Doctoral Dissertation

Content and Privacy Protection Methods for Secure Speech Communication and Its Applications

Candy Olivia MAWALIM

Supervisor: Professor Masashi UNOKI

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
[Information Science]
March, 2022

Abstract

Various forms of speech are utilized throughout social media. Advanced speech technology, such as voice conversion techniques and speech synthesis, can synthesize or clone speech entirely as a human voice. Distributing users' speech publicly on a social network without privacy measures affects the security of speech technology and privacy protection. Without protection, speech samples on the internet could be used for theft of personally identifiable information, fraud, and/or authentication of the ASV system for criminal purposes. Therefore, there must be a solution to the emerging threat of unauthenticated speech signals, such as synthesizing, cloning, and speech conversion.

Speech information hiding (SIH) is one of the approaches for promoting secure speech communication, which is also the main part of this study. Information-hiding-based methods preserve the privacy and security of speech data by imperceptibly embedding particular information that needs to be hidden. SIH has at least three requirements: inaudibility (manipulation does not cause distortion perceivable by the human auditory system), blindness (accurate detection without the original signal), and robustness against common signal processing operations. Although each existing method has advantages, they have shortcomings and need improvement, especially in balancing the trade-off between inaudibility and robustness.

Another approach to improve the trade-off between inaudibility and robustness is considering the features used in speech codecs. Speech codecs are widely applied before speech is transmitted through a communication channel. Thus, using features in speech codecs for speech information hiding improves robustness. Line spectral frequencies (LSFs) are used as features in speech codecs with several speech watermarking methods. LSFs can be directly modified in accordance with a particular speech codec quantization method or manipulated accordingly to control speech formants for representing hidden information.

We investigate a parameter that affects the formation of auditory images, namely the McAdams coefficient, for the feature of SIH in this study. The modification of the McAdams coefficient is useful for adjusting frequency harmonics in audio signals. It has also been introduced for de-identifying or anonymizing speech signals. Since the McAdams coefficient is related to the adjustment of frequency harmonics (related to LSFs), we hypothesize that this coefficient is suitable for speech watermarking.

Another novelty presented in this study is that we propose a speech watermarking method based on a machine learning model. Studies on digital image watermarking based on machine learning models have shown impressive results. However, due to the higher complexity of speech than image data, machine learning models for speech watermarking have not been widely explored. We constructed a machine-learning-based blind detection model by using a binary classification task based on a random forest algorithm (hereafter, we refer to this model as a random forest classifier). The results indicate that our method satisfies the speech watermarking requirements with a 16-bps payload under normal conditions and numerous non-malicious signal processing operations.

Besides the conventional speech codecs, we also analyze a neural vocoder based on the neural source-filter (NSF) model for secure speech communication. We propose a method of improving the primary framework by modifying the state-of-the-art speaker individuality feature (namely, x-vector). Our proposed method is constructed based on x-vector singular value modification with a clustering model. We also propose enhance the proposed technique by modifying the fundamental frequency and speech duration to enhance the anonymization performance. To evaluate our method, we carried out objective and subjective tests. The overall objective test results show that our proposed method improves the anonymization performance in terms of the speaker verifiability, whereas the subjective evaluation results show improvement in terms of the speaker dissimilarity. The intelligibility and naturalness of the anonymized speech with speech prosody modification were slightly reduced (less than 5% of word error rate) compared to the results obtained by the baseline system in Voice Privacy Challenge 2020.

Index Terms: speech information hiding, speaker anonymization, McAdams coefficient, x-vector, speech security and privacy

Acknowledgements

First and foremost, I thank God for His abundant grace and blessings in any circumstances during this study that make this study possible.

Secondly, I would like to express my sincere gratitude to my supervisor, Prof. Masashi Unoki, for his continuous support for my study from my master's until the Ph.D. program. His patience, motivation, and guidance helped me immensely, especially in understanding how to carry out research and write research articles. My completion of this Ph.D. program could not have been accomplished without his support.

Thirdly, I would like to express my special thanks and gratitude to Prof. Masato Akagi as my secondary supervisor. His sincerity and attention in laboratory meetings deeply inspired me to present the research works as clearly and logically as possible.

Next, I would like to thank Prof. Shogo Okada for allowing me to conduct a minor research project under his supervision. This opportunity leads me to have a more diverse and broad perspective of computer science research.

My sincere gratitude goes to the rest of my thesis committee: Prof. Jianwu Dang (JAIST), Prof. Atsuo Yoshitaka (JAIST), and Prof. Akinori Ito (Tohoku University), for their insightful and constructive comments during the preliminary examination and final formal hearing.

I am also overwhelmed in gratefulness to have my supportive fellow labmates and friends for the stimulating discussions and for all the fun we have had. I would like to thank my family for their prayers, caring, and supportive encouragement in any situation.

Last but not least, I appreciate the financial support, travel grants, and research facilities that are crucial for this research project. I would like to express my gratitude for the fellowship granted by the Japan Society for the Promotion of Science (JSPS) along with other supporting grants from a Grant-in-Aid for Scientific Research (B) (No. 17H01761), JSPS KAKENHI Grant (No.20J20580), Fund for the Promotion of Joint International Research (Fostering Joint International Research (B))(20KK0233), and KDDI Foundation (Research Grant Program).

Candy Olivia Mawalim
Ishikawa, Japan

Contents

List of Figures

List of Tables

List Of Symbols/Abbreviations

α	McAdams coefficient
AbS	Analysis-by-Synthesis
AIH	Auditory information hiding
ASR	Automatic speech recognition
ASV	Automatic speaker verification
AWGN	Adding white Gaussian-noise
BER	Bit-error rate
BPF	Band-pass filtering
B1	Primary baseline
B2	Secondary baseline
CD	Cochlear delay
CELP	Code-excited linear prediction
CPS	Cyber-physical system
dB	Decibel
F0	Fundamental frequency
FAR	False acceptance rate
FFT	Fast Fourier transform
FN	False negative
FP	False positive
FRR	False rejected rate
Fs	Sampling frequency
GMM	Gaussian mixture models
HAS	Human auditory system
HMM	Hidden Markov Model
IHC	Information Hiding Criteria
ITU	International Telecommunication Union
LSB	Least-significant-bit
LSD	Log-spectral distance
LP	Linear prediction
LSF	Line spectral frequency
MOS	Mean opinion score

NSF	Neural Source-Filter
PESQ	Perceptual evaluation of speech quality
PCs	Principal components
P-xxxx	Proposed method with xxxx parameter(s)
RMSE	Root-mean-square error
SNR	Signal-to-noise ratio
SPSS	Statistical parametric speech synthesis
SSA	Singular spectrum analysis
SV	Singular value
SVD	Singular value decomposition
UBM	Universal background model
VCTK	Voice Cloning Toolkit
VoIP	Voice over Internet Protocol

Chapter 1

Introduction

Communication in speech is preferable in human communication because of the richness of its content. It conveys not only linguistic information but also para-linguistic and non-linguistic information [38, 39]. Fujisaki defined these three categories of information in [39] (as shown in Fig. 1.1). Linguistic information is symbolic information that is comprised of a set of discrete symbols and their rules. Linguistic information can be represented explicitly in written text or implicitly inferred from the context in a speech. Para-linguistic information is the additional information apart from the linguistic information intentionally added by the speaker, such as speaker intentions and attitudes. Non-linguistic information is the factor that generally cannot be controlled by the speaker and is not directly related to both linguistic and para-linguistic information. The speaker's age, gender, physical, and emotional states are examples of non-linguistic information.

Due to the richness of speech, the development of digital speech technology has significantly advanced to this day. The advancement of speech technology supports the availability of accessing speech or voice data, especially through the Internet. These speech or voice data could improve the performance of existing speech technology with an advanced machine learning approach. For instance, the speech-to-text system, also known as automatic speech recognition (ASR) system, could produce well-written transcription, especially when the input speech is familiar to the system [10].

One of the critical shortcomings of the speech technology advancement is related to the voice privacy and security issue [111, 112]. Exposing the speech to the public causes privacy violation. For example, the currently existing technology could provide us various information from the speech content, deceit intention, until the speaker's mental state with only a few speech utterances of a particular person, we can extract various information. Moreover, we are also able to regenerate new speech utterances (voice cloning)

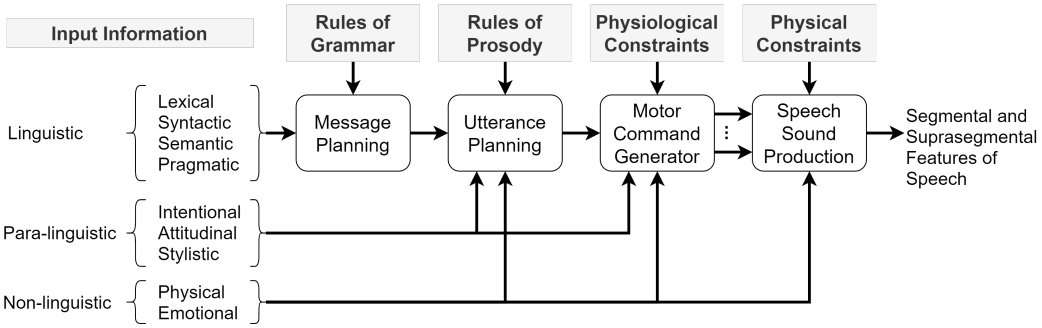


Figure 1.1: Three categories of information that manifested in speech by Fujisaki (1997) [38].

of a particular person with exceptional quality [6]. For example, speech tampering or spoofing techniques are possible with the recent advancement in voice conversion, and text-to-speech (TTS) technology [111].

The ultimate goal of this study is to provide a solution to the privacy violation problems in speech communication. The rest of this chapter states the motivation of this study with regards to secure speech communication. The motivation includes research problems on speech communication, an overview of existing solutions and remaining issues, and the significance of this research. Subsequently, the research objectives will be defined in the succeeding section. Next, the secure speech communication scenario is described as a brief explanation for addressing the proposed solutions to the research problems. Finally, the organization of this thesis is shown at the end of this chapter.

1.1 Research Motivation

Speech communication technology is usually implemented via a communication channel, such as the public switched telephone network (PSTN) and Voice over Internet Protocol (VoIP). The speech communication channel is considerably vulnerable against attacks; thus, protection and prevention countermeasures are indispensable in speech research [111, 129]. Privacy-preserving technology is required in speech communication to protect the speech content and personal profiles of the speaker [112, 129].

In literature, the efforts in privacy-preserving technology can be categorized into two main categories, i.e., cryptography and information hiding. Basically, cryptography, or with regards to speech processing, also known as speech encryption, converts the speech data to another form that only can be accessed by an authorized person with a private key. This approach is useful for specific

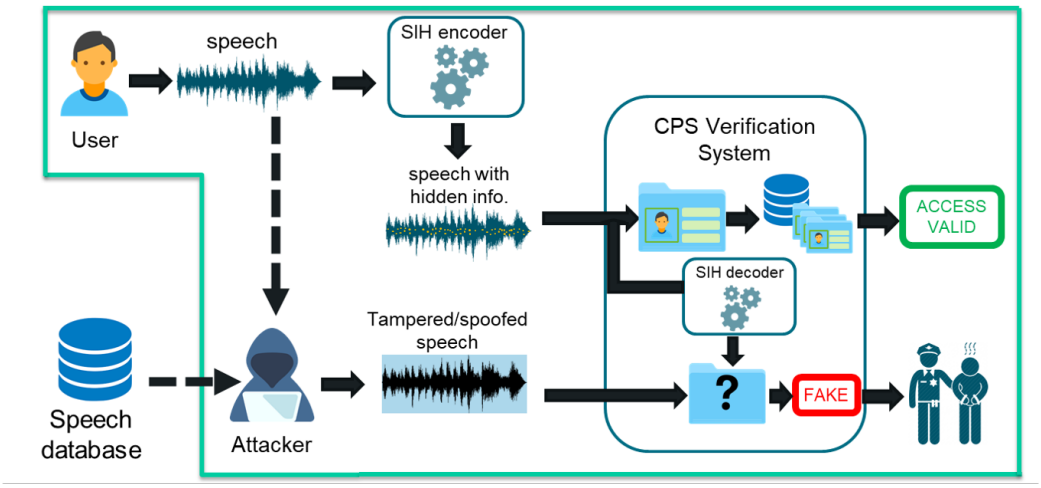


Figure 1.2: Overview of secure speech communication in cyber physical system (CPS) by SIH.

applications that can afford the additional computational time and complexity of the cryptography process. The limitation of cryptography is that it does not protect the speech signal once the content is decrypted [112]. Since the form of the encrypted data is intelligible and increases the suspicious level of the speech content, the attackers may attempt to monitor this communication. Subsequently, they can utilize cryptanalysis methods for decrypting the encrypted data. On the other hand, the second category (information hiding) preserves the privacy and security of speech data by imperceptibly embedding particular information that needs to be hidden [49, 70]. Two information-hiding categories are steganography and watermarking, depending upon the purpose. The application of the information hiding approach can be used to identify original or tampered signals [56, 126].

This study focuses on the information hiding approach to preserve both security and privacy simultaneously in speech communication. The current existing frameworks are generally developed by combining state-of-the-art technologies in various areas of speech technology applications. For instance, for protecting voice privacy, a speaker anonymization system was proposed as the primary baseline system in the Voice Privacy Challenge 2020 based on the state-of-the-art speaker embedding (x-vector) [107], and neural source-filter (NSF) waveform modeling [35, 112]. Unfortunately, the performance still needs further improvement, especially in promoting naturalness and intelligibility. Additionally, the anonymization performance in terms of speaker verifiability was also limited in dealing with attacks scenarios. This study contributes to solving the current essential issues by the proposed framework, which

is developed with consideration of concepts in speech perception, speech production, and signal processing.

The proposed framework integrates the information hiding approach to secure the speaker anonymization, which consists of two main parts, i.e., encoder and decoder (as shown in Fig. 1.2). The encoder is mainly aimed to protect the speaker's identity by using an anonymization approach. In contrast to the other works, the anonymization is conducted with a parameter that will be used to represent watermarks. The result of anonymized speech should be able to conceal the sensitive personal information in speech while maintaining the naturalness and intelligibility of the speech. Meanwhile, the decoder is aimed to protect the authentication of the speech by accurately detecting the embedded watermarks. To ensure reliability and robustness, the proposed framework is evaluated by general datasets and protocols established in the Information Hiding Criteria (IHC) [51] and the Voice Privacy Challenge 2020 [112]. In the application aspect, the proposed framework contributes as an alternative approach for accounting the speech spoofing and tampering detection.

1.2 Research Objectives

The main objective of this study is to propose the framework for preserving both speech security and privacy using an information hiding approach. In order to reach this objective, first, the robust speech properties and their relationship correspond to linguistic, para-linguistic, and non-linguistic information were investigated. Secondly, the analysis and synthesis process based on these features will be studied. This second subgoal aims to study hiding specific vulnerable speech information, e.g., speaking style and speaker identity while maintaining the naturalness and intelligibility of resynthesized speech. Finally, this proposed hiding framework is considered to solve the real-world problem, such as tampering detection, and is evaluated by following the established protocols described in the Information Hiding Criteria (IHC) [51] and the Voice Privacy Challenge 2020 [112].

1.3 Secure Communication Scenario

Recent speech technology, especially speech synthesis, has significantly improved our capability to re-create, clone, or manipulate human voice. The synthesized voice is not only intelligible but also contains personal information, which boosts its naturalness in the human auditory perception. For

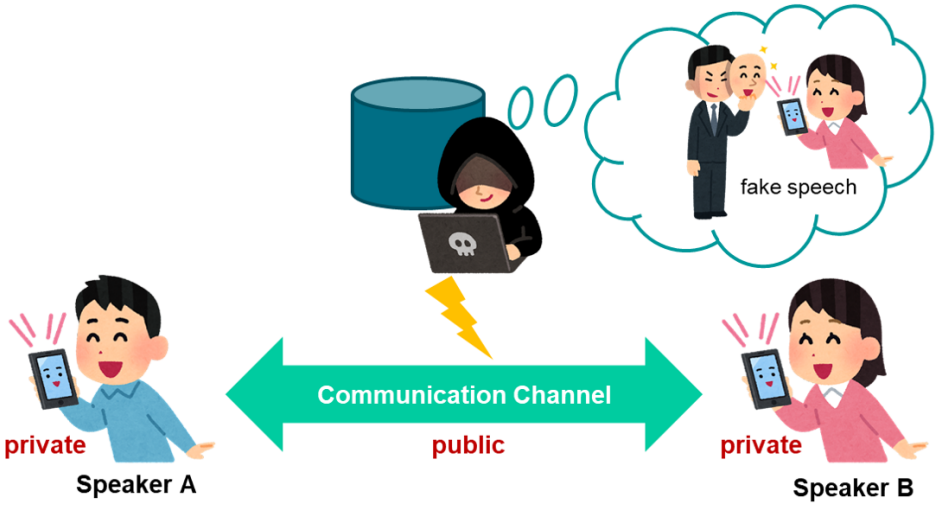


Figure 1.3: Illustration of voice privacy issue in speech communication.

instance, by using the neural speech synthesizers, such as WaveNet [119], we can create relatively high-quality personalized speech. These kinds of speech synthesis technology support many applications positively, such as virtual voice assistants or even facilitating speech-impaired people in daily communication. However, the advancement in speech technology also causes drawbacks, primarily in those issues related to privacy.

As mentioned in the earlier section of this chapter, speech encapsulates various information. Exposing it can cause privacy violations. Figure 1.3 shows the example of voice privacy issue in digital speech communication. When the speech is exposed to the public communication channel, any attackers can easily store and manipulate it to be a fake speech. Fake speech causes a problematical issue when the attackers illegally use it to scam or fraud. Several years ago, these scams were so frequent that they had been warned in ATMs all around the country.

Speaker anonymization or de-identification has been introduced as one of the solutions for privacy preservation in speech communication in the Voice Privacy Challenge 2020 [112]. The goal of speaker anonymization is to suppress the information related to the speaker's identity while maintaining the linguistic aspects and speech naturalness. Besides, it should have low complexity and be flexible so that it can be applied to other existing speech technology. The Voice Privacy Challenge 2020 provides a formal definition for the task, metrics, and protocols of speaker anonymization system. As for the task, speaker anonymization system depends on the following specifications:

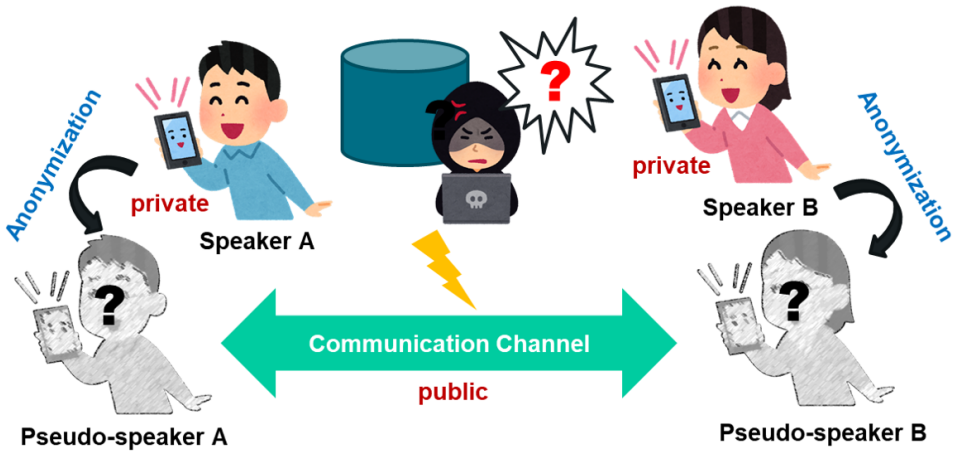


Figure 1.4: Illustration of speaker anonymization for voice privacy preservation.

1. the characteristic of the speech data;
2. the personal identifiable information;
3. the desired goal(s) or application(s);
4. the accessible data by the attacker;
5. the additional information or knowledge of the attackers.

