a comparison study. The LP order and the modification level for the formant enhancement based method are well examined for achieving good performance in inaudibility and robustness. Based on the evaluation results, watermark embedding through formant enhancement does not cause severe degradation to the original speech quality, and the watermark extraction by identifying bandwidth relationship is able to tolerate slight distortions of frequency components caused by other processing. Therefore, the formant enhancement based method can satisfy the requirements of inaudibility, blindness, and robustness, especially the robustness against speech codecs.

Since the formant enhancement based method can satisfy the three basic requirements for watermarking, we apply it to tampering detection scheme of speech signal. Ideally, if the watermarking method can satisfy fragility, tampering can be detected with the mismatched bits between embedded watermarks and the extracted watermarks. The tampering detection ability of the proposed scheme is evaluated against several kinds of tampering. The embedding bit rate of watermarks is 4 bps, each embedded bit is able to account for 0.25 s speech segment when locating the tampering. The evaluation results show that when tampering has been made to the watermarked speech, watermarks in the tampered segment will be destroyed. Therefore, the proposed scheme is fragile against tampering, and it has the ability to detect tampering as well as checking the originality of speech signals. The formant enhancement based watermarking is also applied to hybrid watermarking method, where the formant enhancement based watermarking and cochlear delay based watermarking are combined together. The evaluation results suggest that the robustness of hybrid method can be improved compared with each single method, since the disadvantage of one watermarking method can be concealed by the other watermarking method.

Based on these results, we conclude that the formant enhancement based method can satisfy the first three requirements for general watermarking. It can also satisfy fragility when used for tampering detection. Therefore, it has the ability to solve the problems of speech tampering.

Keywords: Information hiding, speech watermarking, formant enhancement, tampering detection, hybrid watermarking

Acknowledgments

A great many people have contributed to this dissertation. I want to express my gratitude to all these people who have made this dissertation possible.

First of all I would like to express my deepest appreciation to my supervisor, Prof. Unoki for his continuous guidance and support. He gives me the opportunity to pursue a doctor degree and enlightens me the first glance of the research topic in my PH.D program. His broad knowledge is essential for me to enrich my ideas and overcome many difficulties encountered in my research. I am also appreciate for his timely discussions with me, before each deadline of paper submission. In particular, his positive attitude affects and encourages me a lot when I fails to get the acceptance for International conferences. His insightful comments help me make decisions more rationally and finally reach my research goals in my work. Without his guidance and persistent help, this dissertation would not have been possible.

I would also like to give the grateful and sincere thank to my vice supervisor Prof. Akagi for his encouragement and valuable advices. I am thankful to him for reading and commenting on my reports in the weekly and monthly meeting. The valuable discussions help me understand and improve my knowledge better.

My sincere thanks also go to Prof. Dang, who offers me the opportunity to do the minor research with him. He leads me to know more knowledge about speech production and speech perception. I would also like to thank Dr. Miyauchi, who helped and explained me a lot in designing the subjective evaluations of my research.

I also appreciate to my colleagues in our laboratory who always willing to help and give their best suggestions and comments for my research. They always carefully check the grammar and comment on expression in my several manuscripts. They have been always there to listen to my rehearsal and give advices before domestic and International conferences. I am enriched by their experience to design a more vivid and clear poster and presentation. Their contributions are essential for me to successfully express my work in conferences.

I would also want to thank many friends who have helped me stay in Japan through the last three years. Their support and care help me to overcome many difficulties in daily life. They encourage me and stand by me when I have troubles. I miss the days that we went out for tours and enjoys the life in Japan together. I greatly cherish our friendship and the time we spend together no matter good or bad.

Last but not the least, I would like to thank my beloved parents and sister for their love and encouragements in both spiritually and materially through my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The development in digital technologies has significantly impacted the way of human life from communications to social interaction. Benefit from digital technologies, digital multimedia such as digital video, image, audio, and speech can be transmitted universally, and people can access to these massive information in a faster way that cannot be achieved prior to digital technologies. Digital signals offer several advantages over the analog signals. For example, they are in good qualities, easy for storage, and suitable for computing and processing. The modification to digital signal can be operated to the exact location of the whole signal, and a copy of digital signal is exactly the same as the original one. Moreover, digital signals are easily to transmit over the internet and telecommunication systems.

However, since digital technologies also enable digital signals to be delivered in a detached manner crossing time and distances, unforeseen operations associated with content replacement and cropping can be performed when the signals are transmitted. It is known that the digital signals which contain invaluable importance have been widely used for many important occasions, the security for the digital signal has become a tremendously important issue to deal with. To avoid these unauthorized tampering as well as the negative influence that they may cause, the research towards to valid the originality of digital signals should be conducted.

In previous research, the protection of digital signals mainly focus on video, audio, and images which contains commercial values. Currently, as speech signals have been widely used in our daily life, such as mobile and Voice over Internet Protocol (VoIP) communication, digital forensics, and commercial activities, the protection of speech signal has drawn much attention. This work focuses on the protection of speech signal.

1.1 Speech applications and speech protection

1.1.1 Applications of speech signals

Speech is a tool by means of which people can communicate with others to express his/her feelings, emotions, and willingness. In a face to face communication case, there is no doubt that what the listeners hear is what the speakers want to express. With the recent development in the Internet and digital technologies, people can easily record the speech contents (what the speakers say) with modern electronic devices such as a tape recorder and a voice taper. This is a quite powerful way to enable people who have missed a valuable meeting or significant occasions to keep track of what has happened. Apart from these, recorded speech materials can also be found in the following applications such as VoIP communications [1], digital forensics [2], government activities, commercial investigation, where the originality of the speech is extremely important.

- VoIP communications: VoIP refers to making phone calls over the IP network [3, 4]. The technique of VoIP is becoming increasingly popular since people are enabled to make telephone calls at reduced expenses because phones are made over the Internet rather than company's network, just like the email systems. VoIP is available on phones, computers, and other devices. VoIP is not only a way to transport data, but also a foundation for more enriched multimedia communications applications with speech and video. For most business applications, the VoIP calls are managed with private networks so that the information security can be ensured. For common customs, since the VoIP calls connect directly to the Internet, attackers can stole the speech data, record conversations, or spy on the calls. Therefore, there exists a potential threat that the captured speech data may be misused for crime issues. Motivated by this, effective measures should be used to protect the speech data transmitted via VoIP communication.

- In digital forensics, speech signals are usually employed as a kind of digital evidence [5, 6]. The speech evidences may record the criminal activities or the interrogation (also called questioning or interpellation) of the suspects and victims. These evidences need to be recovered from electronic/digital devices and then submitted to the court to support or oppose a hypothesis. Since the judicial proceedings is largely based on these evidences, the integrity and originally of digital speech evidence should be strictly confirmed. If there is tampering motived by malicious intends that try to mislead the listeners, such as cutting or adding some key word in the speech sentences, or transform the individuality of speaker to that of another speaker, unfair results will come out and people will question the fairness of the court and lose the confidence of social justice.

- For the government activities, every element of the government officers' statement greatly affects the society and the human life. If the speech recordings involving official secrets are stolen and attacked, such as content concatenation or replacement, once the modified speech recordings are released publicly, unforeseen effect will be brought [7]. It is tough for the government to address the issue and bring it under control.

- Commercial investigation: forensics may be used in the private section, such as business and intrusion investigations [8, 9]. In a corporation, confidentialities such as negotiation, board meeting, and economic decision are usually recorded for emergency needs and treated with extraordinary secrecy. Once the speech data is the tampered illegally, it may cause serious economic losses.

1.1.2 Protection of speech signals

Undoubtedly, the speech signal contains increasing values. As we can imagine, with the ease of high-quality tampering, speech signal has become vulnerable and difficult to trust. Although we are not sure whether unauthorized tampering to speech has caused serious problems or not, this is indeed a potential threat that we should consider. Investigation about whether the speech has been tampered since its creation should be carried out. This research is aiming to validate the originality of speech. Two main issues involved in whether the speech is original and whether the speech has been tampered with since it was created should be considered to protect speech signal.

1.2 Speech protection based on cryptography

Motivated by speech security problem, many solutions have been proposed. There are generally two categories to prevent the speech signals from being tampered, i.e., active method and passive method. The cryptography [10], as an active method, can date back to the widespread of electronic communications when electronic security is becoming important. It is generally believed that the cryptography realizes a secure transmission over the untruthful medium. Secure transmission indicates that the speech signal (in a particular unreadable format) sent from the sender side cannot be accessed or altered by the third party, and only the legal recipient at the receiver side will be provided with a key to decrypt the speech. Cryptography, therefore, can prevent speech signals from being tampered by setting up a secure transmission from the sender side to the receiver side. Two main processes involved in cryptography are (1) encryption which transforms the data into the unreadable format and (2) decryption which restores the unreadable file to its original format. A secure transmission provided by cryptography generally relates to three requirements including (i) authentication, (ii) privacy, and (iii) integrity [11]:

(i) Authentication: refers to prove and guarantee the identity, i.e. speech signal is not sent by an impostor instead of the specific sender.

(ii) Privacy: concerns with ensuring that any attackers cannot access the transmitted

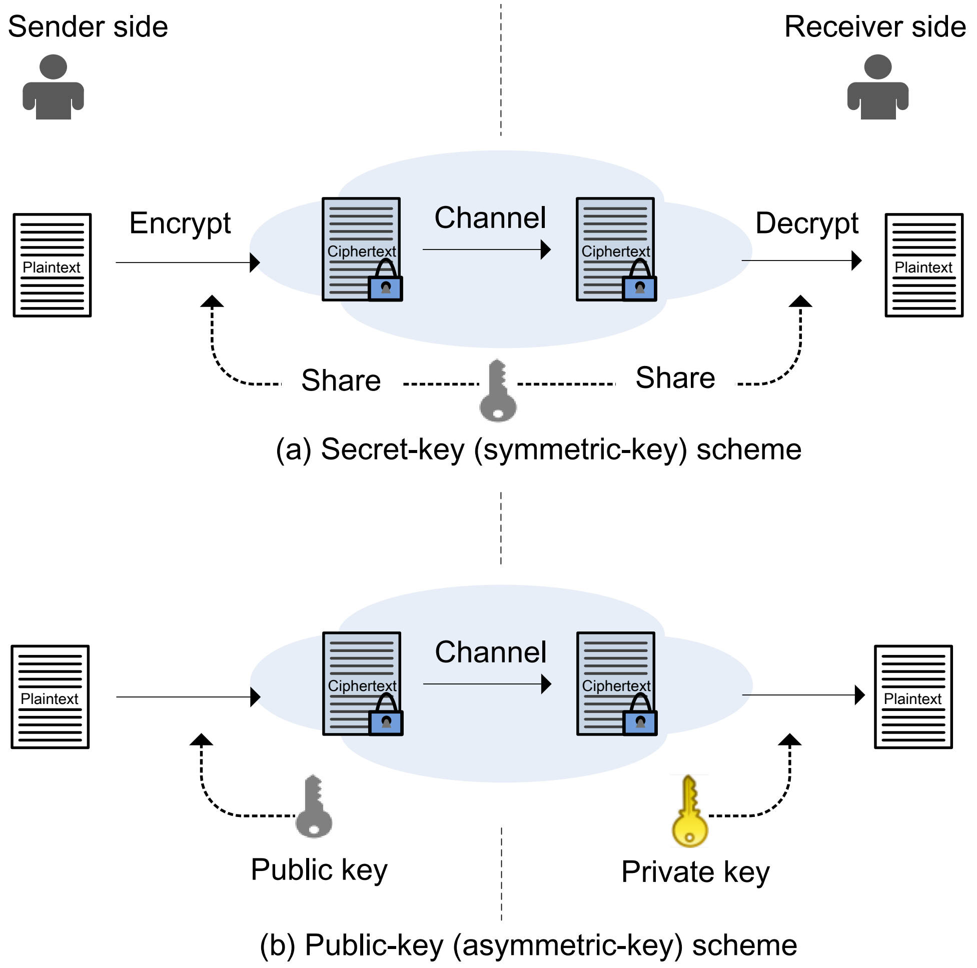


Figure 1.1: Block diagrams of (a) secret-key based cryptography scheme and (b) public-key based cryptography scheme, where "Plaintext" is the original signal and "Ciphertext" is the encrypted signal.

speech except the intended receiver.

(*iii*) Integrity: indicates that the received speech has not been altered prior to receiver in any way from the original.

Three typical cryptography schemes have been designed when accomplish these requirements. These are secret-key (symmetric-key) scheme, public-key (asymmetric-key) scheme, and hash function based scheme. In the secret-key (symmetric-key) scheme, both the sender and receiver use the same single key to encrypt and decrypt the digital signal. In the public-key (asymmetric-key) scheme, two keys are required: the public key is usually known to everybody and the private key is only known to the indented receiver. These two keys are generated in a way that it is impossible to work out the private key based on the public key. The block diagram of these two schemes are illustrated in Fig. 1.1. In the hash function based scheme, an unrecoverable mathematical operations is used to encrypt the signal. The hash algorithms are typically used to ensure that the transmitted signal has not been altered [12, 13].

Benefit from the advance in modern computer, cryptography has become available to everyone for producing the encrypted signal with a complexity that most powerful attack algorithms cannot workout in million years. As an effective security tool, cryptography can meet most of the secure interests in telecommunication, Internet, and confidential message, by preventing them from unauthorized access. However, since the key in cryptography attaches great importance to the security of the whole system, if the decryption key is captured by illegal user, or if the system that used to encrypt the signal is intruded by hackers, cryptography cannot protect the signals anymore. Besides, it is also important to note that although cryptography can ensure the secure transmission of signal, it never examines the signal itself that being protected [10]. That is to say once the decrypted signal is edited or distributed, cryptography cannot (1) protect the decrypted signal against attacks or (2) provide any information to track the signal for its originality. This indicates that cryptography cannot be used to address the tampering issues in the form of decrypted signal. Therefore, more suitable methods should be used to address the tampering issues in speech for originality accounts.

1.3 Data hiding technique and digital watermarking

1.3.1 Overview of information hiding

As a complement technique to cryptography, information hiding technique [14, 15] has been proposed as a passive method. As indicated by the name, this technique hides or embeds additional information (e.g., digital data/ message, serial number, and identification marks) in the digital signals [16, 17]. Compared with cryptography, information hiding technique concentrates on hiding information for particular purposes other than securing the communications [18]. Since information hiding just add additional information to the digital data, it does not prevent the data from being accessed and used.

Two main branches in information hiding can be found in literature, one is steganography [19] and the other is digital watermarking [20, 21]. Steganography refers to embedding confidential information such as message, image, or video, within another digital signal (cover signal) without attracting suspicion. Digital watermarking is an art of embedding digital information (e.g. copyright notice or serial number) into the digital signal, mainly for protection purpose. The embedded data is generally referred as watermarks. Both steganography and watermarking take advantages of the redundant components in digital signal to hiding data. In particular, steganography concentrates on hiding the existence of the embedded information from being discovered by the third party during communication, while watermarking aims to protect the digital signal with the embedded information. Therefore, watermarking can be used for digital signal protection. Compared with cryptography, watermarking does not prevent the users from listening to and using the signal. Moreover, since watermarks are directly embedded within the signal, the embedded information can permanently exist and is difficult to be removed. Therefore, watermarking enables signal to be protected in a more suitable and durable way.

Digital watermarking can provide effective protection of digital signals by preventing them from being unauthorized tamper or identifying the tampering. The hidden watermarks usually under a low energy level so that they won't distort the perceptual quality of these signals. The basic watermarking scheme including watermark embedding and extraction processes are illustrated in Fig. 1.2, where the signal with or without watermarks is referred as watermarked signal and original signal, respectively.

Watermarking was originally motivated when signal which contain commercial values became available in the digital form [20, 21, 22]. In general, the copy of digital signal is completely the same as the original one, watermarking has great concern to music, manufacturers, and publishing companies since unauthorized copies usually lead to great loss.

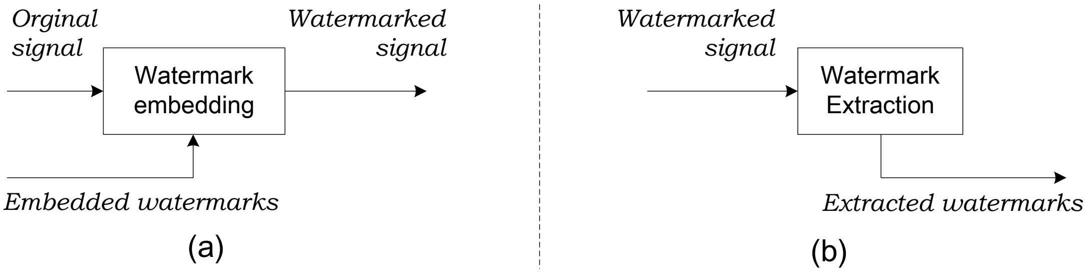


Figure 1.2: Block diagram of (a) watermark embedding and (b) watermark extraction.

In the past few decades, digital watermarking has been found in many applications, such as copyright protection [23, 24], authentication, broadcast monitoring, digital forensics, secure communication, crime investigation, and so on. The details about watermarking will be introduced in the following section.

1.3.2 Digital watermarking for audio

Watermarking technique was originally applied to image and video protection. While as it is getting mature, more attention has been paid to the extensively used audio and speech [25, 26]. To be effective, audio/speech watermarking should satisfy several basic requirements of (a) inaudibility: the embedding of watermarks should not degrade the sound quality of signal for its applications; (b) robustness: allowable processing and common attacks to watermarked signal (e.g., re-sampling and re-quantization) should not destroy the embedded watermarks; (c) blindness: watermarks should be detected without referring to any information of the original signal [27, 28]. The importance of a particular requirement may vary among different applications. For copyright protection of audio signals, watermarks are embedded within the digital products in form of serial numbers or identification code for several purposes such as recording the copyright ownership, tracing the distributions, and identifying the producers, etc. In this case, inaudibility and robustness are controlled as the top priority, since inaudibility keeps the commercial value of the audio product intact and robustness guarantees a reliable extraction of the copyright information after the distribution process. However, since inaudibility and robustness conflict with each other, watermarking that can satisfy both inaudibility and robustness are difficult to be realized.

In previous works, numerous watermarking techniques related to image and video have been studied rigorously [29, 30]. Digital watermarking for audio, however, is more challenging since the human auditory system (HAS) is more sensitive in comparison with human visual system (HVS) due to its wide dynamic range. Nonetheless, many relatively successful watermarking algorithms regarding to audio signals have been proposed and applied in some real situations effectively. Many principles that used in image and video watermarking have been inherited to accommodate audio watermarking. These methods can be divided into several categories.

- Watermarking in time domain. The time-domain methods, such as the least significant bit-replacement (LSB) [31] method took advantage that human was not sensitive to slight modification to the insignificant bits to realize inaudibility. The echo-hiding based methods [32, 33] utilized mask effect in time domain to achieve inaudibility. Most conventional echo hiding methods had the problem concerning the robustness against malicious attacks and security, and the time-spread echo [34] method was then proposed to solve these problems. However, most of the time-domain methods were prone to be not robust.

- Watermarking in transform domain. Another type of methods were implemented in the transform domain for stronger robustness. One typical class of watermarking methods employed the spectrum modulation technique, such as the secure (tamper-resistant) spread spectrum [35], improved spread spectrum (ISS) [36, 37], direct spread spectrum (DSS) [38], and robust spread-spectrum [39, 40]. This kind of watermarking was well known for its robustness since the watermark information was spread over a wide frequency, so it was difficult to destroy or remove it without causing distortion to the original signal. However, inaudibility of these methods could not be always satisfied.

- Watermarking based on human auditory system (HAS). Since the HAS [41] is more sensitive than human visual system (HVS), more watermarking methods tended to exploit the properties of the HAS and applied such knowledge to obtain better performance. In these methods, watermarks were embedded into the perceptually inaudible components while leaving the sensitive components intact to realize inaudibility. For example, Fallahpour and Megas [42] proposed an audio watermarking method in the logarithm domain by utilizing the property of absolute hearing threshold. Swanson et al. [43] suggested to embed copyright protection information into digital audio signal by directly modifying the audio samples after considering the temporal and frequency masking effect. Battisti et al. [44] presented an audio watermarking techniques based on Linear Predictive Coding (LPC) and the frequency masking effect. Unoki and Hamada [45] introduced an audio watermarking technique based on the characteristics of human cochlear delay (CD).

- Watermarking based on quantization index modulation (QIM) [46, 47]. This kind of methods can achieve good balance between embedding capacity and distortion introduced into the host signal, although robustness is improved at the expense of degraded inaudibility.

1.3.3 Digital watermarking for speech

Different from audio signals, speech signals are usually used as a true representation for an event happened at a certain time and place. When verifying this, the originality of speech signals should be validated. Speech watermarking can deal with this issue by addressing two questions: one is whether the speech is original and the other is whether the speech has been tampered since its creation. To accomplish this, a well-designed speech watermarking should not only satisfy three basic requirements of inaudibility, robustness, and blindness, but also have the ability to detect tampering. In this case, speech watermarking is usually designed to satisfy an additional requirement of fragility. Fragility means watermarks should be destroyed and fail to be detected once a slight tampering has been made to the watermarked signal. This kind of watermarking is referred as fragile watermarking

[48, 49]. Fragile watermarking has the ability to identifying positions where tampering has occurred as well as check the originality of speech with the destroyed watermarks.

It is important to note that since speech signals are unavoidably subjected to general speech processing and common attacks, robustness which means watermarks can be reliably extracted although watermarked signals have been processed with these processing, is also very important. Therefore, effective speech watermarking methods should satisfy two conflicting requirements: robustness against general processing and fragility against malicious tampering, to confirm that the failed detection of watermarks could only be caused by tampering, not by general processing. Only then can the watermarking methods provide effective protection of speech signals.

In general, tampering detection schemes for speech come down to two main categories: (1) schemes just verify the originality of speech without localizing the tampering and (2) schemes that can locate the tampering in time domain. The second category is more preferred in practical applications. Although there are many similar characteristics between audio and speech signals, the broadband audio watermarking methods may not be effective for speech since speech is characterized by intermittent voiced and unvoiced period with most information presented below a relative narrow bandwidth of 4 kHz. In the literature, limited tampering detection schemes concerning the above two categories have been found. For example, in [50], Park et al. investigated a scheme with watermarking and pattern recovery to detect tampering. A watermark pattern was attached to speech so that when tampering occurred, destroyed watermark pattern could be used to identify the tampering. In this scheme, tampering was only detected after MP3 (at 16 kbps) and code-excited linear prediction (CELP) (11.5 kbps) compression, and only three tampering, i.e., substitution, insertion, and removal were considered. Celik *et al.* proposed a watermarking method by introducing small changes to pitch (fundamental frequency) [51] with quantization index modulation (QIM) [46, 47]. Insensitivity of human perception to the natural variability of pitch enabled the method to be inaudible. The stability of pitch under low data rate compression (e.g. Global System for Mobile communications coder (GSM) 6.10 and Adaptive Multi-Rate coder (AMR)) also made the method effective for semi-fragile authentication. Nonetheless, the method had not been designed to be robust against attacks that aimed to block the extraction of watermarks. For example, a sophisticated modification of pitch such as re-embedding, would destroyed the watermarks. Wu *et al.* implemented a fragile speech watermarking for tampering detection based on odd/even modulation with exponential scale quantization [52]. Watermarks were embedded as a kind of pseudo-random noise in discrete Fourier transform (DFT) domain by roughly approximating the psychoacoustic model. The time resolution for tampering could be set at 0.5 second or even shorter. Nonetheless, its compatibility with the CELP codecs still needed to be improved. In [45, 53, 54], Unoki and Hamada introduced a watermarking method by employing the characteristics of cochlear delay (CD). Watermarks were embedded by enhancing the phase of the original speech with respect to two kinds of group delays. Based on this concept, a tampering detection scheme was presented in [55]. The performance of this scheme was evaluated at variant embedding bit rate. It was found that the scheme could successfully detect tampering, and the detection precision (in second) could be increased with higher bps. Nonetheless, this scheme did not show strong robustness when subjected to speech codecs of G.726 [56] and G.729 [57] and Conjugate-Structure ACELP (CS-ACELP) [58].

Since the requirements for watermarking usually conflict with each other, it is proven to be difficult for many speech watermarking methods [31], [48] and [55] to satisfy all these requirements simultaneously, such as inaudibility and robustness, and robustness and fragility (fragility may occasionally make the watermarking methods not robust). The performance of these watermarking methods will be much degraded.

1.4 Motivation and purpose

1.4.1 Motivation

Speech signal is actively involved in the human life. Advanced digital techniques have enabled speech signal to be easily accessed and distributed via the Internet and digital devices. These advances have been accompanied by a series of social issues in related to data abuse and authorized tampering. As an important information carrier, speech signal is usually used as a true representation of an event happened at a certain time and place. Therefore, the originality of speech signal is very important, especially for those used in digital forensics and speech communication.

However, since unauthorized tampering to speech can easily delete or replace important information of speech signal, the originality of speech is becoming difficult to be confirmed. Our research is motivated by this issue, and we would like to detect tampering as well as check the originality of speech in this research.

1.4.2 Purpose

Information hiding technique, especially the watermarking method can detect the tampering and validate the originality of speech signals by hiding digital data into speech. To be effective, watermarking methods should satisfy several requirements: (1) inaudibility to human auditory system, (2) blindness for watermark detection, (3) robustness against allowable speech processing and common attacks, and (4) fragility against tampering. However, it is proven to be difficult for watermarking methods to satisfy all these requirements simultaneously.

The main purpose of this research is to solve the problem of unauthorized tampering as well as check the originality of speech with information hiding and watermarking methods that can satisfy all the requirements simultaneously. The first target is to realize a general watermarking method that can satisfy all the first three requirements (inaudibility, blindness, and robustness). After that, the fragility of this method will be investigated for tampering detection. Finally, the watermarking method will be explored for other applications. To achieve these targets, we have deeply studied the knowledge of speech production, the source-filter model, the linear prediction (LP) analysis, and speech perception. The process of speech production is essential for us to construct the framework of watermarking and the knowledge of speech perception helps us to realize inaudible watermarking by taking advantage of insensitive and robust speech parameters.

As we have found, formant is an important acoustic feature for speech perception, and formant can to be tuned when the quality of speech is impaired by environmental noise or other reasons. For example, in the field of speech synthesis, the sound quality of synthesized speech can be improved by tuning the formants. Since formant can be tuned to improve the speech quality, and such modifications do not cause perceptual distortion to the original speech, watermarking based on formant tuning is possible to be inaudible. Moreover, formant can be directly controlled by line spectral frequencies (LSFs) and the LSFs are robust speech parameters. Therefore, if watermarks can be embedded into LSFs as making tuning of formant, the watermarking method is able to realize inaudibility and robustness simultaneously. According to these analysis, we have proposed our main concept of watermarking, i.e., watermarking based on formant tuning. To make our watermarking method effective, the principles of how formants can be produced, controlled, and then tuned by LSFs have been investigated. The whole watermarking scheme is based on the source-filter model of speech production, and the LP analysis is used to separate the information of sound source and the formants so that watermarks can be embedded.

Although, several LP-based information hiding schemes have been proposed in previous works [59, 60, 61], compared with these works, our watermarking method focuses on modifying the formants, i.e., vocal tract filter information for watermark embedding, other than modifying the sound source. Moreover, to the best of our knowledge, our work is the first time to introduce the idea of formant tuning for speech watermarking.

Ideally, if our method can satisfy all the requirements of speech watermarking, it can embed watermarks without degrading the speech quality, and the tampering can be effectively detected. That is to say, it can address the problems of speech tampering and check the originality of speech signal. Some malicious affairs in speech communication, digital forensics, government, and industries can also be avoided.

1.5 Dissertation Outline

The rest of this dissertation is organized as follows, the organization follows the structure in Fig. 1.3.

Chapter 2

In this chapter, the background about speech, speech production, and the source-filter model is talked about. This knowledge is essential for the speech processing based research. The LP analysis which can model the vocal tract filter is also discussed. Speech analysis/synthesis and speech coding based on LP are introduced. Moreover, two important speech parameters, i.e., formants and line spectral frequencies (LSFs), and their properties are given out.

Chapter 3

After knowing the basic knowledge of speech production and the properties of speech parameters, in chapter 2, we introduce our main concept of speech watermarking. The LSFs which have several excellent properties, are selected as the carrier of watermarks. The modifications to LSFs can be considered as make tuning to the formants. Based on this analysis, we propose two concepts of speech watermarking. One is watermarking based on LSFs modifications with QIM, and the other one is watermarking based on formant enhancement.

Chapter 4

In this chapter, we will implement the above two speech watermarking methods. The framework of watermarking based on LSF is firstly constructed. Watermarks are embedded into speech signal by quantizing LSFs with QIM. The overall watermarking scheme consists of watermark embedding and extraction processes are explained. In the formant-enhancement based watermarking, we will introduce the ides of enhancing the formant by directly closing up two LSFs. This idea is then employed for the watermarking embedding. The whole scheme of watermark embedding and extraction based on formant enhancement is also discussed. Additionally, a frame synchronization scheme is also implemented in this chapter.

Chapter 5

In this chapter, we will evaluate the proposed two methods with respect to inaudibility and robustness (both of the two methods are blind methods). The database and evaluation measure are firstly given out. For the LSFs and QIM based method, we will check the inaudibility and robustness performances by using different quantization steps. In the evaluations of formant-enhancement based watermarking, we will give much consideration about the adjustable parameters when balancing the inaudibility and robustness. The obtained results from the watermarking methods will be analyzed. Besides, we will investigate the influence of frame synchronization to watermarking method with respect to speech quality and watermark extraction.

Chapter 6

In this chapter, the formant-enhancement based watermarking will be applied for tampering detection and hybrid watermarking. For tampering detection, the detailed processing at the sender side and the receiver side will be explained. The ability of the proposed method will be evaluated with respect to inaudibility, robustness, and several kinds of tampering. The obtained results will be discussed to check the effectiveness of tampering detection. The formant-enhancement based watermarking is also used to implement a hybrid watermarking by incorporating the cochlear delay based watermarking. The whole scheme of hybrid watermark will be talked about. The performance of the hybrid method will be evaluated and the results will be discussed.

Chapter 7

The contributions and the summary of the dissertation is given out. A simple discussion of future research directions are also talked about.