The main scenario of speaker anonymization based on the Voice Privacy Challenge 2020 [112] is that the speakers intend to suppress their identity while achieving the desired goals. Meanwhile, the attackers have access to an utterance and attempt to identify the corresponding speaker. The attack model is referred to as the model when the attacker has access to various amounts of data, which may or may not have been anonymized. Figure 1.4 shows the illustration of voice privacy preservation using speaker anonymization. Before the real speech is exposed to the public communication channel, the input speech is anonymized. For example, when we have a speaker A and an anonymization technique is applied, we get the anonymized speech (with pseudo-speaker A as the identity). The speaker anonymization knocked off the possibility of attackers faking the voice of the particular speaker. Additionally, hiding the information about the real speaker in the anonymized speech can restore or de-anonymize the speech so that the receiver can perceive the real speaker voice.

Two approaches are commonly utilized for speaker anonymization. The first approach is the approach based on a voice conversion system. The goal of a voice conversion system is to transform the identity of the given speaker to a specific target speaker while maintaining linguistic information. Data from the given speaker and target speaker are required to train a voice conversion model. On the basis of the speaker anonymization task defined in the Voice Privacy Challenge 2020, the speaker anonymization should be able to anonymize any given speakers even without data of target speakers (as in common voice conversion systems). Moreover, a speaker-to-speaker manner is one of the requirements of speaker anonymization, which causes the common voice conversion systems to be inapplicable.

The second approach is the approach based on voice modulation. Voice modulation techniques manipulate the voice features, including pitch, to transform the original voice to a specific speech style. In earlier studies, voice modulation was developed by using acoustic filters to alter spectral information. This technique was often used in the interview session of TV news to mask the voice of suspects, victims, and witnesses. Although using acoustic filters is quite simple and easy to implement, it bears disadvantages, primarily due to the possibility of illegal de-identification by using the corresponding inverse filters. More recent voice modulation techniques are based on vocoders, which are the focus of the speaker anonymization approach in this study.

Speaker-to-speaker manner in speaker anonymization means that the anonymized utterances of a given speaker should not be the same with other speakers. In the real scenario, we can imagine the scenario from an incoming call from known speakers. The anonymized speech should have been able to be recognized by the receiver side. Furthermore, speaker anonymization should be robust in the attack model where the attacker has access to the anonymized speech utterance(s). Due to this requirement, there is no guarantee that the identity of the pseudo-speaker of a given speaker is equivalent over time. Consequently, the difficulty of authenticating an anonymized speech is significantly increased (as shown in Fig. 1.5 (a)).

To deal with the authentication issue, we proposed a SIH framework for preserving voice privacy. We simultaneously performed the anonymization and watermarking in the encoder part of the proposed solution. Figure 1.5 (b) shows the illustration of voice privacy preservation using watermark as the authentication key. By integrating anonymization and watermarking, we can preserve both the speech content and privacy. Furthermore, we can have lower complexity and higher robustness in comparison to conducting both independently.

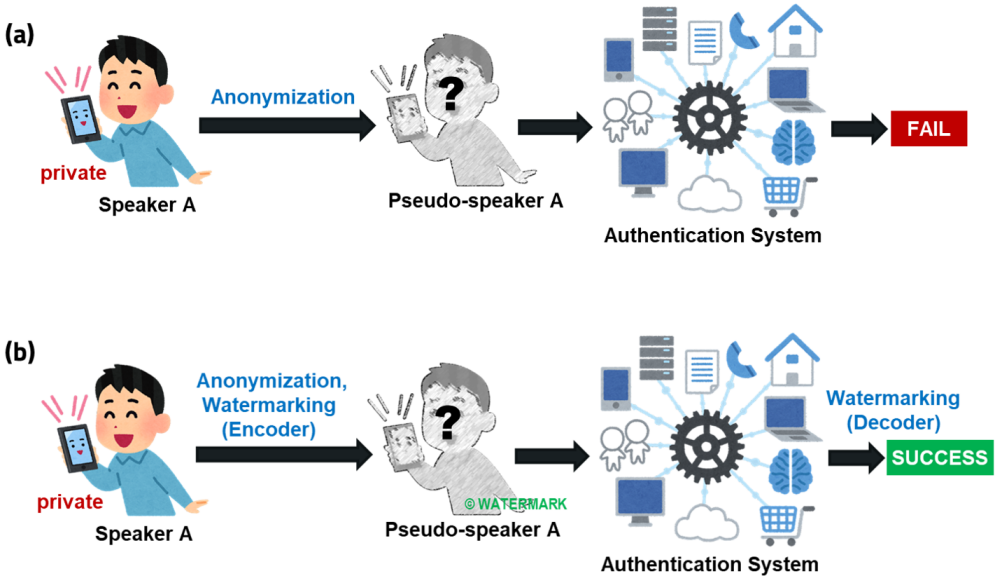


Figure 1.5: Illustration of voice privacy preservation by SIH, e.g. in authentication system: (a) without watermark, (b) with watermark.

1.4 Organization of the Thesis

This thesis is comprised of seven chapters. Figure 1.6 shows the schematic outline of this thesis. Apart from this introduction chapter, the organization of the remaining chapter from Chapter 2 to Chapter 7 is as follows.

Chapter 2 describes the literature review related to this study. Firstly, this chapter explains the general overview of the speech communication system and the problems related to security and privacy. The second chapter introduces the information hiding for secure speech communication. Finally, the last chapter introduces a method for protecting voice privacy, namely speaker anonymization. This chapter will describe the definition, evaluation metrics and criteria, and the existing baseline systems.

Chapter 3 presents the proposed framework for content and privacy protection in speech communication. The first section of this chapter introduces the overview of the proposed framework. Next, the second section describes speech processing based on analysis and synthesis in speech communication. The third section describes the method of securing speech communication based on the information hiding approach. Finally, this chapter ends with using the information hiding approach to improve the security of speaker anonymization methods.

Chapter 4 elaborates the proposed method of content protection using the information hiding approach. The description of the proposed method, including the design of the information hiding method, the embedding, and the detection scheme, will be explained. Additionally, the experiments using the proposed method are explained and discussed.

Chapter 5 elaborates the proposed method of voice privacy protection based on speaker anonymization. We describe the speaker anonymization method by modifying the x-vector singular value and the development by modifying speech prosody. Similar to Chapter 4, we also explain the experiments on speaker anonymization. Finally, we discuss the comparative analysis of speaker anonymization approaches.

Chapter 6 contains the total evaluation of our proposed framework on secure speech communication. Furthermore, the results of the evaluation are analyzed and discussed as the contributions of this thesis.

Chapter 7 contains a summary, highlights of contributions, and the future direction (future works) of this study.

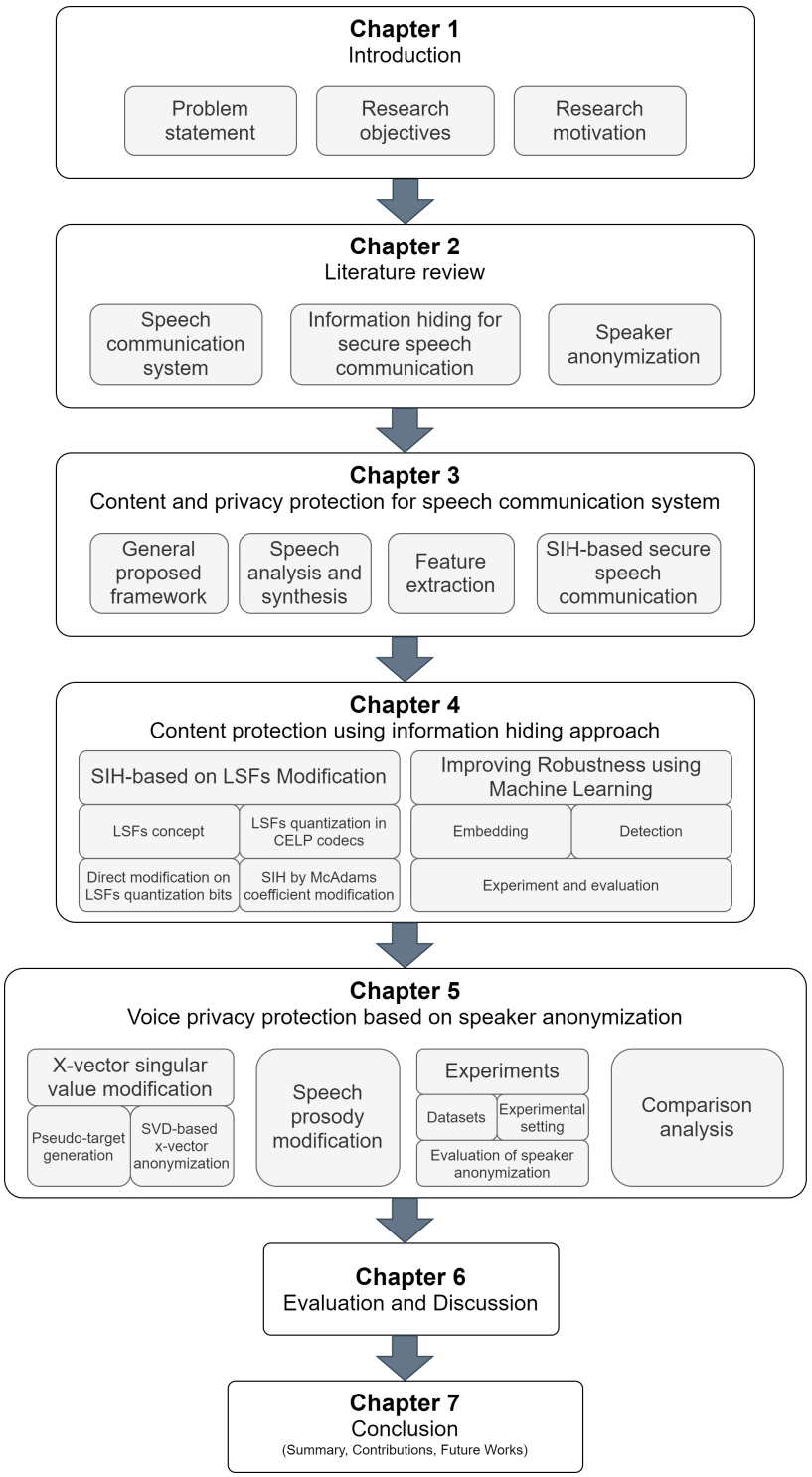


Figure 1.6: Thesis organization.

Chapter 2

Literature Review

This chapter reviews the background knowledge related to secure speech communication systems into three sections. First, the introduction about speech communication system is described, including the elements of speech communication, the analysis-synthesis model for processing speech, and the progress of secure speech communication. Subsequently, the second section explains information hiding for secure speech communication. This section describes the overview of the speech information hiding framework, evaluation metrics, methods, and applications. This chapter ends with an explanation of voice privacy protection via speaker anonymization.

2.1 Speech Communication System

This section provides a description of the speech communication system, i.e., what key elements in speech communication are, how to process speech in the digital communication system (analysis by synthesis model), and the existing approaches for secure speech communication.

2.1.1 Elements of Speech Communication

Typically, basic communication is comprised of four key elements, i.e., source or sender, message, receiver, and feedback. The source speaker or sender is the person who initiates a communication. The receiver is the person who receives the message sent by the sender. The feedback is the response given by the receiver to the sender. In digital communication, message and feedback are encoded and decoded before/after passes through a communication channel. Figure 2.1 shows the illustration of a digital communication system that includes its key elements.

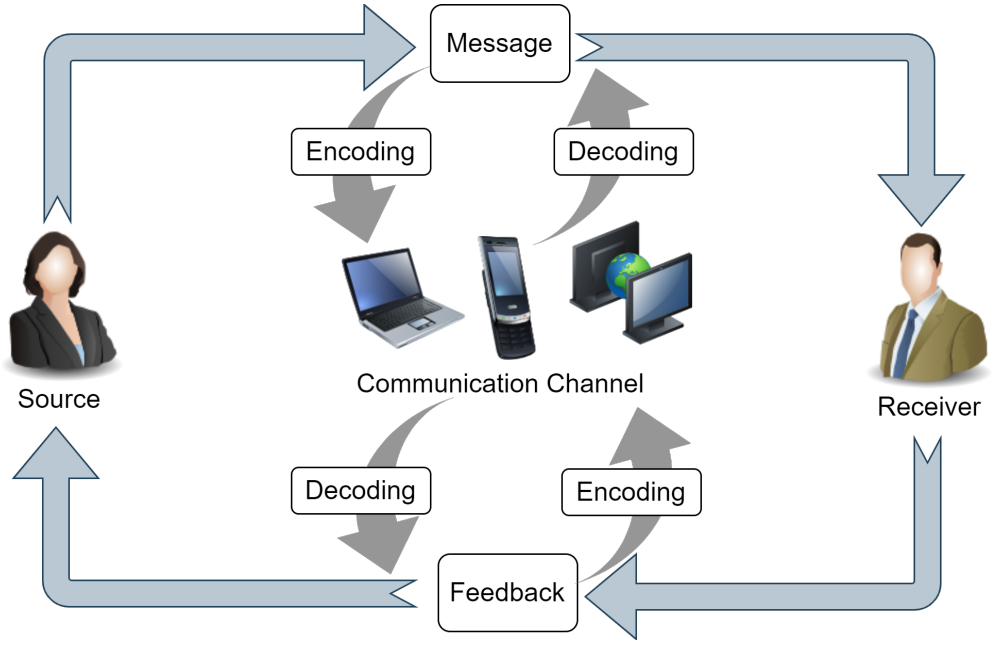


Figure 2.1: Key Elements of Digital Communication System

With regard to speech communication, the transmitted message is in the form of a speech signal. Speech produces by a speech production system, which could be summarized into three mechanisms: (1) respiration by the lungs, (2) phonation at the larynx, and (3) articulation from the movement in the mouth (lips, tongue, jaws, and nasal cavity). Due to these mechanisms and the unique anatomy characteristics, speech produces by each person is unique despite the same linguistic content. As mentioned in Chapter 1, speech signal conveys various information, including the three categories of information described by Fujisaki [39] (also shown in Fig. 1.1).

A communication channel, such as the public switched telephone network (PSTN) and Voice over Internet Protocol (VoIP), mediates the speech communication in the digital system. Before the transmission through a communication channel, a speech signal needs to be encoded so that it can be digitally processed. A speech coder encodes and decodes speech signals into digital information before storing or transmitting them through a communication channel. Most speech coders are based on the analysis by synthesis (AbS) model. The following subsection explains the AbS model in more detail.

2.1.2 Speech Coding: Analysis by Synthesis Model

Speech coding is the process of converting or transforming a speech signal into a more compact digital representation for effective transmission. When transmitting a digital signal, a function of bit rate (bandwidth) is required [62]. The bit rate is the data rate in a given unit of time (commonly, in the unit of bits per second (bps)). A speech coder is a device or technique that performs speech coding, which always comprises an encoder and a decoder (also known as codec). An encoder transforms the input speech to a low-rate bitstream. Meanwhile, a decoder approximates the output bitstream to the original input signal.

Most of the speech coders produce a reconstructed speech signal that differs from the original one in order to remove the redundancy of speech signals. However, sometimes the parameters that represent the speech are reduced and cause a lossy coding. The representation of a vector or a value with reduced precision is referred to as quantization [62]. The distortion caused by this process is called quantization noise. To obtain a reliable speech coder, we have to optimize the perceived similarity between the original speech and the reconstructed speech. Therefore, the study of human auditory perception, as well as the study of human speech production, are widely considered in the development of speech coding.

At an earlier time, eight-bit logarithmic quantizers with a bit rate of 64 kbps are ordinarily applied in network telephony [4]. The speech coders based on these quantizers simply perform the conversion of the analog-to-digital (A/D) and digital-to-analog (D/A). Along with the development of hardware technology, a linear PCM (pulse-coded modulator) system with a bit rate of 128 kbps is used to convert the input speech before passing through an encoder (to obtain a lower bitrate). In 1985, Manfred R. Schroeder and Bishnu S. Atal proposed the code-excited linear prediction (CELP) codec that is based on analysis by synthesis model for speech coding algorithm [8]. The analysis process is performed by encoding the signal with perceptual optimization, and the synthesis process is performed by decoding the signal in a closed-loop. This approach was significantly improved the reconstructed speech than the other existing methods at that time.

Even until recent years, CELP codecs are the most common speech codecs used in digital communication systems due to their low-bitrate high-quality speech representation [8, 102]. These codecs are based on the source-system model that mimics the human speech production mechanism through linear prediction (LP) analysis [8, 96, 102]. In digital speech signal processing, voiced speech is produced by the excitation of the vocal tract filter with quasi-periodic glottal pulses. Meanwhile, unvoiced speech is produced by the

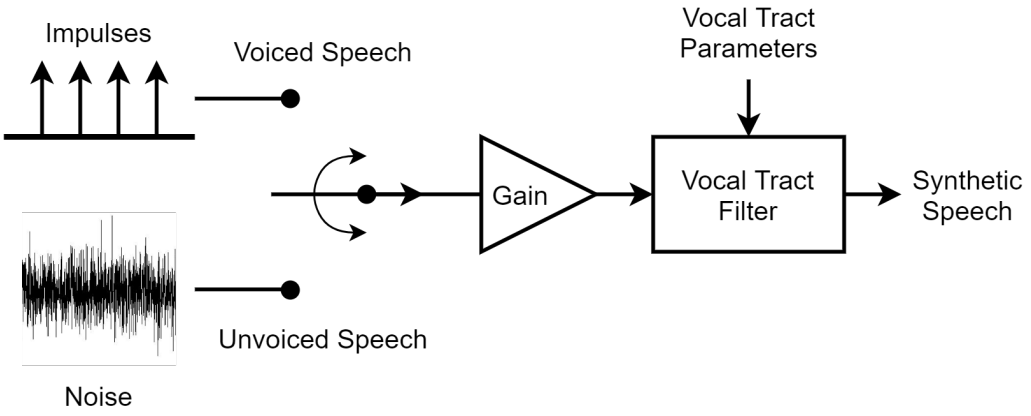


Figure 2.2: Simple Speech Synthesis Based on Source-Filter Model.

air constriction in the vocal tract. Figure 2.2 shows the simplification process of speech synthesis based on source-filter model.

Linear predictive coding attempts to estimate vocal tract parameters by estimating a current speech signal using a linear combination of past samples. The following differential equation characterizes the mathematical form of linear predictive coding:

$$s(n) = \sum_{i=1}^{M} a(i)s(n-i) + e(n) \tag{2.1}$$

where $a(i)$ corresponds to the filter coefficient in i-th order, M is the maximum order of the prediction (typically 10), and $e(n)$ is the prediction error.

The transfer function for the corresponding linear prediction differential equation is represented by tenth-order all-pole autoregressive filters, which is given by:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{10} a(i)z^{-i}} \tag{2.2}$$

Figure 2.3 illustrates the source-system model by AbS linear prediction. In CELP coding, the excitation generator generates an excitation vector codebook $\mathbf{x}$ by minimizing the residual error e, which can be written mathematically as,

$$e^{(i)} = s_w - \hat{s}_w^0 - g^{(i)}\hat{s}_w^{(i)} \tag{2.3}$$

where s_w is a vector of perceptually-weighted input speech, $\hat{s}_w^0$ is the initial filter state output vector, $g^{(i)}$ is the gain factor, and $\hat{s}_w^{(i)}$ is the synthetic speech vector associated with the $\mathbf{x}^{(i)}$ with i as the codebook index.

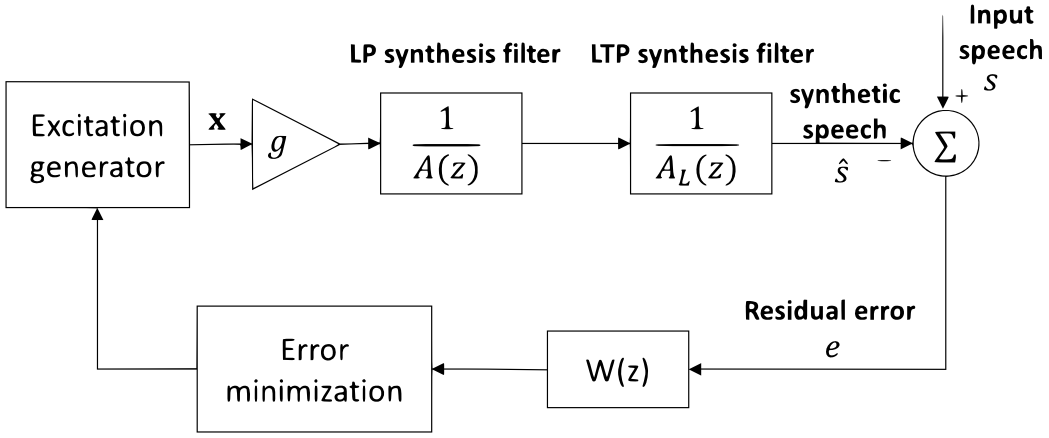


Figure 2.3: Source-filter model in AbS linear prediction.