1.6 Summary

This chapter presents an introduction of this research. An overview of problems in speech protection, data hiding techniques, and speech watermarking methods have been surveyed. Existing watermarking methods and their advantages and disadvantages have

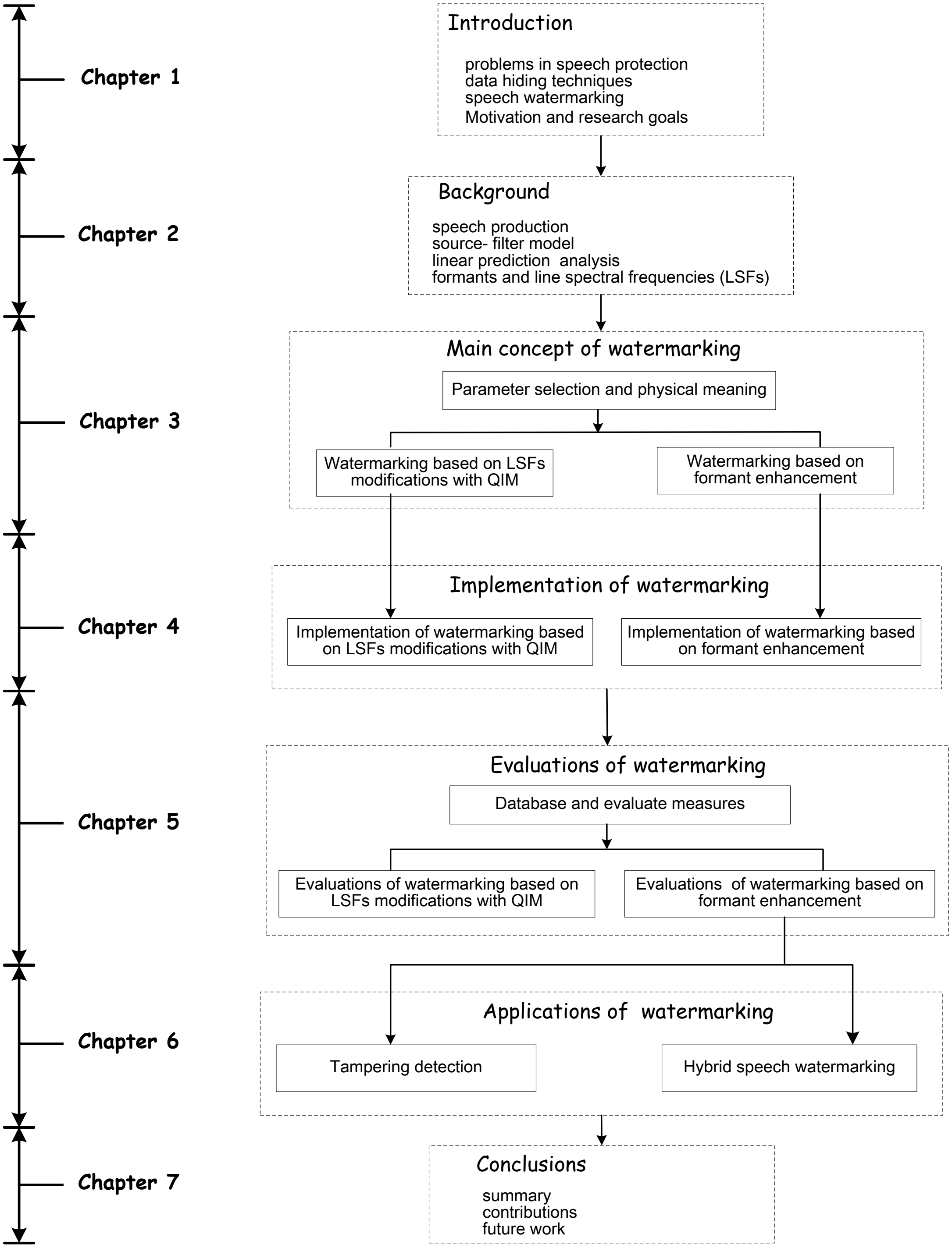

Chapter 1
Chapter 2
Chapter 3
Chapter 4
Chapter 5
Chapter 6
Chapter 7
Introduction
problems in speech protection
data hiding techniques
speech watermarking
Motivation and research goals
Background
speech production
source- filter model
linear prediction analysis
formants and line spectral frequencies (LSFs)
Main concept of watermarking
Parameter selection and physical meaning
Watermarking based on LSFs modifications with QIM
Watermarking based on formant enhancement
Implementation of watermarking
Implementation of watermarking based on LSFs modifications with QIM
Implementation of watermarking based on formant enhancement
Evaluations of watermarking
Database and evaluate measures
Evaluations of watermarking based on LSFs modifications with QIM
Evaluations of watermarking based on formant enhancement
Applications of watermarking
Tampering detection
Hybrid speech watermarking
Conclusions
summary
contributions
future work

Figure 1.3: Organization of the dissertation.

been analyzed and summarized. The motivation and research purpose of this dissertation are clarified. Finally, the organization of this dissertation is outlined.

Chapter 2

Research background

In this chapter, the background knowledge about speech is presented. These includes the basic knowledge of speech production, the source-filter model, the LP analysis, and several acoustical features such as formants and the LSFs.

2.1 Basic knowledge of speech

Speech carries important information for human communication. Most of the research on speech starts from the process of speech production. The mechanism of speech production and the basic parameters involved in speech production are essential for the speech processing based research.

Speech is a phonetic combination of a limited set of vowels and consonants. Speech production concerns to the manners about how the speech organs such as tongue, lips, and jaw work together to make a sound. At the articulatory level, the acoustic properties of phonemes including vowels and consonants are characterized primarily by the manner and constriction location of articulation: consonants are articulated with tight constriction or complete closure in the vocal tract, while constrictions of vowels are not as tight as those of consonants. In the process of speech production, air flow is firstly expelled from the lungs, and then the air flow will pass through the glottis between vocal folds with or without constriction. When the glottis is closed with no air passed and no vocal folds

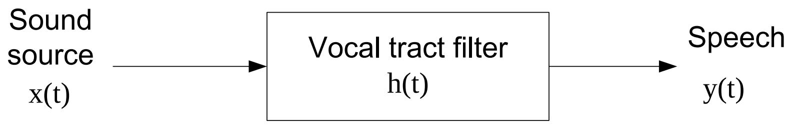


Figure 2.1: The source-filter model for speech production.

vibration, a voiceless sound is produced. If the glottis is opened, the air can pass and the vocal folds will vibrate, then a voiced sound is produced. Finally, the air flow will go to the nasal or the oral cavity for different speech sounds [62, 63].

In speech processing based research, such as speech coding, speech analysis and synthesis, and speech recognition, the physical process of speech production is generally simplified into a source-filter based model. In the next chapter, the source-filter model will be explained.

2.2 Source-filter model for speech production

2.2.1 Theory of source-filter model

Human beings are able to independently control the glottal pulse and vocal tract. The process of speech production can be modelled by a linear system, named as the source-filter model. The basic conception of source-filter model appeared in the work of Chiba and Kajiyama [64], and Fant proposed the systematic theory of source-filter model in 1960 [65, 66]. The source-filter model is of great value to many speech analysis/synthesis methods [67, 68] and low bit-rate speech coders [69, 70] due to its simplicity and good approximation for speech production.

The source-filter model of speech production assumes the glottal pulse is the sound source and the vocal tract acts as an acoustic filter. When we produce speech sounds, as seen in Fig. 2.1, the sound source (also named as excitation signal), $x(t)$, becomes the input signal to the vocal tract filter, $h(t)$. The output speech, $y(t)$, can be expressed by

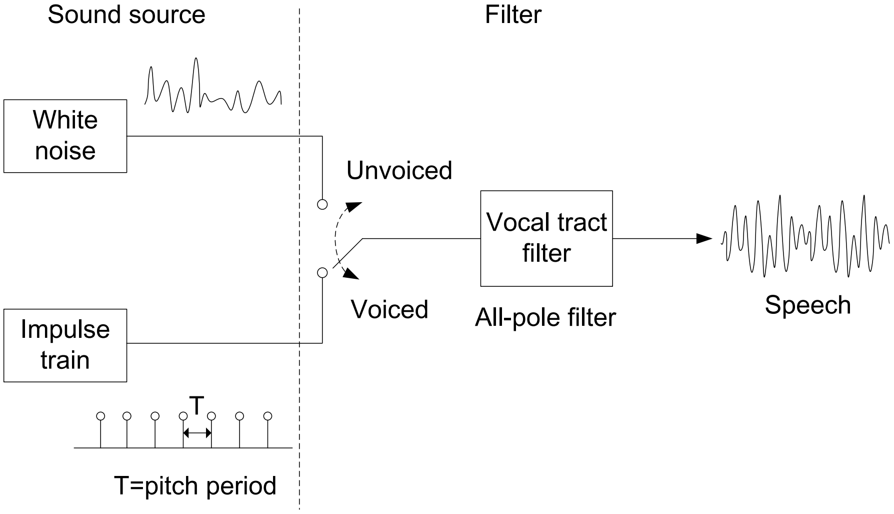


Figure 2.2: The source-filter model of speech production.

the convolution of an input signal and vocal tract filter, that is:

$$y(t) = h(t) * x(t), \tag{2.1}$$

where $*$ denotes the convolution.

In the source-filter model, the sound source and vocal tract filter are assumed to be independent with each other. That is to say we can adjust the properties of the filter without modifying the properties of the sound source. This assumption enables the speech production model to be very practical and accurate [71].

2.2.2 Implementation of source-filter model

In general, the source-filter model can be implemented as follows. For voiced speech, the sound source can be modelled with a periodic impulse train; for unvoiced speech, the sound source can be modelled with a white noise-like signal [72]. This two kinds of state can fit well with true glottal behaviours. The vocal tract filter can be approximated with an all-pole filter. This source-filter model for speech production is illustrated in Fig. 2.2, where the pitch period T varies among different people. The speech sound can be produced by the convolution of the excitation signal and the filter. Adjusting the shape of the vocal tract filter can usually generate the speech sounds with different qualities. Although the actual process of speech production is non-linear and there is an interaction between the sound source and a vocal tract filter, the source-filter model provides us a good and reasonable approximation of speech production for many speech analysis/synthesis applications.

2.3 Linear prediction analysis

2.3.1 Fundamentals of LP analysis

The source-filter model is closely related to LP [73, 74], since the vocal tract filter is an all-pole filter, which can be modelled with the linear prediction. This chapter will talk about the LP model, LP based speech analysis/synthesis, and speech coding.

LP [73, 74, 75, 76] is a technique for time series analysis in linear system. As indicated by its name, LP predicts the current signal (the output of LP system) with a linear combination of its previous samples. A general representation of LP is expressed in Eq. (2.2):

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i), \tag{2.2}$$

where p indicates the LP order, a_i are the LP coefficients, $\hat{x}(n)$ is the prediction value of the output signal $x(n)$, and $x(n-i)$ stands for the i-th previous sample. The prediction error $e(n)$ between the output signal and its prediction value is commonly referred as residue. Residue can be expressed as follows:

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{i=1}^{p} a_i x(n-i). \tag{2.3}$$

The main problem in LP analysis is how to obtain the set of LP coefficients to provide an accurate prediction of signal. Many methods have been proposed to calculate the LP coefficients, e.g., the autocorrelation method. For most methods, LP coefficients are

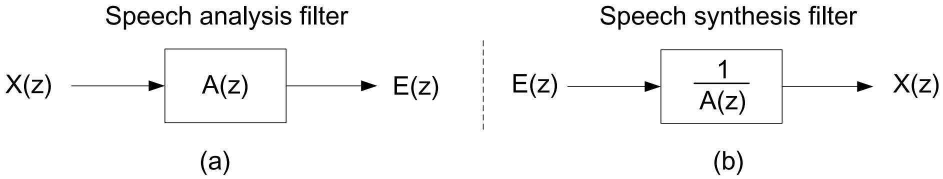


Figure 2.3: Speech analysis and synthesis based on (a) analysis filter and (b) synthesis filter.

calculated by minimizing the squared error E of the residue $e(n)$ as follows:

$$E = \sum_n e(n)^2 = \sum_n \left(e(n) - \sum_{i=1}^{p} a_i x(n-i) \right)^2 \tag{2.4}$$

2.3.2 Speech analysis and synthesis based on LP

The LP model can be used for speech analysis/synthesis based methods. According to Eq. 2.3, the residue $e(n)$ can be represented in z-domain as follows:

$$E(z) = X(z) - \sum_{i=1}^{p} a_i X(z) z^{-i} = X(z) \left(1 - \sum_{i=1}^{p} a_i z^{-i} \right). \tag{2.5}$$

where $X(z)$ and $E(z)$ are the z-transformation of the input signal $x(n)$ and residue $e(n)$, respectively. The above equation can be also written as follows:

$$A(z) = \frac{E(z)}{X(z)} = 1 - \sum_{i=1}^{p} a_i z^{-i}, \tag{2.6}$$

where $A(z)$ is named as transfer function.

With the transfer function $A(z)$, speech analysis process can be realized as Fig. 2.3(a), where LP residue can be calculated by filtering the speech with A(z). This process is also named as inverse filtering, and $A(z)$ is usually called as the speech analysis filter or the inverse filter.

Corresponding to speech analysis process, the speech synthesis process is realized in Fig. 2.3(b) with a similar manner. The output $X(z)$ of the $1/A(z)$ can be excited and reconstructed by the residue signal $E(z)$, where $1/A(z)$ is an all-pole filer, which can be given by

$$\frac{1}{A(z)} = \frac{X(z)}{E(z)} = \frac{1}{1 - \sum\limits_{i=1}^{p} a_i z^{-i}}, \tag{2.7}$$

$1/A(z)$ is usually referred as vocal tract filter or speech synthesis filter.

2.3.3 Speech coding based on LP analysis/synthesis

Nowadays, speech coding algorithms have been widely used in telecommunications, mobile communications, VoIP, and radio system [3, 4]. Speech coding is the process of compressing analog speech waveform into digital form. The main objective of speech coding is to express the speech with fewer bits so they can be efficiently transmitted over the band-limited transmission channel or stored in the digital media devices [77]. The whole process of speech coding can be summarized as follows: at the sender side, speech encoding process will convert the analog speech signal into digital form; at the receiver side, the decoding process will convert the digital speech back to the analog waveform. The perceptual quality and the bit rate are two important criteria that can determine the performance of coding methods. Therefore, efficient speech coding method should provide good quality of reconstructed speech with as few bits as possible.

In general, speech coding algorithms can be divided into two categories: the waveform coders and the voice coders (vocoders). The waveform coders try to directly encode the exact waveform of the speech signal (either temporal or spectra waveform). Therefore, the waveform of reconstructed signal should be as closer as the original one to enable a good perceptual quality. Two kinds of typical waveform coders are Pulse Code Modulation (PCM) [78], and Adaptive Differential Pulse-Code Modulation (ADPCM) [79] in time domain, and sub-band coders and adaptive transform coders in frequency domain. The waveform coders are generally implemented with low complexity but high bit rate (over 16 kbps). When the bit rate is lowered, the quality of reconstructed speech will be drastically degraded.

Compared with waveform coders, the vocoders can realize better speech quality at a lower bit rate since the mechanism of speech production is considered [80]. LP based speech analysis/synthesis method is used in most vocoders, in which the spectral envelope of each speech frame is represented with extracted speech parameters. This kinds of vocoder are referred as LPC. LPC are widely used in the residual excited Linear Prediction (RE-LP), multi-pulse LPC (M-LPC) Vocoder [81], and Code-Excited Linear Prediction (CELP) [82, 58]. In LPC based vocoders, a speech signal $X(z)$ will be passed through the speech analysis filter, as shown in Fig. 2.3 (a). to obtain the residue signal $E(z)$. This process is expressed in Eq. (2.8). The residual signal has very flat spectrum and less redundancy compared with the original speech, so that it can be quantized with fewer bits. The residual signal together with the filter coefficients a_i will be encoded and then transmitted to the receiver.

$$E(z) = X(z)(1 - \sum_{i=1}^{p} a_i z^{-i}). \tag{2.8}$$

At the receiver, the speech $X(z)$ is reconstructed by passing the residual $E(z)$ through the synthesis all-pole filter, as shown in Fig. 2.3 (b). This process is expressed in Eq. (2.9).

$$X(z) = E(z)\left(\frac{1}{A(z)}\right) = \frac{E(z)}{1 - \sum\limits_{i=1}^{p} a_i z^{-i}}. \tag{2.9}$$

2.3.4 Formans and LSFs

Speech analysis and synthesis utilize the residue information and the vocal tract information. The vocal tract information can be represented with one kind of speech parameters, i.e., formants. Formants are concentration of frequencies which are close to the resonance frequencies of the vocal tract. The characteristics that humans require to distinguish vowels can be represented by formant information. Besides, formants are very important parameters to evaluate the quality of speech. The positions of formant frequencies are mainly determined by the shape and length of the vocal tract. People at different ages or with different genders usually have different formant frequencies even for the same speech sound.

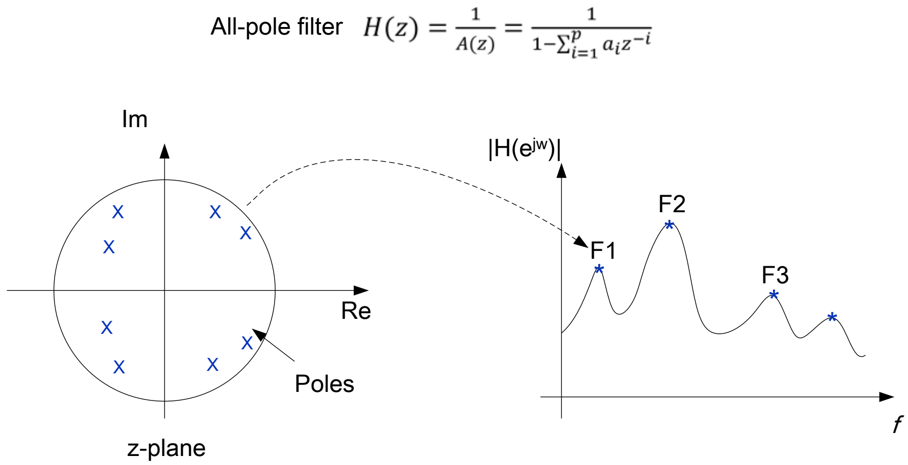

All-pole filter $H(z) = \frac{1}{A(z)} = \frac{1}{1-\sum_{i=1}^{p} a_i z^{-i}}$
Im
Re
Poles
z-plane
|H(e^jw)|
F1
F2
F3
f

Figure 2.4: The source-filter model of speech production.

Based on the source-filter model, the set of LP coefficients constitute a vocal tract filter that has only poles, and these poles correspond to the resonant frequencies of the spectrum, that is the formants. Each pair of complex-conjugate poles in the all-pole filter represents one formant, as seen in Fig. 2.4. All the formants in speech spectrum are consecutively labelled as F1, F2, F3, ..., from the low frequency to high frequency. Based on these analysis, LP coefficients can provide accurate estimate of formants.

However, since LP coefficients are hard to interpolate and sensitive to noise, a small error in LP coefficients can distort the whole spectrum or make the LP filter unstable. In practice, LP coefficients are usually substituted with other advanced representations, such as log area ratios (LARs), line spectral pairs (LSPs) [83, 84], LSFs [85, 86], and reflection coefficients (RCs), to ensure the stability of the predictor. Among these representations, LSPs and LSFs possess several excellent properties over the others. These properties enable them to be widely used in the LP based speech coding and other speech analysis/synthesis based research [87].

LSPs and LSFs are actually a mathematical transformation of the LP coefficients. The transformation from LP coefficients to LSPs can be explained as follow [88, 89]. Usually,

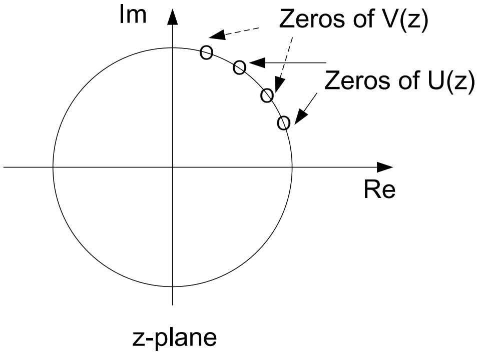


Figure 2.5: Roots of $U(z)$ and $V(z)$ located on the unit circle.

a p-th order LP inverse filter in Eq. (2.6) can be expressed as follows:

$$\begin{aligned} A(z) &= \frac{U(z)+V(z)}{2}, && (2.10) \\ U(z) &= A(z) + z^{-(p+1)}A(z^{-1}), && (2.11) \\ V(z) &= A(z) - z^{-(p+1)}A(z^{-1}). && (2.12) \end{aligned}$$

LSPs are defined as the roots (the zeroes) of $U(z)$ and $V(z)$. All the roots of $U(z)$ and $V(z)$ are alternately located on the half unit circle on the complex plane, as seen in Fig. 2.5. All of roots have a complex conjugate on the z-plane [90].

When p is an even number and greater than two, $U(z)$ has a zero of -1 while $V(z)$ has a zero of 1, along with other $p/2$ pairs of conjugated zeros. Therefore, $U(z)$ and $V(z)$ can also be represented as follows:

$$\begin{aligned} U(z) &= (1+z^{-1})\prod_{i=1}^{p/2}(1-2z^{-1}\cos\phi_i + z^{-2}), && (2.13) \\ V(z) &= (1-z^{-1})\prod_{i=1}^{p/2}(1-2z^{-1}\cos\theta_i + z^{-2}), && (2.14) \end{aligned}$$

where $\cos(\phi_i)$ and $\cos(\theta_i)$ are LSPs in the cosine domain. The ϕ_i and θ_i are LSFs and they interlace with each other and satisfy the ordering property from 0 to π as follows:

$$0 < \phi_1 < \theta_1 < \phi_2 < \theta_2 < \cdots < \phi_{p/2} < \theta_{p/2} < \pi. \qquad (2.15)$$

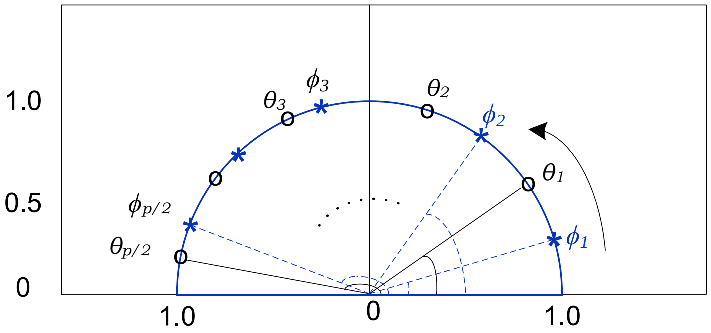


Figure 2.6: The distribution of LSFs on the half unit circle.

The distribution of LSFs on the half unit circle can be observed from Fig. 2.6. The distance of two adjacent LSFs can reflect the resonance information of the speech spectra, i.e, a closer distance indicates a strong resonance. The properties and advantages of LSFs can be summarized as follows:

(1) In comparison to LP coefficients, LSFs are less sensitive to quantization noise [91];

(2) When the roots of $U(z)$ and $V(z)$ are interleaved and monotonically increasing, stability of predictor can be ensured;

(3) LSFs are easy to be interpolated, when used in CELP speech coders, they can be quantized with fewer bits than other LP representations while keeping a good quality of reconstructed speech [92, 93];

(4) The influences caused by deviation of LSFs can be limited to the local spectra;

(5) Except for CELP coders, LSFs are used in almost all of the LP based speech coders. LSFs can be regarded as universal features in different speech codecs.

(6) The LSFs can directly reflect the vocal tract and formant information.

Benefits from these properties, LSFs have been widely used.

2.4 Summary

In this chapter, the background knowledge about speech and the mechanism of speech production are talked about. The physical process of speech production is generally simplified into a source-filter based model. The characteristics vocal tract filter in the source-filter model can be approximated by LP. The fundamentals and principles of LP are introduced. Moreover, two important parameters, i.e., formant and line spectral frequencies (LSFs), and their properties are given out. This knowledge is essential for the speech processing based research, such as speech analysis/synthesis and speech coding.

Chapter 3

Concept of watermarking based on formant tuning

The purpose of this research is to protect speech signals with watermarking methods. The first target is to realize a general watermarking method that can satisfy three basic requirements of inaudibility, blindness, and robustness. After that, some special requirements, such as fragility for tampering detection will be explored to check the originality of speech signals. With this purpose, in this chapter, we would like to introduce our main concept for speech watermarking methods.

3.1 Introduction

Watermarking is an efficient way to protect speech signals without degrading the speech quality and obstructing the speech from being listened or used. In chapter 1, a detailed introduction of data hiding and speech watermarking methods have been surveyed. The advantages and disadvantages of existing watermarking methods have been analyzed and summarized. It is clearer that the demanding requirements of watermarking are hard to satisfy.

To realize desirable watermarking, basic requirements concerning inaudibility, blindness, and robustness, should be considered before selecting the implementation domain and the speech parameters that are used to carry the watermarks. For inaudibility, since all watermarking methods embed the watermarks by artificially modifying the original speech (in either time domain or transformed domain), distortion will be introduced to the original speech signals unavoidably. Therefore, it is important to choose the speech parameters that human is not sensitive to make the embedded watermarks inaudible. Besides, the chosen parameters should be robust enough to guarantee a reliable watermark extraction even watermarked signals have been processed by different speech codecs and speech processing.

3.2 Inaudible speech watermarking

As the sensory system for hearing, the human auditory system is very sensitive [41]. Therefore, understanding the characteristics of human auditory system is helpful for designing inaudible speech watermarking. In the literature, several previous works have exploited the properties of HAS and embedded watermarks to the perceptually inaudible components for inaudibility. These included the work of Celik et al., who proposed a watermarking method by introducing small changes to fundamental frequency [51], the work of Unoki and Hamada, who took advantages of characteristics of cochlear delay (CD) that human is unable to discriminate an enhanced group delay from the original speech [45, 54], and the work of Fallahpour and Megas [42], who proposed a watermarking method in the logarithm domain by utilizing the property of absolute hearing threshold. Besides, there are other properties of HAS, such as the frequency masking effect, temporal masking effect, frequency selectivity, psychoacoustic model, and insensitive speech parameters have been used for embedding watermarks without altering the perceptual quality of the speech signal. In this research, we would like to realize an inaudible speech watermarking by subtly modifying the suitable parameter of speech signal. We believe that if the speech parameter could be slightly modified with suitable method, watermarks are possible to be inaudibly embedded into the speech signals.

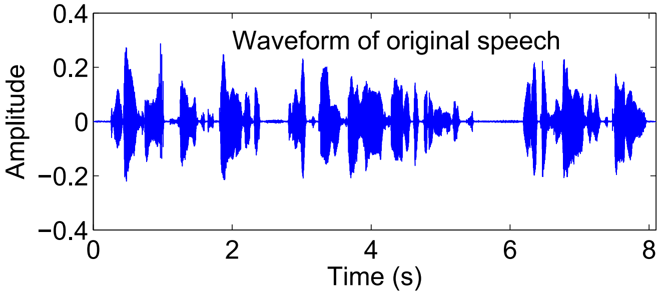


Figure 3.1: The waveform of 8.1 second speech.

3.2.1 Parameter for speech watermarking

In chapter 2, the basic knowledge of speech production, LP analysis, and speech parameters such as formant and LSFs have been reviewed. Based on the source-filter model, the LP coefficients can provide accurate estimation of formants. The LSFs, as substitute parameters of LP coefficients, cannot only represent the formants but also insensitive to noise and easy to ensure the stability of predictor. Moreover, there are other excellent properties can be found about LSFs: (1) the influences caused by the deviation of LSFs can be limited to the local spectral, thus if there is distortion introduced to LSFs, the degradation of speech signal in both spectra and sound quality can be minimized; (2) LSFs are universal features in different speech codecs, it is easy for them to survive from different encoding/ decoding process. Correspondingly, we believe that (1) if LSFs are selected to carry the watermarks, sound distortion introduced to the original speech signal can be minimized; (2) embedding watermarks into LSFs enables the watermarking method to be robust against difficult speech codecs. According to these analysis, the LSFs are selected for watermarking embedding.

3.2.2 Formant tuning: the physical meaning for modifying LSFs

In this chapter, we will talk about the physical meaning for watermarking based on LSFs modifications. Firstly, an example about how to predict a signal and estimate the formants with LP analysis will be given.

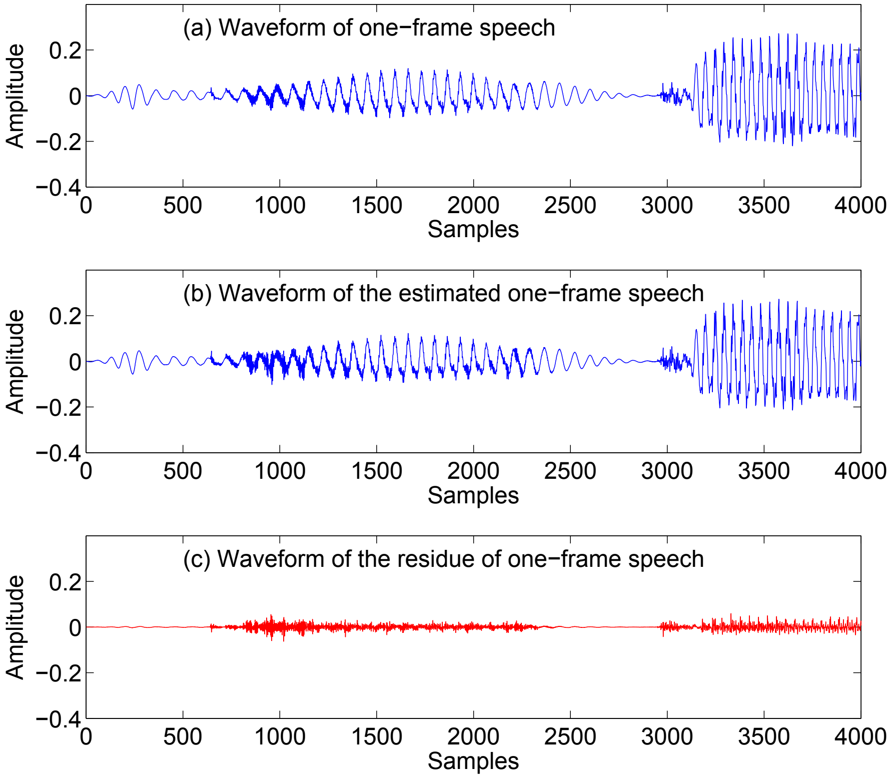


Figure 3.2: The waveform of (a) one speech frame (b) the estimated speech frame with LP analysis, and (c) the residue signal.

One speech stimulus out from the ATR database (B set) [94] was used as example speech signal. This stimulus was 20 kHz sampled, and 16-bit quantized, with a duration of 8.1 second. Figure 3.1 shows the waveform the speech stimulus. One speech frame of 4000 samples (250 ms) was extracted from the speech, as shown in Fig. 3.2(a). The 10-th order LP analysis was applied to this frame. The estimated signal by LP analysis is shown in Fig. 3.2(b). As we can see, LP analysis can provide very accurate estimation of the speech signal. The difference between the original speech frame and the estimated speech is shown in Fig. 3.2(c), this signal is residue.

As we have introduced, the LP coefficients can provide accurate estimation of the vocal tract information, i.e., the formants. The LSFs converted from LP coefficients can also directly reflect the formants. In this example, the order of LP analysis was chosen as

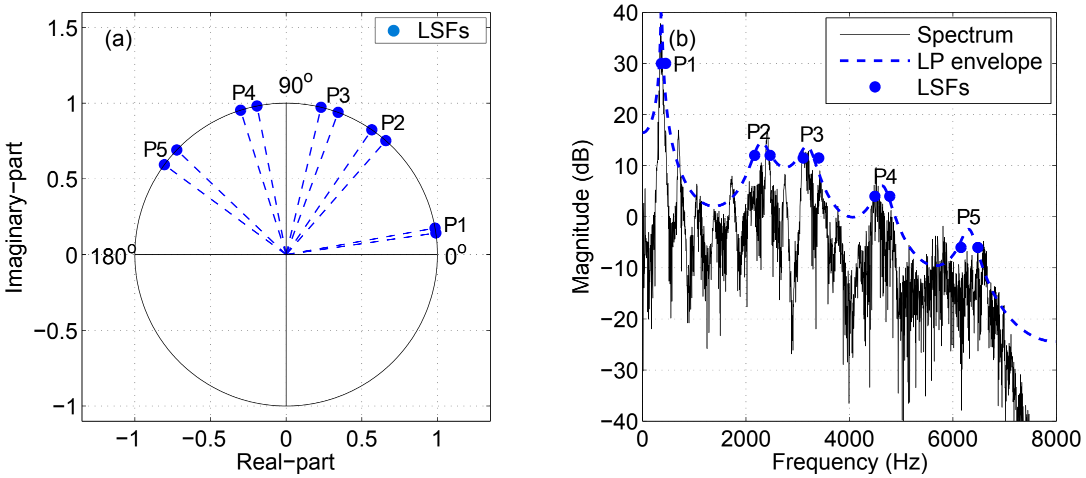


Figure 3.3: (a) LSFs distribution on half unit circle and (b) relationship between LSFs and formants.

10, so we can get 10 LSFs. Figure 3.3 shows the distribution of LSFs on half unit circle and the relationship between LSFs and formants. In theory, two adjacent LSFs (a pair of LSFs) can produce a formant, and the closer two LSFs are, the sharper formant is. As shown in Fig. 3.3(b), five formants are estimated in the LP spectral envelope. These formants correspond to the “P1” to “P5” labelled LSFs pairs in Fig. 3.3(a).

Since LSFs can directly represent the shape of formants, modifications made to LSFs can be considered as make tuning to the formants in physical. In the following, we will talk about our two ideas of how to make tuning to formants by controlling LSFs.

3.3 Concept of watermarking based on LSFs modifications with QIM

QIM is a mathematical operation of quantizing the value of signal to fixed value with a fixed quantization step. A comprehensive and in-depth introduction of the QIM has been presented by Chen and Wornell in [46, 47]. QIM has been considered as a promising method for digital watermarking since the implementation of QIM based watermarking is easy and applicable to any parameters of the digital signals. In addition, for QIM based watermarking, a trade-off among the conflicting requirements of distortion introduced to the speech signal, robustness, and embedding capacity can always be achieved by adjusting the quantization step [95].

Suppose s is the random signal needed to be quantized. The basic form of QIM can be expressed with Eq. (3.1), where $Q(\cdot)$ is the quantization function of QIM, Δ is the quantization step, "$[\cdot]$" stands for the rounding function, and s_w is the quantized value of the signal s. The distance between the values of s_w and s depends on the quantization step Δ.

$$s_w = Q(s) = \Delta \left[\frac{s}{\Delta} \right] \tag{3.1}$$

Considering the excellent properties of LSFs and advantages of QIM, our first concept for watermarking is embedding watermarks by quantizing LSFs with QIM. The framework of this watermarking will be constructed in chapter 4.