In standard AbS linear prediction algorithms, the tenth-order short term linear prediction is used as the linear prediction synthesis filter ($1/A(z)$). $A(z)$ denotes the line spectrum pairs (LSPs) that can be given by:

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_{10} z^{-10} \tag{2.4}$$

where a_i is the i-th order linear prediction coefficients (LPCs).

The long term prediction (LTP) synthesis filter ($1/A_L(z)$) captures the long-term correlation and represents the speech periodicity mechanism. The perceptual weighting filter $W(z)$ models errors by masking the quantization noise with high-energy formants. The perceptual weighting filter $W(z)$ can be written as follows:

$$W(z) = \frac{A(z/\gamma_1)}{A(z/\gamma_2)} = \frac{1 - \sum_{i=1}^{M} \gamma_1(i) a(i) z^{-i}}{1 - \sum_{i=1}^{m} \gamma_2(i) a(i) z^{-i}} \tag{2.5}$$

where γ_1 and γ_2 are the adaptive weights that satisfy $0 < \gamma_1 < \gamma_2 < 1$, and m is the order of the linear predictor. γ_1 ranges between 0.94 and 0.98 and γ_2 ranges between 0.4 and 0.7 depending on the tilt or flatness characteristics of the linear prediction spectral envelope [96, 101].

2.1.3 Progress in Secure Speech Communication

In recent years, speech communication has been rapidly developed along with the development of network communication [86]. Even in our daily life, we often use speech communication networks, including PSTN and VoIP. Unfortunately, the channels utilized in communication networks are not safe (vulnerable) to any threats. For example, the possibility of speech

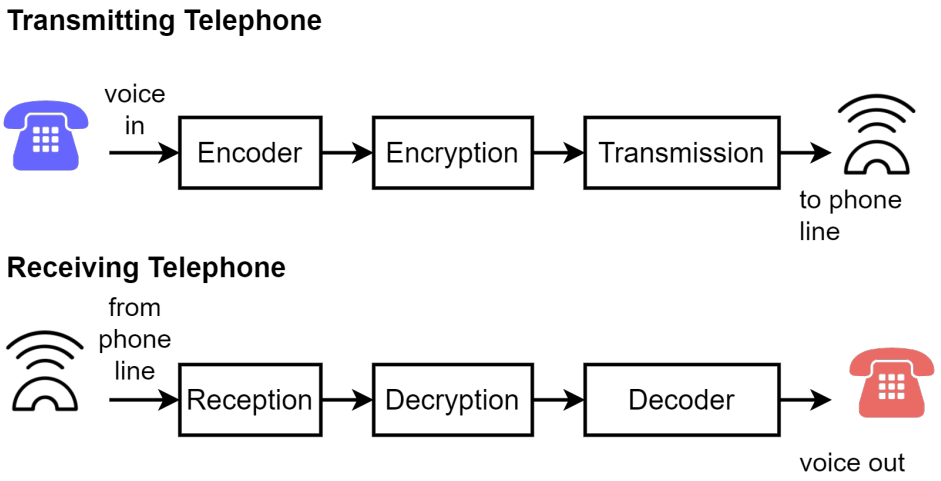


Figure 2.4: Basic system design of cryptography or digital encryption in PSTN [88].

communication being hijacked or attacked by man-in-the-middle is high [129]. Without adequate countermeasures, attackers could easily attack a speech communication [59, 86, 112, 129].

The classical techniques for securing speech communication come from the studies of speech encryption [52, 66, 88, 129]. Speech encryption is the art of securing information by transforming understandable data into another form that is not intelligible for unauthorized parties. Similar to speech coding, speech encryption is generally comprised of a pair of transformations, i.e., encryption and decryption.

Speech encryption can be categorized into two types: digital and analog (scrambling) [128]. In digital speech encryption, a speech signal is digitized before encrypting it into unintelligible signals. Figure 2.4 shows the basic system design of digital encryption in PSTN [88]. Meanwhile, the analog scrambling approach is conducted by segmenting and scrambling the intelligible speech signals in a particular domain(s) to form unintelligible signals [52]. In literature, digital encryption is more popular than analog scrambling in speech secure communication because analog scrambling is poor in the security aspect [98, 129].

Although digital encryption has high confidentiality and easy implementation, it also follows with several drawbacks. First, the strength of most encryption methods highly depends on the encryption procedures and the required capacity. For example, the password (secret key) that we set to encrypt data is highly associated with the security level. Second, the form of encrypted data is intelligible and therefore increases the suspicious level

that it might contain secret and important information. When the attackers monitor this communication, they can easily try to decrypt it with the existing cryptanalysis methods. Another disadvantage of digital speech encryption is that the process of encryption and decryption causes distortion to the speech quality.

To overcome the drawbacks of the classical encryption techniques, the chaotic secure communication methods were developed to improve confidentiality by using the chaos system's sensitivity [67, 129]. These methods emerged along with the improvement in computing power and the advancement in the quantum computer. However, most of the methods have not been used in real applications (under development) and are limited to the prototype implementation [129].

Another approach for securing speech communication is by using a speech information hiding (SIH) technique. Although the idea of information hiding was proposed systematically in the 1990s, its development in speech communication is still limited. This is because SIH has its own challenging points compared with other digital data, such as images or video. The sensitivity and dynamicity level of the human auditory system (HAS) is higher compared with the human visual system, which caused difficulties in achieving the inaudibility requirement in SIH (see Subsection 2.2.1). In addition, the trade-off between the robustness and the inaudibility occurs since the total number of representations for the audio signal is less than video at one particular time. This study focuses on secure speech communication using the information hiding approach.

2.2 Information Hiding for Secure Speech Communication

This section describes the overview of speech information hiding (SIH), evaluation metrics, the existing SIH methods, and SIH applications.

2.2.1 Overview of Speech Information Hiding

SIH literally means the technique of hiding data into a speech signal. A reliable SIH should fulfill the following five requirements:

1. *inaudibility* means keeping the embedded message imperceptible to the human auditory system;

2. *robustness* means keeping the embedded message able to withstand any unintentional (signal processing operations) or intentional attacks;

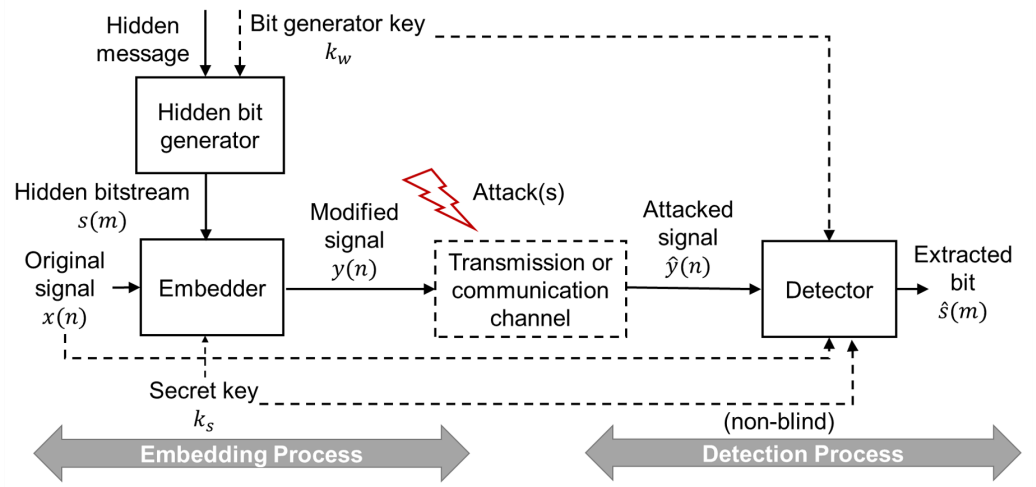


Figure 2.5: Overview of an SIH framework

3. *blindness* means the original signal is not required in the detection process;

4. *high payload* means the output signal conveys a large amount of embedded message;

5. *confidentiality* means secure concealment of embedded data.

Figure 2.5 shows the typical SIH framework. SIH contains three main parts: (1) a hidden bit generator, (2) an embedder, and (3) a detector [70]. Commonly, a hidden bit generator converts the hidden message into the bitstreams $s(m)$. The digital speech to be protected is defined as the original signal $x(n)$, and the signal after the embedding process is defined as modified signal $y(n)$. In order to improve the confidentiality requirements, the additional key k_w is used for generating the hidden bitstream. In contrast, the additional secret key k_s is used in embedder to provide a more secure modified signal $y(n)$. In general, the embedding process can be expressed as follows:

$$y(n) = \text{Embedding}(x(n), s(m), k_w, k_s) \tag{2.6}$$

where n is the number of sample and m is the hidden bit index.

The modified signal $y(n)$ is then passing through the communication or transmission channel. In this stage, the modified signal $y(n)$ may be changed due to unintentional signal processing operations (e.g., compression, resampling) or intentional attacks (e.g., tampering or spoofing attacks). As a result, the input of the detector can be referred to as attacked signal $\hat{y}(n)$. In

Table 2.1: Sound quality description with regards to PESQ (MOS)

PESQ (MOS)	Speech Quality
1	bad (totally unacceptable, poor quality speech)
2	poor (poor speech quality with annoying distortion)
3	fair (acceptable quality with slightly annoying perceptible issue)
4	good (good speech quality with a slight perceptible issue but not annoying)
5	excellent (high speech quality and no perceptible issues)

general, the detector process of the hidden bitstream $\hat{s}(m)$ can be expressed as follows:

$$\hat{s}(m) = \text{Detection}(\hat{y}(n), x(n), k_w, k_s) \tag{2.7}$$

for non-blind approach. On the other hand, for blind approach, the original signal $x(n)$ is not required as the input of the detector which is shown as follows:

$$\hat{s}(m) = \text{Detection}(\hat{y}(n), k_w, k_s) \tag{2.8}$$

2.2.2 Evaluation Metrics

In order to evaluate the effectiveness of various SIH schemes, a standard benchmark for SIH performance evaluation is necessary [51]. The general performance evaluation is comprised of the inaudibility, robustness, and security level assessments.

- **Inaudibility Assessment**

Inaudibility assessment has two main categories: subjective listening tests by human acoustic perception and objective tests by quality measurements based on perception modeling. In subjective listening tests, the human opinion from normal hearing audiences with different backgrounds is considered in most of the applications [7]. On the other hand, the objective tests are conducted by assessing the perceptual evaluation of speech quality (PESQ) [97] and log-spectral distortion (LSD) [42].

PEAQ is one of the standards proposed by ITU-T P.862 to measure speech quality. Principally, PESQ represents the perceptual speech quality of $y(n)$ with $x(n)$ as the reference in mean opinion scores (MOS). The MOS varies from a scale of 1 (bad) to 5 (excellent). The meaning of each scale is shown in Tab. 2.1. Typically, the PESQ threshold for speech watermarking is 3 (fair or slightly annoying). On the other hand, LSD is also important to measure the distance or distortion between two spectral (in SIH case, between the original spectra $X(\omega, k)$ and the modified spectra $Y(\omega, k)$) with k is the frame index.

The calculation of LSD is as follows:

$$\mathrm{LSD}(X,Y) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(10\log_{10}\frac{|X(\omega,k)|^2}{|Y(\omega,k)|^2}\right)^2}, \tag{2.9}$$

where $X(\omega,k)$ and $Y(\omega,k)$ are the short-time Fourier transform of the original ($x(n)$) and modified signal ($y(n)$) of k-th frame, respectively. A lower LSD indicates the better inaudibility (the threshold LSD < 1 dB is usually used in SIH evaluation).

- **Robustness Assessment**

The robustness assessment aims to evaluate the resistance of the detector system's ability not only in normal conditions but also against signal modification in real applications. Practically, the bit error rate (BER) can be used to check whether the embedded information is robust or not. BER is the ratio between the total number of incorrectly extracted bits and the total embedded bits, which is defined in the following equation:

$$\mathrm{BER} = \frac{\text{Number of incorrect bits between } \hat{s}(m) \text{ and } s(m)}{\text{Number of bits of } s(m)} \times 100\% \tag{2.10}$$

A reliable robustness test should comprise unintentional attacks (common signal-processing operations) and intentional attacks. For a brief description, the signal manipulation operations (attacks) can be classified into three categories:

(a) Common signal operations:

e.g., resampling (down-sampling and up-sampling), noise addition (e.g., additive white Gaussian noise (AWGN)), requantization, amplitude scaling, low-pass filtering, echo addition, reverberation, speech coding, compression, digital-to-analog and analog-to-digital (DA/AD) conversion, and the combinations.

(b) Desynchronization attacks:

e.g., random samples cropping, zero inserting, jittering, time-scale, and pitch-scale modification.

(c) Advanced attacks:

e.g., collusion, multiple watermarking.

- **Security Level Assessment**

False acceptance rate (FAR) and false rejection rate (FRR) are metrics to evaluate the security level in biometric systems. FAR represents the error rate when unauthorized people are incorrectly accepted. Meanwhile, FRR represents the error rate when authorized people are incorrectly rejected. The point at which the FAR and FRR lines meet is called EER (Equal Error Rate). In addition to FAR and FRR, we also calculate F1-score to evaluate the overall performance (also describe the harmonic mean between precision and recall of detection).

2.2.3 Speech Information Hiding Methods

On the basis of the hiding medium/domain, SIH techniques can be categorized into two main groups: (1) time-domain methods and (2) transform domain methods [49].

1. Time-domain methods

   In time-domain methods, the hidden information was hidden into time aligned signal [21, 22, 83, 100, 116, 118] or echo-based (not aligned) signal [43, 60, 82, 84, 87]. The generic time-domain SIH techniques can be expressed as:

   $$y(n) = x(n) + \alpha f[x(n), s(m)] \tag{2.11}$$

   where α refers to the strength of embedded signal and $f[\cdot]$ refers to the hidden bit generation by using both original signal $x(n)$ and the hidden information $s(m)$.

   There are several traditional methods for time-aligned SIH techniques, e.g., least significant bits (LSB), phase coding, and cochlear delay (CD) based SIH. In LSB [21], the hidden information is embedded by substituting the least significant bits of the speech samples. The payload which can be achieved by this method can be very high (same rate with the sampling rate, e.g., 44.1 kbps for a speech signal with a sampling frequency of 44.1 kHz). However, it is predictable and vulnerable to any attack. Other improvements for this method were also proposed [22, 100] to improve the robustness, for example, by selecting the low-frequency components only or using repetitive coding. Unfortunately, this method can not fully satisfy the robustness against unintentional and intentional attacks.

   The phase modification techniques, including phase coding and phase modulation, are based on the fact that the human auditory system

(HAS) is not sensitive to the absolute phase difference. HAS can only distinguish the relative phase [9]. In the phase coding technique, the SIH is conducted by substituting the initial phase of the segment of the original signal with a reference phase as the hidden-bit representation. This approach is more effective than LSB in terms of robustness. However, when the phase between each frequency component is drastically changed, sound distortion will occur. CD-based information hiding method is one of the phase modulation techniques. Generally, this method employed the all-pass filter for controlling the phase to represent the bit information [83, 116, 118].

In echo hiding, the hidden information is embedded via an echo from the original signal. It works by controlling three parameters of the echo, including initial amplitude, decay rate, and offset [9, 43]. The embedding process is conducted by setting two delays in time as the representation of two binary bits ("0" or "1"). These delays should be chosen carefully so that the quality of the modified signal can be reliable. Subsequently, the extraction process is based on complex cepstrum analysis [49]. A magnitude peak will occur at the appointed delay time in the cepstrum spectrum of the modified signal. By comparing this peak, the hidden bit information can be successfully extracted. There are various techniques proposed to improve the traditional echo hiding [60, 84, 87]. However, some remaining issues still remain, including the difficulties in determining the exact delay time of low-magnitude echoes when the attacks happen. The robustness and inaudibility trade-off is also a big problem due to the echo magnitude and delay time.

2. Transform domain methods

   Besides of time-domain, the hidden information can be embedded into other domains [48, 50, 61, 63, 69, 73], for instance, in spread spectrum (SS), wavelet domain, and cepstrum domain. The generic model for transform domain methods is as follows:

$$Y(k) = X(k) + \alpha m(k) \tag{2.12}$$

   where $X(k)$ and $Y(k)$ are the transform domain representations of original signal $x(n)$ and modified signal $y(n)$, respectively; whereas $m(k)$ refers to either non-modulated or modulated hidden bit stream.

   In the traditional SS technique, the hidden bitstream is spread across the frequency spectrum of the original signal [9]. There is various SS communication. One of them is the direct sequence spread spectrum

(DSSS). This method works by introducing a *chip* (a modulated pseudorandom sequence) that is multiplied by the signal. In other words, it spreads the pseudorandom sequence into the original signal). For the detection process, the *chip* signal is multiplied by the modified signal and passes through both the band-pass filter and the phase detector sequentially. Several approaches were also proposed to improve the inaudibility performance of this technique [61, 73].

In the wavelet domain, the hidden information is embedded into wavelet coefficients of original signal [50, 63]. Wavelet transform produces a better time-frequency representation of a non-stationary signal, including speech signals. The decomposition and reconstruction of this technique require multiresolution analysis and synthesis. The wavelet domain methods are generally lacking in robustness [70]. Alternatively, the cepstrum domain SIH used the cepstral coefficient as embedding medium [48, 69]. The complex cepstrum is referred to as the inverse Fourier transform of the complex logarithm of the Fourier transform of an input signal. Generally, the performance of the method based on the cepstrum domain is similar to the one in wavelet domain [70].

2.2.4 Applications of Speech Information Hiding

SIH can be utilized in a wide variety of applications. In general, SIH can be divided into two categories, i.e., watermarking and steganography [70]. The primary purpose of speech watermarking is copyright protection. On the other hand, steganography is mainly used for providing secret communication. However, the applications of SIH are not only limited to those two primary purposes but also for broadcast monitoring, usage tracking, tampering detection, and spoofing detection.

- **Copyright protection**

During the past few years, protecting intellectual properties has become a social issue due to the development of digital technology. SIH, especially speech watermarking, is the most popular alternatives for providing the solution for copyright protection [64]. This technique works by embedding the additional data (called a watermark) into the speech contents that must be protected before distributing it. The proof of ownership can be claimed by detecting the watermark from this distributed speech data.

In order to reach the ideal copyright protection, the embedded watermark should be robust and inaudible. The robustness is required since there may be any unintentional or intentional signal processing affected in the distribution

process. For example, when a user wants to reduce the file size by compressing it to an MP3 type, the watermark detector should be able to detect it. The watermark is also expected to be robust when the speech data gets uploaded into a file-sharing website. Furthermore, it is essential to provide a highly inaudible watermark because bad inaudibility implies the bad degradation in the quality of the watermarked data.

- **Secret communication**

As mentioned previously, secret communication is the primary purpose of steganography. The word "steganography" refers to the practice of concealing a secret message (i.e., the hidden information that nobody apart from the sender and the appointed recipient notices it) into a cover signal [20]. Generally, speech steganography hides a message within a cover speech signal using a stego key, which is equal at both the transmitter and receiver sides. The output of the speech steganography embedder in the transmitter side is referred to as stego speech signal. On the other hand, on the receiver side, the embedded message is retrieved from the cover speech signal using the stego key by extracting it via the speech steganography detector.

To achieve the "secret" communication by definition, the inaudibility requirement is the most important that expected to be fulfilled. Only the trusting parties are allowed to notice the presence of the hidden message. In addition, the payload of the information to be embedded is also important. The more information can be hidden, the more composite messages (e.g., not only text, but also images, videos, etc.) can be used for communication.

- **Broadcast monitoring**

Utilizing SIH to the speech data at the time of production or broadcast allows the content owners to identify with granular precision which is broadcasting the data, when and where content is broadcast, and how long is the duration for the data to be produced. SIH works by making slight modifications to the original data by adding some bits of data (hidden information, e.g., watermark) disseminated throughout the content. The modifications are expected to be inaudible to the human auditory system but can easily be extracted and decoded using a dedicated SIH detector [7, 20, 70].

- **Usage tracking**

SIH can also be used to trace the source of illegal access, for example tracing the camcorder piracy in theatre, which proposed in [81]. SIH enables the content owner to monitor the usage of digital speech (whether it is legal

or not). If the illegal copy is found, the appropriate amercement can be made to compensate for the commercial loss.

By embedding the identity to the authorized copy (watermark or fingerprint), the owner can identify and track the source of the file. The robustness requirement in the usage tracking application is the most essential since the illegal copies are usually created by using several specific speech processing tools, e.g., a speech recorder.

- **Tampering detection**

The negative impact on the development of speech processing technology allows the illegal users to use the speech forgery techniques to conduct piracy over the Internet or falsify court evidence. The speech forgery techniques are performed by secretly modifying the digital speech to fabricated evidence. SIH can be used by digital forensics to verify whether the digital speech data is the original or fabricated/tampered signal.

Several methods in SIH proposed to detect the tampered signal [56, 117, 126, 128]. In tampering detection, limited robustness is required for distinguishing the tampered signal and the location of tampering. The limited robustness refers to the robustness only against common signal processing operations but not against vicious attacks.

- **Spoofing detection**

Spoofing attacks have become one of the challenging issues in speech communication, especially in speaker verification systems [130, 131]. It attracts special attention to many researchers from not only inside but also outside the speaker recognition community to develop a speech anti-spoofing system [111]. SIH can be utilized in speaker verification systems to detect the spoofed signal by embedding specific information (e.g., watermark) to verify the authenticity of the transmitter, including the sensor and feature extractors. SIH approach can also justify the integrity of the authentication mechanism [37].

2.3 Speaker Anonymization: Voice Privacy Challenge 2020

2.3.1 Definition of Speaker Anonymization

Speaker anonymization (also known as de-identification) is a method of protecting voice privacy. It works by concealing the personally identifiable

information of uttered speech without degrading the linguistic information [113].

A speaker anonymization system must meet four requirements in accordance with the Voice Privacy Challenge 2020:

1. output should be a speech waveform,
2. speaker identity should be hidden,
3. output speech should be natural and intelligible, and
4. anonymized utterances of a given speaker should be different from those of other speakers.

Several open-source corpora are introduced in the Voice Privacy Challenge 2020 to develop a speaker anonymization system, as follows:

(a) LibriSpeech [89], a corpus of English read speech designed for automatic speech recognition (ASR). This corpus contains a total of approximately 1,000 hours of 16-kHz speech.

(b) LibriTTS [133], a corpus of approximately 585 hours of 24-kHz speech that derived from LibriSpeech corpus and designed for text-to-speech (TTS).

(c) VCTK [120], a corpus of approximately 44 hours of 48-kHz English read speech spoken by 109 native speakers with various accents and initially designed for TTS.

(d) VoxCeleb-1,2 [19, 80], an audiovisual corpus designed for speaker verification research. This corpus contains approximately 2,770 hours of 16-kHz speech spoken by 7,360 speakers in various accents and languages.

These corpora were divided into several subsets for training, development, and evaluation. The detail description and statistics of these subsets was explained in the Voice Privacy Challenge 2020's evaluation plan [112].

2.3.2 Evaluation Metrics

The privacy and utility metrics are assessed objectively and subjectively based on the requirements of a speaker anonymization system mentioned in Subsection 2.3.1.

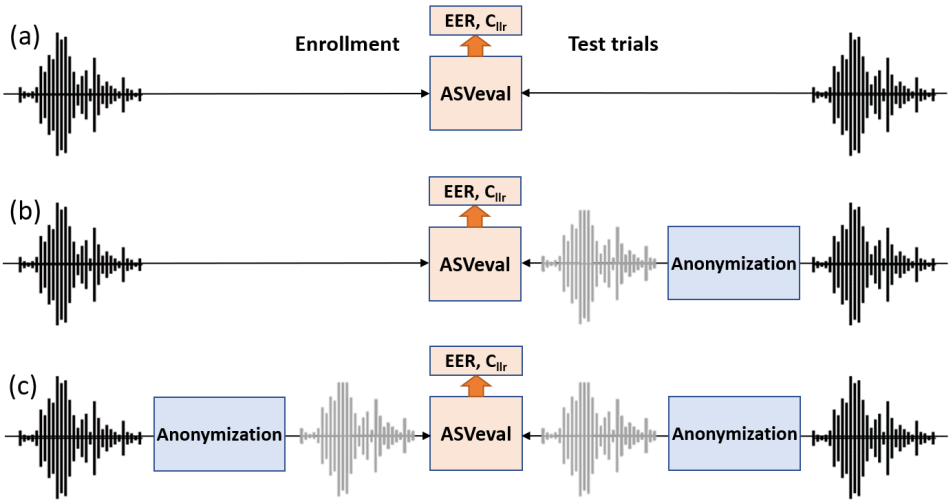


Figure 2.6: ASV evaluation for (a) clean trial and enrollment (o-o), (b) anonymized trial and clean enrollment (o-a), and (c) anonymized trial and enrollment (a-a) [112].

Objective Assessment

The privacy metric should measure the speaker verifiability, whereas the utility metric should measure the ability to preserve the linguistic content. An ASV system is deployed to assess the speaker verifiability metric and an ASR system is deployed to assess utility metric [112]. Both ASV and ASR systems for assessing an anonymization system (hereafter, we refer to these systems as ASVeval and ASReval) are trained on a subset of the LibriSpeech dataset (LibriSpeech-train-clean-360) using a Kaldi toolkit [94].

An evaluation using ASVeval is conducted utilizing probabilistic linear discriminant analysis (PLDA) on the x-vector (state-of-the-art speaker embedding) [107], under the three conditions shown in Fig. 2.6. In ASVeval, the equal error rate (EER) and log-likelihood-ratio cost function (C_{llr} and $C_{\text{llr}}^{\text{min}}$, proposed in [14]), are computed as the objective verifiability metrics.

EER is commonly used in a biometric system to predetermine the threshold values of false acceptance rate (FAR) and false rejection rate (FRR) when they are equals (EER $= P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{fr}}(\theta_{\text{EER}})$). In a speaker anonymization, the FAR ($P_{\text{fa}}(\theta)$) and FRR ($P_{\text{fr}}(\theta)$) are computed as follows [112]:

$$P_{\text{fa}}(\theta) = \frac{\#\{\text{impostor trials with score} > 0\}}{\#\{\text{total impostor trials}\}}, \tag{2.13}$$

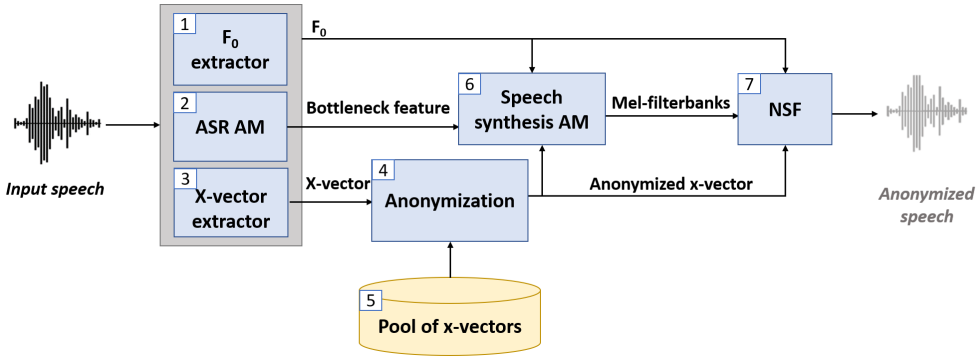


Figure 2.7: Schematic diagram of the primary baseline speaker anonymization system (B1) in Voice Privacy Challenge 2020 [112].

$$P_{\text{fr}}(\theta) = \frac{\#\{\text{target trials with score} \leq 0\}}{\#\{\text{total target trials}\}}. \tag{2.14}$$

Mathematically, log-likelihood-ratio cost function (C_{llr}) in the evaluation set is computed as follows:

$$C_{\text{llr}} = \frac{1}{2}\left(\frac{1}{N_{\text{tar}}}\sum_{i\in\text{targets}}\log_2(1+e^{-\text{LLR}_i}) + \frac{1}{N_{\text{imp}}}\sum_{j\in\text{impostors}}\log_2(1+e^{-\text{LLR}_j})\right), \tag{2.15}$$

where N_{tar} and N_{imp} are the LLR values of the number of target and impostor, respectively. The $C_{\text{llr}}^{\text{min}}$ is referred as a discrimination loss which is estimated by calibration based on monotonic transformation and LLR algorithm.

On the other hand, the evaluation using ASReval is conducted based on a factorized time delay neural network (TDNN-F) acoustic model (AM) [35, 91] and a trigram language model using a Kaldi recipe for a LibriSpeech dataset. The word error rate (WER) is computed to identify the intelligibility of the anonymized speech in comparison with the original speech only in the trial. WER is calculated as the following formula:

$$\text{WER} = \frac{N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}}{N_{\text{ref}}}, \tag{2.16}$$

where N_{sub}, N_{del}, N_{ins}, and N_{ref} are the number of substitution error, deletion error, insertion error, and words in reference, respectively.

Subjective Assessment

In the Voice Privacy Challenge 2020 evaluation plan [113], subjective assessment was conducted to evaluate four metrics, i.e., speaker verifiability, speaker linkability, speech naturalness, and speech intelligibility. This assessment was conducted using large-scale crowd-sourced listening tests with 10-scale rating system. The scenarios for evaluating each metric are as follows:

1. *Speaker verifiability.* The subjective assessment for this metric follows the protocol in ASV spoof countermeasure [111]. It aims to measure the similarity of the speaker between clean enrollment utterance and anonymized trial utterance given the scenario when receiving a phone call. The participants were instructed to judge whether the given voice from the incoming call is similar to the clean utterance.

2. *Speaker linkability.* The subjective assessment for this metric is based on clustering concept. The participants were instructed to place several anonymized utterances from various speakers in dimensional space based on their perceived speaker similarity.

3. *Speech naturalness.* The subjective assessment for this metric aims to measure the naturalness of the speech. The participants were instructed to evaluate the naturalness of a set of speech utterances, which contains of either clean speech or anonymized speech.

4. *Speech intelligibility.* The subjective assessment for this metric is similar to naturalness but with different purpose. In this assessment, the participants evaluated the intelligibility of the given speech.

2.3.3 Baseline Systems

In the Voice Privacy Challenge 2020, two anonymization techniques were introduced as the baseline systems [112]. The primary baseline (B1) system was developed using x-vectors and an NSF model [35]. The second baseline (B2) system was developed based on linear prediction analysis using McAdams coefficient [77]. In this paper, we focus on developing an anonymization system based on the B1.

Speaker anonymization using x-vectors and an NSF model (B1)

The B1 system is primarily built on the idea of separating linguistic content and speaker individuality features from the input speech. The anonymized speech is then synthesized by the extracted linguistic content (to preserve

the linguistic information) and the modified speaker individuality feature. Figure 2.7 shows the block diagram of the B1 system, which consists of seven components: an F_0 extractor, an ASR AM, an x-vector extractor, an anonymization model, a pool of x-vectors, a speech synthesis AM, and an NSF model. The anonymization process is subdivided into the following three main steps:

(i) Feature extraction: extraction of the F_0, a bottleneck feature (as linguistic feature representation using an ASR acoustic model (AM) model [35, 91]), and a speaker individuality feature (x-vector based on [107]);

(ii) X-vector anonymization: modification of the extracted x-vector by averaging a set of candidate x-vectors from the pool of x-vectors; and

(iii) Speech synthesis: speech synthesis using the F_0, the bottleneck features, and the modified/anonymized x-vector based on the speech synthesis AM [35] and NSF [127] models.

The Kaldi toolkit [94] is used in the feature extraction step. The YAAPT algorithm is used as the F_0 extractor. Subsequently, an ASR AM model is built based on factorial time delay neural network (TDNN-F) model architecture [35, 91] and trained using the training data of the LibriSpeech dataset [89] to extract the bottleneck feature. The output of an x-vector extractor constructed using a time delay neural network (TDNN) model [107] and trained using the VoxCeleb-1,2 dataset is used to represent the speaker individuality feature.

In the x-vector anonymization step, the x-vector of a given input speaker is modified by a new pseudo x-vector obtained by averaging a set of candidate x-vectors determined by a given similarity distance range. The candidate x-vectors belong to the pool of x-vectors extracted from the train-other-500 subset of the LibriTTS dataset [133]. The cosine similarity, or probabilistic linear discriminant analysis (PLDA), is used as the similarity distance measure. A smaller set of x-vectors is randomly chosen from a set of most farthest x-vectors as the candidate x-vectors.

As the last step, the anonymized speech is resynthesized using a speech synthesis AM model and an NSF model. Both models were trained using the train-clean-100 of the LibriTTS dataset. The speech synthesis AM model was constructed based on an autoregressive network [35]. This model transforms the input F_0, bottleneck features, and anonymized x-vector into Mel-filterbanks features. Subsequently, the NSF model [127] is used to generate the anonymized speech from the F_0, Mel-filterbanks features, and the anonymized x-vector. Table 2.2 shows the description of B1 with the models and training corpora.

Table 2.2: Description of the primary baseline (B1) of speaker anonymization system: models and corpora [113]. Subscript numbers represent the feature dimensions.

#	Model	Description	Output features	Training dataset
1	ASR AM	TDNN-F Input: MFCC_{40} + $\text{i}-\text{vectors}_{100}$ 17 TDNN-F hidden layers Output: 6032 triphone ids LF-MMI and CE criteria	BN_{256} features extracted from the final hidden layer	Librispeech: train-clean-100 train-other-500
2	X-vector extractor	TDNN Input: MFCC_{40} 7 hidden layers + 1 stats pooling layer Output: 7232 speaker ids CE criterion	speaker $\text{x}-\text{vectors}_{512}$	VoxCeleb:1,2
3	Speech synthesis AM	Autoregressive (AR) network Input: F0 + BN + x-vectors FF * 2+ BLSTM + AR + LSTM * 2 + highway-postnet MSE criterion	$\text{Mel}-\text{filterbanks}_{80}$	LibriTTS: train-clean-100
4	NSF model	sinc1-h-NSF Input: F0 + Mel-fbanks + x-vectors STFT criterion	speech waveform	LibriTTS: train-clean-100
Pool of speaker x-vectors				LibriTTS: train-other-500

Speaker anonymization using McAdams coefficient (B2)

The secondary baseline system (shown in Fig. 2.8) was developed on the basis of modifications to the McAdams coefficient [90]. The McAdams coefficient is related to the adjustment of harmonic frequency distributions, which affects the perception of timbre [77]. Although the results of the secondary baseline were not as good as the first baseline, it requires no training data and is much less complex.

The McAdams coefficient proposed in [77] is a parameter derived on the basis of the additive synthesis method in music signal processing [29]. This method is applied to timbre generation by resynthesizing multiple harmonic cosinusoidal oscillations. by an inverse Fourier series with magnitude and phase shift. Mathematically, the additive synthesis process is expressed as

$$y_{syn}(t) = \sum_{h=1}^{H} r_h(t) \cos(2\pi (h f_0)^{\alpha} t + \Phi_h), \tag{2.17}$$

where $y_{syn}(t)$ is the synthesized signal, h is the harmonic index, $r_h(t)$ is the amplitude, Φ_h is the phase, and α is the McAdams coefficient [77].

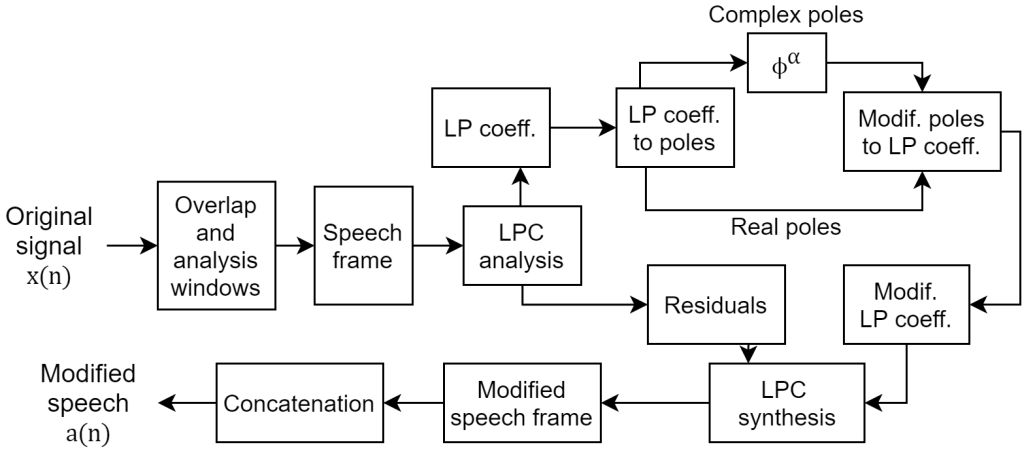


Figure 2.8: Block diagram of speaker anonymization based on McAdams coefficient [90]. "LP coeff." is referred to as linear prediction coefficients. "LPC" is referred to as linear predictive coding. "ϕ" is the angle of poles with a non-zero imaginary part. "α" is the McAdams coefficient.

The McAdams coefficient is used for adjusting frequency harmonics to non-harmonics components that affects the perception of timbre. Prior work on speaker anonymization [90], has shown that the McAdams coefficient can transform the spectral envelope of speech signals and affect timbre perception. The McAdams coefficient was manipulated to alter the formant position of original speech at the frame level on the basis of linear predictive coding (LPC) analysis and a synthesis technique. The procedures of speaker anonymization using McAdams coefficient are as follows.

The original signal in the time-domain ($x(n)$) is first divided into several overlap frames. Each speech frame is then passed through a linear prediction (LP) analysis filter, which is an all-pole filter that mimics the source-filter analysis model of a speech production system. In this study, we used the LP order of 12 ($M = 12$). The LPC analysis is characterized by the following differential equation:

$$s(n) = \sum_{i=1}^{M} c(i)s(n-i) + e(n), \tag{2.18}$$

where $s(n)$ is the speech frame, $c(i)$ is the i-th order LP coefficient, M is the order of LP, and $e(n)$ is the prediction error (residuals). The corresponding transfer function ($H(z)$) for Eq. 2.18 is represented using all-pole

autoregressive filters as follows

$$H(z) = \frac{1}{1 - \sum_{i=1}^{M} c(i) z^{-i}}. \tag{2.19}$$

The LP coefficients ($c(i)$) obtained from the LPC analysis are used to derive the poles (ϕ). The derived poles can be categorized into complex and real poles. Complex poles have non-zero imaginary values, whereas real poles have a zero-valued imaginary term. The McAdams coefficient (α) corresponds to the power of complex poles. The manipulation of alpha results in angle shifting of complex pole positions (ϕ^{α}) either clockwise or counter-clockwise depending on α and ϕ [90]. When $\alpha < 1$, ϕ^{α} is in the counter-clockwise direction when $\phi < 1$ radian and in the clockwise direction when $\phi > 1$ radian. The opposite direction applies when $\alpha > 1$. We investigate McAdams coefficient manipulation when $\alpha < 1$ in this study. Figure 2.9 shows the poles location and frequency-response envelopes obtained from original signal and McAdams coefficient manipulation when $\alpha = \{0.85, 0.9, 0.95\}$.

After shifting complex pole locations by manipulating the McAdams coefficient, both complex and real poles are converted to new LP coefficients ($c'(i)$). These LP coefficients and the original residuals ($e(n)$) are resynthesized as modified speech frames. Finally, the modified speech frames are concatenated using the overlap and add technique to generate the modified speech signal ($a(n)$).

In speaker anonymization, the McAdams coefficient is manipulated as far as possible from the original speech ($\alpha = 1$) with consideration of the sound distortion caused. For example, the anonymization introduced as the secondary baseline in the Voice Privacy Challenge 2020[1] used fixed $\alpha = 0.8$ [112]. Those results indicate the degree of McAdams coefficient manipulation on our perception of speaker individuality [90].