3.4 Concept of watermarking based on formant enhancement

As a crucial acoustic feature for speech perception, formant needs to be enhanced when the quality or intelligibility of speech is impaired by noise or other reasons. The method of re-shaping the formant to make it sharper is commonly referred as formant enhancement. This kind of method was originally developed in the adaptive post-filtering of speech codec to alleviate the perceptual effect caused by quantization noise [96]. Similar approaches that deal with formant to achieve better speech quality are widely found in the speech recognition system where the speech quality is reduced by noise [97, 98], and the hidden Markov model (HMM) based speech synthesis [99] where speech is muffled by the over-smoothed spectral envelope. In speech synthesis, the post-filtering technique for mel-cepstrum based and all-pole spectrum based spectra can be applied to increase the dynamics between formant peak and the spectral valley [99, 100, 101]. For example in [101], speech spectrum is modified so that the low-energy parts of the spectrum (valleys) can be additionally reduced and spectral peak is kept unmodified, this modification can be considered as making formant much sharper. Most of these methods try to obtain a more prominent formant structure by enhancing formant without shifting the center frequency of formant to maintain and optimize the speech quality [102].

Since formant can be enhanced to improve speech quality [103], and such modifications do not cause perceptual distortion to the original speech, watermarking based on formant enhancement is possible to be imperceptible to human. Therefore, we employ this concept to achieve inaudibility for watermarking. Watermarks will be embedded through formant enhancement. This watermarking method will be implemented in chapter 4.

3.5 Summary

In this chapter, we introduce our main concept of speech watermarking methods. In general, human is insensitive to tiny changes of speech parameters, watermarks are possible to be inaudibly embedded by subtly modifying speech parameters. Since LSFs can not only represent the formants but also have several excellent properties such as easy to ensure the stability of predictor and robust against different speech codecs, LSFs are selected as the carrier of watermarks. We investigate how the formant can be estimated by LP analysis and the relationship between LSFs and formants. The modifications to LSFs can be considered as make tuning to the formant. Then two concepts of speech watermarking are proposed: one is watermarking based on LSFs modifications with QIM, and the other is watermarking based on formant enhancement. In the next chapter, we will implement these two watermarking methods.

Chapter 4

Implementation of speech watermarking

The purpose of this research is to protect speech signals with watermarking methods. In the previous chapter, we have introduced the our main concept of speech watermarking, one is watermarking based on LSFs modifications with QIM, and the other is watermarking based on formant enhancement. In this chapter, we will implement these two speech watermarking methods.

4.1 Watermarking based on LSFs modifications with QIM

4.1.1 QIM based watermark embedding and extraction

QIM-based watermark embedding

QIM has been considered as a promising method for digital watermarking since it is easy to implement and applicable to any parameters of the digital signals. In QIM based watermarking, dither modulation-quantization index modulation (DM-QIM) which can provide two quantizers are usually used to embed different watermarks bits “0” and “1”

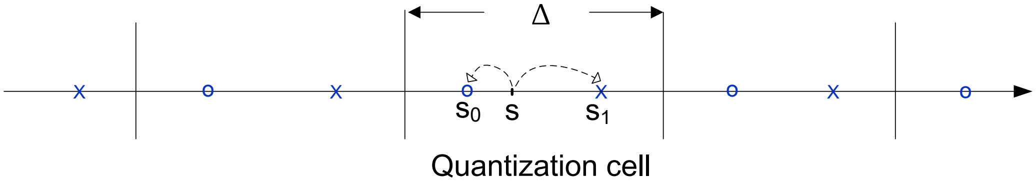


Figure 4.1: An illustration of QIM based watermark embedding.

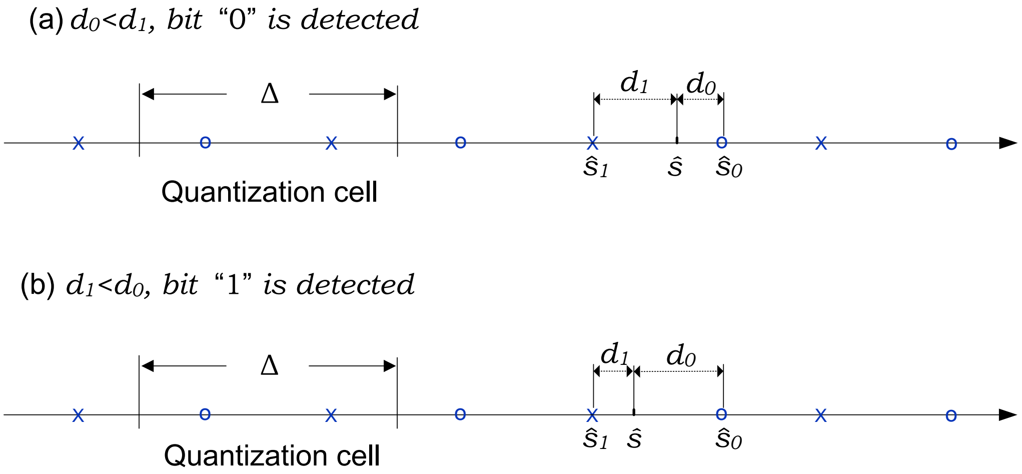
[104]. A general embedding function of DM-QIM can be expressed as $Q(s, w)$ with two variables of s and w, where s stands for the signal needed to be quantized for watermark embedding and w stands for the indexes of different quantizers for different watermark bits. Before embedding watermark bits, the quantization step Δ, which can decide the size of quantization cell, should be chosen. Based on Δ, two quantizers of DM-QIM, $Q_0(s, 0)$ and $Q_1(s, 1)$, that have different parameters for embedding watermarks "0" and "1" are decided. Eqs. (4.1) and (4.2) give the embedding functions for bits "0" and "1" in our QIM based watermarking:

$$s_0 = Q_0(s,0) = Q(s-b_0)+b_0 = \Delta\left[\frac{s-b_0}{\Delta}\right]+b_0, \quad b_0=-\frac{\Delta}{4}, \tag{4.1}$$

$$s_1 = Q_1(s,1) = Q(s-b_1)+b_1 = \Delta\left[\frac{s-b_1}{\Delta}\right]+b_1, \quad b_1=\frac{\Delta}{4}, \tag{4.2}$$

where b_w ($w = 0$ or 1) denotes the dither vector corresponding to $Q_0(s, 0)$ or $Q_1(s, 1)$, s_0 and s_1 are the quantized values of s that carry watermark "0" and "1", respectively. In general, there is not strict rules for how to fix the embedding functions, the only criterion is to make sure that there is no overlap quantized values can be generated by both Q_0 and Q_1.

Figure 4.1 illustrates the QIM based watermark embedding. Suppose s lies somewhere in one quantization cell. By using the embedding function, s will be uniquely mapped to the o−point (labelled as s_0) after embedding bit "0" or the x−point (labelled as s_1) after embedding bit "1" in the same cell.


(a) $d_0 < d_1$, bit "0" is detected

(b) $d_1 < d_0$, bit "1" is detected

Figure 4.2: An illustration of QIM based watermark extraction.

QIM based watermark extraction

Watermark extraction means extract the embedded bit "0" or "1" from the received watermarked signal. The watermarked signal is renamed as $\hat{s}$, and the value of $\hat{s}$ is not completely equal to the quantized value, s_0 or s_1, since it may be affected by channel noise or other turbulences.

To extract the embedded bit, we re-quantize $\hat{s}$ with two quantizers $Q_0(s,0)$ and $Q_1(s,1)$ and obtain two quantized values of $\hat{s}_0$ and $\hat{s}_1$. The extracted bit can be decided by comparing the distances between $\hat{s}_0$ and $\hat{s}$, and the distance between $\hat{s}_1$ and $\hat{s}$. The shorter distance indicates the embedded bit. These calculations can be expressed with Eqs. (4.3) to (4.5). The illustration of this process is shown in Fig. 4.2.

$$d_0 = |\hat{s} - Q_0(s,0)| = \left|\hat{s} - \left(\Delta\left[\frac{s-b_0}{\Delta}\right] + b_0\right)\right|, \quad b_0 = -\frac{\Delta}{4} \tag{4.3}$$

$$d_1 = |\hat{s} - Q_1(s,1)| = \left|\hat{s} - \left(\Delta\left[\frac{s-b_1}{\Delta}\right] + b_1\right)\right|, \quad b_1 = \frac{\Delta}{4} \tag{4.4}$$

$$w = \begin{cases} 0, & d_0 < d_1 \\ 1, & \text{otherwise} \end{cases} \tag{4.5}$$

We can find that in the watermark extraction process, only the watermarked signal and the quantization step are needed to extract the watermark bit. Therefore, QIM based

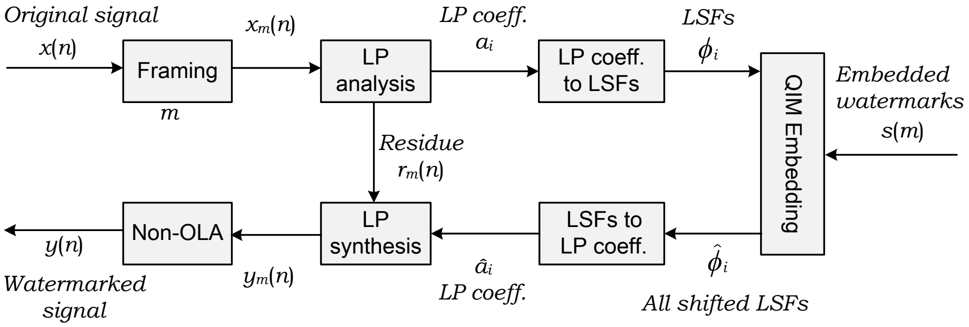


Figure 4.3: Block diagram of LSFs-QIM based watermark embedding process.

watermarking can realize blind watermark extraction.

4.1.2 Scheme of watermark embedding and extraction

In this chapter, we will construct the framework of watermarking by quantizing LSFs with QIM. This watermarking method (LSFs-QIM based watermarking) utilizes the source-filter model for speech production. The overall watermarking scheme consists of watermark embedding and extraction processes. Watermarks $w(m)$ are embedded into the LSFs of original speech signal $x(n)$ to construct the watermarked signal $y(n)$. The embedded watermarks are then extracted from watermarked signal blindly.

Watermark embedding process

Figure 4.3 has the block diagram of the watermark embedding process. Watermarks are embedding to LSFs of original speech signal as follows:

Step 1 Original signal, $x(n)$, is segmented into non-overlapping frames, and frame number is labelled as "m". The framed signal is referred as $x_m(n)$.

Step 2 Each frame is analyzed with a p-th order LP analysis to extracted the LP coefficients, a_i $(i = 1, 2, \cdots, p)$ and LP residue, $r_m(n)$.

Step 3 The LP coefficients, a_i $(i = 1, 2, \cdots, p)$ within one frame are converted to LSFs, ϕ_i $(i = 1, 2, \cdots, p)$. Converted LSFs are expressed in the angle domain (°), and all

LSFs within one frame satisfy the ordering property from 0 to π as: $0 < \phi_1 < \phi_2 < \cdots < \phi_{p-1} < \phi_p < \pi$.

Step 4 Current watermark $w(m)$ ($w(m) =$ "0" or "1") for frame $x_m(n)$ will be duplicated p times for the watermark embedding.

Step 5 All LSFs within the speech frame $x_m(n)$ are quantized with one of the DM-QIM quantizers $Q_w(\phi_i, w)$ ($w = 0$ for embedding "0" and $w = 1$ for embedding "1") as follows:

$$\hat{\phi}_i = Q_w(\phi_i, w), \quad w = 0 \text{ or } 1,\ i = 1,\ 2, \cdots, p, \tag{4.6}$$

where $Q_w(\phi_i, w)$ are defined as follows:

$$\begin{aligned} Q_0(\phi_i, 0) &= Q_0(\phi_i - b_0) + b_0 = \Delta \left[\frac{\phi_i - b_0}{\Delta}\right] + b_0, \quad b_0 = -\frac{\Delta}{4}, & (4.7)\\ Q_1(\phi_i, 1) &= Q_1(\phi_i - b_1) + b_1 = \Delta \left[\frac{\phi_i - b_1}{\Delta}\right] + b_1, \quad b_1 = \frac{\Delta}{4}. & (4.8) \end{aligned}$$

In these equations, Δ is also expressed in angle (°). After this process, all the LSFs in one frame are mapped to fixed points, $\hat{\phi}_i$ ($i = 1, 2, \cdots, p$).

Step 6 Modified LSFs, $\hat{\phi}_i$, that contain watermarks are converted back to LP coefficients, $\hat{a}_i$ ($i = 1, 2, \cdots, p$).

Step 7 Current frame $y_m(n)$ is then synthesized with the LP coefficients, $\hat{a}_i$ and the residue $r_m(n)$ which is obtained in **step 2**.

Step 8 The whole watermarked signal $y(n)$ is finally reconstructed with all watermarked frames using non-overlapping and adding function.

Watermark extraction process

Figure 4.4 illustrates the watermark extraction process, where six steps are involved. Watermarks can be extracted from $y(n)$ as follows:

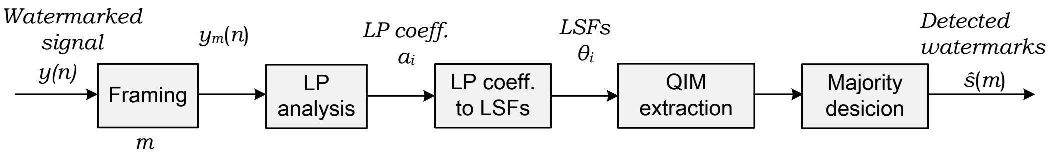


Figure 4.4: Block diagram of LSFs-QIM based watermark extraction process.

Step 1 Watermarked signal $y(n)$ is segmented into non-overlapping frames of the same size in the embedding process, and frame number is labelled as "m". The framed watermarked signal is referred as $y_m(n)$.

Step 2 The p-th LP analysis is applied to each frame to obtain LP coefficients, $\hat{a}_i$ ($i = 1, 2, \cdots, p$).

Step 3 LP coefficients are converted to LSFs, θ_i ($i = 1, 2, \cdots, p$). Since we embed the same bits in all LSFs of one frame in the embedding process, and there exists the possibility that not all the LSFs can be correctly detected. Thus, all the LSFs are associated together to determine the embedded bit for one frame with a majority decision in the following step.

Step 4 Each LSF within one frame is re-quantized with both two quantizers in Eq. (4.9).

$$\hat{\theta}_{iw} \quad = \quad Q_w(\hat{\theta}_i, w), \quad w = 0 \text{ and } 1, \; i = 1, \; 2, \cdots, p \tag{4.9}$$

where $\hat{\theta}_{iw}$ is the quantized value of $\hat{\theta}_i$. The distances between two quantized results $\hat{\theta}_{iw}$ (w =0 and 1) and $\hat{\theta}_i$ are calculated as follow:

$$d_{iw} = \hat{\theta}_{iw} - \hat{\theta}_i, \quad w = 0 \text{ and } 1, \; i = 1, \; 2, \cdots, p \tag{4.10}$$

Each LSF can indicate an embedded bit ("0" or "1") with the quantizer that provides shorter distance using Eq. (4.11).

$$\hat{w}(m)_i = \begin{cases} 0, & d_{i0} < d_{i1} \\ 1, & \text{otherwise} \end{cases}, \; i = 1, \; 2, \cdots, p \tag{4.11}$$

Table 4.1: Original LSFs and modified LSFs in one frame.

Modifications to LSFs in one frame										
No.	1	2	3	4	5	6	7	8	9	10
Original LSFs (∘)	8.21	10.08	48.80	55.52	69.90	76.64	101.11	107.53	136.30	143.59
Modified LSFs (∘)	9.00	9.00	49.00	57.00	69.00	77.00	101.00	109.00	137.00	145.00

Step 5 The final decision on the embedded bit of current frame is obtained by

$$\hat{w}(m) = \begin{cases} 0, & \sum_{i=1}^{p} \hat{w}(m)_i < p/2 \\ 1, & \text{otherwise} \end{cases} \tag{4.12}$$

Step 6 The extracted bit from each frame can construct the whole extracted watermarks, $\hat{w}(m)$.

4.1.3 Example

This section illustrates an example of the watermarking based on LSFs modifications with QIM. One speech stimulus from the ATR database (B set) [94] was used as the original signal. This stimulus had a sampling frequency of 20 kHz, and 16-bit quantization, with a duration of 8.1 second. The order of LP analysis was chosen as 10, so we could get 10 LSFs from each frame. One bit watermark "1" was embedded into one frame. This watermark was duplicated to ten times ("1111111111") and then embedded into ten LSFs with a quantization step of 4.0°. The values of original LSFs and modified LSFs are given in Tab. 4.1. The distribution of LSFs on half unit circle and the LP spectral envelope before and after DM-QIM modifications have been shown in Fig. 4.5.

Clearly, this modifications can make original LSFs shifted on half unit circle, and the positions of formants will be moved. From Fig. 4.5(b), we can also notice that the QIM based modifications to LSFs are quite unintentional, since some of LSFs have been positioned closer while some of them have been positioned further. This will easily

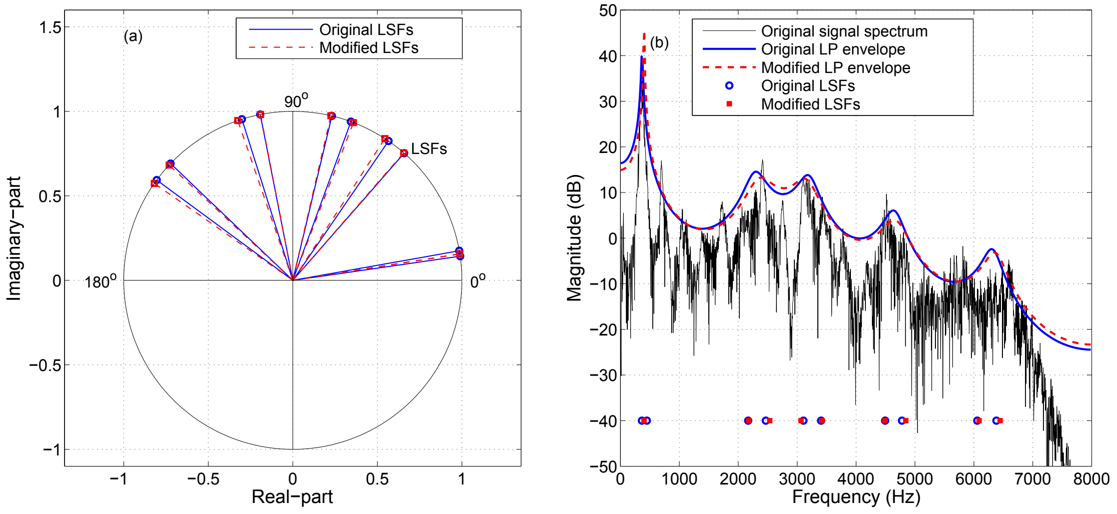


Figure 4.5: Example of (a) distribution of LSFs on half unit circle before and after modifications and (b) LP spectral envelope before and after modifications.

disrupt the original formant structure of speech signals and degrade the sound quality of watermarked signals.

In general, the performance of QIM based watermarking is characterized by the size of quantization step. A large quantization step will aggravate the distortion introduced to the digital signal, however, a stronger robustness can be obtained, and vice versa. Since robustness is increased at the expense of reduced sound quality, this may indicate that it is difficult for this method to achieve a trade-off between inaudibility and robustness.

4.2 Watermarking based on formant enhancement

Since formant enhancement can improve the sound quality of speech, modifications introduced by formant enhancement may not cause perceptual distortion to the original speech. Therefore, watermarking based on formant enhancement is possible to be imperceptible to human to realize inaudibility. In this chapter, we take advantage of formant enhancement to achieve inaudibility for watermarking. Watermarks will be embedded through formant enhancement.

In most speech synthesis methods [99, 100, 102], formants are enhanced with complicated methods so that the dynamics between formant peaks and spectral valleys can be increased. However, such complicated methods are not suitable for watermark embedding and robust watermark extraction. As to inherited formant enhancement for watermarking, we investigate a simple but effective formant enhancement method. The following subsections will separately talk about how the formant can be enhanced and then applied for formant enhancement based watermarking.

4.2.1 Formant enhancement by controlling LSFs

The positions of LSFs on half unit circle can reflect the formants of speech: the closer two LSFs are, the sharper the formant is. Therefore, formant can be effectively enhanced by directly closing up two LSFs. Figure 4.6 illustrates how this idea can be implemented. In Fig. 4.6, original formant (dotted curve) is produced by a pair of LSFs, ϕ_l and ϕ_r. Its sharpness can be mathematically measured by the tuning level, that is Q-value defined in Eq. (4.13), where f_c is the center frequency of formant, BW is the bandwidth between f_l and f_r that converted from ϕ_l and ϕ_r with Eq. (4.14), in which F_s is the sampling frequency of signal.

$$\begin{aligned} Q &= \frac{f_c}{BW} = \frac{f_c}{f_r - f_l} && (4.13) \\ f_r &= \frac{\phi_r}{2\pi} \times F_s \quad \text{and} \quad f_l = \frac{\phi_l}{2\pi} \times F_s && (4.14) \end{aligned}$$

To enhance this formant, as seen in Fig. 4.6, two LSFs ϕ_l and ϕ_r are symmetrically shifted to be closer to each other, that is ϕ_l to ϕ_{lw} and ϕ_r to ϕ_{rw}. This process can be expressed with Eq. (4.15), where Δ is used to control the degree of shift, a bigger Δ indicates a more severe shift of LSFs as well as a much enhanced formant.

$$\phi_{lw} = \phi_l + \Delta \quad \text{and} \quad \phi_{rw} = \phi_r - \Delta, \quad 0 < \Delta < (\phi_r - \phi_l)/2 \qquad (4.15)$$

After obtaining two shifted LSFs ϕ_{lw} and ϕ_{rw}, a narrower bandwidth BW_{ew} is produced. According to Eq. (4.16), the tuning level of original formant has been increased to Q_{ew},

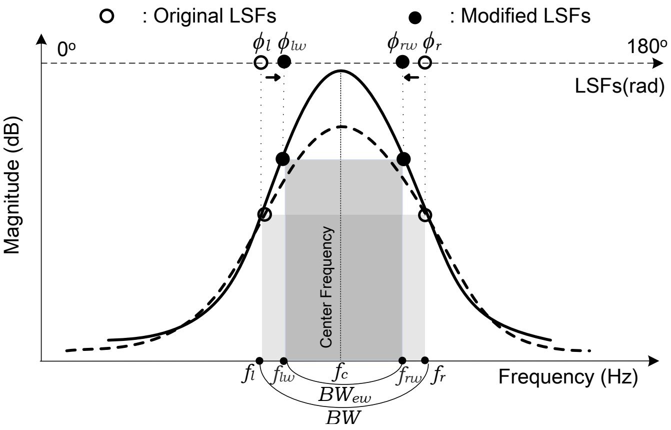


Figure 4.6: Formant enhancement by controlling a pair of LSFs.

and the enhanced formant (solid curve in Fig. 4.6) has become much sharper.

$$Q_{ew} = \frac{f_c}{BW_{ew}} = \frac{f_c}{f_{rw} - f_{lw}} \tag{4.16}$$

where f_{lw} and f_{rw} are calculated as follows:

$$f_{rw} = \frac{\phi_{rw}}{2\pi} \times F_s \quad \text{and} \quad f_{lw} = \frac{\phi_{lw}}{2\pi} \times F_s \tag{4.17}$$

Note that in the above manipulation, two LSFs are symmetrically shifted, so there is no deviation between the center frequency of the original formant and the enhanced formant which furthest maintains the sound quality of the original signal.

4.2.2 Formant enhancement for watermarking

Preliminary analysis

We employ the concept of formant enhancement for watermarking. Watermarks can be embedded into the original signal when LSFs are shifted for formant enhancement. Before embedding, several issues should be clarified to make the watermarking method effective.

(i) Selection of the suitable formant for enhancement. Several formants can be estimated from the speech segment in each frame, we should select the suitable formant for enhancement. As we have surveyed, the distortion caused by enhancing formants in the lower and higher frequencies can be easily perceived by human, we thus leave the first formant and last formant unmodified. Only one formant in the middle region will be enhanced for watermark embedding.

(*ii*) Embedding and blind extraction mechanism. Formants in each frame can be consecutively indexed with F1, F2, F3, $\cdots$, from the low frequency to high frequency. For different frames, if the watermarks are embedded into the same indexed formants, it will be easy for the attackers to destroy them with simple rule. As to well hide watermarks, the formant for embedding will be randomly selected from each frame according to watermark "0" or "1". Moreover, since formant structures vary widely with different speech frames, it is preferable to enhance the selected formants according to their original tuning characteristics (self-adaptive enhancement) to achieve inaudibility.

However, the above embedding mechanism concerning random formant selection and self-adaptive enhancement results in a serious problem for blind watermark extraction since it is so difficult to detect watermarks just relying on the irregular formant structure extracted from the watermarked signal when any prior knowledge about which formant has been enhanced and how it has been enhanced is not available. As we have considered, one solution for both inaudibility and blind extraction is we can enhance the selected formant and hence to establish an internal relationship between the enhanced formant and another formant in current frame, where the relationship is used to reflect the position of enhanced formant and how the formant is enhanced. In extraction process, two formants can make a cross-reference. Watermarks can be extracted by identifying the relationship.

Embedding concept

The chapter talks about how to embed watermark and extract watermark for one speech frame by formant enhancement. For each frame, one bit watermark will be embedded. "0" is embedded by enhancing the sharpest formant and "1" is embedded by enhancing the second sharpest formant. Since the closer two LSFs are, the sharper the formant is, these two formants can be easily found from the speech frame by checking the bandwidths of each formant and selected two smallest ones. The reason why these two formants are selected to carry the relationships of watermarks will be explained later. Note that the

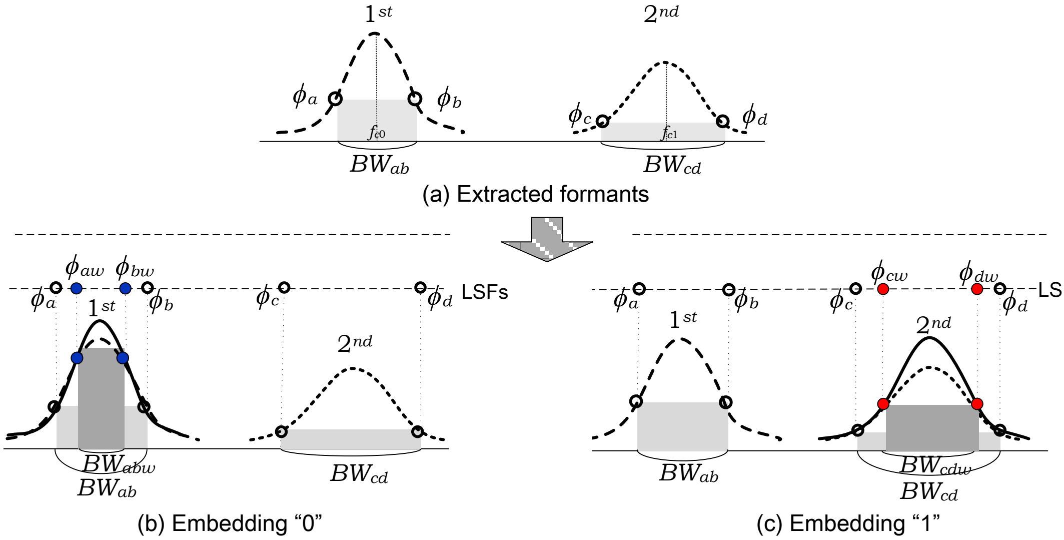


Figure 4.7: Concept of watermark embedding: (a) extracted two formants (b) embedding "0", and (c) embedding "1".

first formant in low frequency and the last formant in high frequency are not involved in the selection process, since enhancing them will drastically distort the sound quality of the original signal.

Figure 4.7(a) illustrates the extracted two formants, where the sharpest formant (labelled as 1^{st}) is produced by ϕ_a and ϕ_b and the second sharpest formant (labelled as 2^{nd}) is produced by ϕ_c and ϕ_d. According to previous chapter, the sharpest formant has the smallest bandwidth BW_{ab} and its tuning level is $Q_0 = f_{c0}/BW_{ab}$, the second sharpest formant has the second smallest bandwidth BW_{cd}, and its tuning level is $Q_1 = f_{c1}/BW_{cd}$. The rules for embedding are as follow:

A. Rule of embedding "0": To embed "0", as seen in Fig. 4.7(b), the sharpest formant will be enhanced. An enhancing factor Ω_{e0} ($\Omega_{e0} > 1$) in Eq. (4.18) is used to control how much the formant is enhanced. According to Eq. (4.18), BW_{ab} has to be reduced to its $1/\Omega_{e0}$ for the enhancement, that is the newly obtained bandwidth BW_{abw} equals BW_{ab}/Ω_{e0}. To achieve this, original LSFs ϕ_a and ϕ_b will be shifted to ϕ_{aw} and ϕ_{bw} with the modification degree Δ_{e0} in Eq. (4.19), where Δ_{e0} is calculated by ϕ_a, ϕ_b, and Ω_{e0} with Eq. (4.20). Since BW_{cd} is originally bigger than BW_{ab}, after enhancing the sharpest formant, an updated relationship, $BW_{cd} > BW_{abw} \times \Omega_{e0}$, has been established in the current frame.

$$\begin{aligned} Q_0 \times \Omega_{e0} &= \frac{f_{c0}}{BW_{ab}} \times \Omega_{e0} = \frac{f_{c0}}{BW_{ab}/\Omega_{e0}} = \frac{f_{c0}}{BW_{abw}}, \Omega_{e0} > 1 && (4.18) \\ \phi_{aw} &= \phi_a + \Delta_{e0} \quad \text{and} \quad \phi_{bw} = \phi_b - \Delta_{e0} && (4.19) \\ \Delta_{e0} &= \frac{1}{2}\left[(\phi_b - \phi_a) \times \left(1 - \frac{1}{\Omega_{e0}}\right)\right] && (4.20) \end{aligned}$$

B. Rule of embedding “1”: To embed “1”, as seen in Fig. 4.7(c), the second sharpest formant will be enhanced. An enhancing factor Ω_{e1} ($\Omega_{e1} = \frac{BW_{cd}}{BW_{ab}}$) in Eq. (10) is used for the enhancement. With this factor, BW_{cd} will be reduced to the same as BW_{ab}. This is achieved by shifting ϕ_c and ϕ_d to ϕ_{cw} and ϕ_{dw} with Eq. (11), where Δ_{e1} is calculated by ϕ_c, ϕ_d and Ω_{e1} with (4.23). Therefore, after embedding “1”, the bandwidth relationship $BW_{cdw} = BW_{ab}$ has been established in the current frame.

$$\begin{aligned} Q_1 \times \Omega_{e1} &= \frac{f_{c1}}{BW_{cd}} \times \Omega_{e1} = \frac{f_{c1}}{BW_{cd}/\Omega_{e1}} = \frac{f_{c1}}{BW_{ab}} = \frac{f_{c1}}{BW_{cdw}}, \Omega_{e1} = \frac{BW_{cd}}{BW_{ab}} && (4.21) \\ \phi_{cw} &= \phi_c + \Delta_{e1} \quad \text{and} \quad \phi_{dw} = \phi_d - \Delta_{e1} && (4.22) \\ \Delta_{e1} &= \frac{1}{2}\left[(\phi_d - \phi_c) \times \left(1 - \frac{1}{\Omega_{e1}}\right)\right] && (4.23) \end{aligned}$$

In summary, different watermarks are embedded by establishing different bandwidth relationships between the sharpest and the second sharpest formants via formant enhancement. The different bandwidth relationships enable watermarks to be blindly extracted. Note that this watermarking method can be applied for both voiced/unvoiced speech frames, while the formants extracted from unvoiced speech segment are just pseudo-formants. Especially, when all the samples in speech frame are completely 0, the samples will be added very tiny values (like white noise) to enable the LP analysis work.