[1] `https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020`

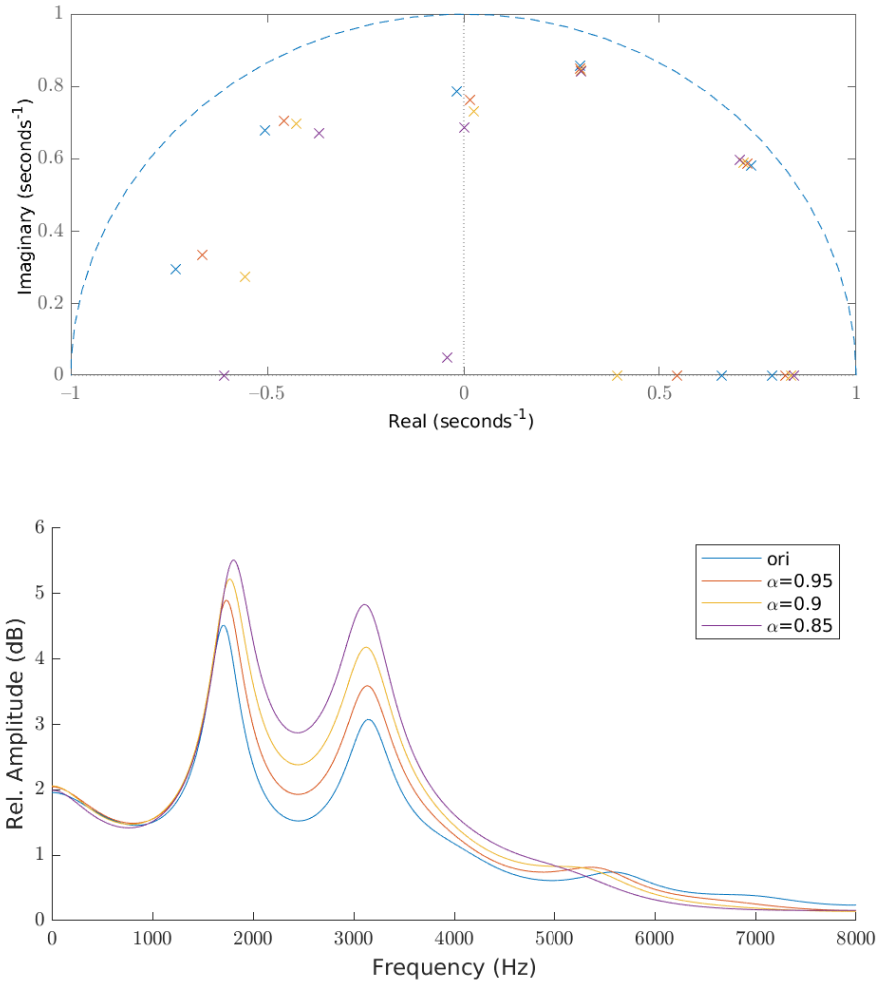


Figure 2.9: Pole locations and frequency-response envelopes of original signal (ori) and modified signals with McAdams coefficients ($\alpha = \{0.85, 0.9, 0.95\}$).

Chapter 3

Content and Privacy Protection for Speech Communication System

As explained in Chapter 2, there are two categories that are mainly adopted for content and privacy protection for speech processing, i.e., cryptography and information hiding. This study adopts the information hiding approach as the focus. This chapter introduces the general proposed framework to secure speech communication which is based on information hiding. Subsequently, the methods to protect the information in speech will be described in the succeeding sections.

3.1 General Proposed Framework of Secure Speech Communication

Speech signals contain various information, including linguistic content, speaker intention, behavior, etc. Although it is difficult to definitely manifest this information, the categorization by Fujisaki [38] highly contributes to model analysis and synthesis processes in speech. Figure 1.1 shows three categories of information expressed in speech: linguistic, para-linguistic, and non-linguistic information. This categorization is mainly based on two conditions: (1) whether the speech is intentionally spoken and (2) whether continuous intensity changes are allowed. Linguistic information is the discrete and categorical content that is intentionally spoken (which can be represented using a set of discrete symbols and grammatical rules). Contrarily, para-linguistic as well as non-linguistic information can be either discrete or continuous. Para-linguistic information and non-linguistic information are distinguish-

able based on the first condition (para-linguistic information is intentionally spoken; otherwise, it is categorized into non-linguistic information).

Although the categorization by Fujisaki has been widely accepted and practically used, the boundaries between those three categories may not always be definite and independent [38]. For example, there is a dispute in defining emotion as para-linguistic information. A speaker can deliberately express emotion in the spoken utterances (para-linguistic), but the emotional state of a speaker is not controllable (non-linguistic) [132]. Subsequently, the factors that influence each category of information in speech are entangled. For instance, the vocabularies chosen as in the linguistic information may infer not only the speaker's emotion but also the speaker's behavior.

Due to those reasons, speech signal processing tasks are complex. In earlier studies, the spectrum analysis, such as the Fourier transform [12], is very useful to analyze the frequency-dependent energy distribution of a waveform (including speech signal). However, a speech signal has its own characteristics that are not considered in the Fourier transform. The source-filter model is then proposed to improve speech analysis and synthesis based on the excitation and modulation mechanisms in the speech production system [36]. Speech signals are processed by imposing several assumptions based on the process of the speech production system (also known as physical assumptions), i.e., (i) speech is produced by a source-filter process, (ii) source is white noise (unvoiced) or pulse train (voiced), and (iii) the vocal tract filter is an all-pole filter (linear prediction) [68, 110] (see Fig. 2.2). This source-filter model has been widely utilized in many works of speech signal processing.

In this study, a framework based on the source-filter model is proposed to protect the content and privacy of speech signals simultaneously. As aforementioned, we know that there is entanglement in the factors that influence each category of information in speech. Accordingly, content and speaker-related information in speech share common features or attributes. We have to investigate these common features to modify the speech signals effectively. If we carried out double modifications for content and privacy protection, not only the speech quality will be drastically reduced, but also each modification will affect the unwanted changes to other information.

The excitation and modulation decomposition in the source-filter model is generally used in speech codecs standards, such as CELP codecs [8]. Consequently, controlling the features related to this composition may cause better robustness of the proposed method. We investigate the features in CELP codecs that are related to both content and speaker traits.

In spite of the advantage of the source-filter model, the excitation and modulation decomposition was reported to work better for low-level signals but not directly associated with high-level information, such as speaker traits [110].

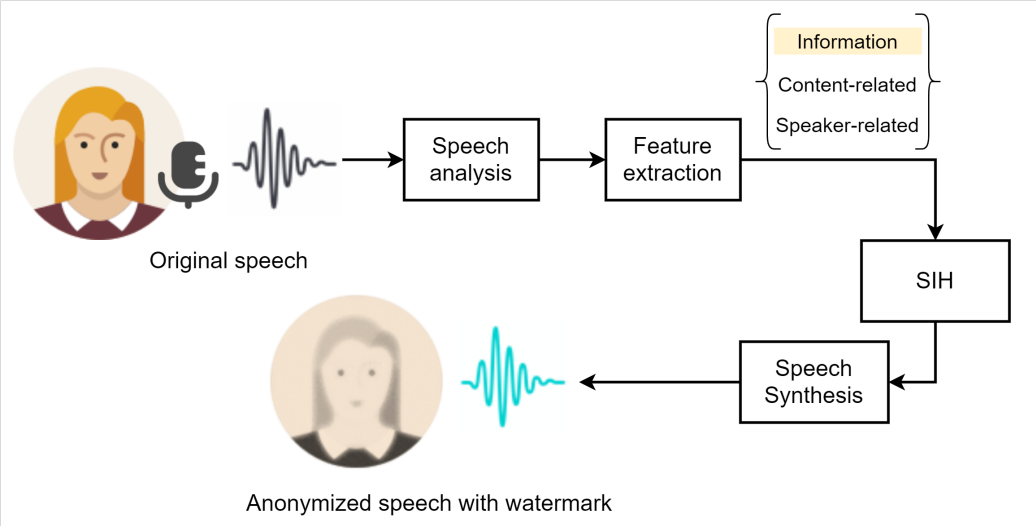


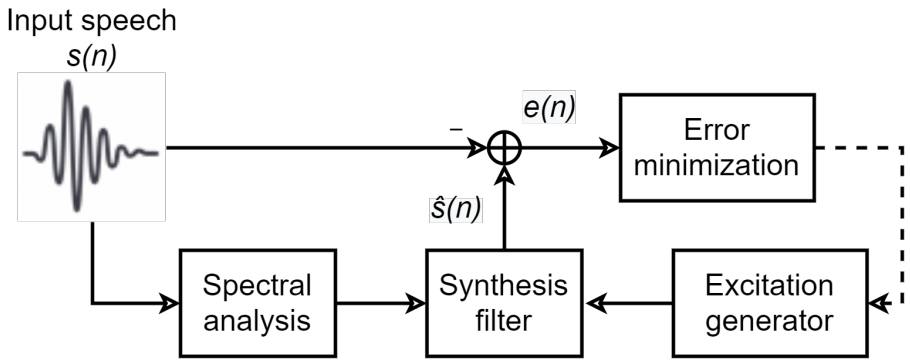
Figure 3.1: General abstraction of proposed framework for content and privacy protection in speech communication.

Hence, we also investigate the current state-of-the-art speaker individuality feature, namely x-vector speaker embedding [107]. More detail explanation about this feature can be found in Section 3.3.

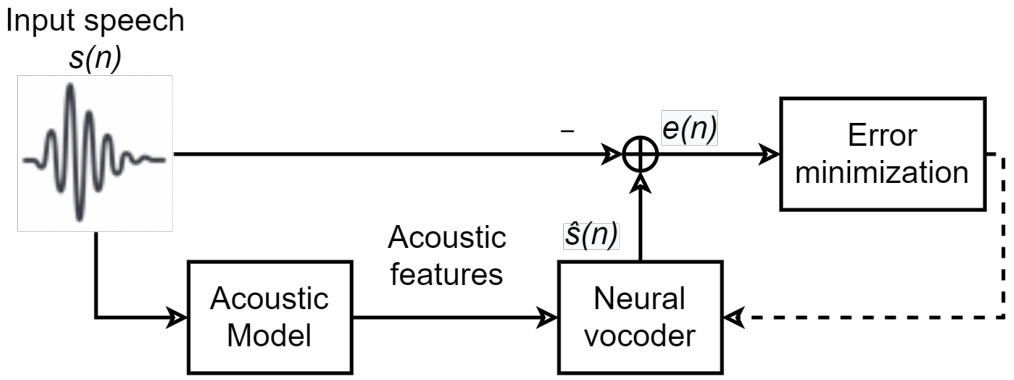
Figure 3.1 shows the general abstraction of our proposed framework for content and privacy protection in speech communication. The proposed framework is mainly comprised of three components, i.e., speech analysis and synthesis, feature extraction, and speech information hiding. We investigated and utilized the existing typical analysis and synthesis methods for speech signals. Subsequently, we conducted an analysis on the robust features used in the analysis and synthesis methods for SIH. These features were then extracted and manipulated for protecting content and privacy in speech which is the main contribution of this thesis. A brief description of each component is explained in the following sections.

3.2 Speech Analysis and Synthesis

One of the substantial components in any speech processing tasks are speech analysis and synthesis method. Input speech has to be convertible into the form that can be processed by a computer. Recently, we can classify speech analysis and synthesis methods into two main groups, including the conventional speech coder and neural vocoder (as shown in Fig. 3.2).



(a) Conventional speech analysis and synthesis coder.



(b) Speech analysis and synthesis based on neural vocoder.

Figure 3.2: Typical Speech Analysis and Synthesis Methods.

The conventional speech coder can be found in various speech codecs standards, especially the one from the CELP family. As explained in Section 2.1, the CELP speech codecs are based on mechanisms in a speech production system with linear predictive coding (LPC). The generic idea is shown in Fig. 3.2a. First, the input speech is processed using spectral analysis. In spectral analysis, the spectral envelope and harmonic structure of the input speech are observed. The spectral envelope is related to short-term correlations, whereas the harmonic structure is related to long-term correlations. By using a linear prediction (LP) analysis filter, we can observe the short-term correlations. Consequently, filtering a speech signal with an LP filter renders a residual signal. The speech coder represents this residual signal in a quantized version (excitation parameters) which is used to synthesize the reconstructed output speech. The error minimization optimizes the quantization noise in

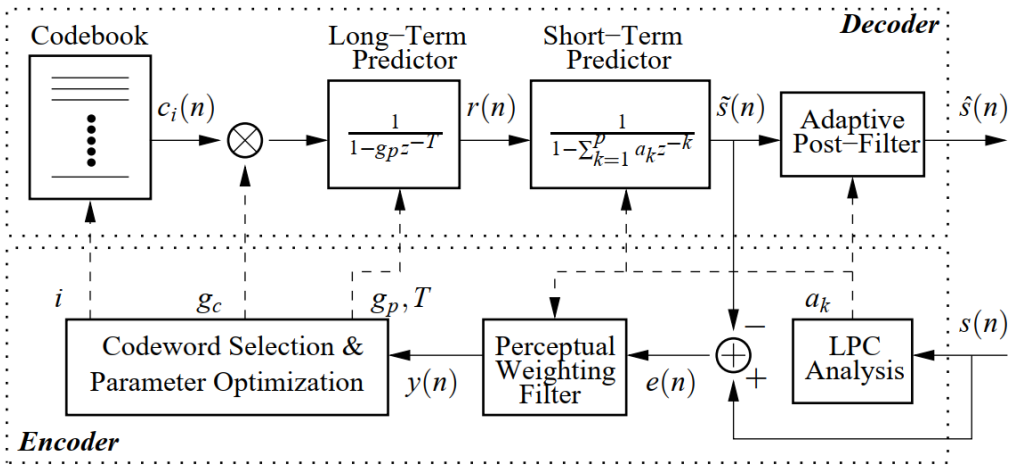


Figure 3.3: General model of the CELP codec [122].

the excitation signal.

When the pulse excitation in the conventional speech analysis and synthesis coder is replaced with a codebook excitation, the coder is referred to as CELP codecs [8]. CELP codecs are based on vector quantization techniques [62]. There are many variants of CELP codecs depending on several parameters, such as pre-processing techniques, long-term and short-term predictors, weighting filter, and post filter. Figure 3.3 shows the general model of the CELP encoder and decoder.

In the encoding process, the input speech is segmented into frames (typically around 20 ms to 30 ms long) and subframes (typically around 5 ms to 7.5 ms long). Next, a short-term LP analysis is applied to obtain the LP coefficients (LPCs). Subsequently, a long-term LP analysis is performed on each subframe with the short-term prediction error as its input. Afterward, the perceptual weighting filter, pitch synthesis filter, and/or modified formant synthesis filter are utilized. The output from those filters is the excitation signal ($y(n)$) which will be converted to a codebook based on a particular codebook searching algorithm and parameter optimization. The CELP output stream obtained from the encoding processes is mainly comprised of the index of excitation codebook, LPCs, gain, and long-term LP parameters. The decoder, on the other hand, decodes and unpacks the CELP output stream as the resynthesized speech ($\hat{s}(n)$). An adaptive post-filter is utilized to enhance the resynthesized signal quality.

Along with the development of machine learning research and the availability of big data, speech synthesis techniques also transformed from parametric synthesis (such as CELP codecs) to neural synthesis [103]. With the statis-

tical approach, the succeeding parametric synthesis approaches utilize the availability of recorded human voices as a function with a set of parameters in the training process of the synthesizer. The parametric synthesis based on a statistical approach is also known as statistical parametric speech synthesis (SPSS). SPSS approaches are reported to have several advantages, including unnecessarily storing audio samples, language in-dependency, and voice characteristics flexibility. The vocoders that developed using statistical parametric synthesis could be categorized into three groups [2]:

1. sinusoidal vocoders: the vocoders that are typically that synthesize speech as "a sum of sinusoidal signals", such as HMPD (harmonic model + phase distortion) [24] and HNM (harmonic + noice model) [33];

2. glottal vocoders: the vocoders that parameterize speech signal into the glottal excitation and vocal tract, such as GlottHMM [95] and GSS (glottal spectral separation) [16];

3. mixed excitation with a spectral envelope: the vocoders that combined the spectral analysis with the reconstruction method of time-frequency components of speech signal, such as STRAIGHT [57, 58] and WORLD [79].

Although those SPSS approaches have improved significantly, the quality of output synthesized speech is relatively not satisfying yet. Recent speech synthesis approaches are based on deep learning methods, which are also referred to as neural vocoders. The term neural vocoder originates from the use of neural networks in the encoding and decoding processes of the vocoder. WaveNet [119] is one of the examples of neural vocoder that has achieved great attention due to its ability to generate new high-quality speech-like waveforms. The core idea of WaveNet is based on an autoregressive model which mathematically expresses the joint probability of the waveform equals to a product of conditional probabilities of the previous time steps. The autoregressive model in WaveNet was constructed using a fully convolutional network with dilated convolutions. Subsequently, some other neural vocoders were also proposed, including WaveGAN [30], neural source-filter (NSF) model [127], etc.

This study investigated the analysis and synthesis model from both the conventional approach and the one using neural networks to propose robust methods for protecting speech content and privacy. As a study case for the conventional approach, we utilized the conventional CELP codecs, which were based on LPC (as shown in Fig. 3.3). Meanwhile, as a study case for the analysis and synthesis based on neural networks, we utilized the

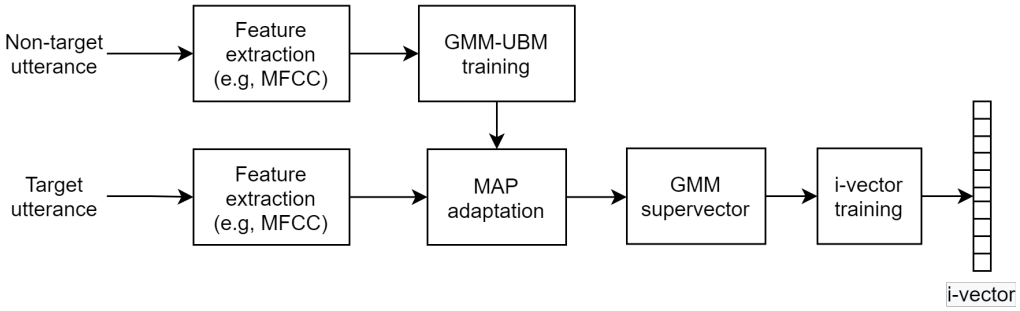


Figure 3.4: Simplified block diagram of the i-vector extraction process.

neural vocoder, which is based on an NSF model (as shown in Fig. 2.7). Consequently, the observed features for SIH are based on the corresponding vocoder.

3.3 Feature Extraction for Secure Speech Communication

In order to modify the speech signal in an effort to secure speech communication, we have to determine the important features or attributes of speech that are related to content and speaker characteristics. Speech signal contains a lot of information, and some are irrelevant to the task that we aimed to accomplish [5]. Accordingly, feature extraction is required as the preceding process in most speech processing tasks. This section explains the features that we observed for protecting speech content and speaker individuality.

Human auditory perception can understand speech from different speakers easily, but machine speech recognition struggles with this task [28, 121]. The environment and voice properties changes affect so much in the performance of machine speech recognition. Meanwhile, the reasons that human auditory perception is so robust against these variations are still unclear [85, 121]. The prior studies showed that the main hurdle is the difficulty in pulling apart the analysis of formant frequencies that contains entangled information of speech content (i.e., the type of speech sound) and speaker-related information (i.e., vocal tract parameters) [121].

This study addresses the analysis of features related to formant frequencies in conventional speech codecs for robust content and privacy protection in speech communication. In the standardized CELP codecs, a speech signal is analyzed based on LPC 3.3. From the LPC analysis, the direct-form of LPCs and the residual signal were obtained. LPCs are often derived into

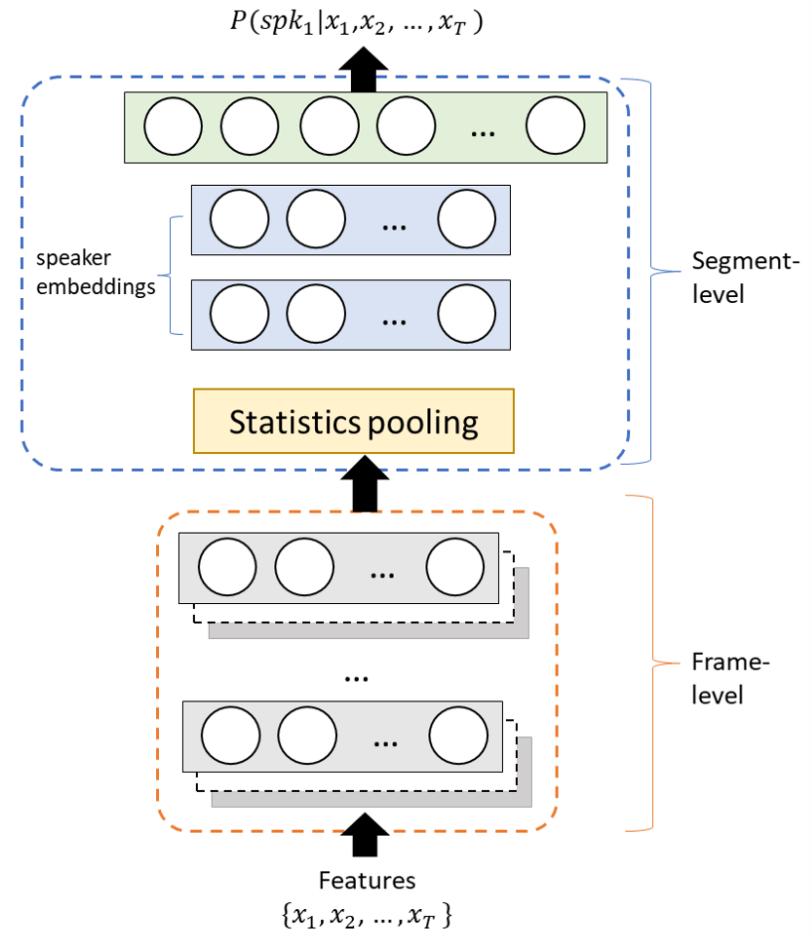


Figure 3.5: A deep neural network (DNN) with an embedding layer architecture as an x-vector extractor [105].

line spectrum pairs (LSPs) or line spectral frequencies (LSFs) for robust representation in the quantization of the excitation codebook. LSFs are highly related to formant frequencies. Consequently, the modification of LSFs is promising to simultaneously affects the content-related and speaker-related information. We described the method for SIH based on LSFs modification in Chapter 4, i.e., SIH by direct modification on LSFs quantization bits and SIH by modifying the McAdams coefficient.

Furthermore, we also investigate the method for privacy protection based on a particular neural vocoder, specifically the NSF model (as shown in Fig. 2.7). The feature extraction for this approach utilizes the state-of-the-art speaker individuality feature in speaker recognition systems, namely x-vector [107]. X-vector is derived from an identity vector (i-vector) modeling approach

with speaker embedding. In the i-vector approach for speaker recognition [26], a low-dimensional vector that is extracted using joint factor analysis (JFA) represents a speech segment. This approach has been reported to reduce high-dimensional sequential speech data to a lower-dimensional fixed-length vector representation that contains more relevant information. In an earlier study, the i-vector model was formed by stacking the mean vectors of the speaker and channel/session subspaces using a Gaussian mixture model with a universal background model (GMM-UBM) [25], as follows:

$$M = m + Vy + Dz + Ux \tag{3.1}$$

where M is a supervector representing a speaker utterance and m is a speaker- and session-independent supervector. V and D represent a speaker subspace, i.e., the eigenvoice matrix and diagonal residual, respectively. Furthermore, U defines the session subspace (eigenchannel matrix). The vectors x, y, and z are the speaker- and session-dependent factors that are assumed to be random variables with normal distributions. Figure 3.4 shows the simplified block diagram of the i-vector extraction process.

In the former i-vector modeling approach, the assumption of a Gaussian feature distribution was made; however, this is not always applicable in practice. Thus, a DNN model was developed to address this issue [104]. Subsequently, to improve the robustness of the i-vector obtained with the DNN model, the process of obtaining an i-vector from a DNN with embedding layers was proposed by Snyder et al. [106, 105]. This i-vector is also known as an x-vector [105]. The architecture of the x-vector extractor is shown in Figure 3.5. We utilized the pretrained VoxCeleb [19, 80] x-vector model provided by David Snyder that are available in the Kaldi toolkit [94, 105]. Chapter 5 shows our proposed method for protecting voice privacy by x-vector modification.

3.4 Secure Speech Communication Based on SIH

This study aims to propose a SIH framework that can simultaneously protect speech content and voice privacy. The protection of speech content is expected to be achieved by speech watermarking. Speech watermarking aims to protect the security in a speech signal by imperceptibly embedding within it a particular message, such as a signature that indicates the speech's ownership. Meanwhile, voice privacy is preserved by using speaker anonymization methods. Such methods are ideally expected to suppress the leakage of personally

identifiable information (PII) while maintaining the linguistic information of the speech signal.

Table 3.1 show the general comparison of alternative solutions for securing speech communication. Although some points may not be applicable to a particular method, this table is expected to overview the common existing solutions and how they differ from the proposed solution. The proposed solution aims to simultaneously protect voice privacy and authenticate the content for any possible user in a widely used digital speech communication system.

Speech watermarking should fulfill at least three requirements: inaudibility (not perceivable by the human auditory system), blindness (detection without the availability of original signal), and robustness against common signal processing operations. The trade-off between inaudibility and robustness has been the most pressing issue in existing speech watermarking techniques [49]. On the other hand, four requirements were determined for the speaker anonymization technique in the Voice Privacy Challenge 2020 [112]: (1) the output had to be a speech waveform, (2) it must maximize the suppression of speaker individuality information, (3) it must preserve speech naturalness and intelligibility, and (4) it must ensure the distinction of voices of different speakers. To evaluate our proposed framework, we conducted the objective evaluation tests based on the standardization in speech watermarking [51] and speaker anonymization [112].

Table 3.1: General Comparison of Alternative Solutions for Secure Speech Communication.

Criteria	**Alternative Solutions**			
	Conventional SIH	**Speaker Anonymization**	**Voice Conversion**	**Proposed Solution**
Advantage(s)	Easy to implement, based on signal processing	Privacy protection to any user	Relatively high performance with enough source data	Privacy and content protection (any user), tampering and spoof detection
Disadvantage(s)	Mostly fragile against speech codecs	Speech quality and naturalness distortion, authentication issue	Limited to some users, can be used to fake speech	Currently has limited payload, more requirements to be considered
Challenge(s) (e.g., most important metrics)	Robustness-inaudibility trade-off, sensitivity of HAS	Utility-privacy trade-off attack models	Naturalness and intelligibility	Inaudibility, robustness, verifiability, intelligibility
Application(s)	Limited (depending on the robustness and payload)	Widely use for secure speech communication	Limited use for privacy protection, assist speaking disabilities	Widely use for secure speech communication

Chapter 4

Content Protection Using Information Hiding Approach

This chapter provides an explanation of content protection using the speech information hiding approach. We propose the SIH method by modifying line spectral frequencies (LSFs) that are mostly found in CELP-based speech codecs. Subsequently, we improve the proposed method by machine learning technique.

4.1 SIH Based on Line Spectral Frequencies Modification

Line spectral frequencies (LSFs) are one of the parameters derived by linear predictive coding that is commonly used in speech technology, including information hiding. It provides strong robustness for information hiding in dealing with speech coding algorithms compared with other typical methods [123, 124, 125, 129]. For example, a direct modification of LSFs for a speech watermarking method using dither modulation-quantization index modulation (DM-QIM) was proposed in [123]. Unfortunately, this method is weak against several signal processing operations. To improve the robustness, Wang et al. proposed an LSFs modification-based speech watermarking technique based on the concept of formant tuning [124, 125]. A linear prediction analysis was conducted to estimate the formants of the speech signal in each frame. Subsequently, the formant tuning was performed by controlling the formant bandwidth with regard to the desired watermark bit.

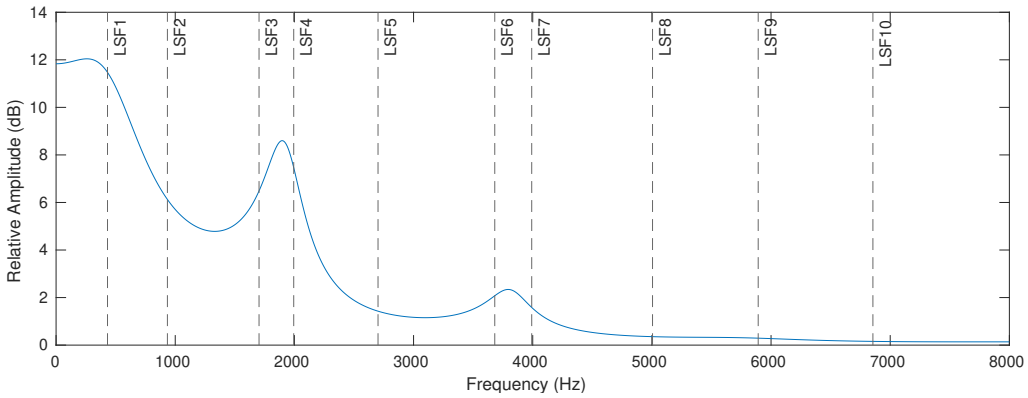


Figure 4.1: Example of the frequency response of a linear predictive filter overlaid with the corresponding LSFs obtained from the tenth-order linear predictive analysis of a 25-ms-long voiced speech segment.

4.1.1 LSFs Concept

Direct quantization of LPCs, $a(i)$, is commonly not applicable in standardized coding algorithms due to its sensitivity. A slight modification to LPCs can cause a significant distortion in the speech since it raises loss to the filter stability. In other words, directly altering the LPCs will most likely causes the poles to be positioned outside the unit circle. Due to this reason, another quantization method is preferable. In the CELP-based speech coding algorithm, LSPs are generated due to their superior quantization characteristics [109, 78].

As described in Eq. (2.4), the LSPs are typically a tenth-order polynomial. This polynomial is computed using two auxiliary polynomials $P(z)$ and $Q(z)$, which are given by:

$$P(z) = A(z) + z^{-11}A(z^{-1}) \tag{4.1}$$

$$Q(z) = A(z) - z^{-11}A(z^{-1}) \tag{4.2}$$

where $P(z)$ is a symmetric polynomial, and $Q(z)$ is an anti-symmetric polynomial. $P(z)$ and $Q(z)$ consist of five complex conjugate pairs of zeros that typically lie on the unit circle. These two polynomials can be regarded as an interconnected tube representation of the vocal tract in a speech production system [78]. The linear combination of these two polynomials represents the actual resonance $A(z)$, which is given by:

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{4.3}$$

Table 4.1: LSF quantization matrix in FS-1016 CELP codec [76].

		Quantization Index															
LSF Index		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	100	170	225	250	280	340	420	500	0	0	0	0	0	0	0	0
	2	210	235	265	295	325	360	400	440	480	520	560	610	670	740	810	880
	3	420	460	500	540	585	640	705	775	850	950	1050	1150	1250	1350	1450	1550
	4	620	660	720	795	880	970	1080	1170	1270	1370	1470	1570	1670	1770	1870	1970
	5	1000	1050	1130	1210	1285	1350	1430	1510	1590	1670	1750	1850	1950	2050	2150	2250
	6	1470	1570	1690	1830	2000	2200	2400	2600	0	0	0	0	0	0	0	0
	7	1800	1880	1960	2100	2300	2480	2700	2900	0	0	0	0	0	0	0	0
	8	2225	2400	2525	2650	2800	2950	3150	3350	0	0	0	0	0	0	0	0
	9	2760	2880	3000	3100	3200	3310	3430	3550	0	0	0	0	0	0	0	0
	10	3190	3270	3350	3420	3490	3590	3710	3830	0	0	0	0	0	0	0	0

The roots of the two polynomials $P(z)$ and $Q(z)$ are referred to as LSFs, which are associated with speech formants [78]. The relationship between LSFs and the frequency response of a linear prediction filter is shown in Fig. 4.1. Speech formants are important aspects of speech perception. Due to this fact, the importance level of formants is considered in the quantization process in the coding algorithm [78]. For example, on the basis of the example in Fig. 4.1, the LSF lines 5 and 6 may be related to the formant F2. However, since the formant F2 is less important than formant F1 (represented by lines 3 and 4), the quantization representation in CELP codecs for the LSF lines 3 and 4 is more detailed than LSF lines 5 and 6.

4.1.2 LSFs Quantization in CELP Codec

We investigate one of the classical standard CELP codec, namely the Federal Standard-1016 (FS-1016) codec, to observe the quantization process of LSFs. In standardized CELP codecs, three to four bits are allocated as quantization bits to represent each LSF extracted from Eqs. 4.1 and 4.2. The FS-1016 CELP codec is one of the first-generation CELP codecs that operate at a bitrate of 4.8 kb/s. This standard configuration is based on gain-shape vector quantization and is designed for 8-kHz sampled speech segmented into 30-ms intervals. We chose this algorithm because the core element of the CELP codec (AbS linear prediction coding) is clearly represented, and thus, adapting the current technique to other more advanced algorithms is possible. Furthermore, the simplicity of the FS-1016 CELP quantization process is derived from using perceptual criteria, and good interpolation properties [17, 96]. These criteria are based on the special properties when the LSPs $A(z)$ is in the minimum phase condition. In the minimum phase condition, the zeros lie on the unit circle, and the zeros of the two polynomials are interlaced [109]. These properties are perceptually meaningful, which should be preserved after quantization [108].

The FS-1016 CELP quantization algorithm uses an independent, non-uniform scalar quantization procedure. The quantization of LSFs is based on the quantization matrix (as shown in Table. 4.1). There are 34 bits per frame that represent the LSFs. Three bits are used for representing LSF 1 and LSFs 6 to 10. Four bits are allocated for representing (LSFs 2 to 5). The quantization procedure may result in non-monotonicity, which leads to the loss of the minimum phase condition (ill-conditioned case) [55]. Accordingly, after the quantization process, an adjustment process is required to restore the monotonicity.

4.1.3 SIH by Direct Modification on LSFs Quantization Bits

On the basis of LSFs quantization in FS-1016, we proposed a speech information hiding method by modifying the least significant LSF quantization bit. In this subsection, we explain the LSF quantization bit modification, embedding, and detection process in detail.

LSFs Quantization Bits Modification

Instead of direct quantization of LSFs, we follow the FS-1016 gain-shape vector quantization to preserve the robustness of the information hiding method for this specific speech codec. The adjustment process in the FS-1016 CELP codec, as mentioned in Subsection 4.1.2, applies a slight modification to the LSB of the allocated bits for LSFs. Since the speech distortion caused by this adjustment is minor, it is promising to obtain an inaudible speech information hiding method by modifying the most insignificant bit of the allocated LSF quantization index.

Figure 4.2 shows the impact in frequency response spectra changing caused by the LSF modification on the basis of the FS-1016 CELP quantization algorithm. From this figure, we can see that the frequency response spectra are shifted when the quantization process is performed. Despite this shifting, the impact is less significant because this quantization algorithm is based on perceptual criteria properties (e.g., the higher formant is less meaningful in perception). Moreover, this figure also shows that since the embedding is based on the standardized CELP quantization method, the different spectra between the quantized LSFs and modified LSFs are insignificant (potential for inaudible modification).

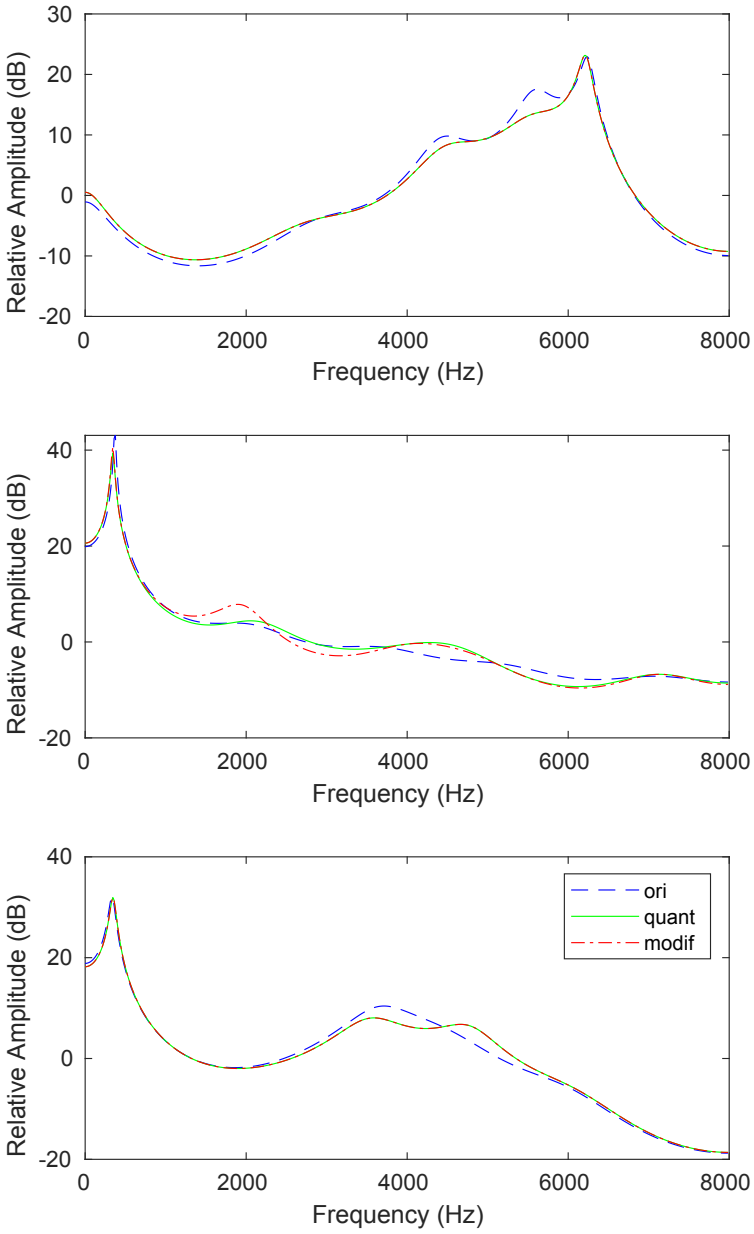


Figure 4.2: Frequency response spectra from actual LSFs (ori), quantized LSFs (quant), and modification of least significant quantized LSFs (modif).

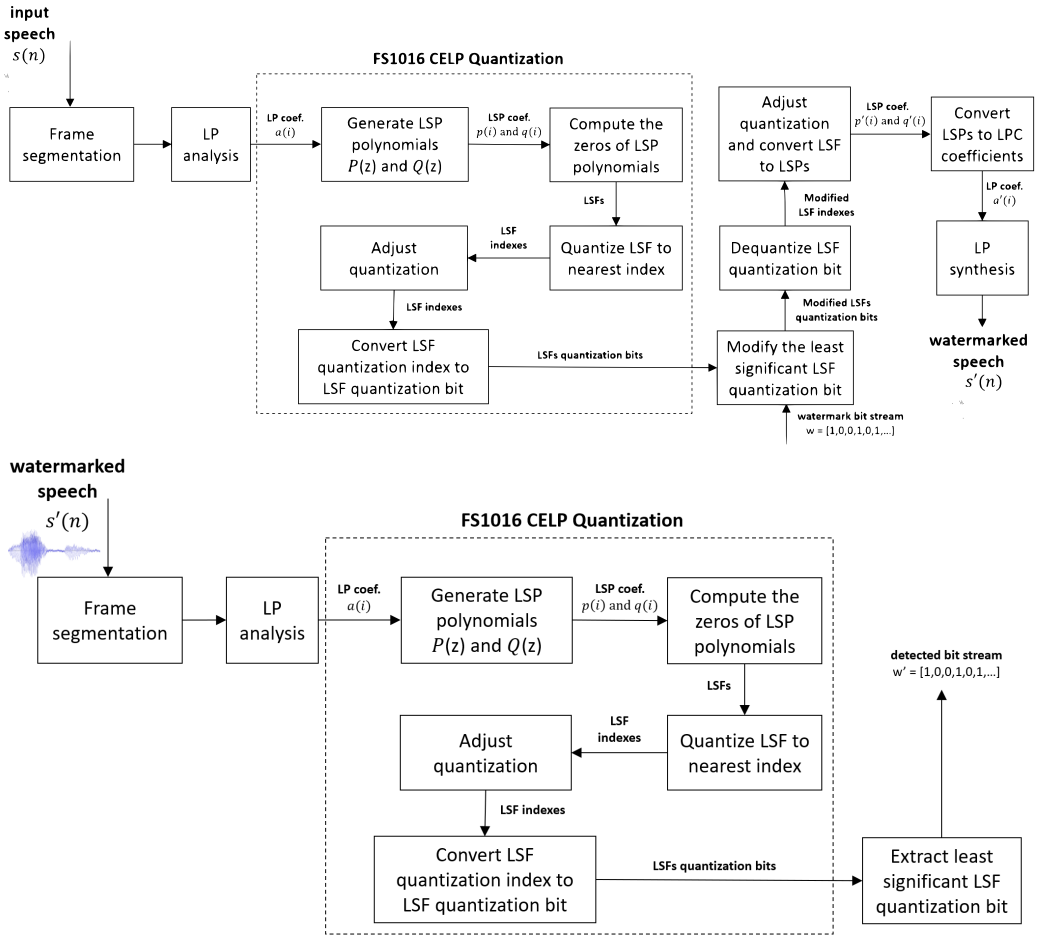


Figure 4.3: Block diagram of proposed SIH based on direct modification on LSFs quantization bits: (top) embedding process and (bottom) detection process.