Extraction concept

According to the embedding rules, bandwidth relationships always exist in the sharpest and the second sharpest formants no matter for embedding “0” or “1”. Therefore, in extraction process, for each frame of watermarked signal, we extract these two formants

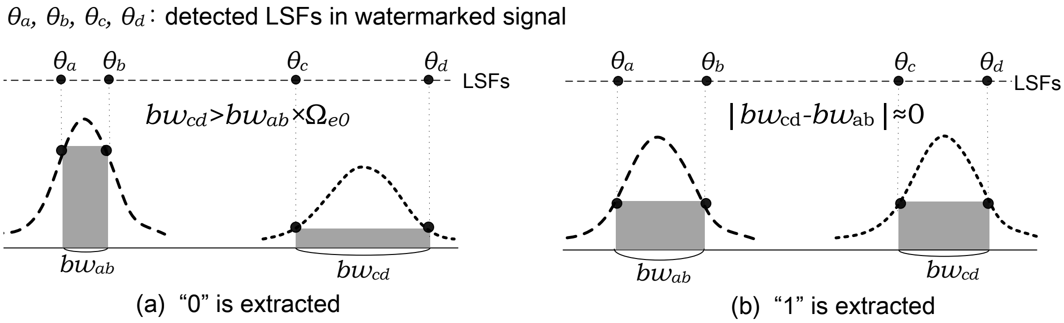


Figure 4.8: Concept of watermark extraction: (a) "0" is extracted and (b) "1" is extracted.

respectively. As seen in Fig. 4.8, the sharpest formant should have the smallest bandwidth, we name it as bw_{ab} (produced by θ_a and θ_b). The second sharpest formant should have the second smallest bandwidth, we name it as bw_{cd} (produced by θ_c and θ_d). If "0" has been embedded, according to Fig. 4.8(a), the relationship between bw_{ab} and bw_{cd} should be $bw_{cd} > bw_{ab} \times \Omega_{e0}$, an equivalent representation is given in Eq. (4.24); if "1" has been embedded, bw_{cd} in Fig. 4.8(b) should be similar to bw_{ab}, an equivalent representation is given in Eq. (4.25). Since LP analysis calculates LP coefficients (or LSFs) based on the criterion that the mean-squared error is always minimized, the LP coefficients (or LSFs) that are derived from watermarked frame are not exactly the same as those after embedding process even there is no modifications. Therefore, as shown in Eq. (4.26), we set a threshold (half of the difference between two extracted bandwidths) to discriminate two cases of embedding "0" or "1", and enable the method to be error-tolerant.

$$\text{embedding "0":} \quad bw_{cd} - bw_{ab} > bw_{ab} \times (\Omega_{e0} - 1) \tag{4.24}$$

$$\text{embedding "1":} \quad bw_{cd} - bw_{ab} \approx 0 \tag{4.25}$$

$$\hat{s}(m) = \begin{cases} 0, & bw_{cd} - bw_{ab} > bw_{ab} \times \frac{\Omega_{e0}-1}{2} \\ 1, & \text{otherwise} \end{cases} \tag{4.26}$$

Embedding and extraction analysis

Now we discuss why the sharpest and the second sharpest formants are selected to carry the relationship for watermarks. For the example in Fig. 4.9, three sharpest formant that

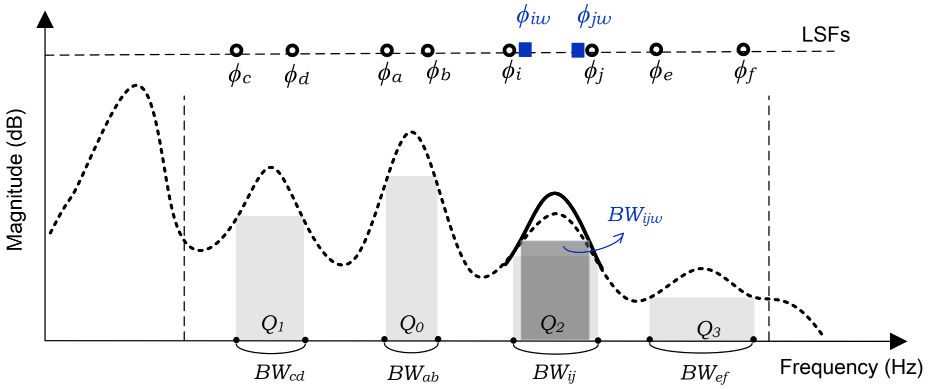


Figure 4.9: Problem when enhancing a smooth formant for watermarking.

labelled as Q_0 (the sharpest formant), Q_1 (the second sharpest formant), and Q_2 (the third sharpest formant) originally follow the bandwidth relationship that $BW_{ij}>BW_{cd}>BW_{ab}$. Consider one case that Q_0 and Q_2 labelled formants are selected for watermark embedding. To embed "1", BW_{ij} will be made to the same as BW_{ab} for formant enhancement. Since $BW_{ij}>BW_{cd}>BW_{ab}$, the modification to BW_{ij} will be severer in comparison with enhancing the Q_1 labelled formant. Therefore, sound quality will be much degraded. Alternatively, if we slightly reduce BW_{ij} to embed "1" and if BW_{ijw} in Fig. 4.9 is still larger than BW_{cd} after enhancement, it will be difficult or even impossible to recognize bandwidth relationship for watermark extraction. Although this phenomenon can be alleviated by setting bandwidth bounds for extraction, formant enhancement in embedding process, however, will be much hampered and complicated.

In comparison, establish bandwidth relationships in the sharpest and the second sharpest formant can effectively avoid the above problem. This is because these two formants always possess two smallest bandwidths no matter before or after watermarking, so the bandwidth relationships in the extraction process can be extracted for watermark extraction without any ambiguity. Besides, the distortion introduced by formant enhancement in this case can be minimized compared with enhancing other formants.

In addition, in the embedding process, the enhanced formant is selected according to the frequency characteristics of each frame and the watermark "0" or "1", therefore, the enhanced formant is possible to exist in any frequency range, which enables the watermarks to be well hidden. Moreover, since watermarks are embedded into the intrinsically irregular formant structures, it is difficult for the attackers or the third party to confirm whether the formant structure was formed by artificial manipulation, since the embedded bandwidth relationship is also possible in a rough speech. Especially when the LP order for estimating formants is unknown, bandwidth relationship is unable to discover.

4.2.3 Scheme of watermark embedding and extraction

In this chapter, we will construct the framework of watermarking by enhancing formants. The overall scheme consists of watermark embedding and watermark extraction processes. Watermarks $w(m)$ are embedded into the LSFs of original speech signal $x(n)$ to construct the watermarked signal $y(n)$. The embedded watermarks are then extracted from the watermarked signal blindly.

Watermark embedding process

Figure 4.10 has a block diagram of embedding process. Watermarks are embedded as follows.

Step 1 Original signal, $x(n)$, is segmented into non-overlapping frames, and frame number is labelled as "m". The framed signal is referred as $x_m(n)$.

Step 2 Each frame is analyzed with a p-th order LP analysis to extracted the LP coefficients, a_i $(i = 1, 2, \cdots, p)$ and LP residue, $r_m(n)$.

Step 3 The LP coefficients, a_i $(i = 1, 2, \cdots, p)$ within one frame are converted to LSFs, ϕ_i $(i = 1, 2, \cdots, p)$.

Step 4 Each frame will be embedded with one bit watermark "0" or "1" (according to $s(m)$), after which, a pair of shifted LSFs (ϕ_{aw} and ϕ_{bw} for embedding "0", or ϕ_{cw} and ϕ_{dw} for embedding "1") are generated.

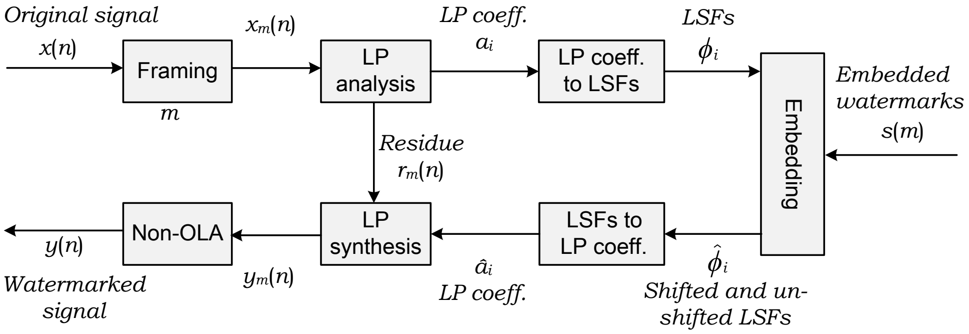


Figure 4.10: Block diagram of of formant enhancement based watermark embedding process.

Step 5 All LSFs $\hat{\phi}_i$ including the shifted LSFs and the other un-shifted LSFs will be converted back to LP coefficients $\hat{a}_i$ $(i = 1, 2, \cdots, p)$.

Step 5 Current frame $y_m(n)$ is then synthesized with the LP coefficients, $\hat{a}_i$ and the residue $r_m(n)$.

Step 8 The whole watermarked signal $y(n)$ is finally reconstructed with all watermarked frames using non-overlapping and adding function.

Watermark extraction process

The watermark extraction process is illustrated in Fig. 4.11, watermarks are extracted as follows:

Step 1 Watermarked signal $y(n)$ is segmented into non-overlapping frames of the same size in the embedding process, and frame number is labelled as "m". The framed watermarked signal is referred as $y_m(n)$.

Step 2 The p-th LP analysis is applied to each frame to obtain LP coefficients, $\hat{a}_i$ $(i = 1, 2, \cdots, p)$.

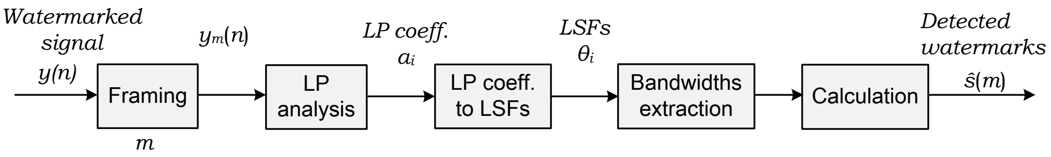


Figure 4.11: Block diagram of of formant enhancement based watermark extraction.

Step 3 LP coefficients are converted to LSFs, θ_i $(i = 1, 2, \cdots, p)$, and two smallest bandwidths are then extracted for representing the sharpest formant and the second sharpest formant.

Step 4 The watermark in one frame is extracted with the method introduced in chapter 4.2.2.

Step 5 The extracted bit from each frame can construct the whole extracted watermarks, $\hat{w}(m)$.

4.3 Frame synchronization

In our proposed two methods, the watermark extraction process works when the frame positions are known. In practice, the speech signal may be cropped or incomplete, in this case, we have to automatically find the starting point of each frame. Therefore, an automatic frame synchronization scheme should be implemented. Frame synchronization indicates that without knowing any information about the frame size and the total frame numbers, all the frames can be correctly segmented.

We have implemented the frame synchronization scheme for the proposed two methods. The frame synchronization is separated with the watermark embedding and extraction processes. As shown in Fig. 4.12, after the watermark embedding process, the synchronization information will be embedded into the watermarked signal. In our method, a random 0-1 sequence of length 20 is used as the synchronization information. This information is embedded to the first 20 speech samples of each frame of the watermarked

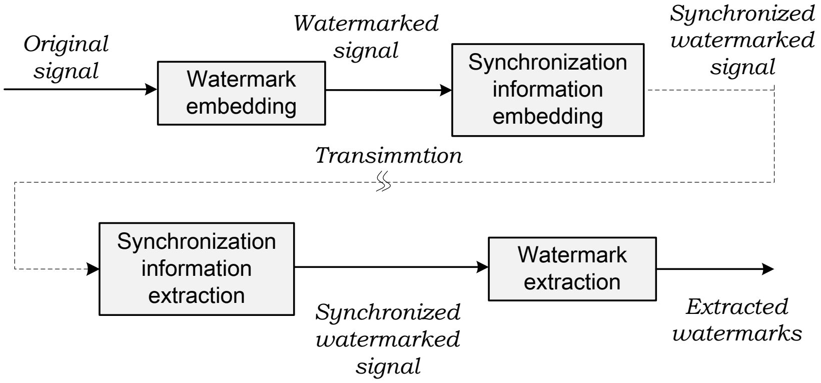
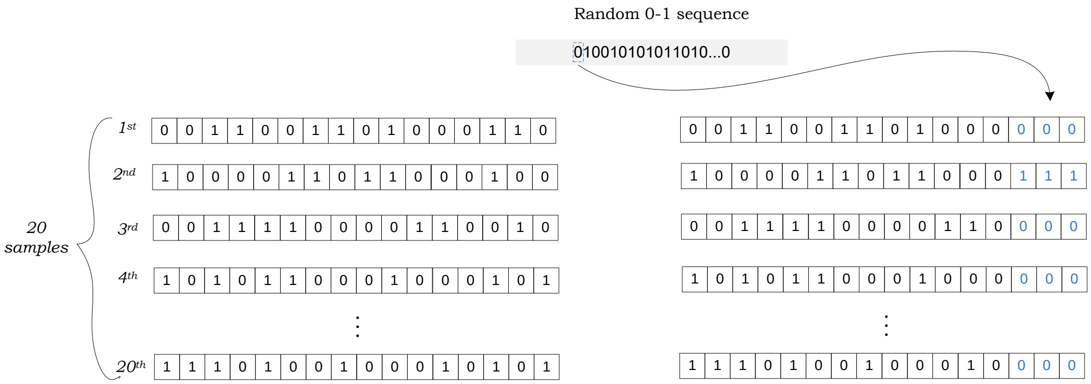

Original signal
Watermark embedding
Watermarked signal
Synchronization information embedding
Synchronized watermarked signal
Transimmtion
Synchronization information extraction
Synchronized watermarked signal
Watermark extraction
Extracted watermarks

Figure 4.12: Frame synchronization scheme.



1st	0 0 1 1 0 0 1 1 0 1 0 0 0 1 1 0	0 0 1 1 0 0 1 1 0 1 0 0 0 0 0 0
2nd	1 0 0 0 0 1 1 0 1 1 0 0 0 1 0 0	1 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1
3rd	0 0 1 1 1 1 0 0 0 0 1 1 0 0 1 0	0 0 1 1 1 1 0 0 0 0 1 1 0 0 0 0
4th	1 0 1 0 1 1 0 0 0 1 0 0 0 1 0 1	1 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0
⋮	⋮	⋮
20th	1 1 1 0 1 0 0 1 0 0 0 1 0 1 0 1	1 1 1 0 1 0 0 1 0 0 0 1 0 0 0 0

Figure 4.13: Embedding the 0-1 sequence for frame synchronization.

signal. The detailed embedding process is as follows. Firstly, 20 speech samples will be expressed with 16-bit binary code. Then we got 20 16-bit binary codes. Secondly, the last three bits of the 20 binary codes will be sequently replaced the 0-1 sequence, as shown in Fig. 4.13. Finally, the 20 binary codes will be expressed in decimal to express the watermarked speech.

The watermarked signal can be transmitted, and it may be cut of some samples and incomplete. Before watermark extraction, all the speech frames can be segmented with the embedded synchronization information. As the pattern (0-1 random sequence) which indicates the beginning of each frame, is embedded into the watermarked signal, by applying the correlation technique between the received signal and the 0-1 random sequence, the beginning of each frame can be found. Figure 4.14 shows an example for frame syn-

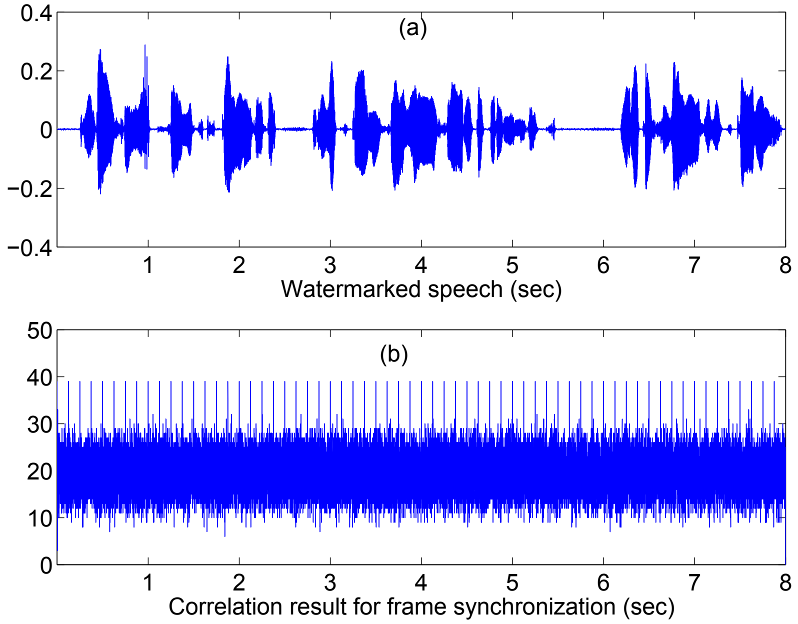


Figure 4.14: Result of frame synchronization.

chronization. As shown in Fig.4.14(a), the speech watermarked signal is 8.0 second, and we set the frame size is 0.125s, so there would be 64 frames. In Fig. 4.14(b), after using correlation, there are totally 64 peaks can be observed. These peaks correspond to the beginning of frames. The watermarks can be extracted separately for each segmented frame.

The frame synchronization does not consume much time, because only the correlation is used. In this method, the 0-1 sequence replaces the last three bits of the 20 samples of each frame for reliable frame segmentation. Actually, replace the last one bit of the speech samples is enough for frame segmentation. Furthermore, embedding the 0-1 sequence to the last several bits of 20 samples of each speech frame will not greatly degrade the speech quality.

4.4 Summary

In this chapter, we have separately implement two speech watermarking methods. In the LSFs and QIM based watermarking, the basic knowledge of DM-QIM which can provide two quantizers to embed different watermarks bits “0” and “1” is introduced. Then, we construct the framework of watermarking by quantizing LSFs with QIM. The overall watermarking scheme consists of watermark embedding and watermark extraction. An example of the watermarking based on LSFs modifications with QIM is illustrated. From this example, the influences to the LP spectral envelope and the formants by modifying LSFs with QIM can be observed.

In the formant enhancement based watermarking, we firstly introduce the ides of enhancing the formant by directly closing up two LSFs. This idea is then employed for the watermarking embedding. Several issues are clarified before watermark embedding to make the method effective, these include (i) how to select the most suitable formant for enhancement and (ii) how to realize the blind watermark extraction. To solve these problems, different watermarks are embedded by enhancing different formants, after which different bandwidth relationships between the sharpest and the second sharpest formants are established. These different bandwidth relationships can be used to blindly detect watermarks in the extraction process. Finally, we construct the whole scheme of watermark embedding and extraction.

Additionally, an automatic frame synchronization scheme is implemented. With this scheme, the frames of speech can be automatically segmented. Moreover, this scheme is designed beyond the watermarking methods, it is applicable to any watermarking methods.

Chapter 5

Evaluations of proposed methods

The purpose of this research is to protect speech signals with watermarking methods. In the previous chapter, we have implemented two speech watermarking methods. In this chapter, we will evaluate these two methods with respect to inaudibility and robustness (both of the two methods are blind methods).

5.1 Database and conditions

We conducted several experiments with respect to inaudibility and robustness to evaluate the proposed methods. Twelve speech stimuli (Japanese sentences, uttered by six males and six females) in the ATR speech database (B set) [94] were used as the original speech signals. All stimuli were clipped into 8.1 second duration, sampled at 20 kHz, and quantized with 16 bits. All of the evaluations were conducted on Linux operating system with kernel 3.4.87-2vl6. The CPU is Intel (R) Core (TM) i7-4771 with frequency of 3.50 GHz, and the memory is 15.6 GB.

5.1.1 Measurements for inaudibility

Inaudibility can be checked by objective and subjective tests. The log spectrum distortion (LSD) [105] and the perceptual evaluation of speech quality (PESQ) [106] are objective measures. They can estimate the degradation between the original speech and the watermarked speech.

LSD defined in Eq. (5.1) can measure the spectral distance between the original speech and the watermarked speech, where m indicates the frame index, M is the total numbers of frames, $X(\omega, m)$ and $Y(\omega, m)$ are the spectra of m-th frame in the original speech and the watermarked speech, respectively. LSD of 1.0 dB is chose as the criterion, and a lower value indicates a less distortion.

$$\mathrm{LSD} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(10\mathrm{log}_{10}\frac{|Y(\omega,m)|^2}{|X(\omega,m)|^2}\right)^2} \quad (\mathrm{dB}) \tag{5.1}$$

PESQ that recommended by ITU-T recommendation P.862 evaluates the speech quality with Objective Difference Grades (ODG) that range from -0.5 (very annoying) to 4.5 (imperceptible). ODG of 3.0 (slightly annoying) was set as the criterion, and a higher value indicates a better speech quality.

5.1.2 Measurements for robustness

Watermarking method should be robust against allowable speech processing (e.g., speech codecs ad general speech processing, such as re-sampling and re-quantization) to guarantee the effectiveness of the embedded watermarks. Robustness can be indicated by Bit Detection Rate (BDR), i.e., the ratio between correctly extracted watermarks and all embedded watermarks. The BDR can be calculated with Eq. (5.2), where $s(m)$ represents embedded watermarks, $\hat{s}(m)$ is the detected watermarks, and M is the total length of $s(m)$. The symbol "$\oplus$" denotes the operation of "exclusive-OR", that is, if the bit values of $s(m)$ and $\hat{s}(m)$ are different ($s(m) = 1$ and $\hat{s}(m) = 0$, or $s(m) = 0$ and $\hat{s}(m) = 1$), "$s(m) \oplus \hat{s}(m)$" equals 1; otherwise, "$s(m) \oplus \hat{s}(m)$" equals 0. We chose BDR of 90% as the criterion, and a higher BDR indicates a stronger robustness.

$$\mathrm{BDR} = \frac{M - \sum\limits_{m=0}^{M} s(m) \oplus \hat{s}(m)}{M} \times 100 \quad (\%) \tag{5.2}$$

5.2 Evaluations for LSFs and QIM based watermarking

In this chapter, we evaluate the LSFs and QIM based watermarking method with respect to inaudibility and robustness. The embedded information was "JAIST-IS". The LP order was 10-th. For extended use of watermarking as information hiding method, we evaluated the performance of the proposed method as a function of embedding bit rate. The bit rates were ranged from 4, 8, 16, 32, 64, 128, 256, 512 and 1024 (Here, one effective bit information ("0" or "1") was embedded into one frame. If 1 second original signal were segmented into four frames, then the bit rate was 4 bps). Quantization steps of QIM were adopted as 0.5°, 1.0°, and 3.0°.

5.2.1 Evaluations for inaudibility

Figure 5.1 plots the results for the evaluations of inaudibility. These results were calculated on the average of twelve stimuli. The straight blue dashed-lines in each sub-figure indicated the criteria for LSD ($\leq$ 1dB) and PESQ ($\geq$ 3.0). As we can see from Fig. 5.1 (a), sound quality got worse when bit rate increased, nevertheless, for all bit rates from 4 bps to 1024 bps, watermarked signals with all quantization steps could satisfy the criterion of LSD. In Fig. 5.1(b), only results for the quantization steps of 0.5° and 1.0° reached the criterion of PESQ. This indicated smaller quantization steps could lead to better sound quality of watermarked signals.

Especially, in this figure, an additional result labelled as "ResynOrg" was also given out. This result was calculated between the original signal and the resynthesized original signal (LP analysis/synthesis of original signal without watermarking) for checking whether sound distortion could be caused by speech analysis/synthesis in spite of the embedding of watermarks. Based on the obtained results, the resynthesized signal had almost the same sound quality as the original signal, which suggested sound distortion caused by speech analysis/synthesis was imperceptible.

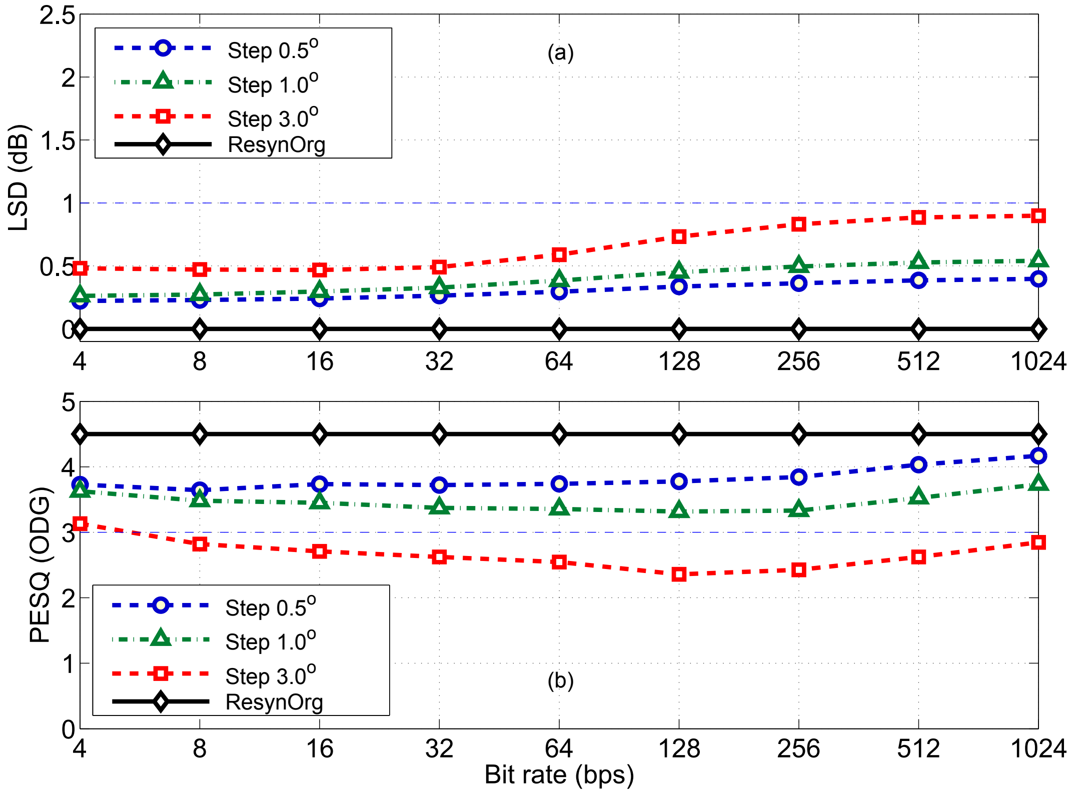


Figure 5.1: Evaluations of inaudibility for LSFs and QIM based watermarking: (a) LSD and (b) PESQ.

5.2.2 Evaluations for robustness

Watermarking method should be robust against allowable speech processing to guarantee the effectiveness of the embedded watermarks. In this section, the robustness of proposed method was evaluated from two aspects: (a) robustness against different speech codecs, and (b) robustness against general speech processing.

Evaluations against speech codecs

Speech codec is a kind of necessary processing for speech transmission over the Internet and telecommunication systems. Currently, there are many types of speech codecs to compress speech data for transmission, and watermarking methods should be robust against these different speech codecs. Speech codecs can generally be classified into waveform-

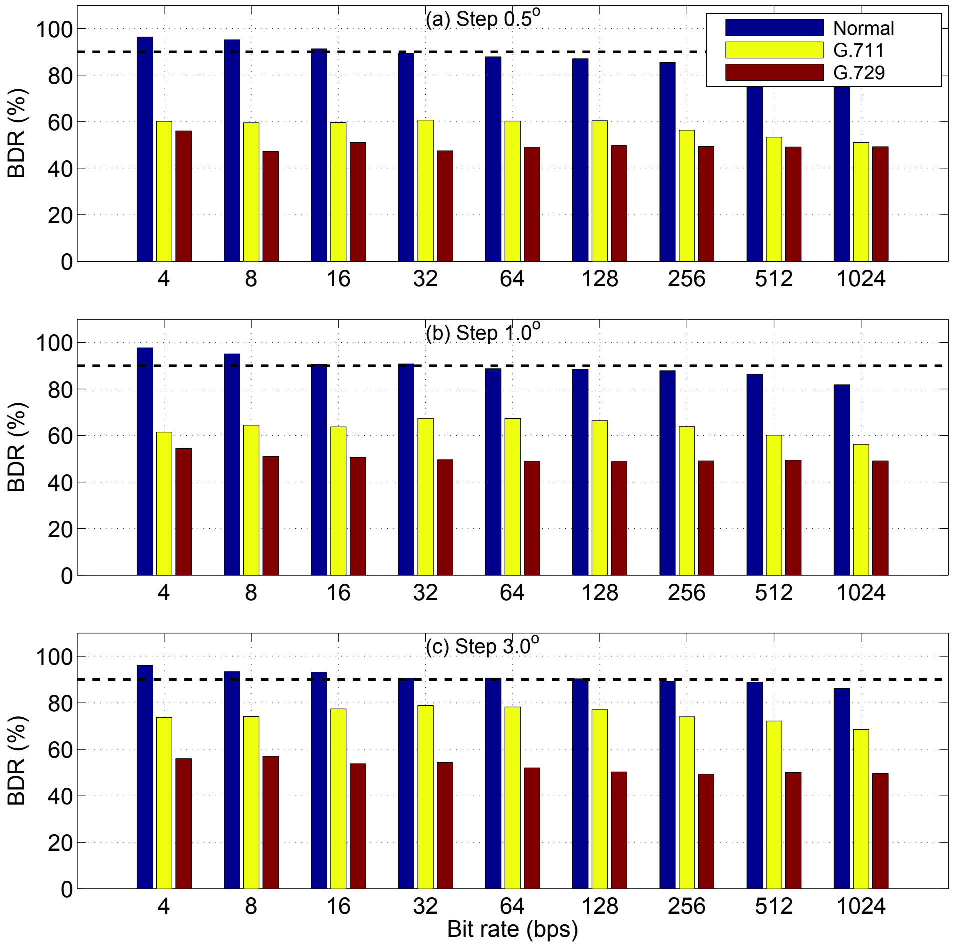


Figure 5.2: Evaluations of robustness for LSFs and QIM based watermarking against speech codecs.

based and parameter-based schemes. Accordingly, we applied two typical speech codecs of G.711 (waveform-based, PCM) and G.729 (parameter-based, CELP) to the watermarked signals to evaluate the robustness of the proposed method.

The BDR results for normal detection and detection after G.711 and G.729 are presented in Fig. 5.2. The straight dashed line in each sub-figure indicated the criteria for BDR ($\geq 90\%$). It is shown that, for normal detection, the proposed had good bit detection rate for bit rate increased from 4 to 512 bps with all quantization steps. In contrast, BDR results deteriorated a lot after watermarked signals have been processed by G.711, especially for the small quantization step. This indicated that larger quantization could

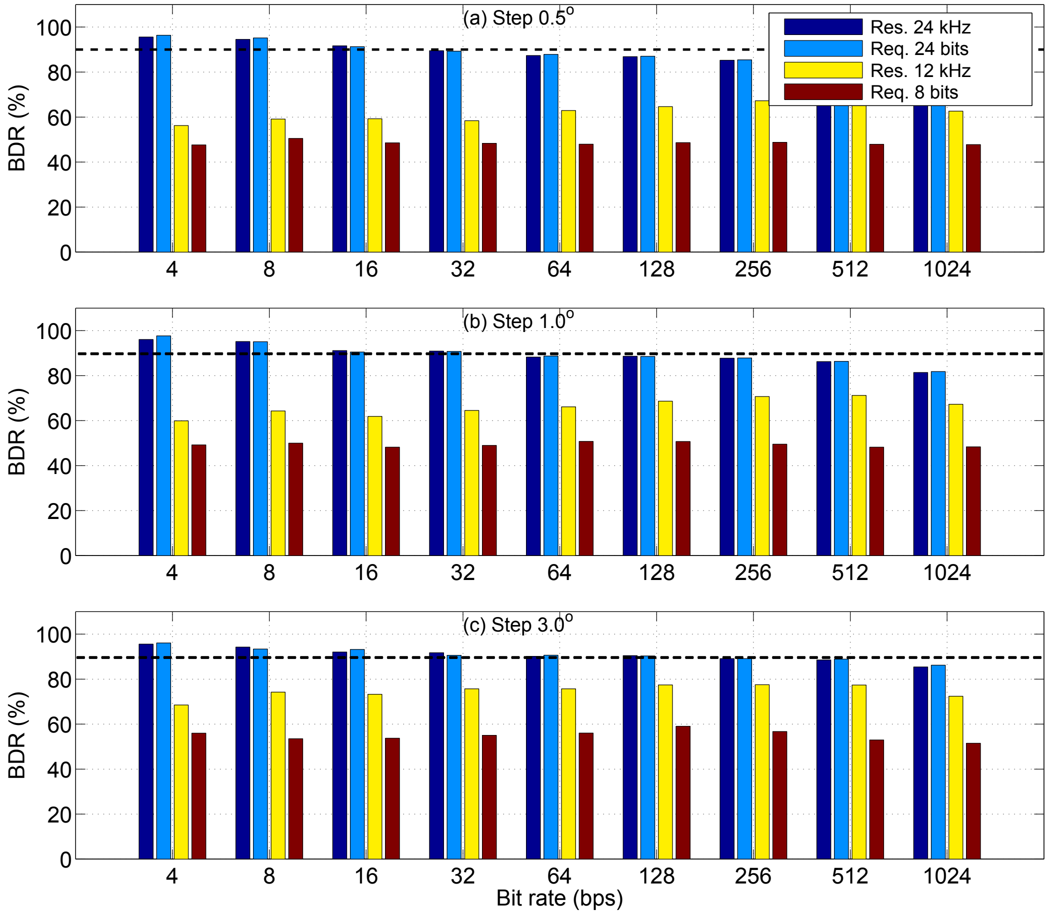


Figure 5.3: Evaluations of robustness for LSFs and QIM based watermarking against re-sampling and re-quantization.

lead to stronger robustness. However, based on the BDR results after G.729, we can basically conclude that this method was not robust against this speech codec, although LSFs is a kind of speech parameter employed by this speech codec.