Embedding

Figure 4.3 (top) shows the embedding process of our proposed method. There are five main steps as follows:

1. The input speech $s(n)$ is segmented into non-overlapping t-length-frames. t denotes the time length in ms (which we will use as our independent variable in Section 4.1.3).

2. A 10-th order linear prediction (LP) filter is used to analyze the framed input signal to obtain the 10 LPCs $a(i)$, where $i = 1, 2, ..., 10$.

3. The LPCs $a(i)$ obtained from the previous step are converted to LSF quantization bits on the basis of the FS-1016 CELP quantization mechanism by using the following substeps:

   (a) generating the LSP polynomials $P(z)$ and $Q(z)$ on the basis of Eqs. (4.1) and (4.2) with regard to the LPCs $a(i)$;

   (b) computing both zeros from symmetrical and anti-symmetrical polynomials on the basis of Descartes' rule to obtain the LSFs;

   (c) quantizing the LSFs on the basis of the LSF quantization matrix in Table. 4.1 to obtain the LSF quantization indexes;

   (d) adjusting the LSF quantization indexes to preserve monotonicity by checking and correcting the ill-conditioned cases;

   (e) converting the adjusted LSF quantization indexes to a binary form as LSF quantization bits.

4. The least significant LSF quantization bits are manipulated in accordance with the watermark bit stream w. After the modification, the dequantization process is performed to obtain the modified LSP coefficients $p'(i)$ and $q'(i)$. Next, these coefficients are converted to LPCs $a'(i)$.

5. Finally, the watermarked speech $s'(n)$ is obtained by using LP synthesis in accordance with the modified LPCs $a'(i)$.

Detection

Figure 4.3 (bottom) shows the detection process of our proposed method. It begins with the first three steps of our proposed embedding process with the watermarked signal $s'(n)$ as the input. Subsequently, we extract the least significant LSF quantization bit as the detected watermarks w'.

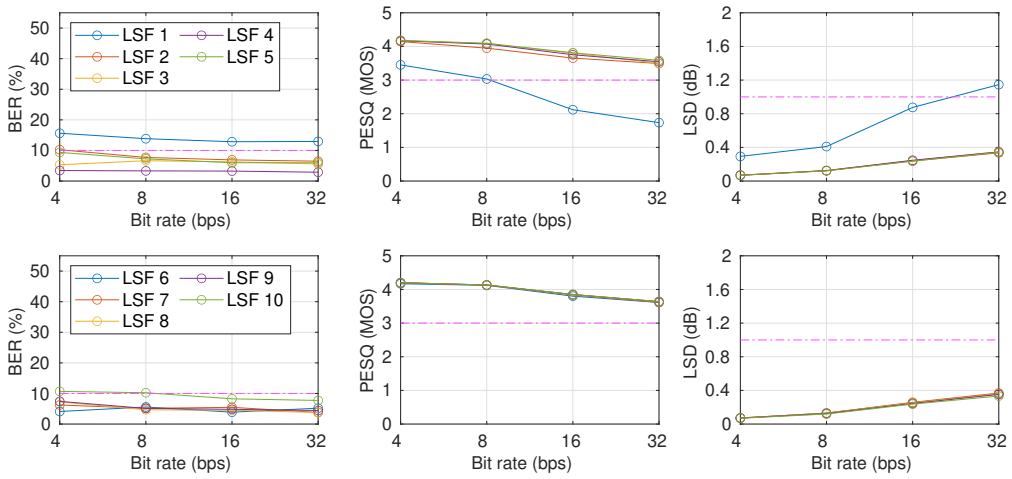


Figure 4.4: Objective evaluation of our proposed method in each LSF quantization bit by using BER, PESQ, and LSD in the original FS-1016 CELP quantization algorithm configuration. The input signal is sampled at 8 kHz and its frame segmentation length t is 30 ms.

Evaluation and Discussion

We evaluated our proposed method using several scenarios to check the feasibility and robustness of our proposed method. First, we investigated our method's feasibility by using the designated configuration (input signals and analysis parameters) of the FS-1016 CELP codec. Then, we utilized another speech dataset with a different configuration to investigate our method's flexibility despite the various input and analysis parameters. We also investigated the possibility of enhancing the robustness and payload of our method. Finally, we compared our method with a typical speech information hiding method, such as LSB and DSS [49], under normal and several signal processing attacks conditions. We performed an objective evaluation to measure the robustness and inaudibility of our proposed method. We calculated the bit error rate (BER) in % for the robustness evaluation, and calculated the LSD [42] and perceptual evaluation of speech quality (PESQ) [97] ITU-T P.862 for the inaudibility evaluation.

Basic Evaluation

The basic evaluation follows our first evaluation scenario. This evaluation aims to check the feasibility of hiding information in the least significant LSF quantization indexes using the FS-1016 CELP codec. As per the aforemen-

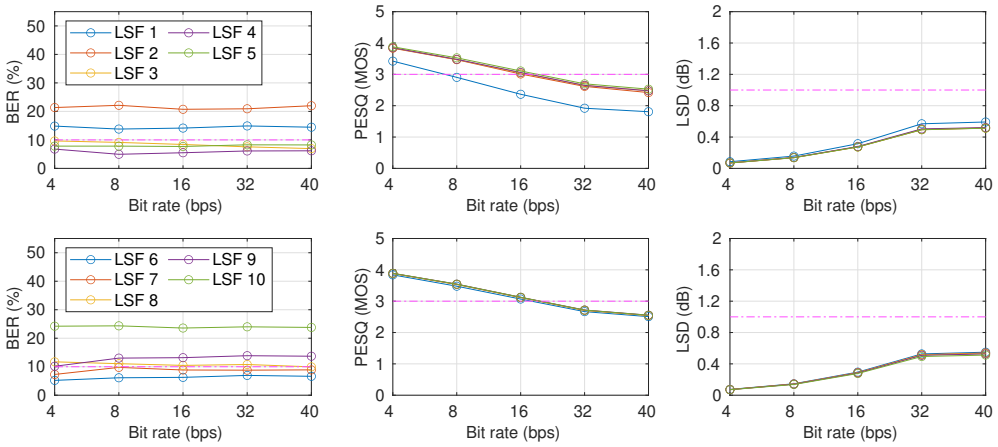


Figure 4.5: Objective evaluation of our proposed method in each LSF quantization bit by using BER, PESQ, and LSD in the adapted quantization configuration. The input signal is sampled at 16 kHz and its frame segmentation length t is 25 ms.

tioned description of the FS-1016 CELP algorithm, an opensource dataset (VoxForge) with ten selected English-spoken speech stimuli was used in the first evaluation scenario. Each stimulus in this dataset is sampled at 8 kHz with 16-bit quantization. The duration of each stimulus ranges between five and ten seconds. The frame length parameter is 30 ms, which is the same as that in the coding algorithm. Since a fixed frame length and one LSF channel is used, the maximum available payload is only 33 bps. Due to this limitation, we analyzed the performance of our proposed method in various bitrates (4, 8, 16, and 32 bps).

Figure 4.4 shows the result of the basic evaluation. This figure confirmed the feasibility of hiding information in the speech by the proposed method. The adequate detection rate could be obtained despite the watermark position in any LSF, except LSF 1. The modification of LSF 1 caused a significant distortion to the watermarked signal. LSF 1 often represents the first formant that is significantly meaningful for speech perception. Thus, changing this parameter is not recommended for information hiding.

The inaudibility of our proposed method can be represented in Fig. 4.4 at the second and the third columns. The perceptual quality of the watermarked signal is good (PESQ score almost around four4, even in the at a high-bitrate). Along with the perceptual quality, the sound distortion is also small enough (LSD is less than 1 dB).

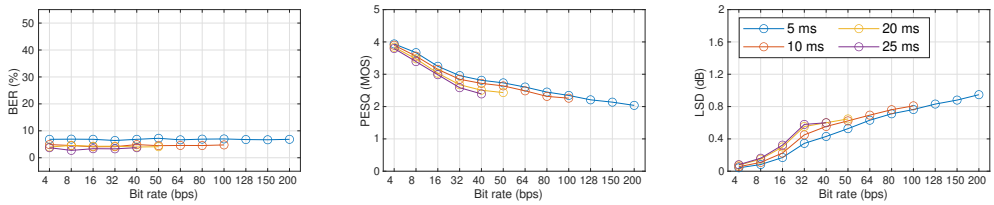


Figure 4.6: Objective evaluation results of the proposed method in comparison with several frame segmentation lengths (5, 10, 20, and 25 ms).

Robustness Evaluation

Unlike the input parameter in the FS-1016 CELP codec, we utilized the ATR Japanese speech dataset (B set) [65], which is sampled at 16 kHz, to investigate our proposed method's robustness. Twelve stimuli were selected from this dataset for our evaluation. Each signal in this dataset has an 8.1-sec duration length. In this subsection, we aim to investigate whether our method can work regardless of the different input and analysis parameters.

Figure 4.5 shows the objective evaluation results of our proposed method with a 25-ms-long analysis-synthesis frame. Although there is a slight drop in performance, the overall result in this scenario ties well with that shown in Fig. 4.4. In most cases, the robustness and inaudibility when hiding in each LSF are sufficient (BER around 10%, PESQ around 3, and LSD less than 1 dB), except for LSFs 1, 2, and 10. Thus, LSFs 1, 2, and 10 are not recommended as embedding mediums.

In summary, this result highlights that our proposed method is robust enough to deal with different segmentation lengths and input signals sampled at the different sampling frequencies. The compression in the quantization process does not cause significant defects in the embedded watermarks. Moreover, due to the fact that the process in our proposed method is based on AbS with the FS-1016 CELP codec, the robustness for this coding algorithm can be assured.

Further Potential Improvement

One of the straightforward methods to improve the robustness and payload of our proposed method is by using multiple embedding or reducing the duration of the analysis-synthesis frame. In this subsection, we investigate the ways to improve the robustness and payload considering the impact of degradation in sound quality.

On the basis of the detection accuracy evaluation results in Figs. 4.4 and

Table 4.2: Evaluation results for multiple embedding in three selected LSFs (LSF 4, 6, and 7)

Variable	**Evaluation Score**	**bit rate (bps)**				
		12	**24**	**48**	**96**	**120**
Payload	BER %	8.54	8.59	8.7	8.56	8.54
	PESQ (MOS)	3.81	3.39	3.00	2.57	2.42
	LSD (dB)	0.08	0.16	0.33	0.59	0.61
		4	**8**	**16**	**32**	**40**
Robustness	BER (%)	3.73	3.73	3.36	3.31	3.73
	PESQ (MOS)	3.80	3.39	2.99	2.58	2.39
	LSD (dB)	0.08	0.16	0.32	0.58	0.60

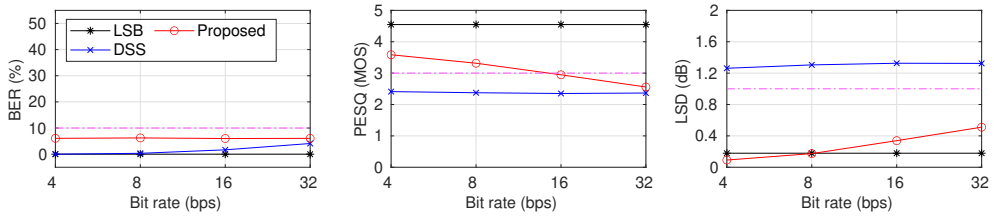


Figure 4.7: Objective evaluation result of comparative methods under normal conditions based on detection accuracy (BER) and inaudibility (PESQ and LSD).

4.5, we selected the three most robust least significant LSF quantization bits (LSFs 4, 6, and 7). We checked the improvement of payload and robustness performance of multiple bit embedding in consideration of the sound quality degradation impact. The evaluation for payload improvement was conducted by inserting three different watermark binaries into the selected LSFs. Thus, the payload could be improved threefold. Another evaluation for robustness improvement followed the repetitive coding concept. A watermark bit is duplicated into three watermarks, which were then embedded into the selected LSFs. The detected watermark bit was determined by calculating the mean value of those three watermarks and classifying them into binary 0 or 1 with a threshold of 0.5.

Table 4.2 shows the evaluation results for multiple embedding. The results suggest that multiple embedding could improve both the payload and robustness of the proposed method with an almost similar sound quality with single embedding. By embedding three different watermark bits into three LSFs, the detection accuracy is also similar to that of single embedding (BER is less than 10%). This result shows that we can also attempt to embed

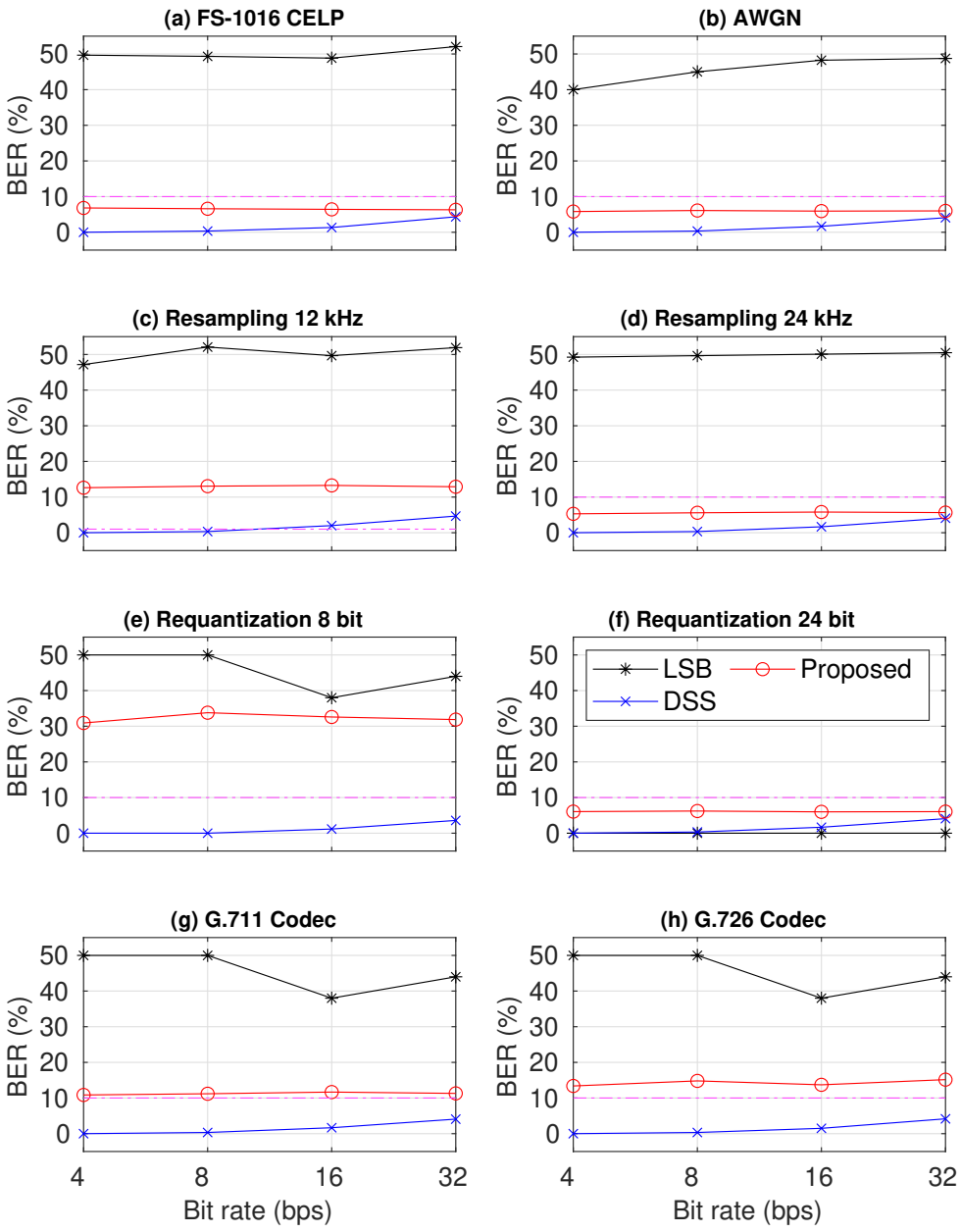


Figure 4.8: Comparative robustness evaluation of our proposed method (single embedding), LSB, and DSS against signal processing attacks: (a) FS-1016 CELP codec, (b) Gaussian noise addition (AWGN), (c) down-sampling to 12 kHz, (d) up-sampling to 24 kHz, (e) requantization to 8 bit, (f) requantization to 24 bit, (g) G.711 codec, and (h) G.726 codec.

Table 4.3: Optimization result using a combination of multiple embedding and varying frame lengths.

Bit rate (bps)	**50**	**100**	**200**	**400**	**800**	**1600**
BER(%)	4.033	5.144	6.626	10.761	11.461	18.816
LSD	0.144	0.292	0.313	0.636	0.901	1.065
PESQ	3.499	3.042	2.996	2.479	2.065	1.833

the watermark stream into other LSF quantization bits (LSF 3 5 8, and 9) as a further prospective improvement. Moreover, the evaluation result of multiple embedding also shows that we could use repetitive coding if our system requires a higher detection accuracy (BER is less than 5%).

Improving the payload robustness is also likely to be achieved by reducing the fixed frame segmentation length of t. Figure 4.6 shows the comparison result of the objective evaluation using BER, PESQ, and LSD where the frame segmentation varies from 5 to 25 ms. This result indicates that the high detection accuracy could be achieved at a high bitrate, although the frame segmentation length is short (BER is less than 10%). As for the inaudibility evaluation result, our proposed method could satisfy the threshold of the LSD score even at a higher bitrate. In contrast, the result of the PESQ evaluation shows the constraint in speech quality degradation at a higher bitrate.

Improving payload and detection accuracy could be achieved using either multiple embedding or a shortened frame length (as shown in Table. 4.2 and Fig. 4.6). On the basis of these results, we optimized the embedding capacity by using both these methods. First, we improved the detection accuracy by assigning weights to each LSF on the basis of the detection accuracy obtain in the basic evaluation. Subsequently, on the basis of the weights, we performed majority voting to determine a detected watermark. Finally, we preserved the speech quality by not embedding to LSF 1 and 2, and optimizing the repetitive embedded bits (minimizing the embedded bits but preserving the accuracy) for a watermark.

Comparison with Typical Speech Information Hiding Methods

We performed a comparative evaluation between our proposed method (single embedding in LSF 4) and two typical methods (LSB and DSS) with the objective evaluation of robustness and inaudibility. The traditional LSB method alters the most insignificant quantization bits of the speech signal with watermarks to maintain the inaudibility of the distortion. In contrast, the DSS method spreads the desired watermarks over the whole frequency

band to ensure robustness. The comparative evaluation was conducted by using the ATR dataset. The bitrate ranges from 4 to 32 bps.