Evaluations against general speech processing

We also evaluated the proposed method against several speech processing [45]. These included re-sampling at 24 kHz and 12 kHz, re-quantization with 24 bits and 8 bits. The BDR results for these evaluations are presented in Fig. 5.3. The straight dashed line in each sub-figure indicated the criteria for BDR ($\geq 90\%$). From this figure, we can see that for re-sampling at 24 kHz and re-quantization at 24-bit, the proposed method was robust. However, for down-sampling at 12 kHz, and low-bit quantization at 8-bit, the embedded watermarks could not be correctly extracted, i.e., the proposed method was not robust. This was because once the speech signal was processed with these operations, the shape of waveform changed. Correspondingly, we could not accurately estimate the modified LSFs, and extract watermarks.

5.2.3 Discussion

In this chapter, we evaluated the watermarking method by carrying out two kinds of objective evaluations with respect to inaudibility (LSD and PESQ) and robustness against different speech codecs and general processing. We took into consideration of the inaudibility and robustness that were influenced by minor modifications to LSFs with different quantization steps. The results from inaudibility evaluation revealed that the proposed method could satisfy inaudibility when quantization step was small. The results from robustness evaluation revealed that the proposed method had good bit detection rate for normal detection and some of general speech processing. However, the weak robustness of current method against speech codecs, down-sampling, and low-bit re-quantization greatly restricted its performance and effectiveness since it was very sensitive to modifications that could change the shape of waveform or the value of signal.

5.3 Evaluations for formant enhancement based watermarking

In this chapter, we conducted several experiments with respect to inaudibility and robustness to evaluate the formant enhancement based watermarking. In comparison to LSFs and QIM based watermarking, formant enhancement based watermarking in which the original formant structure of speech signal is considered to embed watermark, should be able to achieve better performance in inaudibility and robustness. The same twelve

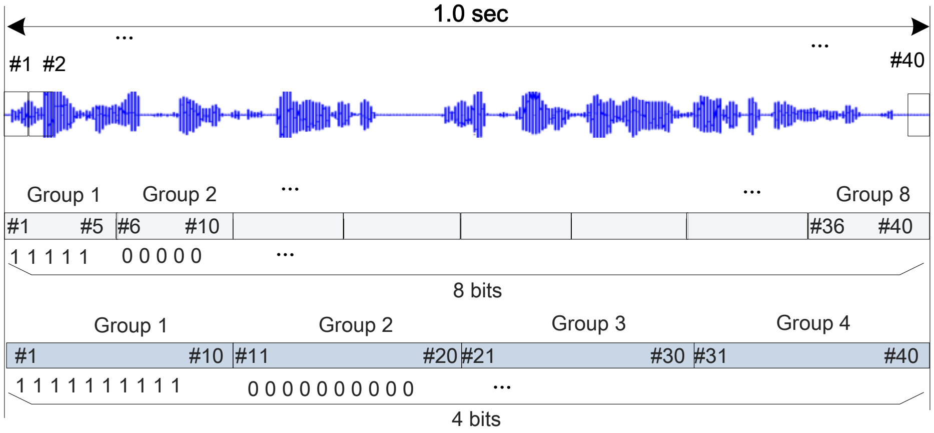


Figure 5.4: Group separation for bit rates of 4 bps and 8 bps.

speech stimuli in the ATR speech database (B set) were used as the original signals. The embedded watermarks was "JAIST-IS-Acoustic". Since our method is based on speech analysis/synthesis, the frame size was fixed at 25 ms (40 frames in 1.0 second) to attain better sound quality. The performance of the proposed method as a function of bit rate. To construct the bit rates, all frames within 1.0 second speech segment were separately divided into 4, 8, 20, and 40 groups. Frames within the same group were embedded with the same watermark and then extracted the watermark with a majority decision. Thus, the bit rates for the proposed method were 4, 8, 20, and 40 bps. An example of frame separation at 4 bps and 8 bps is shown in Fig. 5.4.

5.3.1 Parameter analysis

In the proposed method, two adjustable parameters, i.e., LP order and Ω_{e0} for embedding "0" affect the performance of inaudibility and robustness (Ω_{e1} for embedding "1" is automatically fixed according to bandwidth characteristics of each frame). These two parameters should be optimized for the proposed method.

LP analysis The order of LP analysis is important to determine the characteristics of formant structure. High LP order is beneficial to follow the details of spectrum contour, and more finer formants can be estimated under high LP order. Low LP

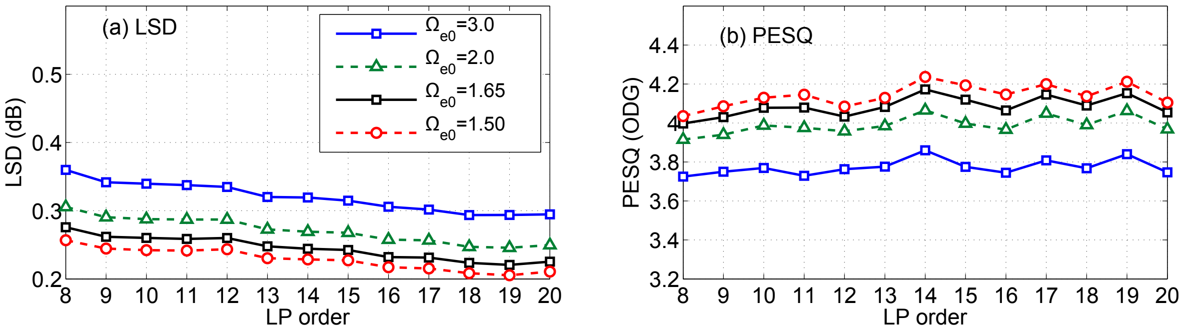


Figure 5.5: Inaudibility affected by LP order and Ω_{e0}: (a) LSD and (b) PESQ.

order can just provide global frequency information, only a few global formants can be provided in this case. Under low order LP analysis, each estimated formant will carry more information in comparison with the formant that estimated by high order LP analysis. That is to say the sound distortion brought by tuning one formant that estimated with low order LP analysis will be more severe. Therefore, to achieve inaudibility, LP order should be as higher as possible. On the other hand, since most processing will bring distortion to the formant structure of watermarked signal, if LP order is so high to follow all the spectral details, any distortion will result in LSFs deviation, which will obstruct the watermark extraction. In this case, LP order should be low to achieve robustness.

Modification degree Ω_{e0} According to Eqs. (4.24) to (4.26), bigger Ω_{e0} will increase the bandwidth difference between the sharpest formant and the second sharpest formant which makes it easier to discriminate "0" or "1". However, bigger Ω_{e0} also means severe modification to the sharpest formant in the original signal which will degrade the sound quality severely.

The inaudibility and robustness are conflicting, and affected by LP order and Ω_{e0}. To select the optimal parameters, we tentatively checked the inaudibility and robustness performance (at 4 bps) as a function of LP order and Ω_{e0}. LP order was selected as 8, 10, 12, 14, 16, 18, and 20. Ω_{e0} was selected as 1.50, 1.65, 2.0, and 3.0. Since objective measures enable quick results, we evaluated inaudibility with LSD and PESQ. Robustness

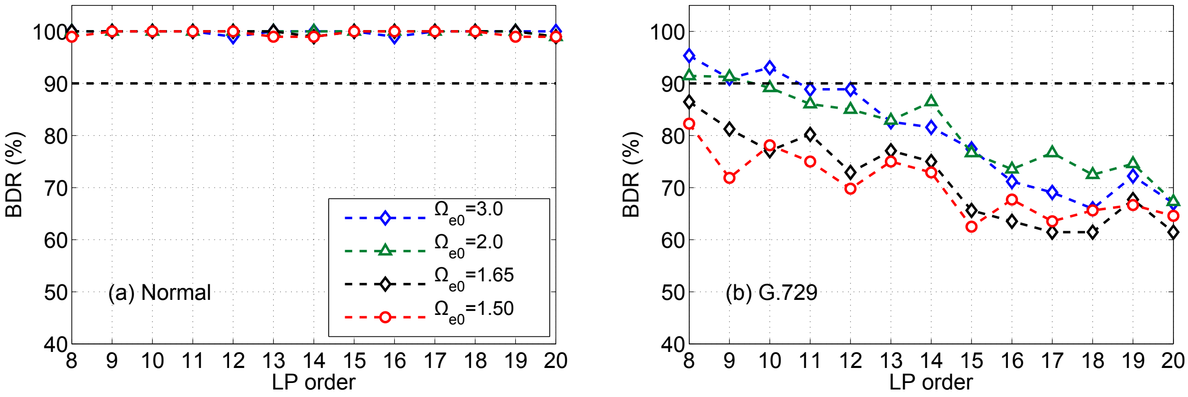


Figure 5.6: Robustness affected by LP order and Ω_{e0}: (a) normal detection and (b) detection after G.729.

was checked by normal extraction with BDR results. We also checked the watermark extraction after G.729 codec. This was because many watermarking methods failed to extract watermarks after this codec. Therefore, robustness against G.729 was one of the most difficult criterion, which could typically check whether the method was robust or not.

According to the LSD and PESQ results in Fig. 5.5, we can find: (*i*) under the same Ω_{e0}, inaudibility was not obviously affected by different LP orders; (*ii*) under the same LP order, when Ω_{e0} was increased to 3.0, there was an obvious distortion in inaudibility. Therefore, Ω_{e0} should be less than 3.0 for inaudibility.

Figure 5.6 shows the robustness results. We can see that with all different LP orders and Ω_{e0}, normal watermark extraction had almost the same BDR results, close to 100%. On the other hand, these two parameters greatly influenced the robustness against G.729, since BDR results drastically increased when LP order was increased. Therefore, it would be prefer to choose lower LP order for robustness. According to these results, we finalized Ω_{e0} as 2.0 for inaudibility and LP order as 10 for robustness (where BDR after G.729 at $\Omega_{e0} = 2.0$ could be controlled over 90%).

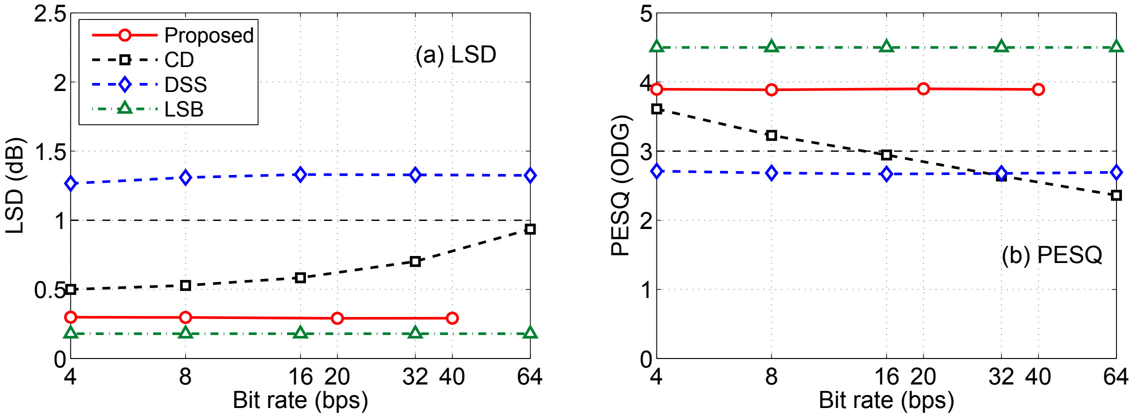


Figure 5.7: Evaluations of inaudibility for formant enhancement based watermarking: (a) LSD and (b) PESQ.

5.3.2 Evaluations for inaudibility

We follow the above parameters to evaluate the proposed method. Evaluations were also done to three typical methods: the least significant bit-replacement (LSB) method [31], direct spread spectrum (DSS) method [38], and cochlear delay (CD) method [54] These methods have separately exhibited excellent performance in inaudibility, robustness, and both inaudibility and robustness. A quick review of these methods is as follows: LSB replaces the least significant bits with watermarks at the quantization level so that replacement in less perceptible component does not cause distortion to human perception; DSS spreads watermarks over many (possibly all) frequency bands so that watermarks cannot be easily destroyed; CD embeds watermarks by enhancing the phase information of the original signal with respect to two kinds of cochlear delay (one is for “0” and the other one is for “1”). The embedding bit rates for LSB, DSS, and CD were 4, 8, 16, 32, and 64 bps according to their original implementations. The inaudibility of the proposed method was evaluated with objective measures, i.e., LSD and PESQ, and a subjective listening test. All evaluation results were calculated on the average of twelve stimuli.

Objective evaluations

Figure 5.7 plots the LSD and PESQ results for the proposed method, CD, DSS, and LSB. As we can see, LSB had the best performance among all the four methods. CD could satisfy inaudibility when the bit rate was no more than 16 bps. DSS could not satisfy the criteria for either LSD or PESQ. The proposed method could satisfy criteria for both LSD and PESQ, which indicated it could objectively satisfy the inaudibility requirement.

Subjective evaluations

Inaudibility of the proposed method was also investigated through a listening test in which all twelve stimuli were involved. The experiment conditions referred to those in [45]. For each stimulus, five test pairs were set up. Each test pair contained two speech tracks, one was the original (Org) stimulus and the other was the same original (Org) stimulus, or the resynthesized original (ResynOrg) stimulus, or the watermarked stimulus (at 4 bps) that was realized by the proposed method (Pro), CD, or DSS, where the test pair consisted of the original (Org) stimulus and resynthesized original (ResynOrg) stimulus was evaluated for the proposed method to check whether sound distortion could be caused by speech analysis/synthesis in spite of watermarks. Three male subjects and one female subject with normal hearing ability have attended to the listening test. Each subject was presented with one test pair in a trial and then asked to report the similarity between two tracks by choosing a subjective score from 0 (completely the same), 1 (probably the same), 2 (probably different), and 3 (completely different). Each subject was totally presented with 60 test pairs (twelve stimuli × five pairs (Org-Org, Org-ResynOrg, Org-Pro, Org-CD, Org-DSS)).

The mean subjective scores on five test pairs for each stimuli are given out in Fig. 5.8. These results revealed that it was difficult for subjects to tell the difference between two tracks in the Org-Org, Org-ResynOrg, and Org-Pro test pairs, which suggested that the sound distortion caused by speech analysis/synthesis and watermarks embedding in proposed method was perceptually insignificant. In comparison, CD was slightly perceptible

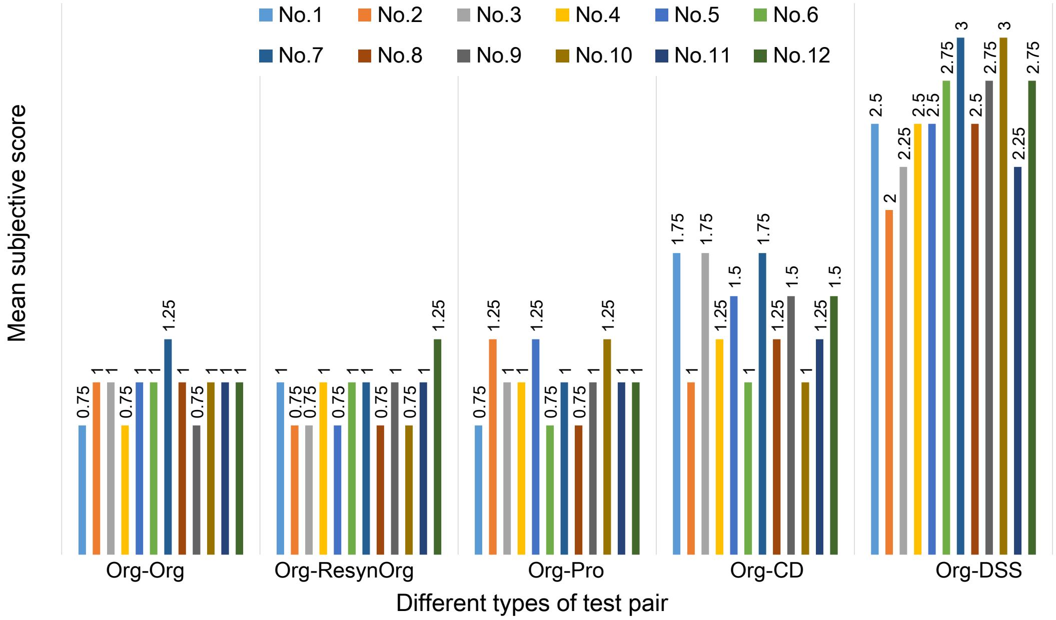


Figure 5.8: Evaluations of inaudibility for formant enhancement based watermarking with listening test.

for a few stimuli, and DSS introduced obvious distortion to the original signals.

5.3.3 Evaluations for robustness

Evaluations against speech codecs

We applied some typical speech codecs to the watermarked speech to evaluate the robustness. These included G.711 and G.729 that we used in LSF-QIM based watermarking, G.726 (waveform-based, ADPCM), G.723.1 [107], and Mixed Excitation Linear Predictive (MELP) [108]. G.723.1 is a dual rate speech coder for communications at 5.3 and 6.3 kbit/s. The G.723.1 is an audio codec for voice which is completely different codec from G.723. G.723.1 is mostly used in VoIP due to its low bandwidth requirement. The MELP speech codec was proposed for low bit rate speech coding in 1995. MELP is based on the traditional LPC vocoder, and it can generate the parametric representation of speech signals and provide the speech with good quality. The MELP is mainly used in military

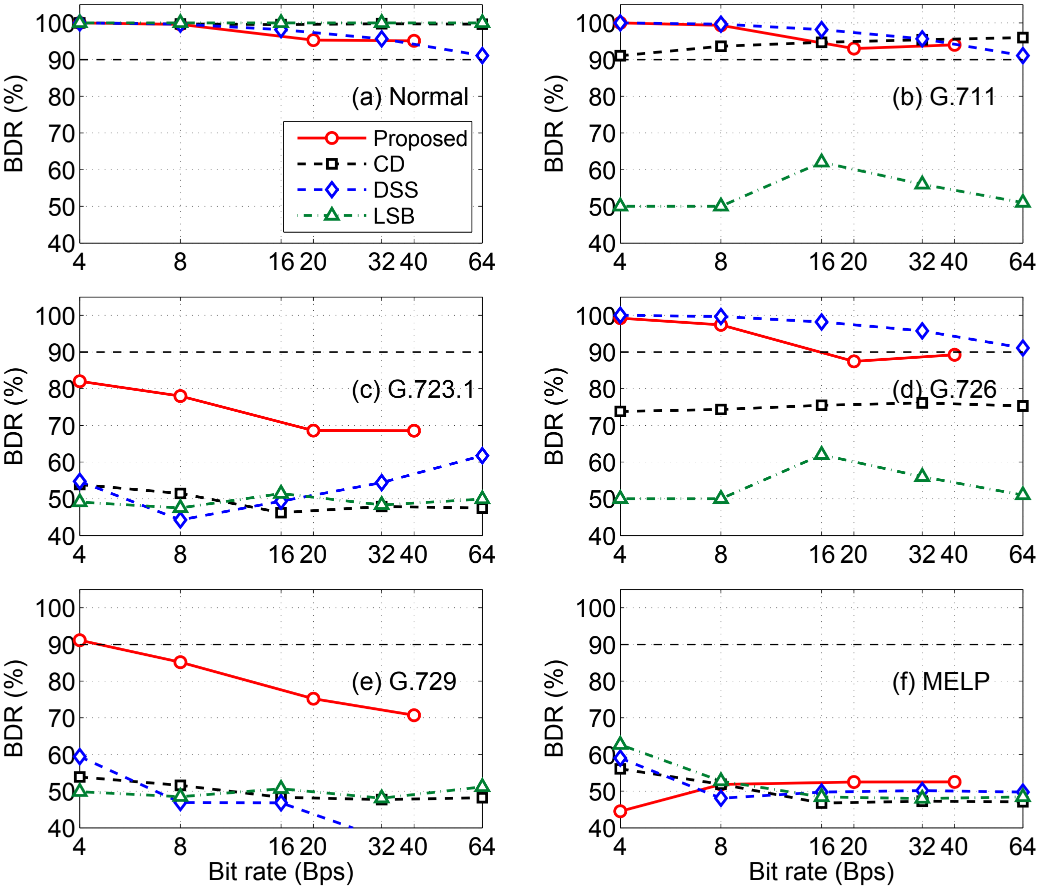


Figure 5.9: Evaluations of robustness for formant enhancement based watermarking against (a) normal extraction, (b) G.711, (c) G.723.1, (d) G.726, (e) G.729, and (f) MELP.

application, satellite communications, and secure voice devices.

As shown in Fig. 5.9, the LSB method was not robust against any speech codec; CD method was only robust against normal extraction and G.711; DSS method was robust against normal extraction, G.711, and G.726. The proposed method was robust against normal extraction, G.711, G.726, G.729, and its performance against G.723.1 was better than the other methods. These results implied that the proposed method had better robustness than the other watermarking methods. Nonetheless, it was not robust again MELP, and none of the watermarking could survive from this speech codec.

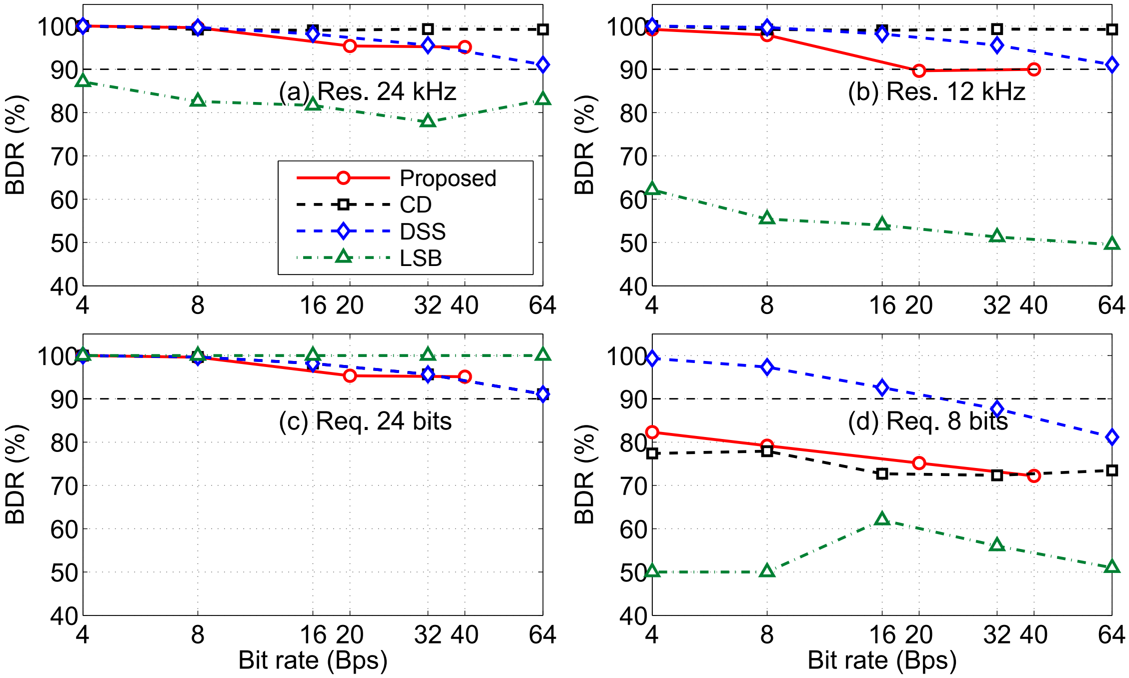


Figure 5.10: Evaluations of robustness for formant enhancement based watermarking against (a) 24 kHz and (b) 12 kHz and re-quantization with (c) 24 bits and (d) 8 bits.

Evaluations against general speech processing

First, we evaluated the proposed method against several processing: (a) re-sampling at 12 kHz and 24 kHz, (b) re-quantization with 24 bits and 8 bits. Figure 5.10 plots all results. DSS obviously performed the best. LSB was only good for re-quantization with 24 bits. The proposed method and CD provided good performance except for re-quantization with 8 bits. The reason for this with the proposed method was re-quantization at lower rate compared with signal's original sampling rate introduced some distortions to the watermarked signal, which destroyed the bandwidth relationship for watermark extraction.

Second, we evaluated the proposed method with other practical speech processing. These included (a) signal amplifying by 2.0 and 0.5, speech analysis/synthesis by (b) short-time Fourier transform (STFT), and (c) gammatone filterbank (GTFB). The BER results (in %) at 4 bps and 8 bps are listed in Tab. 5.1. From these results, we can conclude that the proposed method and DSS were robust against these processing.

Table 5.1: BDR (%) results of robustness against practical processing.

Bit rate	Processing	CD	DSS	LSB	Proposed
4 bps	Ampl. by 2.0	96.43	100.00	50.00	99.22
	Ampl. by 0.5	96.43	100.00	47.42	99.22
	STFT	96.43	100.00	100.00	99.22
	GTFB	66.57	100.00	53.42	98.96
8 bps	Ampl. by 2.0	96.41	100.00	47.79	99.14
	Ampl. by 0.5	94.48	100.00	49.33	99.55
	STFT	93.09	100.00	100.00	99.22
	GTFB	60.53	100.00	51.27	98.75

Evaluations against common attacks

We also evaluated the proposed method against several attacks. These included (a) signal flip, (b) signal sample repletion, (c) signal flanger, (d) signal jitter, (e) echo addition, and (f) Gaussian noise addition. In signal flip, the values of two randomly chosen samples in each frame (25ms) were exchanged, thus in one second, 40 sample pairs were exchanged. In speech sample repetition, one randomly chosen sample in each frame was repeated. In the evaluation, 40 samples were repeated in one second, therefore, the duration of the signal was increased. Signal flanger was an operation to create a signal by mixing a slightly delayed copy of itself. In the evaluation, the delay time was 9.375 ms, i.e., around one third of the frame size was delayed for each frame. In signal jitter, the randomly chosen samples of each frame was set to be 0. In echo attacks, a single 100 ms echo addition of -6 dB was added to the watermarked signals. In noise addition, a Gaussian noise addition with an overall average SNR of 36 dB was added to the watermarked signals. The first four attacks were referred to [109], and the last two attacks were recommended by the Information Hiding and its Criteria (IHC) committee [110]). The BDR results were plotted in Fig. 5.11, it is shown that proposed method was basically robust against these attacks.

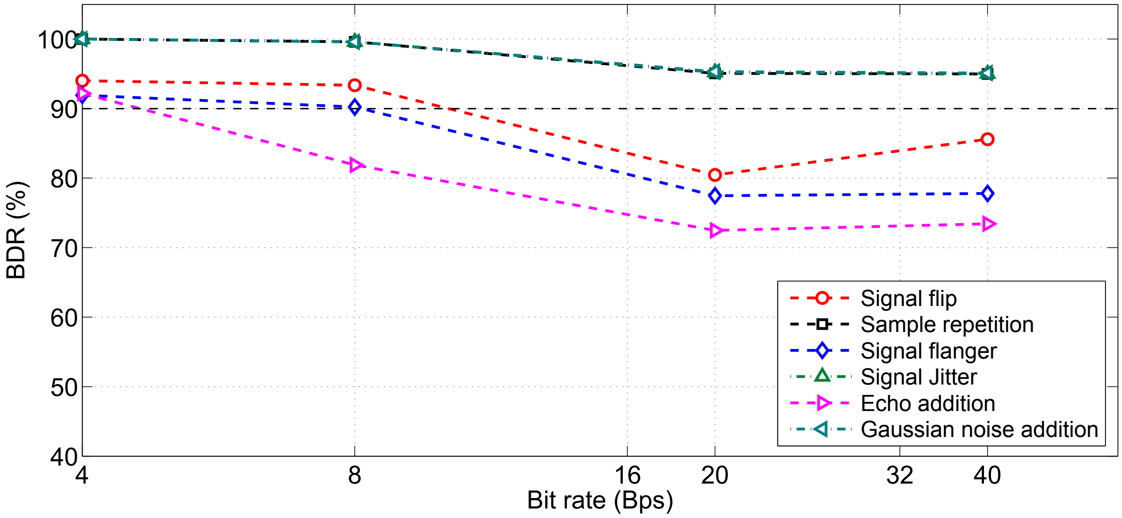


Figure 5.11: Evaluations of robustness for formant enhancement based watermarking against common attacks.

5.3.4 Discussion

In this chapter, we evaluated the formant enhancement based watermarking by carrying out evaluations with respect to inaudibility (LSD, PESQ, and a listening test) and robustness against different speech codecs, general processing, and common attacks. The LP order and Ω_{e0} were well examined for achieving good performance in inaudibility and robustness. The evaluation suggested the proposed method had two advantages.

(1) Since formant enhancement is capable to improve the sound quality of synthesized speech, watermarks embedded as formant enhancement was almost inaudible. Therefore, the proposed method can satisfy the inaudibility requirement.

(2) Watermarks embedded in LSFs and extracted by identifying bandwidth relationship can tolerate small modifications of frequency components that are caused by speech codecs, general processing, and common attacks. Therefore, the proposed method shows stronger robustness than the other compared watermarking methods.

Moreover, from the watermark embedding and extraction mechanism, several superiorities can also be found in the proposed method, such as:

(3) Each frame has its own frequency characteristic, the enhanced formant (the sharpest

formant or the second sharpest formant) is possible to exist in any frequency range. Therefore, when small proportion of frequency component is changed, where watermarks are not contained, watermarks can be still extracted.

(4) From the point of security, embedding watermarks into the intrinsically irregular formant structures makes the watermarks confidential. This is because various formant structures make it difficult for the attackers or the third party to confirm whether the formant structure is formed by artificial manipulation or not, since embedded bandwidth relationship is also possible in a rough speech.

(5) When the LP order for estimating formants is unknown, bandwidth relationship is unable to discover.

It is also important to note that although LSFs in the proposed method were shifted so that watermarks could be embedded, the proposed method was essentially different from QIM based watermarking. This is because QIM based watermarking modify embedding parameter without physical meaning, while the modification to LSFs in our method was motivated by formant enhancement.

5.4 Influence of frame synchronization to watermarking method

The frame synchronization scheme has been implemented for the proposed methods in chapter 4. Since the information for frame synchronization needs to be embedded into the watermarked signal, it is necessary to investigate whether the synchronization information will additionally degrade the sound quality, and whether it will obstruct watermark extraction. In the following, we will carry out evaluation about the above two questions. The synchronization information was embedded into the watermarked signal that obtained by the formant enhancement based watermarking, and the frame size is 25ms.

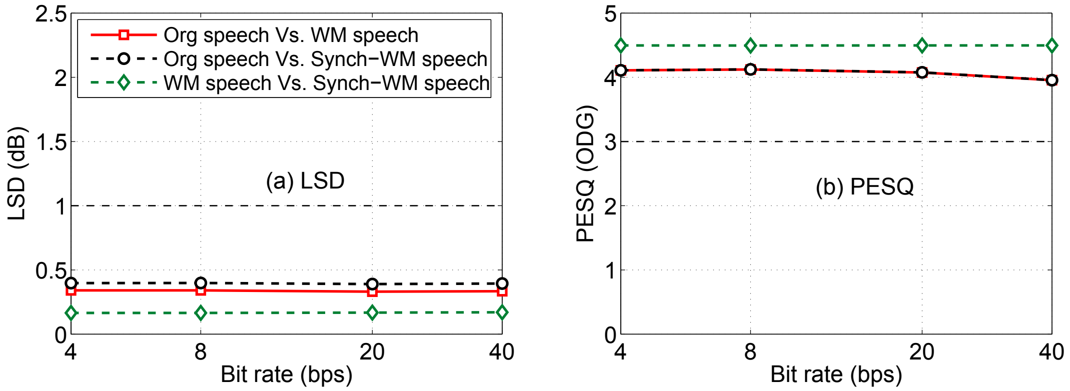


Figure 5.12: Influence of frame synchronization to the inaudibility of watermarked speech: (a) LSD and (b) PESQ.

5.4.1 Influence to inaudibility

The LSD and PESQ were used as the evaluation measures. Figure 5.12 plots the results for inaudibility. These results were calculated on the average of twelve stimuli. As seen in this figure, three kinds of inaudibility results were given out: the red line shows the result between the original speech and the watermarked speech, the black line shows the result between the original speech and the watermarked speech that embedded with synchronization information, and the green line shows the result between the watermarked speech with and without synchronization information. For LSD, the red line and the black line are closer to each other, and the green line is around 0.18 dB, which indicated that the synchronization information only introduced very slight distortion to the watermarked signal. For PESQ, the red line and the black line are overlapped, and the green line is reached to the full score of 4.5 ODG. This suggested for PESQ evaluation, the watermarked speech with and without synchronization information had the same speech quality, and distortion introduced by synchronization information was negligible.

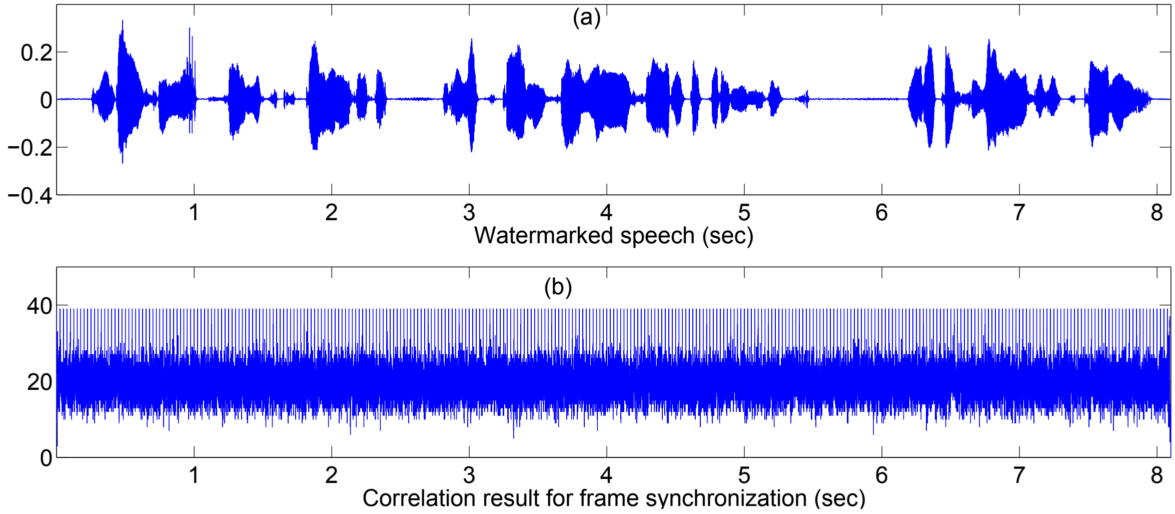


Figure 5.13: Frame synchronization result of watermarked signal.

5.4.2 Influence to watermark extraction

Before the watermark extraction process, the correlation technique was applied to segment the speech frames. Figure 5.13 shows the correlation result, the peaks correspond to the beginning of each frame. After this process, watermarks were extracted from the watermarked signal. Figure 5.14 has present the two BDR results: the red line is the result calculated from the watermarked signal without synchronization information and the black line is the result calculated from the watermarked signal embedded with synchronization information. It is found that the two lines are overlapped, this indicated the synchronization information did not affect the watermark extraction of the watermarking method.

5.5 Summary

In this chapter, we have separately evaluated two speech watermarking methods with inaudibly and robustness. For the LSFs and QIM based method, we took into consideration of the inaudibility and robustness that were influenced by the different quantization steps. The results from inaudibility evaluation revealed that the proposed method could

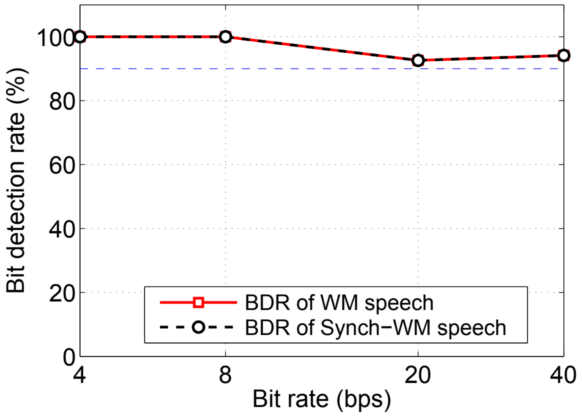


Figure 5.14: Influence of frame synchronization to watermark extraction.

satisfy inaudibility when quantization step was small. However, when subjected to the processing that could change the shape of waveform or the value of signal such as speech codecs, down-sampling, and low-bit re-quantization, this method was not robust.

In the evaluations of formant enhancement based watermarking, we gave much consideration about the parameters such as the LP order and the Ω_{e0} when balancing the inaudibility and robustness. In this method, watermark embedding through formant enhancement did not cause severe degradation to the original speech quality, and the watermark extraction by identifying bandwidth relationship was able to tolerate slight distortions of frequency components caused by other processing. Therefore, in comparison with other watermarking methods, this method could achieve a good trade-off between inaudibility and robustness. Nonetheless, the robustness of the proposed method against some speech codecs such as G.723.1 and MELP needs to be improved.

Besides, we also investigated the influence of frame synchronization to watermarking method. From the evaluations, we found that the embedding of synchronization information did not cause severe distortion to the speech quality, and the synchronization information did not affect the extraction of watermarks.

Chapter 6

Applications of formant enhancement based watermarking

In previous chapter, we have evaluated two watermarking methods. The LSFs and QIM based watermarking cannot satisfy the inaudibility and robustness requirements simultaneously. In comparison, the formant enhancement based watermarking has better performance. In this chapter, this watermarking method will be applied for tampering detection and hybrid watermarking.

6.1 Tampering detection with formant enhancement based watermarking

6.1.1 Introduction

Rapid development in digital technologies has greatly facilitated the speech signals to be reduplicated and edited at high fidelity. These advances lead to new social issues related to malicious attacks and unauthorized tampering to speech. For example, by using free editing software, ordinary people are allowed to alter speech without leaving perceptual clues. Some specialized speech analysis/synthesis tools such as STRAIGHT [111], voice conversion [112, 113] , and speech morphing [114], are professional to produce high naturalness and intelligibility of tampered speech, although important information has been changed. As these progresses enable speech to be tampered in a more realistic and credible way, it is becoming difficult to identify the tampering and confirm the originality of speech.

Speech signals can be used for a variety ways. The criminal investigation [115, 116] is a major one. As a kind of digital evidence, speech can record (1) what happened in a certain place and time and (2) the information provided by the victims or the suspects. However, in most cases, speech is not immediately used after being recorded. They have to pass through a series of judicial procedures in which different people may be involved. Since improper actions taken to handle, examine, and store the speech are possible to destroy the originality of it (intentionally or unintentionally), and not everyone involved is trustful, it is difficult to ensure the originally of speech after the complicated processing. Besides, if the people who are responsible for handling the speech have malicious intent to mislead the listener, the originally of speech cannot be ensured. For example, by using voice conversion, speech content (what is the speaker saying) can be tampered, e.g., a word replacement from "YES" to "NO"; by using speech morphing, the individuality of speaker (who is saying) can be deliberately transformed to that of another speaker. These tampering are able to conceal important information or covering up the reality. As the speech content and speaker identity plays a key role in the criminal investigation, any single word change or forged speaker will result in serious problem for judgment.

To confirm the speech is best suited to the unique acquisition environment and the truth, investigation about whether the speech has been tampered since its creation should be carried out. Digital watermarking can effectively check if the original speech signals have been tampered by embedding digital data into them. To be effective, watermarking method should be implemented according to four requirements of (1) inaudibility to human auditory system, (2) blindness to extract watermarks without referring to the original signal, (3) robustness against speech processing, and (4) fragility against tampering. In chapter 5, the formant enhancement based watermarking has been evaluated, and it is found that this method can satisfy the first three requirements. In this chapter, this

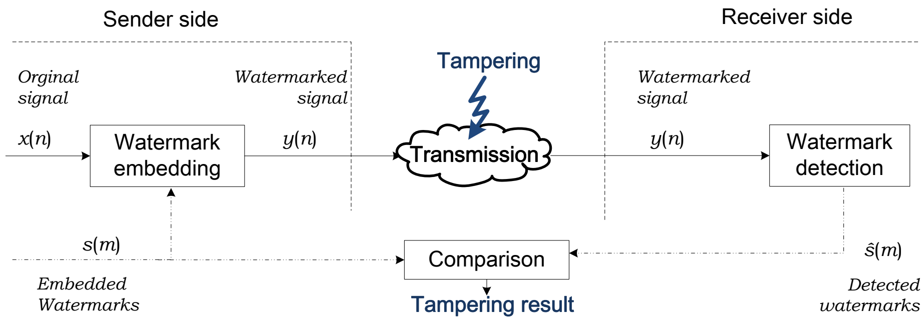


Figure 6.1: Scheme for tampering detection.

watermarking method is employed to detect tampering in speech signal by exploring its fragility properties.

6.1.2 Scheme of tampering detection

To check whether tampering has occurred to speech signals during the transmission, an overall block diagram for tampering detection is given out in this chapter, where speech watermarking is employed. Watermarks will be embedded at the sender side and then extracted at the receiver side for detecting tampering. If the speech watermarking method can satisfy both robustness and fragility, tampering could be detected by the mismatched bits between the embedded watermarks and the detected watermarks. The whole process can be explained as follows:

Process at the sender side At the sender side, we have the original speech signal $x(n)$ and the watermarks $s(m)$. As shown in Fig. 6.1, before sending the original signal $x(n)$ to the receiver, watermarks $s(m)$ will be embedded into it to construct the watermarked signal $y(n)$. The detailed embedding processes are the same as those in chapter 4.2.3. Finally, the watermark signal $y(n)$ will be transmitted.

Process at the receiver side After receiving $y(n)$ at the receiver side, watermarks will be extracted from received $y(n)$. The detailed watermark extraction processes are the same as those in chapter 4.2.3. The extracted watermarks, named as $\hat{s}(m)$, will be compared with $s(m)$ to check whether tampering has occurred.

Verification of tampering Ideally, if the watermarking method could satisfy fragility, once tampering occurred, watermarks in tampered segment will be destroyed. Therefore, tampering could be detected by the mismatched bits between $s(m)$ and $\hat{s}(m)$. If there is no mismatch, it means the received signal is the original signal and no tampering occurred; otherwise, each mismatch indicates that the corresponding frame in received signal has been possibly tampered. For example, if $s(m)$=01001101... while the detected $\hat{s}(m)$=01101101..., this indicates the third frame may have been tampered.

6.1.3 Evaluations for tampering detection scheme

In this chapter, we evaluated the proposed tampering detection scheme with respect to inaudibility, robustness, and fragility (the proposed scheme is a blind method). The database is the same as those used in previous chapters. For tampering detection scheme, the top priority is whether the proposed scheme has the ability to detect tampering and there is no requirement for embedding bit rate. Therefore, we evaluate the proposed scheme at the fixed embedding bit rate, 4 bps. The order of LP analysis was chosen as 10 based on our previous analysis. Ω_{e0} for embedding "0" was also adopted as 2.0 (Ω_{e1} for "1" was automatically fixed based on bandwidth characteristics of each frame). Embedded watermarks was a single word "GOOD". Evaluations were also done to two other methods: LSB and CD.

Evaluations for inaudibility

Inaudibility was checked by LSD and PESQ. The evaluation results of LSD and PESQ for three methods are plotted in Fig. 6.2, where the straight dashed-lines in each sub-figure indicate the criteria for LSD ($\leq$ 1.0 dB) and PESQ ($\geq$ 3.0 OGD). As we can see, all three methods could satisfy the criteria for LSD and PESQ. The LSB method performed the best and the proposed method was a little better than CD method. These results

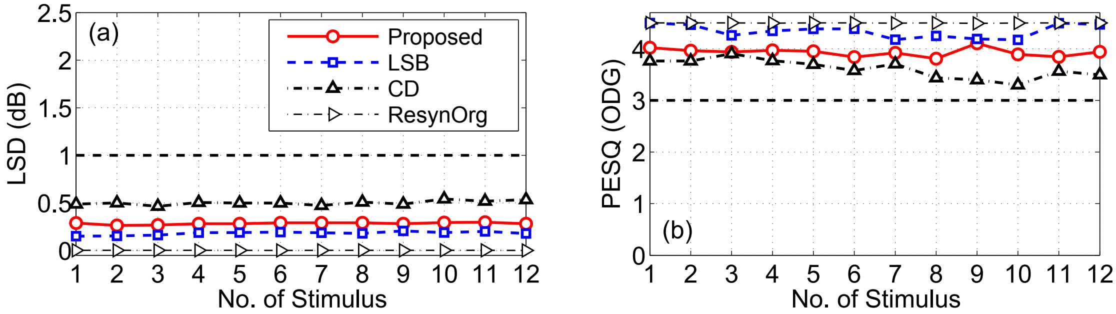


Figure 6.2: Evaluations of inaudibility for tampering detection scheme: (a) LSD and (b) PESQ.

indicated that these methods could objectively satisfy the inaudibility requirement. The result labelled as "ResynOrg" was also given out for checking whether sound distortion could be caused by speech analysis/synthesis in spite of the embedding of watermarks in the proposed method. Based on the obtained result, the resynthesized original signal had almost the same sound quality as the original signal.

Evaluations for robustness

The robustness of proposed was firstly evaluated against speech codecs. The speech codecs were chosen as G.711, G.723.1, G.726, G.729, and MELP. The BDR results calculated after speech codecs are presented in Fig. 6.3, where normal extraction is also given out in Fig. 6.3(a). The straight dashed line in each sub-figure indicated the criteria for BDR ($\geq 90\%$). It is clear that LSB was not robust against any speech codec except for normal extraction, CD was only robust against normal extraction and G.711. In contrast, the proposed method could survive from normal extraction and three kinds of speech codecs (100% for G.711 and G.726, around 90% for G.729). These implied the proposed method was more robust against these speech codecs compared with LSB and CD. However, the robustness of proposed method against G.723.1 and MELP needed to be improved.

We also evaluated the proposed method against several speech processing [45]. These included re-sampling at 24 kHz and 12 kHz, re-quantization with 24 bits and 8 bits, signal amplifying by 2.0 times, a single 100 ms echo addition of -6 dB, speech analysis/synthesis

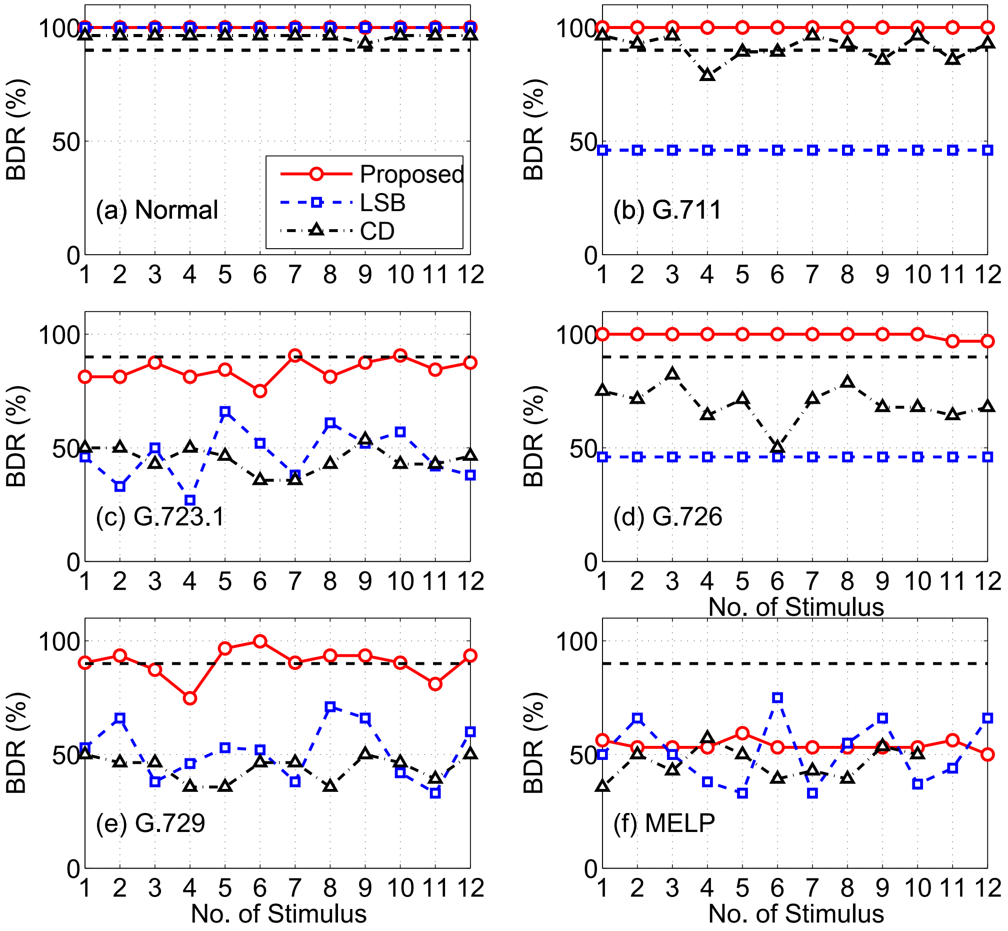


Figure 6.3: Evaluations of robustness for tampering detection scheme against (a) normal extraction, (b) G.711, (c) G.723.1, (d) G.726, (e) G.729, and (f) MELP.

by STFT, and GTFB. The BDR results after each processing have been plotted in Fig. 6.4. LSB was only robust against re-sampling at 24 kHz, re-quantization with 24 bits, and STFT; CD was robust against most processing except for re-quantization with 8 bits, echo addition, and GTFB. In comparison, the proposed method could correctly extract watermarks after these processing, which meant it was more robust than LSB and CD.

Evaluations for fragility

Many previous works, e.g., [51] and [55], have confirmed the fragility of their methods by carrying out various types of tampering. However, there is no consistent definition

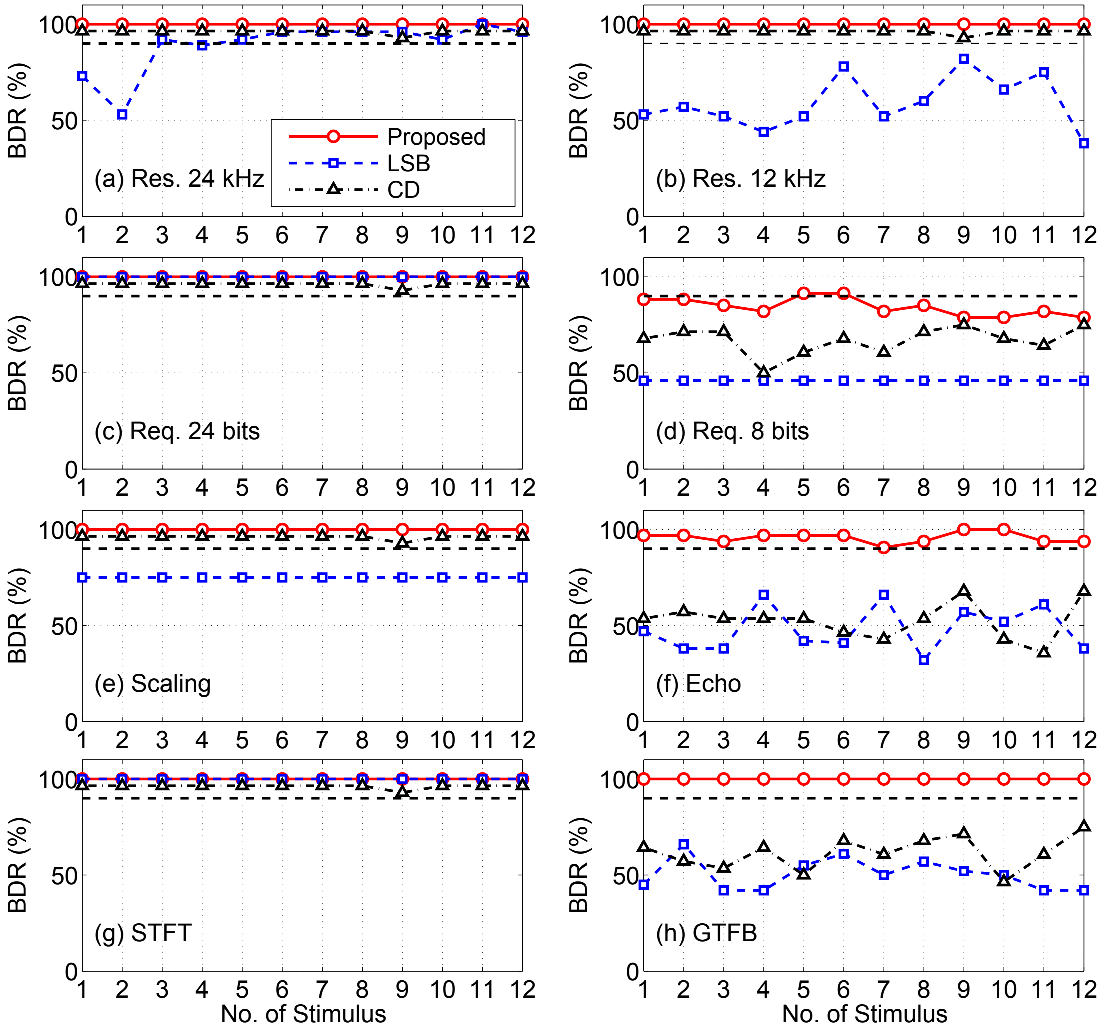


Figure 6.4: Evaluations of robustness for tampering detection scheme against speech processing.

for tampering among these works. In general, tampering are performed based on the motivation of the attackers. In this case, any operation that can be used to tamper a speech should be evaluated for watermarking method when verifying its fragility and ability for tampering detection. Therefore, we evaluated the fragility of the proposed method against several possible tampering in this chapter. Since LSB and CD are not completely robust, even they are fragile against tampering, they are unable to tell whether the failed extraction of watermarks is caused by speech processing or tampering. That

Table 6.1: BDR (%) results in fragility evaluations of tampering detection scheme.

No.	Tampering type	Description	BDR (%)
(b)	No tampering	$--$	100.0
(c)	Add white noise	normal distribution, $N(0.01, 1)$	45.19
(d)	Reverberation	real impulse response of 0.3 s	68.80
(e)	Concatenation	concatenate with un-watermarked speech	42.86
(f)	Low-pass filtering	order: 32-th, normalized cut-off frequency: 0.99	41.98
(g)	High-pass filtering	order: 32-th, normalized cut-off frequency: 0.01	49.85
(h)	Speed up	speed up the whole speech by +4%	71.56
(i)	Speed down	speed down the whole speech by -4%	79.51
(j)	Pitch shift	change the pitch of speech -4% in real time	68.12

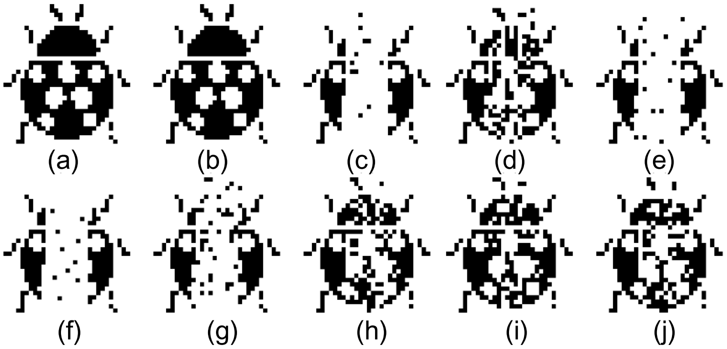


Figure 6.5: Evaluations of fragility for tampering detection scheme against tampering.

is to say, they cannot successfully detect tampering unless robustness being improved. Therefore, fragility evaluation was only conducted to the proposed method.

As to intuitively reflect fragility, a 32×32 bitmap image in Fig. 6.5(a) was used as watermarks. Since bit rate was 4 bps, as to embed the complete image, 12 speech tracks are repeatedly connected to construct a long original signal (256 second). After embedding the image to the original signal, the middle segment of watermarked signal was separately tampered with the tampering listed in Tab. 1 (Line 2 to Line 9). These evaluations referred to [55]. Adding white noise and reverberation are channel distortion, tampering speech with these operations can be considered as disturbing the speech. Concatenat-

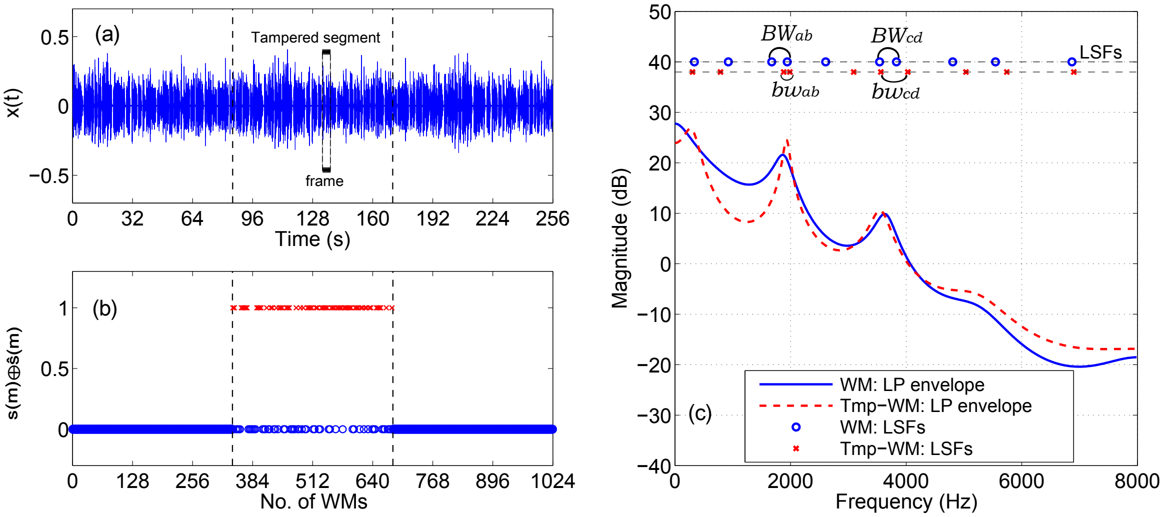


Figure 6.6: Example on fragility analysis: (a) original signal with middle segment tampered by adding white noise, (b) extraction errors densely appear in the tampered segment, and (c) one frame analysis: bandwidth relationship for watermark "1" has been destroyed due to tampering.

ing the watermarked signal with un-watermarked speech can be considered as content replacement. Filtering with low-pass and high-pass filters is regarded as removing specific frequency information of speech. Speed change (speech up and speed down) can modify the duration and tempo of speech without affecting its pitch. Pitch shift is to proportionally shift frequency components while preserving the duration of speech, which can be regarded as manipulating the individualities of the speaker.

The extracted image from un-tampered watermarked signal is shown in Figs. 6.5(b). where watermarks could be correctly extracted. The extracted images from other tampered watermarked signals are separately shown in Figs. 6.5(c) to (j). It is noticeable that watermarks in the tampered segment were destroyed. Tab. 1 gives out the accurate BDR results. Since the BDR calculated from the tampered segment were quite low compared with no tampering (normal extraction), we can conclude that the proposed method was fragile against the evaluated tampering.

Figure 6.6 illustrates an example of how detection errors happened, for example, after tampering by adding white noise. In Fig. 6.6(b), detection errors (shown in red cross) densely appeared in the tampered segment. In contrast, watermarks in the un-tampered segment could be correctly detected due to the robustness the proposed method. Therefore, tampering could be indicated with the destroyed watermarks. Figure 6.6(c) examined the bandwidth relationships before and after tampering of one tampered frame, where "1" has been embedded. Before tampering, bandwidth relationship, $BW_{ab} = BW_{cd}$, could be easily observed to correctly extracted the watermark. After tampering, BW_{ab} has been much narrowed to bw_{ab}, that is, bw_{ab} is much narrower than bw_{cd}. Therefore, it would be easily taken as that "0" has been embedded.

These obtained results suggested that the proposed method was fragile against tampering, and the destroyed watermarks could provide an evidence that signal has been tampered. As we found, after the tampering, the highest BDR rate was still lower than 80%, and the average of BDR was around 52%, so we would like to set 65% as the criteria for the real tampering. A lower BDR indicated strong confirmation of tampering. A high BDR indicated suspected tampering, in this case a deep investigation needed to be carried out to confirm whether this was a real tampering or not. In the evaluation, the embedding bit rate of watermarks was 4 bps, each embedded bit was able to account for 0.25 s speech segment when locating the tampering, although 0.25 s was too short to make a meaningful tampering of speech content. Therefore, as shown in Fig. 6.5, even some correct bits could still be intermittently extracted from the tampered speech segment, tampering could also be detected by checking several adjacent bits where detection errors densely appeared. Additionally, the detection precision was possible to be improved by increasing the embedding bit rate.

Ability for tampering detection

The above chapter evaluated the robustness and fragility of the proposed tampering detection scheme. Evaluation results indicated the proposed scheme was enough robust and fragile. To investigate the tampering detection ability of proposed method in more realistic situation, we considered the following evaluations.

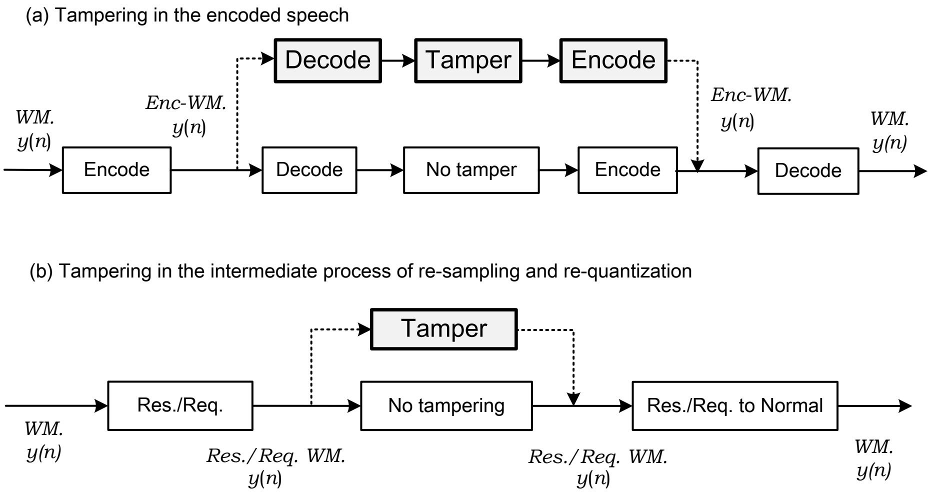


Figure 6.7: Flowchart of tampering in more realistic situation.

In realistic situation, encoding process is generally performed for watermarked signal at the sender side, and the decoding process is performed before watermark detection at the receiver side. To tamper the transmitted speech, as seen in Fig. 6.7(a), attackers should firstly decode the speech to raw data, make tampering, and then encoded back with the original coder. Likely, tampering also possibly happens to watermarked signals which are in the intermediate process of re-sampling and re-quantization. To investigate whether the proposed method could identify tampering under the situations that speech processing (speech codecs, re-sampling, and re-quantization) also exist, we followed the tampering process in Fig. 6.7 and then extract watermarks. Note that, to make a fair comparison, encoded watermarked signal in Fig. 6.7(a) was decoded and encoded even no tampering occurred. This process was made to compensate the speech codecs caused extraction error in the tampering case. In these evaluations, speech codecs of G.711, G.726, and G.729 were used, since the proposed method was not robust against G.723.1 and MELP. Twelve speech stimuli were embedded with the watermarks “GOOD”. The types of tampering were the same as those used in fragility evaluations. All evaluation results were calculated on the average of twelve signals.

Figure 6.8(a) compares the BDR results of normal extraction (1st bar) and those

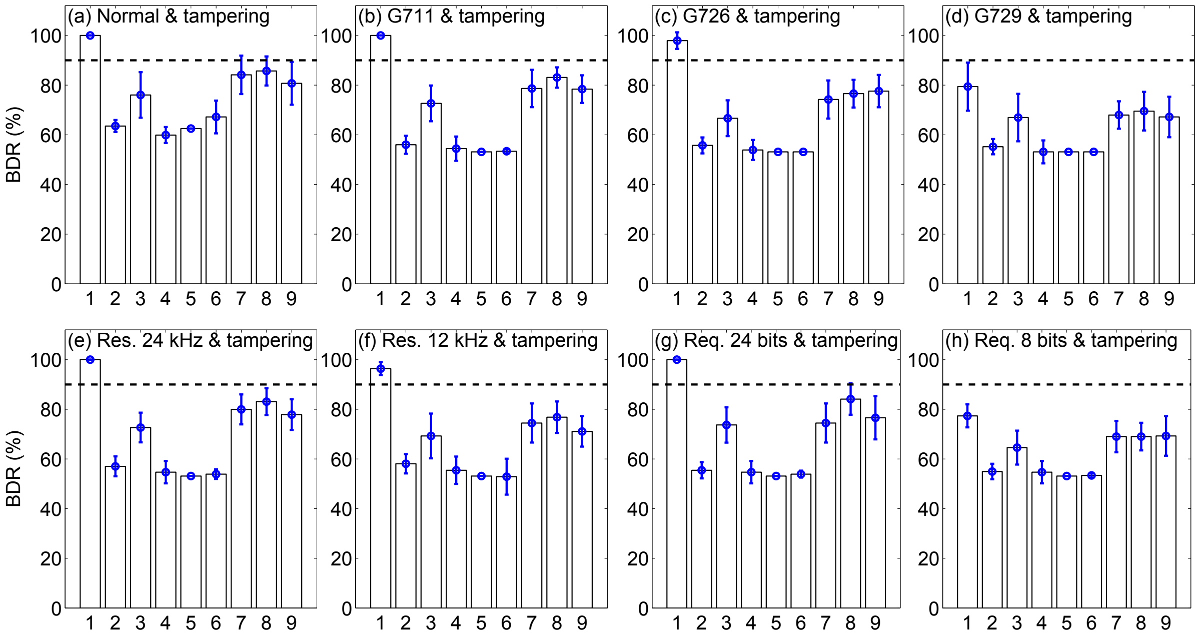


Figure 6.8: Evaluations for the tampering detection ability of the tampering detection scheme: (a) BDR comparison between normal extraction and after tampering, (b) to (h) BDR comparison after one kind of speech processing (G.711, or G.726, or G.729, or re-sampling at 24 kHz, or re-sampling at 8 kHz, or re-quantization with 24 bits, or re-quantization with 8 bits) and after both speech processing and tampering.

after different tampering (2nd bar: addind white noise, 3rd bar: reverberation, 4th bar: concatenation, 5th bar: low-pass filtering, 6th bar: high-pass filtering, 7th bar: speed up, 8th bar: speed down, 9th bar: pitch shift). We got the similar results that when tampering occurred, BDR drastically reduced which enabled tampering to be easily figured out. Figures 6.8(b) to (h) compare the BDR results between two cases, one is BDR after one kind of speech processing (1st bar), and the other one is BDR after both the speech processing and different tampering (2nd bar: adding white noise, 3rd bar: reverberation, 4th bar: concatenation, 5th bar: low-pass filtering, 6th bar: high-pass filtering, 7th bar: speed up, 8th bar: speed down, 9th bar: pitch shift). For Figs. 6.8(b), (e), and (g), BDR was quite high when only speech processing applied, while after tampering, BDR was reduced. In Figs. 6.8(c) and (f), speech processing slightly introduced some bit extraction errors, while compared with those after tampering, the discrepancy in BDR could be equivalently kept as that in Fig. 6.8(a) (normal extraction & tampering). This was because speech processing had the same influence to watermarked signal no matter there was tampering or not. These results suggested that speech processing did not affect the detection of tampering and tampering could be detected no matter there is speech processing or not. However, in Fig. 6.8(d), BDR after G.729 were deteriorated even without tampering, this was because watermarked signal was encoded and decoded twice by G.729 which doubly introduced bit detection errors, thus it would be easily mistaken G.729 as tampering. Similarly, in Fig. 6.8(h), BDR after re-quantization with 8 bits was also deteriorated and made it difficult to distinguish it from tampering. To overcome these problems, robustness of the proposed method should be continually improved in the next step.

6.1.4 Discussion

The above chapter evaluated the performance of the proposed tampering detection scheme with respect to inaudibility, robustness, and fragility. In inaudibility evaluations, the proposed scheme can satisfy the criteria of both LSD and PESQ, which indicates it can objectively satisfy inaudibility. In robustness evaluations, performances of the proposed scheme, LSB, and CD are evaluated against speech codecs and speech processing. LSB method and CD method cannot show strong robustness when subjected to several processing. The proposed method exhibits stronger robustness compared with other methods.

Based on the results from robustness evaluations, fragility evaluations are only conduced to the proposed scheme. In these evaluations, a series of tampering are performed to the watermarked signals, due to which watermarks cannot be correctly extracted. Therefore, the destroyed watermark can function as a sign to indicate that tampering has occurred. Additionally, to check the detection ability of the proposed scheme under the situation that speech processing also exist, an in-depth evaluation is also carried out. By comparing the BDR results obtained from watermarked signal processed by speech processing, and the results from watermarked signal processed by both speech processing and tampering, tampering can be distinguish from most speech processing. These results further verify the tampering detection ability of the proposed scheme.

In summary, the proposed scheme has good performance in inaudible, robustness, and fragility. Moreover, it can detect tampering with its fragility. The embedding capacity of the proposed scheme, although relatively low, is still sufficient for locating tampering in time domain. Moreover, since an automatic frame synchronization scheme has been implemented for the proposed method, even if the attacker tries to add or crop segment to the transmitted signal, this kind of tampering will not disturb the watermark extraction, and it is easy to judge how long the watermarked signal has been added or cropped by checking the length of extracted watermarks. Nonetheless, more types of tampering should be investigated to verify the detection ability of the proposed scheme.

6.2 Hybrid speech watermarking based on formant enhancement and cochlear delay

6.2.1 Introduction

It is known that the requirements for speech watermarking are conflicting with each other, and it is difficult for most methods to get a trade-off between them. To realize desired watermarking, hybrid watermarking method which can combine two watermarking methods together, or can be implemented beyond two domains has been explored for image [117], [118], and video protection [119]. The hybrid watermarking was motivated to improve the performance of each single method by taking advantage of both of them. For example, in [117], a hybrid watermarking method benefited from the genetic programming and particle swarm optimization to achieve both robustness and imperceptibility; a visual-audio hybrid watermarking [119] embedded the error correcting information of the video watermarks as audio watermarks to refine the retrieved watermark during watermark extraction. In the literature, limited hybrid watermarking methods have been found for audio signals. For example, an audio watermarking that combined spread spectrum (SS) and singular value decomposition (SVD) has been proposed for copyright protection [120]. This method took advantage that the destroyed watermarks in one domain, SS or SVD, was likely to be recovered from the other domain to ensure the robustness of the whole scheme. Another hybrid audio watermarking [121] has been investigated based on SVD, quantization, and chaotic encryption.

In comparison with single watermarking, the hybrid watermarking method possesses the superiority in robustness that watermarks embedded with one method (or in one domain) can assist or refine the watermark detection of the other method (the other domain). Moreover, since two methods can mutually complement each other, hybrid watermarking can benefit from each method for improved performance. To the best of our knowledge, there is no work that has dealt with hybrid watermarking for speech signal to achieve good performance. According to our evaluations in previous chapter, the FE (short for formant enhancement) based watermarking is inaudible and robust, the cochlear delay (CD) based watermarking [54] is basically inaudible and robust in comparison with the LSB (not robust) and DSS (not inaudible). Therefore, we believe that if these two methods can be incorporated to realize a hybrid watermarking, the robustness of the hybrid method can be improved than each single watermarking, and the inaudibility can be kept.

6.2.2 Scheme of hybrid watermarking

The process of speech production can be simplified as the source-filter model, in which the sound source (excitation signal or residue) and the vocal tract filter (characterized by the formants), are assumed to be independent with each other. The hybrid watermarking method employs the source-filter model of speech production so that the CD and FE methods can be separately applied. As shown in Fig. 6.9, the CD method is applied to the excitation signal and the FE method is applied to formants. A brief introduction of the CD method is given as follows.

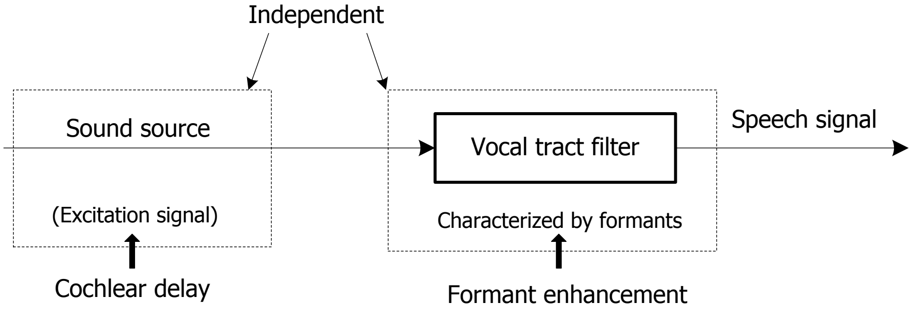


Figure 6.9: Concept of hybrid watermarking with FE and CD watermarking.

Cochlear delay based watermarking

A. Embedding

The CD watermarking utilizes that human cannot distinguish enhanced delay from original speech to embed inaudible watermarks. Different watermarks "0" and "1" in this method are embedded as two kinds of group delays that related to the human CD characteristics. Two 1st-order IIR all-pass filter, $H_m(z)$ ($m = 0$ and 1), in Eq. (6.1) are designed to generated the group delays for watermarks "0" and "1" with different values of b_m. Based on the subjective experiments in CD method, b_0=0.795 for embedding "0" and b_0=0.865 for embedding "1" were determined to achieve inaudibility.

$$H_m(z) = \frac{-b_m + z^{-1}}{1 - b_m z^{-1}}, \quad 0 < b_m < 1 \tag{6.1}$$

As shown in Fig. 6.10(a), different watermark "0" or "1" is embedded by filtering the original speech with $H_0(z)$ or $H_1(z)$, after which watermarked speech is obtained.

B. Blind extraction

In the CD method, different watermarks "0" and "1" are embedded with the two filters that carrying different poles b_m and zeros $1/b_m$ of $H_m(z)$ ($m = 0$ and 1). In the extraction process, watermarks can be extracted by analyzing the watermarked speech with two types of chirp-z transforms (CZTs) with the parameters of r=$1/b_0$ and r=$1/b_1$, as shown in Fig.

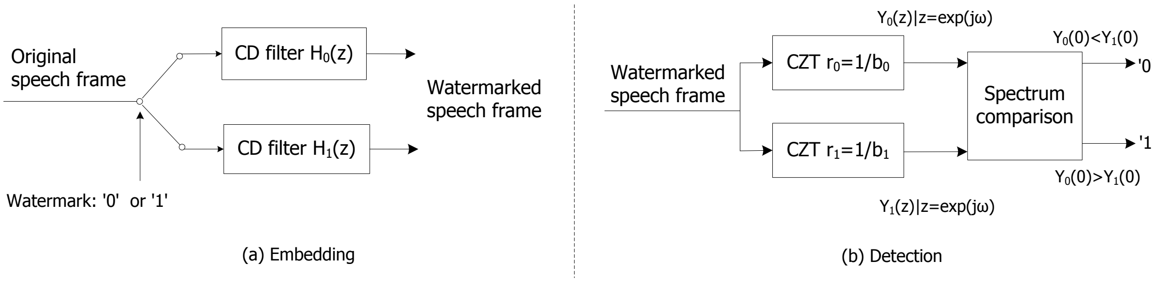


Figure 6.10: CD based watermarking: (a) embedding and (b) extraction.

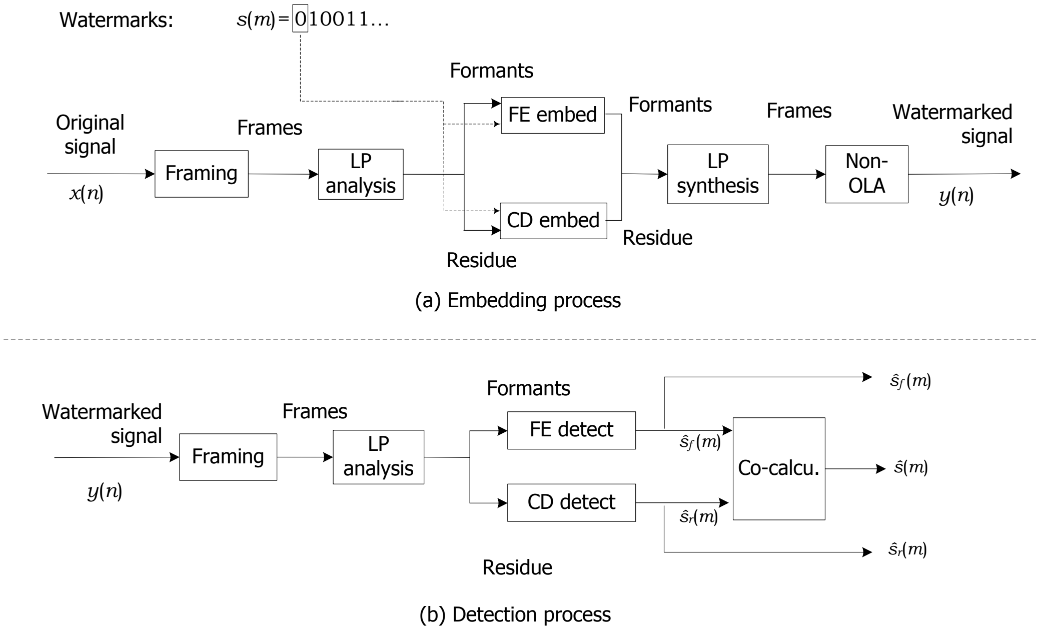


Figure 6.11: Proposed scheme of hybrid watermarking.

6.10(b). Watermarks can be decided by comparing the lowest spectra of $Y_0(0)$ and $Y_1(0)$.

Hybrid watermarking scheme

The hybrid watermarking is based on the source-filter model. The LP analysis is used to separate the formants and sound source so that two watermarking methods can be separately applied.

Figure 6.11(a) has a block diagram of the watermark embedding process. Watermark signal, $s(m)$, is embedded into original signal, $x(n)$, as follows.

Step 1 $x(n)$ is first segmented into non-overlapping frames. For each frame, LP is applied to extract formants and residue.

Step 2 One bit watermark will be separately embedded into the formant with the FE method and into the residue with the CD method.

Step 3 All the formants including the enhanced formant and the other formants will be synthesized with the residue that containing watermark to obtain current frame.

Step 4 Watermarked signal, $y(n)$, is constructed by all watermarked frames using non-overlapping and adding function.

Figure 6.11(b) has a block diagram of the watermark extraction process. Watermarks are extracted as follows.

Step 1 We apply the same procedures as those in the embedding process to the watermarked signal, $y(n)$, to obtain the formants and residue of each frame.

Step 2 One bit watermark will be separately detected from the formants with the FE method and residue with the CD method.

Step 3 The above procedure is repeated for all frames so that the whole extracted watermark signal, $\hat{s}_f(m)$, of the FE method, and, $\hat{s}_r(m)$, of the CD method can be obtained.

Step 4 The detected watermark signal, $\hat{s}(m)$, for the hybrid watermarking can be co-calculated with $\hat{s}_f(m)$ and $\hat{s}_r(m)$.

6.2.3 Evaluations for hybrid watermarking method

In this chapter, we evaluated the proposed hybrid scheme with respect to inaudibility and robustness. The database is the same as those used in previous chapters. The order of LP analysis was chosen as 10 and Ω_{e0} for embedding "0" was 2.0. Embedded watermarks was

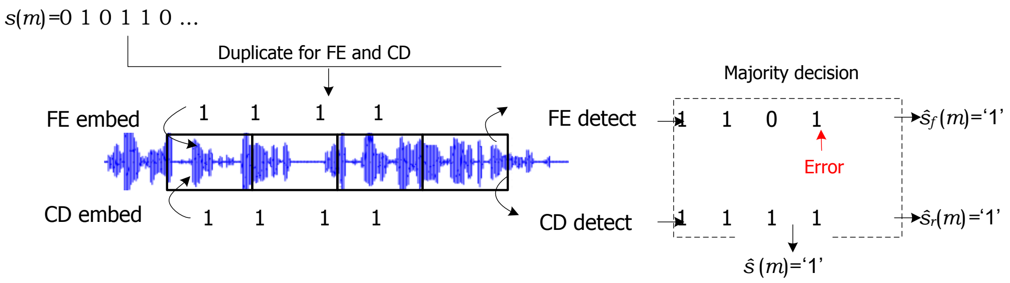


Figure 6.12: Extraction of one bit watermark for FE, CD, and hybrid (FE-CD) based on majority decision, where extraction error may happen.

one sentence "How are you". To increase robustness, each one bit watermark was duplicated for every 4 frames for the FE and CD embedding, and then decided the watermark for the FE, CD, and hybrid (FE-CD) detection with a majority decision. An example of the embedding and detection for FE, CD, and hybrid (FE-CD) of one bit watermark is illustrated in Fig. 6.12. The bit rates for the proposed method were set as 1, 2, 4, 8, 16, 32, 64, 128, 256 bps based on the above embedding and detection rules.

Evaluations for inaudibility

In this chapter, we evaluated the inaudibility of the proposed hybrid method with LSD and PESQ. The evaluation results are plotted in Fig. 6.13. As we can see, the hybrid method could satisfy the criteria for LSD and PESQ at low bit rate. Additionally, the results of sound quality for single watermarking ("CD": embed watermarks to residue with cochlear delay method and leave the formants un-modified, "FE": embed watermarks to formants with formant enhancement method and leave the residue un-modified) are also given out. These results try to show how are the influences to sound when only residue or formants are modified. As we can see, formant enhancement method could satisfy the criterion (LSD and PESQ), and the hybrid method has almost the same sound quality when only the residue is embedded with watermarks with CD method. This indicated that the sound distortion of the watermarked signal was mainly caused by embedding watermarks to the residue, not by formant enhancement. However, we cannot conclude

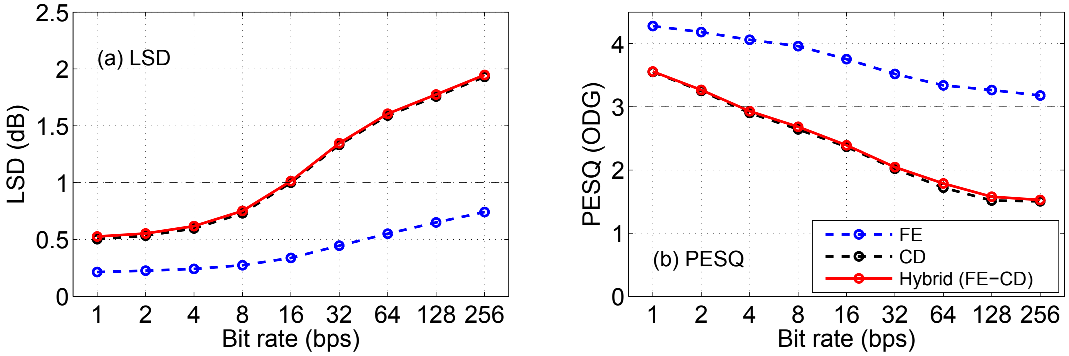


Figure 6.13: Evaluations of inaudibility for hybrid watermarking scheme: (a) LSD and (b) PESQ.

that cochlear delay method is audible [54], since CD was currently applied to residue, and residue contained so important informant for the synthesized speech that any modification to it would drastically degrade the sound quality of synthesized speech.

Evaluations for robustness

We first evaluated the robustness of the proposed method against normal extraction and speech codecs of G.711, G.726, and G.729. As shown in Fig. 6.14(a), watermarks could be successfully extracted with FE method and CD methods, that is to say, the hybrid method was feasible since one method did not affect the performance of the other method. From Figs. 6.14(b) and (c), since both FE and CD methods were robust against G.711 and G.726, the hybrid (FE-CD) method was also robust against these two speech codecs. For G.729, only FE method provided satisfactory BDR. Nonetheless, the BDR of hybrid (FE-CD) method could be refined with FE despite the low BDR of CD method. This further verified that the disadvantage of one watermarking method can be concealed by incorporating another watermarking method, and the robustness of the hybrid method could be increased in comparison with singe watermarking.

Second, we we evaluated the robustness of the proposed method against other processing. These were re-sampling at 24 kHz and 12 kHz, re-quantization with 24 bits and 8 bits, Scaling by 2.0 times, single 100-ms echo addition of 6 dB, speech analysis/synthesis

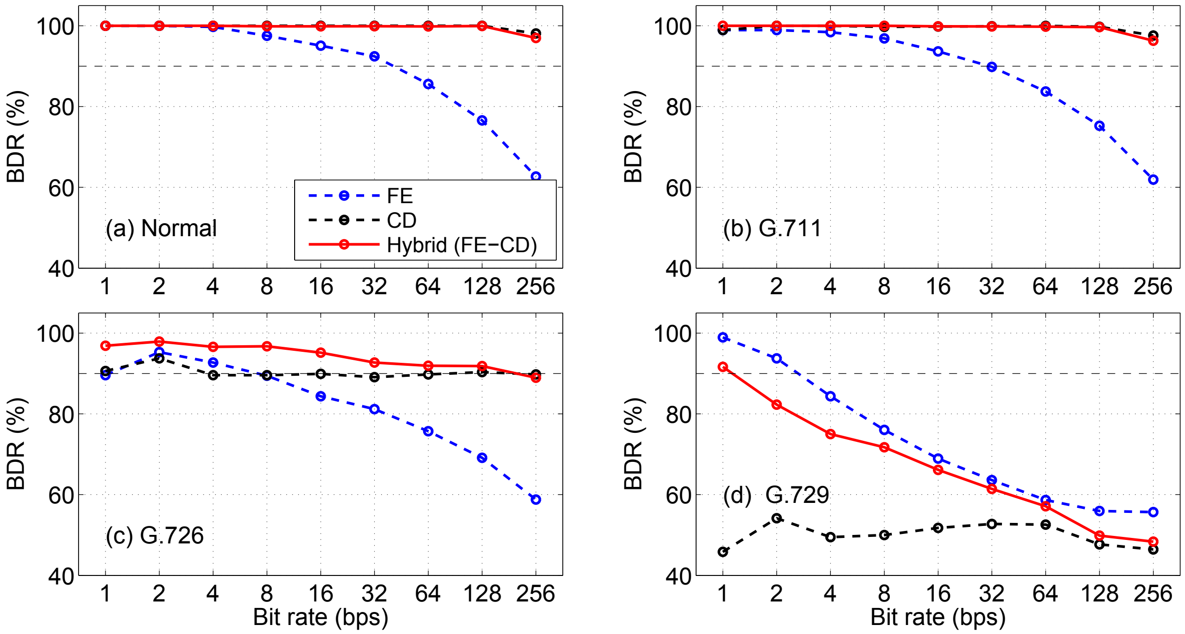


Figure 6.14: Evaluations of robustness for hybrid watermarking scheme against (a) Normal extraction, (b) G.711, (c) G.726, and (d) G.729.

by STFT and GTFB . The BDR results have been separately plotted in Figs. 6.15 and 6.16. Since the destroyed watermarks in one method could be recovered from the other method, the proposed hybrid watermarking method demonstrated good robustness against all these processing.

6.2.4 Discussion

In this chapter, a brief introduction and advantages of hybrid speech watermarking method were given. Since the hybrid watermarking method can achieve better performance by benefiting each employed single watermarking method, we proposed a hybrid watermarking method for speech signals based on the concepts of FE and CD. This hybrid method utilizes the source-filter model of speech production to separate the speech into the sound source and vocal tract filter so that the CD and FE watermarking methods can be separately applied. We investigated the inaudibility and robustness of the proposed hybrid method. The results showed that the proposed method could satisfy inaudibility.

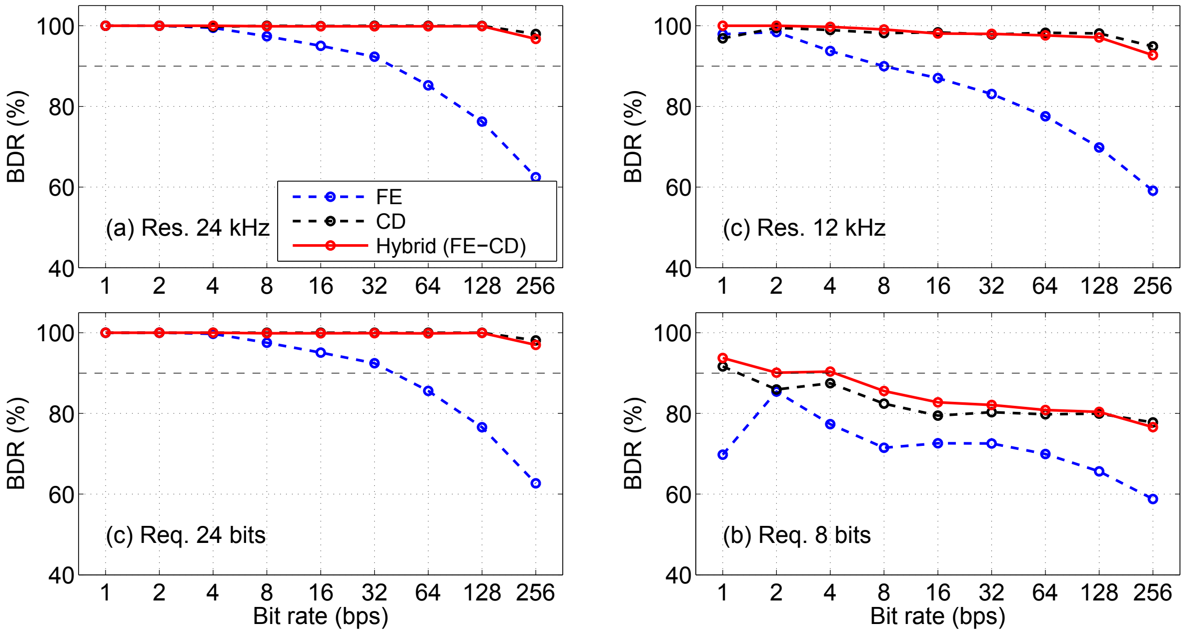


Figure 6.15: Evaluations of robustness for hybrid watermarking scheme against (a) re-sampling at 24 kHz and (b) 12 kHz and re-quantization with 24 bits and (c) 24 bits.

Moreover, the combination of FE and CD enabled the proposed method to benefit from both two methods for stronger robustness since watermarks could always be detected even one of them failed. These results verified that the proposed method could successfully achieve inaudibility and robustness.

6.3 Summary

In this chapter, the formant enhancement based watermarking was applied for tampering detection and hybrid watermarking.

For tampering detection, we constructed the whole tampering detection scheme. The detailed processing at the sender side and the receiver side were explained. The ability of the proposed scheme was evaluated according to inaudibility, robustness, and several kinds of tampering. The first two evaluations revealed that the proposed scheme could not only satisfy inaudibility but also provide good robustness. The evaluation against

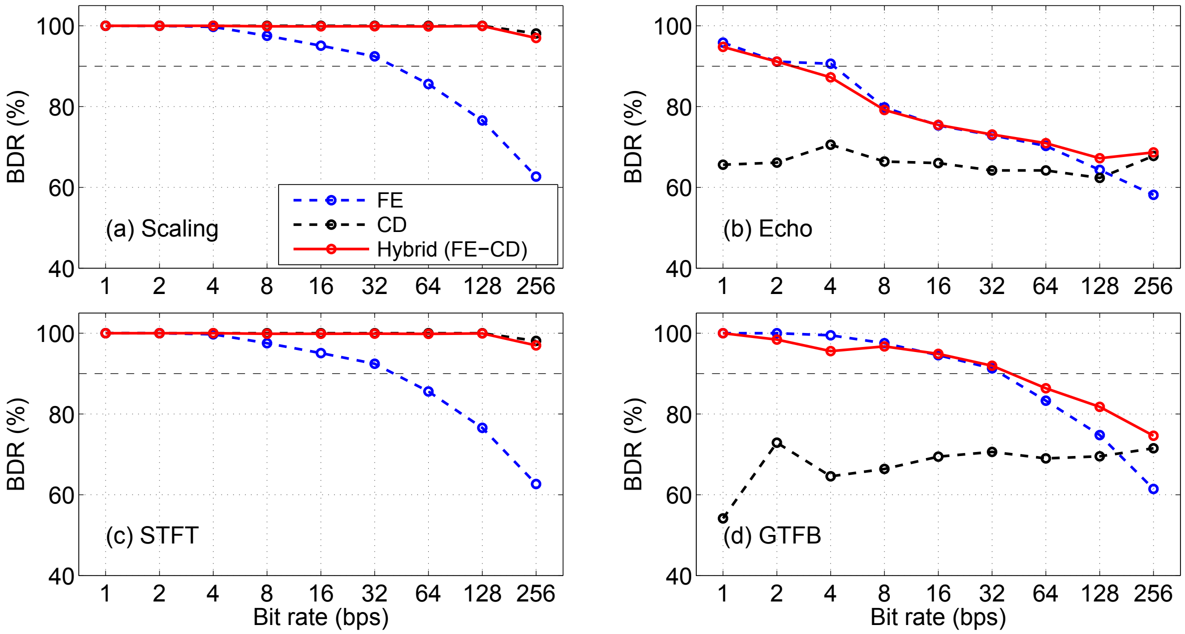


Figure 6.16: Evaluations of robustness against (a) Scaling (b) Echo, (c) STFT, and (d) GTFB.

several kinds of tampering results showed that when tampering has been made to the watermarked speech, watermarks in the tampered segment were destroyed. Moreover, the embedding bit rate of watermarks was 4 bps, each embedded bit was able to account for 0.25 s speech segment when locating the tampering. Based on the obtained results, the proposed scheme had the ability to detect tampering as well as check the originality of speech signals.

For hybrid method, the formant enhancement based watermarking and cochlear delay watermarking are combined together to be a hybrid method. This method was evaluated with respect to inaudibility and robustness. The evaluation results suggested that the robustness of the hybrid method can be improved compared with each single method, since the hybrid watermarking method could achieve better performance by benefiting each employed single watermarking method.

Chapter 7

Conclusions

In this chapter, we will conclude the work in this dissertation and discuss the future works.

7.1 Summary

Development in digital technologies have enabled speech to be used in many applications such as VoIP communications, digital forensics, and commercial investigation. As an important information carrier, speech signal contains significant value. However, with some speech processing tools, speech signal can be easily tampered, and now it is becoming difficult to confirm the originality of speech signals. The main motivation of this research is to protect speech signal and check the originality of speech by identifying whether there is tampering happened to the speech signal. Information hiding technique has been proposed as an efficient way to protect the speech signals. Watermark is considered as a kind of special information hidden in the speech signals. Our work focuses on protecting speech signals with watermarking methods.

Since watermarking method directly embeds watermarks into the speech signal, and the embedded watermarks can permanently exist and difficult to remove, tampering can be reliably detected with the watermarks. To be effective, watermarking methods should satisfy several requirements: (1) inaudibility to human auditory system, (2) blindness for watermark detection, (3) robustness against allowable speech processing and common attacks, and (4) fragility against tampering. The first three requirements are required for general watermarking methods, and the last one is an additional requirement when watermarking methods are used for tampering detection. However, it is proven to be difficult for watermarking methods to satisfy all these requirements simultaneously.

Our research aim is to solve the problem of tampering with information hiding and watermarking methods that can satisfy all the requirements. Under this research aim, our several objectives as specialized as follows:

First target Realize a general speech watermarking method that can satisfy all the first three requirements: inaudibility, blindness, and robustness.

Second target Employ the general watermarking method for speech tampering detection by exploring the fragility of watermarking method.

Third target Apply speech watermarking in other applications.

To achieve these targets, the basic knowledge of speech production, LP analysis, and speech parameters such as formant and LSFs have been reviewed in chapter 2. We believed that if the speech parameter could be slightly modified with suitable method, watermarks were possible to be inaudibly embedded into the speech signals. Therefore, the concept of formant tuning was introduced to our watermarking methods. Based on the source-filter model, the LP coefficients could provide accurate estimation of formants. In comparison to LP coefficients, the LSFs could not only represent the formants but also had several excellent properties, such as less sensitive to noise, deviation could be limited to the local spectral. Moreover, they were universal features in different speech codecs. Therefore, LSFs were selected as the carrier of watermarks, and the medications to LSFs for watermark embedding could be considered as a kind of tuning of formant. That's the reason why we named our whole watermarking concept as formant tuning.

To make the watermarking effective, we investigated how the formant could be estimated by LP analysis and the how was the relationship between LSFs and formants. After that, we proposed two concepts of speech watermarking. One was watermarking based on LSFs modifications with QIM. The QIM based watermarking has been considered as a promising method for digital watermarking since the implementation was easy, and a trade-off among the requirements of distortion introduced to the speech signal, robustness, and embedding capacity could always be achieved by adjusting the quantization step. The other concept was watermarking based on formant enhancement, this concept was mainly motivated by the research in the field of speech synthesis, where formants could be enhanced to improve speech quality, and such modifications did not cause perceptual distortion to the original speech. Therefore, if watermarks could be embedded through formant enhancement, it was possible to be imperceptible to human.

In chapter 4, we separately implemented the above two speech watermarking methods. In the LSFs and QIM based watermarking, we constructed the framework of watermarking by quantizing LSFs with QIM. Two quantizers were used to embed different watermarks bits "0" and "1". In the formant enhancement based watermarking, we investigated how to enhance a formant by directly closing up two LSFs. Different watermarks were embedded by enhancing different formants, after which different bandwidth relationships between the sharpest and the second sharpest formants were established. These different bandwidth relationships could be used to blindly extract watermarks. Since both methods were frame-based methods, we also implemented a frame synchronization scheme in this chapter.

In chapter 5, we evaluated two speech watermarking methods with inaudibly and robustness. For the LSFs and QIM based method, we took into consideration of the inaudibility and robustness that were influenced by the different quantization steps. The results from inaudibility evaluation revealed that the proposed method could satisfy inaudibility when quantization step was small. However, when subjected to the processing that could change the shape of waveform or the value of signal such as speech codecs, down-sampling, and low-bit re-quantization, this method was not robust. In the evaluations of formant enhancement based watermarking, we gave much consideration about the parameters such as the LP order and the Ω_{e0} when balancing the inaudibility and robustness. In this method, watermark embedding through formant enhancement did not cause severe degradation to the original speech quality, and the watermark extraction by identifying bandwidth relationship was able to tolerate slight distortions of frequency components caused by other processing. Therefore, in comparison with other watermarking methods, this method could achieve a good trade-off between inaudibility and robustness. Besides, we also investigated the influence of frame synchronization to watermarking method. From the evaluations, we found that the embedding of synchronization information did not cause severe distortion to the speech quality, and the synchronization information did not affect the extraction of watermarks.

Since the formant enhancement based watermarking method could satisfy the requirements of inaudibility, blindness, and robustness. We employed it for tampering detection and for hybrid watermarking in chapter 6. The tampering detection ability of the proposed scheme was evaluated against several kinds of tampering. The evaluation results showed that when tampering has been made to the watermarked speech, watermarks in the tampered segment were destroyed. Therefore, the proposed scheme was fragile against tampering, and it had the ability to detect tampering as well as check the originality of speech signals. In the hybrid watermarking method, evaluation were carried out concerning inaudibility and robustness. The results suggested that the robustness of the hybrid method could be improved compared with each single method, since the disadvantage of one watermarking method could be concealed by the other watermarking method.

Based on all these evaluations, we can conclude that the three targets in this research can be realized as follows. Firstly, we realized a general speech watermarking method, i.e., the formant enhancement based watermarking that could satisfy all the first three requirements: inaudibility, blindness, and robustness. Secondly, when we applied the formant enhancement based watermarking for tampering detection, our method was fragile and it could successfully detect the tampering, which meant our method could effectively prevent the tampering and check the originally of speech signal. Thirdly, we also applied the formant enhancement based watermarking for hybrid watermarking, the evaluation results suggested that the performances, especially robustness, of the hybrid method could be improved compared with each single method.

7.2 Contributions

This research focuses on solving the problems of tampering in speech signals with information hiding and watermarking methods. Four requirements of (1) inaudibility, (2) blindness, (3) robustness, and (4) fragility were addressed when designing the effective watermarking method. As mentioned before, three objectives have been achieved in this research. Corresponding to these objectives, the main contributions of this research can be summarized as follow:

(1) The source-filter model for speech production and the relationship between formant and LSFs was investigated. The concept of formant enhancement in the field of speech synthesis was introduced for speech watermarking.

(2) A formant enhancement based watermarking was proposed. The evaluation results showed that compared with other methods, this method could simultaneously satisfy inaudibility, blindness, and robustness. Besides, this watermarking method explored the source-filter model of speech production. Since the source-filter model of speech production is widely used in many voice coders, such as algebraic CELP (A-CELP), multi-pulse CELP (M-CELP), and conjugate structure CELP (CS-CELP), this watermarking method can be easily transplanted to these speech codecs by directly embedding watermarks into LSFs, which means it is possible for these speech codecs to automatically realize the protection of speech signals with watermarking inside.

(3) A tampering detection scheme based on the above watermarking method was proposed. The evaluation results reveal that the proposed scheme could not only satisfy inaudibility but also provided good robustness. Moreover, the proposed method was capable of locating the tampering in time-domain at sufficient precision with its fragility, and its detection ability was not degraded even speech processing exist. Therefore, the proposed method could effectively detect tampering in speech signals, which means it can effectively check the originality of speech signals.

(4) A hybrid speech watermarking method was proposed by incorporating the formant enhancement based watermarking and the cochlear delay based watermarking. The

evaluation results showed that the destroyed watermarks in one method can be recovered from the other method. Therefore, the proposed hybrid method demonstrated very good robustness. Besides, in previous works, there is no work that has dealt with hybrid watermarking for speech signal, the proposed hybrid method provided a basic model to show how two watermarking can be implemented together by utilizing the intrinsic characteristics of the process of speech production.

7.3 Future work

According to the summary of this research, we can conclude the effective of our work. However, our work left something to be desired:

(1) In the evaluations of formant enhancement based watermarking, robustness against G.723.1, MELP, and re-quantization at lower bits needs to be improved.

(2) Our proposed hybrid watermarking is a combination of two watermarking methods. The watermark extraction for the hybrid method is based on the extraction results of two single watermarking methods. Currently, we use majority decision to calculate the results. However, since each single watermarking method has different robustness against the same speech processing, more suitable method should be used to decide the extraction result for the hybrid method.

(3) Formants are essential for the speech quality and speech perception. In our method, we enhance the formants for watermarking embedding. It is still not known whether the valleys in speech envelope affect the speech quality a lot and how they affect the speech quality. In the next step, we will check the influence of valleys to speech quality and we will consider if the valley could be used for inaudible speech watermarking.

References

[1] *http://en.wikipedia.org/wiki/Voice-over-IP*

[2] S. L. Garfinkel, "Digital forensics research: The next 10 years," *Digital Investigation: The International Journal of Digital Forensics & Incident Response*, vol. 7, pp. 64–73, 2010.

[3] R. G. Cole and J. H. Rosenbluth, "Voice over IP performance monitoring," *ACM SIGCOMM Computer Communication Review*, vol. 31, no. 2, pp. 9–24, 2001.

[4] S. Lingfen and E.C. Ifeachor, "Voice quality prediction models and their application in VoIP networks," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 809–820, 2006.

[5] R. Poisel and S. Tjoa, "Forensics Investigations of Multimedia Data: A Review of the State-of-the-Art," *Proc. IT Security Incident Management and IT Forensics (IMF)*, pp. 48–61, 2011.

[6] S. Milani , M. Fontani , P. Bestagini , M. Barni , A. Piva , M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Trans. Signal and Information Processing*, vol. 1, pp. 1–18, 2012.

[7] I. W. Evett, G. Jackson, J. A. Lambert, and S. McCrossan, "The impact of the principles of evidence interpretation on the structure and content of statements," *Science & Justice*, vol. 40, pp. 233–239, 2000.

[8] J. G. Rodrguez, A. Drygajlo, D. R. Castro, M. G. Gomar, and J. O. Garca, “Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition,” *Computer Speech and Language*, pp. 331–355, 2006.

[9] I. W. Evett, “Towards a uniform framework for reporting opinions in forensic science case-work,” *Science & Justice*, vol. 38, pp. 198–202, 1998.

[10] I. J. Cox, G. Dorr, and T. Furon, “Watermarking is not cryptography,” *Lecture Notes in Computer Science*, vol. 4283, pp. 1–15, Springer, 2006.

[11] *http://www.garykessler.net/library/crypto.html/hash*.

[12] C. D. Roover, C. D. Vleeschouwer, F. Lefebvre, and B. Macq, “Robust video hashing based on radial projections of key frames,” *IEEE Trans. Signal Processing, Supplement on Secure Media* vol. 53, pp. 4020–4037, 2005.

[13] V. Monga and K. Mhcak, “Robust image hashing via non-negative matrix factorizations,” *Proc. ICASSP*, vol. II, pp. 225–228, 2006.

[14] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, “Techniques for Data Hiding,” *IBM Systems Journal*, vol. 35, no. 3/4, pp. 313–336, 1996.

[15] A. P. F. Petitcolas, R. J. Anderson, and M. G. Kuhn, “Information hiding-a survey,” *Proc. IEEE*, vol. 87, no. 7, pp. 1062–1078, 1999.

[16] M. D. Swanson, M. Kobayashi, and A. H. Tewfik, “Multimedia data-embedding and watermarking technologies,” *Proc. IEEE*, vol. 86, no. 6, pp. 1064–1087, 1998.

[17] C. I. Podilchuk and E. J. Delp, “Digital watermarking: algorithms and applications,” *IEEE Signal Processing Magazine*, vol. 18, no. 4, pp. 33–46, 2002.

[18] S. Khurana, “Watermarking and Information-Hiding,” *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol. 2, no. 4, pp. 1679–1681, 2011.

[19] B. Li, J. He, J. Huang, and Y. Shi, " A Survey on Image Steganography and Steganalysis," *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 2, no. 2, pp. 142–172, Apr. 2011.

[20] R. G. van Schyndel, A. Z. Tirkel, and C. F. Osborne, "A Digital Watermark," *Proc. IEEE International Conference on Image*, vol. II, pp. 86–90, 1994.

[21] F. Hartung and M. Kutter, "Multimedia Watermarking Techniques," *Proc. IEEE*, vol. 87, pp. 1079–1107, 1999.

[22] F. Cayre, C. Fontaine, and T. Furon, "Watermarking security: Theory and practice," *IEEE Trans. Signal Processing, Supplement on Secure Media*, vol, 53, pp. 3976–3987, 2005.

[23] P. K. Dhar, M. I. Khan, and J. Kim, "A new audio watermarking system using discrete Fourier transform for copyright protection," *International Journal of Computer Science and Network Security*, vol. 10, no. 6, 2010.

[24] Y. H. Chen and H. C. Huang, "Coevolutionary genetic watermarking for owner identification," *Neural Computing and Applications*, vol. 26, no. 2, pp. 291–298, Feb. 2015.

[25] S. Wu, J. Huang, D. Huang, and Y. Q. Shi, "Efficiently self-synchronized audio watermarking for assured audio data transmission," *IEEE Trans. broadcasting*, vol. 51, no. 1, pp. 69–76, 2005.

[26] B. Lei, I. Y. Soon, F. Zhou, Z Li, and H. Lei, "A robust audio watermarking scheme based on lifting wavelet transform and singular value decomposition," *Signal Processing*, vol. 92, no. 9, pp. 1985–2001, 2012.

[27] X. Wang and H. Zhao, "A novel synchronization invariant audio watermarking scheme based on DWT and DCT," *IEEE Trans. Signal Processing*, vol. 54, no. 12, pp. 4835–4840, 2006.

[28] S. V. Dhavale, R. S. Deodhar, D. Pradhan, and L. M. Patnaik, " High payload adaptive audio watermarking based on cepstral feature modification," *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 5, no. 4, pp. 586–602, Oct. 2014.

[29] G. C. Langelaar, I. Setyawan, and R. L. Lagendijk, "Watermarking digital image and video data. A state-of-the-art overview," *IEEE Signal Processing Magazine*, vol. 17, no. 5, pp. 20–46, 2000.

[30] E. T. Lin, A. M. Eskicioglu, R. L. Lagendijk, and E. J. Delp, "Advances in Digital Video Content Protection," *Proc. IEEE*, vol. 93, no. 1, pp. 171–183, 2005.

[31] P. Bassia and I. Pitas, "Robust audio watermarking in the time domain," *Proc. EUSIPCO*, pp. 25–28, 1998.

[32] H. J. Kim and Y. H. Choi, "A novel echo-hiding scheme with backward and forward kernels," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 885–889, 2003.

[33] Y. Erfani and S. Siahpoush, "Robust audio watermarking using improved TS echo hiding," *Digital Signal Processing*, vol. 19, no. 5, pp. 809–814, 2009.

[34] B. S. Ko, R. Nishimura, and Y. Suzuki, "Time-spread echo method for digital audio watermarking," *IEEE Trans. Multimedia*, vol. 7, no. 2, pp. 212–221, 2005.

[35] I. J. Cox, J. Killian, T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Trans. Image Processing*, vo1. 6, no. 12, pp. 1673–1687, 1997.

[36] H. S. Malvar and A. F. Florncio, "Improved spread spectrum: A new modulation technique for robust watermarking," *IEEE Trans. Signal Processing*, vol. 51, no. 4, pp. 898–905, 2003.

[37] F. Hartung, J. K. Su, and B. Girod, “Spread spectrum watermarking: Malicious attacks and counter attacks,” *Proc. SPIE Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 147–158, 1999.

[38] L. Boney, H. H. Tewfik, and K. H. Hamdy, “Digital watermarks for audio signals,” *Proc. International Conference on Multimedia Computing and Systems (ICMCS)*, pp. 473–480, 1996.

[39] D. Kirovski and H. Malvar, “Robust spread spectrum audio watermarking,” *Proc. ICASSP*, vol. 3, pp. 1345–1348, 2001.

[40] Q. Cheng and J. Sorensen, “Spread spectrum signaling for speech watermarking,” *Proc. ICASSP*, vol. V, pp. 1337–1340, 2001.

[41] B. C. J. Moore, “An introduction to the psychology of hearing,” *Academic Press*, Sixth edition, 1997.

[42] M. Fallahpour and D. Megias, “High capacity logarithmic audio watermarking based on the human auditory system,” *IEEE International Symposium on Multimedia (ISM)* , pp. 28-31, 2012.

[43] M. D. Swanson, B. Zhu, A. H. Tewfik, and L. Boney, “Robust audio watermarking using perceptual masking,” *Signal Processing*, vol. 66, no. 3, pp. 337–355, 1998.

[44] F. Battisti, M. Carli, and C. Rinaldi, “Perceptual audio watermarking driven by Human Auditory System,” *Proc. International Symposium on Signals, Circuits and Systems (ISSCS)*, pp. 1–4, 2013.

[45] M. Unoki and D. Hamada, “Method of digital-audio watermarking based on cochlear delay characteristics,” *International Journal of Innovative Computing, Information and Control*, vol. 6, no.(3(B)), pp. 1325–1346, 2010.

[46] B. Chen and G. W. Wornel, “Quantization index modulation: a class of provably good methods for digital watermarking and information embedding,” *IEEE Trans. Information Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[47] B. Chen and G. W. Wornell, “Preprocessed and postprocessed quantization index modulation methods for digital watermarking,” *Proc. SPIE Security and Watermarking of Multimedia Contents*, vol. 3971, 2000.

[48] C. Wu and C. Jay Kuo, “Fragile speech watermarking based on exponential scale quantization for tamper detection,” *Proc. ICASSP*, vol. IV, pp. 3305–3308, 2002.

[49] J. Sang and M. S. Alam, “Fragility and robustness of binary phase only filter based fragile/semi-fragile digital image watermarking,” *IEEE Trans. Instrumentation and Measurement*, vol. 57, no. 3, pp. 595–606, 2008.

[50] C. M. Park, D. Thapa, and G. N. Wang, “ Speech authentication system using digital watermarking and pattern recovery,” *Pattern Recognition Letters*, vol. 28 pp. 931–938, 2008.

[51] M. Celik, G. Sharma, and A. M. Tekalp, “Pitch and duration modification for speech watermarking,” *Proc. ICASSP*, vol. II, pp. 17–20, 2005.

[52] C. P. Wu and C. C. J. Kuo, “Fragile speech watermarking based on exponential scale quantization for tamper detection,” *Proc. ICASSP*, vol. IV, pp. 3305–3308, 2002.

[53] M. Unoki, K. Imabeppu, D. Hamada, A. Haniu, and R. Miyauchi, “Embedding limitations with digital-audio watermarking method based on cochlear delay characteristics,” *Journal of Information Hiding and Multimedia Signal Processing (JIHMSP)*, vol. 2, no. 1, pp. 1–23, 2011.

[54] M. Unoki and R. Miyauchi, “Reversible watermarking for digital audio based on cochlear delay characteristics,” *Proc. International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP)*, pp. 314–317, 2011.

[55] M. Unoki and R. Miyauchi, “Detection of tampering in speech signal with inaudible watermarking technique,” *Proc. International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP)*, pp. 118–121, 2012.

[56] ITU-T, “40, 32, 24, 16 kbit/s Adaptive Differential Pulse Code Modulation (ADPCM) Technical Report G.726,” *International Telecommunications Union*, Geneva, 1990.

[57] *http://www.itu.int/rec/T-REC/en.*

[58] ITU-T, “Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP) Technical Report G.729,” *International Telecommunications Union*, Geneva, 1996.

[59] L. J. Xin, L. Z. Ming, and L. Hao, “A CELP-speech information hiding algorithm based on vector quantization,” *Proc. Information Assurance and Security (IAS)*, pp. 75–78, 2009.

[60] B. Geiser, P. Jax, and P. Vary, “Artificial bandwidth extension of speech supported by watermark transmitted side information,” *Proc. INTERSPEECH*, pp. 1497–1500, 2005.

[61] B. Geiser and P. Vary, “Backwards compatible wideband telephony in mobile networks: CELP watermarking and bandwidth extension,” *ICASSP*, pp. 533–536, 2007.

[62] G. Fant, “Speech Sounds and Features,” *MIT Press*, 1973.

[63] J. Butler and H. Wakita, “Articulatory constraints on vocal tract area functions and their acoustic implications,” *STL Research Reports*, pp. 88–94, 1982.

[64] T. Chiba and M. Kajiyama, “The Vowel: Its nature and structure,” Tokyo: Tokyo-Kaiseikan Pub. Co., Ltd, 1942.

[65] G. Fant, “Acoustic theory of speech production: with calculations based on X-Ray studies of russian articulations,” Volume 2 of Description and analysis of contemporary standard Russian. Hague, The Netherlands: Mouton. pp. 15–90. ISSN 0070-3826 (1960).

[66] G. Fant, “Acoustic theory of speech production,” Mouton De Gruyter. ISBN 90-279-1600-4.

[67] F. Itakura and S. Saito, “Digital filtering techniques for speech analysis and synthesis,” *Proc. International Congress on Acoustics (ICA)*, 25(C-1), 261–264, 1971.

[68] R. McAuley and T. F. Quatiery, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, pp. 744–754, 1986.

[69] M. R. Schroeder and B. S. Atal, “Code-excited Linear Prediction (CELP): High-quality Speech at Very Low Bit Rates,” *Proc. ICASSP*, vol. 10, pp. 937–940, 1985.

[70] B. S. Atal, “Predictive coding of speech at low bit rates,” *IEEE Trans. Communications*, vol. 30, pp. 600–614, 1982.

[71] P. Mermelstein, “Articulatory model for the study of speech production,” *Journal of the Acoustical Society of America*, vol. 53, pp. 1070–1082, 1973.

[72] B. Galantucci, C. A. Fowler, and M. T. Turvey, “The motor theory of speech perception reviewed,” *Psychon. Bull. Rev.*, vol. 13, pp. 361–377, 2006.

[73] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[74] B. S. Atal, “The history of linear prediction,” *IEEE Signal Processing Magazine*, vol. 23, pp. 154–161, 2006.

[75] J. D. Markel and A. H. Gray, “Linear prediction of speech,” *Springer Verlag*, 1976.

[76] B. S. Atal and M. R. Schroeder, “Predictive coding of speech signals and subjective error criteria,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 27, pp. 247–254, 1979.

[77] N. Jayant, J. Johnston, and R. Safranek, “Signal compression based on models of human perception,” *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1421, 1993.

[78] ITU-T, “Pulse Code Modulation (PCM) of Voice Frequencies, Technical Report G.711,” *International Telecommunications Union*, Geneva, 1993.

[79] ITU-T, “5-, 4-, 3- and 2-bits per Sample Embedded Adaptive Differential Pulse Code Modulation (ADPCM), Technical Report G.727,” *International Telecommunications Union*, Geneva, 1990.

[80] B. S. Atal and J. R. Remde, “A new model of LPC excitation for producing natural-sounding speech at low bit rates,” *Proc. ICASSP*, pp. 614–617, 1982.

[81] B. S. Atal, “High-quality speech at low bit rates, Multi-pulse and stochastically excited linear predictive coders,” *Proc. ICASSP*, pp. 1681–1684, 1986.

[82] R. Salami et al., “Design and description of CS-ACELP, A toll quality 8 kb/s speech coder,” *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 116–130, 1998.

[83] F. Soong and B. H. Juang, “ Line spectral pair (LSP) and speech data compression,” *Proc. ICASSP*, pp. 37–40, 1984.

[84] T. Bckstrm and C. Magi, “Properties of line spectrum pair polynomials A review,” *Elsevier Signal Processing*, vol. 86, pp. 3286–3298, 2006.

[85] F. Itakura, “Line spectrum representation of linear predictor coefficients of speech signals,” *Journal of the Acoustical Society of America*, vol. 57. no. 537(A), pp. 35–35, 1975.

[86] B. Kleijn, T. Bckstrm, and P. Alku, “On line spectral frequencies,” *IEEE Signal Processing Letter*, vol. 10, pp. 75–77, 2003.

[87] N. Sugamura and N. Farvardin, " Quantizer design in LSP speech analysis-synthesis," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 432–440, 1988.

[88] M. H. Johnson, "Line spectral frequencies are the poles and zeros of a discrete matched-impedance vocal tract model," *Journal of the Acoustical Society of America*, vol. 108, no. 1, pp. 457-46, 2000.

[89] T. Bckstrm, P. Alku, T. Paatero, and B. Kleijn, "A time-domain interpretation for the LSP decomposition," *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 554–560, 2004.

[90] S. W. Lang and J. H. McClellan, "A simple proof of stability for all-pole linear prediction models," *Proc. IEEE* , vol. 67, pp. 860–861, 1979.

[91] F. Soong and B. -H. Juang, "Optimal quantization of LSP parameters," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 15–24, 1993.

[92] N. Sugamura and F. Itakura, "Speech data compression by LSP speech analysis-synthesis technique," *Trans. IEICE*, vol. J64, no.8, pp. 599–606, 1981, (in Japanese).

[93] G. S. Kang and L. J. Fransen, "Application of line-spectrum pairs to low-bit-rate speech coders," *Proc. ICASSP*, pp. 731–734, 1995.

[94] K. Takeda et al, "Speech database user's manual," *ATR Technical Report* TR-I-0028, 2010.

[95] B. Chen and G. W. Wornell, "An information-theoretic approach to the design of robust digital watermarking systems," *Proc. ICASSP*, pp. 2061–2064, 1999.

[96] J. H. Chen and A. Gersho, "Adaptive postfiltering for quality enhancement of coded speech," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 59–71, 1995.

[97] H. Brouckxon, W. Verhelst, and B. D. Schuymer, “Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments,” *Proc. INTERSPEECH*, pp. 557–560, 2008.

[98] Y. Ueda, S. Hario, and T. Sakata, “Formant based speech enhancement for listening speech sound in noisy place,” *Proc. ICSV*, pp. 515–522, 2008.

[99] HTS, “HMM-based speech synthesis system,” *http://hts.sp.nitech.ac.jp.*

[100] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, “The HMM-based speech synthesis system (HTS) version 2.0,” *Proc. ISCA Workshop on Speech Synthesis*, pp. 294-299, 2007.

[101] Z. H. Ling, Y. J. Wu, Y. P. Wang, L. Qin, and R. H. Wang, “USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method,” *Proc. Blizzard Challenge Workshop*, 2006.

[102] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, “Comparison of formant enhancement methods for HMM based speech synthesis,” *Proc. ISCA Speech Synthesis Workshop*, 2010.

[103] Recommendation ITU-T P.800, “Methods for subjective determination of transmission quality,” *International Telecommunication Union*, 1996.

[104] B. Chen and G.W. Wornell, “Dither Modulation: A new approach to digital watermarking and information embedding,” *Proc. SPIE Security and Watermarking of Multimedia Contents*, vol. 3657, pp. 342–353, 1999.

[105] A. Gray, Jr., and J. Markel, “Distance measures for speech processing,” *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 380–391, 1976.

[106] Y. Hu and P. C. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[107] *https://en.wikipedia.org/wiki/G.723.1*

[108] *https://en.wikipedia.org/wiki/Mixed-excitatio-linear-prediction*

[109] M. Steinebach , F. A. P. Petitcolas , F. Raynal , J. Dittmann , C. Fontaine , S. Seibel , N. Fates and L. C. Ferri, "Stirmark benchmark: Audio watermarking attacks," *Proc. Int. Conf. on Information Technology: Coding and Computing*, pp. 49–54, 2001.

[110] Information hiding and its criteria for evaluation
http:// www.ieice. org/iss/emm/ihc/en/index.php

[111] H. Kawahara, I. Masuda-Kasuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a reptitive Structure in Sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[112] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[113] T. Saitou, M. Goto, M. Unoki, and M. Akagi, "Vocal Conversion from speaking voice to singing voice using STRAIGHT," *Proc. Synthesis of Singing Challenge, Special Session at INTERSPEECH*, 2007.

[114] H. Kawahara, H. Banno, T. Irino and P. Zolfaghari, "ALGORITHM AMALGAM: Morphing waveform based methods, sinusoidal models and STRAIGHT," *Proc. ICASSP*, pp. 13–16, 2004.

[115] *http://en.wikipedia.org/wiki/Digital-forensics*

[116] M. C. Stamm and K. J. R. Liu, "Forensic detection of image manipulation using statistical intrinsic fingerprints," *IEEE Trans. Information Forensics and Security*, vol. 5, no. 3, pp. 492–506, 2010

[117] F. Golshan and K. Mohammadi, "A hybrid intelligent SVD-based digital image watermarking," *Proc. Systems Engineering (ICSEng)*, pp. 137–141, 2011.

[118] E. Chrysochos, V. Fotopoulos, M. Xenos, and A. N. Skodras, "Hybrid watermarking based on chaos and histogram modification," *Journal of Signal, Image and Video Processing*, vol. 8, no. 5, pp 843-857, 2014.

[119] P. W. Chan, M. R. Lyu, and R. T. Chin, "A novel scheme for hybrid digital video watermarking: approach, evaluation and experimentation," *IEEE Trans. Circuit and system for video technology*, vol. 15, no. 12, pp. 1638–1649, 2005.

[120] A. Dhawan and S. K. Mitra, "Hybrid audio watermarking with spread spectrum and singular value decomposition," *Proc. India Conference*, pp. 11–16, 2008.

[121] B. Y. Lei, K. T. Lo, and H. j. Lei, "Hybrid SVD-based audio watermarking scheme," *Proc. Communications, Circuits and Systems (ICCCAS)*, pp. 428–432, 2010.

Publications

Journal papers

[1] Shengbei Wang and Masashi Unoki, “Speech Watermarking Method based on Formant Tuning,” IEICE Trans. INF. & SYST., Special Section on Enriched Multimedia, vol. E98-D, no. 1, pp. 29-37, Jan., 2015.

[2] Shengbei Wang, Ryota Miyauchi, Masashi Unoki, and Nam Soo Kim, “Tampering Detection Scheme for Speech Signals using Formant Enhancement based Watermarking,” Journal of Information Hiding and Multimedia Signal Processing (JIHMSP), 2015, (accepted).

International conferences

[3] Erick Christian Garcia Alvarze, Shengbei Wang, and Masashi Unoki, “An Automatic Watermarking in CELP Speech Codec based on Formant Tuning,” Proc. 11^{th} Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2015), Australia, 2015, (accepted).

[4] Shengbei Wang, Masashi Unoki, and Nam Soo Kim, “Formant Enhancement based Speech Watermarking for Tampering Detection,” Proc. 15^{th} Annual Conference of International Speech Communication Association (Interspeech2014), pp. 1366-1370, Singapore, 2014.

[5] Shengbei Wang and Masashi Unoki, "Watermarking of Speech Signals based on Formant Enhancement," Proc. 22^{nd} European Signal Processing Conference (EUSIPCO2014), pp. 1257-1261, Portugal, 2014.

[6] Masashi Unoki, Jessada Karnjana, Shengbei Wang, Nhut Minh Ngo, and Ryota Miyauchi, "Comparative Evaluations of Inaudible and Robust Watermarking for Digital Audio Signals," Proc. 21^{st} International Congress on Sound and Vibration (ICSV2014), China, 2014.

[7] Shengbei Wang and Masashi Unoki, "Hybrid Speech Watermarking based on Formant Enhancement and Cochlear Delay," Proc. 10^{th} Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2014), pp. 272-275, Japan, 2014.

[8] Shengbei Wang and Masashi Unoki, "Watermarking Method for Speech Signals based on Modifications to LSFs," Proc. 9^{th} Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP2013), pp. 283-286, China, 2013.

[9] Nhut Minh Ngo, Shengbei Wang, Masashi Unoki, "Method of Digital-audio Watermarking Based on Cochlear Delay in Sub-bands," Proc. 27^{th} International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC 2012), D-W1-03, Japan, 2012.

Domestic conferences

[10] Shengbei Wang and Masashi Unoki, "Study on Hybrid Speech Watermarking based on Formant Enhancement and Cochlear Delay," IEICE Tech. Rep., vol. 114, no. 316, EMM2014-65, pp. 53-58, Nov., 2014.

[11] Shengbei Wang and Masashi Unoki, "Study of Watermarking of Speech Signals based on Formant Enhancement," Proc. 2014 ASJ, 1-R5-2, pp. 379-382, Mar., 2014.

[12] Jessada Karnjana, Masashi Unoki, Shengbei Wang, Nhut Minh Ngo, and Ryota Miyauchi, "Comparative Evaluations of Inaudible and Robust Watermarking Methods for Digital Audio Signals," IEICE Tech. Rep., vol. 113, no. 480, EMM2013-110, pp. 63-68, Mar., 2014.

[13] Shengbei Wang and Masashi Unoki, "Study on Speech Watermarking based on Formant Enhancement," IEICE Tech. Rep., vol. 113, no. 415, EMM2013-98, pp. 57-62, Jan., 2014.

[14] Shengbei Wang and Masashi Unoki, "Study of Speech Watermarking based on Modifications to LSFs by DM-QIM," Proc. 2013 ASJ Autumn meeting, 1-P-31c, pp. 359-362, Sep., 2013.

[15] Shengbei Wang and Masashi Unoki, "Study on Speech Watermarking based on Modifications to LSFs for Tampering Detection," IEICE Tech. Rep., vol. 113, no. 138, EMM2013-34, pp. 233-238, Jul., 2013.

[16] Shengbei Wang and Masashi Unoki, "Study on Digital Watermarking for Speech Signal based on LSFs Modification," IEICE Tech. Rep., vol. 113, no. 66, EMM2013-7, pp. 37-42, May, 2013.

[17] Masashi Unoki, Shengbei Wang, and Ryota Miyauchi, "Detection of Tampering in Speech Signals with Digital-audio Watermarking Technique based on Cochlear Delay," IEICE Tech. Rep., vol. 112, no. 420, EMM2012-102, pp. 65-70, Jan., 2013.

[18] Nhut Minh Ngo, Shengbei Wang, and Masashi Unoki, "Method of Digital-audio Watermarking based on Cochlear Delay in Sub-bands," IEICE Tech. Rep., vol. 112, no. 171, EA2012-57, pp. 19-24, Aug., 2012.

[19] Shengbei Wang, Nhut Minh Ngo, and Masashi Unoki, "Digital-audio Watermarking based on Cochlear Delay in Sub-bands," Proc. 2012 ASJ Autumn meeting, 2-9-2, pp. 629-623, Sep., 2012.