Figure 4.7 shows the comparative evaluation result under the normal condition (without considering any attacks). This result indicates that the LSB method could achieve a high accuracy and inaudibility even at a high bitrate. Contrarily, the DSS method caused a significant distortion to the watermarked signal despite the high detection accuracy. Our proposed approach works in between the LSB and DSS methods. Although it could not achieve perfect accuracy (BER is less than 10%), our proposed method could achieve better inaudibility compared with the DSS method. In other words, we could say that our proposed method is reliable (robust and inaudible) at low bitrates (up to 16 bps for single embedding).

In the actual speech communication system, several signal processing operations often invaded the transmitted speech. Figure 4.8 denotes the robustness evaluation result of our comparative methods against several signal processing operations. In most cases, the LSB method (black line) is very fragile against any attacks, whereas the DSS method (blue line) is very robust. Even though it is not as robust as the DSS method, our proposed method (red line) could provide robustness against several operations. Figure 4.8(a) confirmed our hypothesis that our method is robust against the specific FS-1016 CELP codec. Moreover, our proposed method is also robust against noise addition (AWGN) (Fig. 4.8(b)), resampling (Fig. 4.8(c–d)), and requantization to higher bits (Fig. 4.8(f)). The requantization to lower bits (Fig. 4.8(e)) remains as a limitation robustness in our proposed method. However, our proposed method is somewhat robust against other speech codecs, e.g. G.711 and G.726 (Fig. 4.8(g–h)).

4.1.4 SIH by McAdams Coefficient Modification

Instead of direct modification on LSFs quantification bits, we also proposed a novel SIH method based on LSFs modification by using McAdams coefficient. McAdams coefficient is a parameter that controls the harmonic structure of speech spectral information which is reported to have relationship with speaker individuality information [77, 90, 112]. By modification of McAdams coefficient, we hypothesized that our proposed method can simultaneously protect both content and privacy of the speaker (as our goal of this study).

Embedding

Figure 4.9 shows the block diagram of our embedding process. The detail embedding process is as follows:

1. As the first step, we generated the anonymized signals from the original signal ($x(n)$) using two different McAdams coefficients (α_0 and α_1). The anonymization procedure follows the steps in Fig. 2.8.

2. Next, the original speech was segmented into speech frames with frame length depending on the watermarking payload. Subsequently, the speech frame was analyzed using linear predictive coding (LPC) with an order of 20 ($M = 20$).

3. From LPC analysis, we obtained the linear prediction coefficients (LP coefficients) and residuals. These LP coefficients ($c(i)$) were then used to derive the pole positions. The derived poles were comprised of complex poles (poles with non-zero-valued imaginary terms) and real poles (poles with zero-valued imaginary terms). The shift of complex poles position (ϕ^{α}) was resulting in the angle shifting to either clockwise or counter-clockwise of the complex positions [90]. After the modification of McAdams coefficients α, the modified complex poles and the real poles were converted back to LP coefficients. The anonymized speech frame was resynthesized from these LP coefficients and the original residuals.

4. After obtaining the anonymized signals, we normalized them to be in similar relative loudness (fixed a target peak level in decibel relative to full scale (dBFS)) and range of frequency components (using a bandpass filter (BPF)). The cut-off frequencies for the BPF were 125 Hz and 4 kHz.

5. Finally, we constructed the watermarked signal ($y(n)$) by frame-by-frame concatenation of the anonymized signals obtained by bit inverse according to watermarked bit-stream.

Detection

We found that the anonymized signals from different McAdams coefficients carried different amounts of power spectrum, specifically in the lower frequency components. Using a higher McAdams coefficient results in a higher power spectrum in the lower frequency. On the basis of this characteristic, we determined a power threshold for the blind detection process (shown in Fig. 4.10). The detection process was conducted by comparing the power spectrum obtained by fast Fourier transform (FFT) of the watermarked signal ($|Y(\omega)|$) in a specific frequency range with the designated threshold θ. The detail process of detection is as follows:

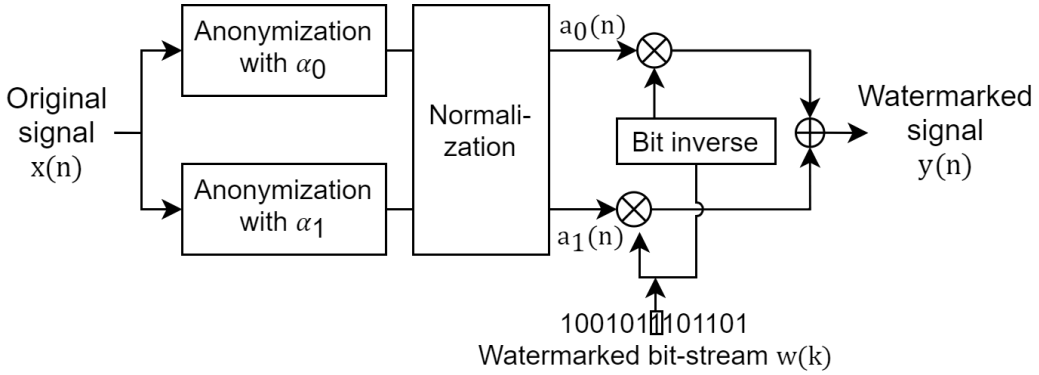


Figure 4.9: Block diagram of embedding process. α_0 and α_1 are the McAdams coefficients for representing binary bit "0" and "1". $a_0(n)$ and $a_1(n)$ are the output anonymized speech in time domain.

1. First, the watermarked signal $y(n)$ was segmented into speech frames with frame length depending on the watermarking payload.

2. Then, each watermarked speech frame was passed through a band-pass filter with cut-off frequencies 125 Hz (lower) and 4 kHz (upper).

3. The filtered speech frame was then analyzed by FFT to derive the power spectrum ($|Y(\omega)|$). This power spectrum was compared to a threshold (θ) to determine the detected bit information in each speech frame. For example, if the $|Y(\omega)|$ is lower than θ, the detected bit from the observed speech frame is assigned as '0'. On the other hand, if $|Y(\omega)|$ is higher than θ, the detected bit is assigned as '1'.

Evaluation and Discussion

We used LibriSpeech [89] and VCTK [120] datasets (development and testing sets) that were provided in VP2020. LibriSpeech is an English speech corpus designed for automatic speech recognition (ASR) research sampled at 16 kHz [89]. VCTK is an English speech corpus that contains 109 native speakers with various accents and was designed for text-to-speech (TTS) research sampled at 48 kHz. The development and training data of both datasets in VP2020 consists of more than 20,000 utterances from almost 200 speakers. The sampling rate of the speech data is set to 16 kHz. Similar to the secondary baseline [90], we do not need any training data for our proposed method. The McAdams coefficient used to represent bit "0" was 0.6 ($\alpha_0 = 0.6$) and bit "1" ($\alpha_1 = 0.8$) was 0.8.

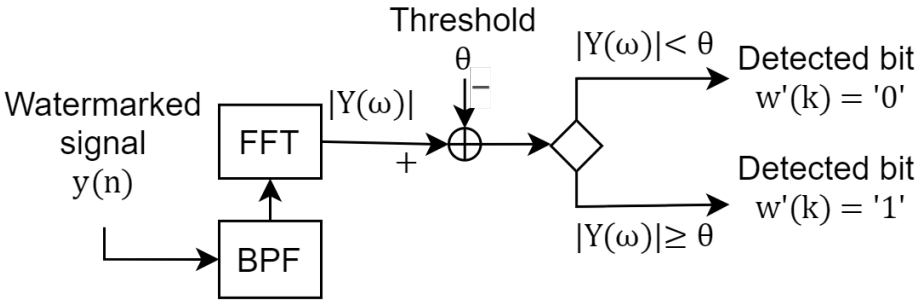


Figure 4.10: Block diagram of blind detection process. "BPF" stands for the band-pass filtering. "FFT" stands for fast Fourier transform. $|Y(\omega)|$ is the power spectrum of the watermarked signal $y(n)$ obtained by FFT. θ is the power spectrum threshold for blind detection process. $w'(k)$ is the detected watermark bit of the k-th frame.

For speech watermarking, we also evaluated the speech quality and robustness of our method with a total 100 randomly selected utterances from the LibriSpeech and VCTK datasets. Since the original signal was not available, we used MOSNet, the pretrained mean opinion score (MOS) predictor proposed in [71]. MOSNet is an objective evaluation tool based on deep learning approach for predicting human MOS ratings in a voice conversion system. Subsequently, we evaluated the robustness of our propsed method as suggested in [51] by calculating the bit error rate (BER) and balanced F1-score during normal (no attack) operations along with several signal processing operations, including noise addition, resampling, requantization, compression, and speech codecs. We also examined the security level by calculating the false acceptance rate (FAR) and false rejection rate (FRR). The maximum acceptable BER threshold as the robustness indication is 10% [51]. We embedded random binary streams with payloads of 2, 4, 8, 16, and 32 bps and varied the detection threshold in the order of lower to higher payloads (0.15, 0.09, 0.05, 0.02, and 0.01, respectively). Due to the space limitation, the results reported in this paper are mainly in mean value.

Table 4.4 shows the MOSNet results of original signal, anonymized signal with McAdams coefficients $\alpha = 0.8$, and the output signal of our proposed method with various payloads. We can see that there was a speech quality degradation (MOS degraded from 3.15 to 2.70) caused by the McAdams coefficient-based anonymization method, while in contrast, the proposed method could maintain a similar MOS even with a relatively high payload.

We carried out a robustness test by calculating the detection accuracy from the output speech after several common signal processing operations. Figures 4.11 and 4.12 show the robustness test results. We examined nine cases: no

Table 4.4: MOSNet evaluation results.

	payload (bps)	MOS
original	-	3.15 ± 0.49
anonymized	-	2.70 ± 0.18
proposed method	2	2.73 ± 0.20
	4	2.73 ± 0.21
	8	2.70 ± 0.19
	16	2.67 ± 0.18
	32	2.60 ± 0.18

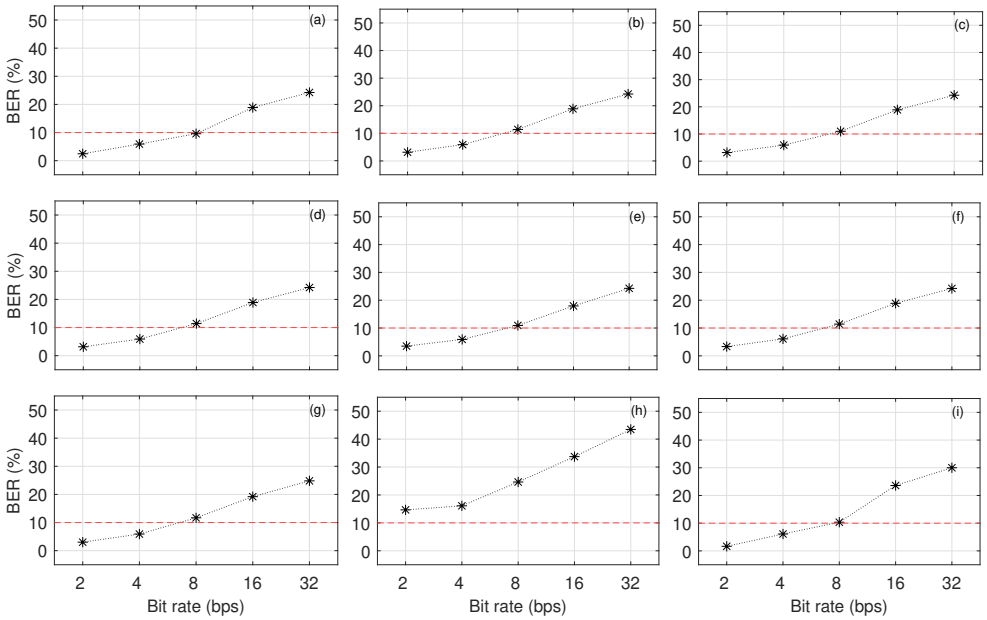


Figure 4.11: Robustness test results in terms of BER (bit error rate) in nine cases: (a) normal, (b) AWGN, (c) resample-8, (d) resample-24, (e) requant-8, (f) requant-24, (g) mp3, (h) flv, and (i) G723.1.

attack (normal), addition of white Gaussian noise (AWGN), downsampling to 8 kHz (resample-8), upsampling to 24 kHz (resample-24), bit compression to 8 bits (requantize-8), bit extension to 24 bits (requantize-24), MP3 compression (MP3), flash video format (flv), and G723.1 codec. For AWGN processing, the signal to noise ratio used is 40 decibel (dB). Meanwhile, the bitrate range for MP3 compression was from 220-260 kbps (kilo bits per second). The bitrate of G723.1 codec was 5.3 kbps with algebraic code-excited linear prediction (ACELP) algorithm. As we can see, our proposed method was robust against

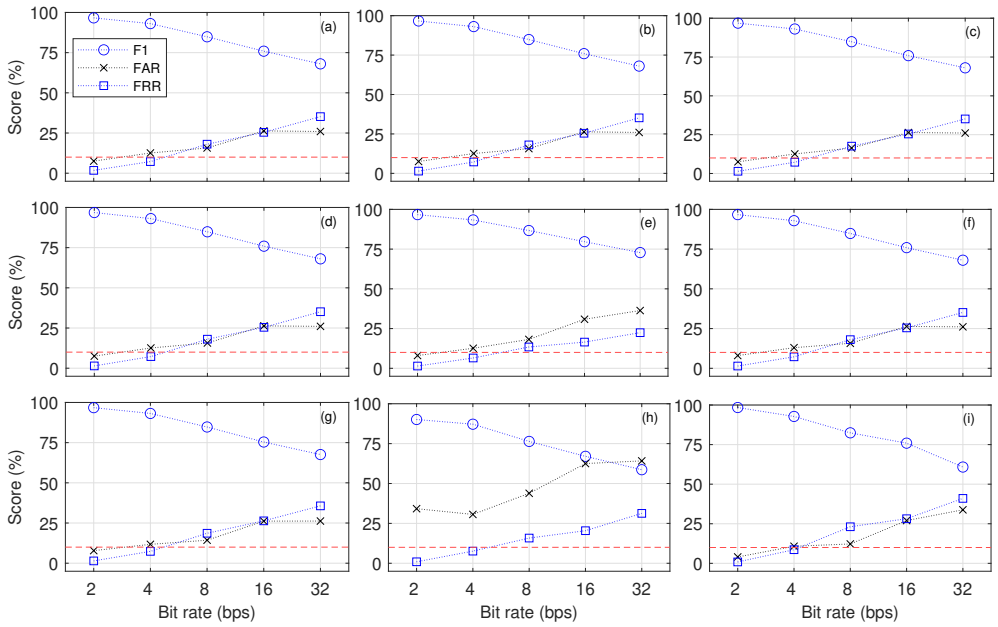


Figure 4.12: Robustness test results in nine cases: (a) normal, (b) AWGN, (c) resample-8, (d) resample-24, (e) requant-8, (f) requant-24, (g) mp3, (h) flv, and (i) G723.1. For metrics were used for the robusness evaluation, including F1 (F1-score), FAR (false acceptance rate), and FRR (false rejection rate).

other operations (the BER was similar to the normal case), except for the conversion to video codec (flv). The results here demonstrate that our method is suitable for watermarking purposes, since the BER for 4 bps satisfied the robustness criteria (BER $< 10\%$). The results also suggest that the security level indicated by the FAR, FRR, and F1-score is adequate for payloads up to 4 bps.

4.2 Improving Robustness of SIH using Machine Learning

We previously proposed a watermarking framework for improving the security of McAdams-coefficient-based speaker anonymization [75]. Two fixed McAdams coefficients were used to represent the binary bit information for speech watermarking. These values were chosen on the basis of the second baseline speaker anonymization system in the Voice Privacy Challenge 2020 [112] and the optimal gap for stochastic McAdams-coefficient-based speaker watermarking [90]. The further away the McAdams coefficient is from the

original speech ($\alpha = 1$), the greater level of anonymization (better performance in reducing speaker verifiability metrics) [90]. However, this advantage results in more speech distortion (degrades speech intelligibility and naturalness). Too much distortion is non-compensable for speech watermarking since speech quality relates to one of the most important requirement in speech watermarking, i.e., inaudibility. Our watermarking processes were conducted in the manner similar to signal modulation. The watermark detection process was based on the threshold of a certain parameter. The experimental results indicated that our watermarking framework could be applied to improve the security of speaker anonymization with a limitation of relatively low payload.

In contrast to the related studies on speaker anonymization [75, 90, 112], we consider a McAdams coefficient closer to the original speech and a smaller shift to maintain the inaudibility criteria on speech watermarking. A smaller shift means that the McAdams coefficient for representing bit-0 (α_0) is close to that representing bit-1 (α_1). We developed a random forest classifier to detect embedded watermarks. We then investigated the optimal McAdams coefficient that can balance inaudibility with the blind-detectability robustness.

4.2.1 McAdams coefficient manipulation

The manipulation of the McAdams coefficient follows the block diagram shown in Fig. 2.8. The original signal in the time domain ($x(n)$) is first divided into several overlap frames. Each speech frame is then passed through a linear prediction (LP) analysis filter, which is an all-pole filter that mimics the source-filter analysis model of a speech production system. In this study, we used the LP order of 12 ($M = 12$).

The LP coefficients ($c(i)$) obtained from the LPC analysis are used to derive the poles (ϕ). The derived poles can be categorized into complex and real poles. Complex poles have non-zero imaginary values, whereas real poles have a zero-valued imaginary term. The McAdams coefficient (α) corresponds to the power of complex poles. The manipulation of alpha results in angle shifting of complex pole positions (ϕ^α) either clockwise or counter-clockwise depending on α and ϕ [90]. When $\alpha < 1$, ϕ^α is in the counter-clockwise direction when $\phi < 1$ radian and in the clockwise direction when $\phi > 1$ radian. The opposite direction applies when $\alpha > 1$. We investigate McAdams coefficient manipulation when $\alpha < 1$ in this study. Figure 2.9 shows the poles location and frequency-response envelopes obtained from original signal and McAdams coefficient manipulation when $\alpha = \{0.85, 0.9, 0.95\}$.

After shifting complex pole locations by manipulating the McAdams coefficient, both complex and real poles are converted to new LP coefficients ($c'(i)$). These LP coefficients and the original residuals ($e(n)$) are resynthesized as

modified speech frames. Finally, the modified speech frames are concatenated using the overlap and add technique to generate the modified speech signal ($a(n)$).

4.2.2 Data-embedding process

Figure 4.9 the block diagram of the data-embedding process. This embedding process is based on signal modulation (similar to our previous study [75]). Generally, we use two McAdams coefficients to represent binary information (bit "0" and bit "1"). The watermarked bit-stream ($w(k)$) which is comprised of binary information, is embedded in the following steps:

Step 1: The original signal ($x(n)$) is modified on the basis of the McAdams coefficient manipulation process explained in Subsection 4.2.1. Two McAdams coefficients are used in the embedding process to represent binary bit-"0" (α_0) and bit-"1" (α_1). The gap between α_0 and α_1 can be regarded as the scaling factor of watermarking. A larger gap creates a stronger watermark but increases distortion. Next, a normalization method is applied based on the ratio of the power spectral density of both modified signals to compensate for the gap in spectral shift at frame transition. The results of this step are two speech signals with new McAdams coefficients ($a_0(n)$ and $a_1(n)$).

Step 2: The watermarked bit-stream ($w(k)$) is set in accordance with the hidden information in a binary stream representation, e.g., $w(k) = \{1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1\}$.

Step 3: The watermarked signal ($y(n)$) is determined by bit-inverse switching between the modified signals (a_0 and a_1) and the watermarked bit-stream ($w(k)$). For example, if the current bit is "1", the current speech frame is set to the speech frame derived using α_1. We concatenate all the speech frames obtained from this process to be $y(n)$.

4.2.3 Data-detection process

As shown in Fig. 2.9, McAdams coefficient manipulation causes the shifting in pole locations and frequency-response envelopes. We thus investigated the statistical properties of these cues for the data-detection process. In contrast to the common watermark detection methods that are based on a threshold or fixed set of rules, a machine learning model was constructed to blindly detect watermarks on the basis of those cues as features and is based on a random forest algorithm [13] (as shown in Fig. 4.13). The random forest classifier generates a number of decision tree classifiers on random sub-samples of the training dataset to control overfitting and improve prediction accuracy.

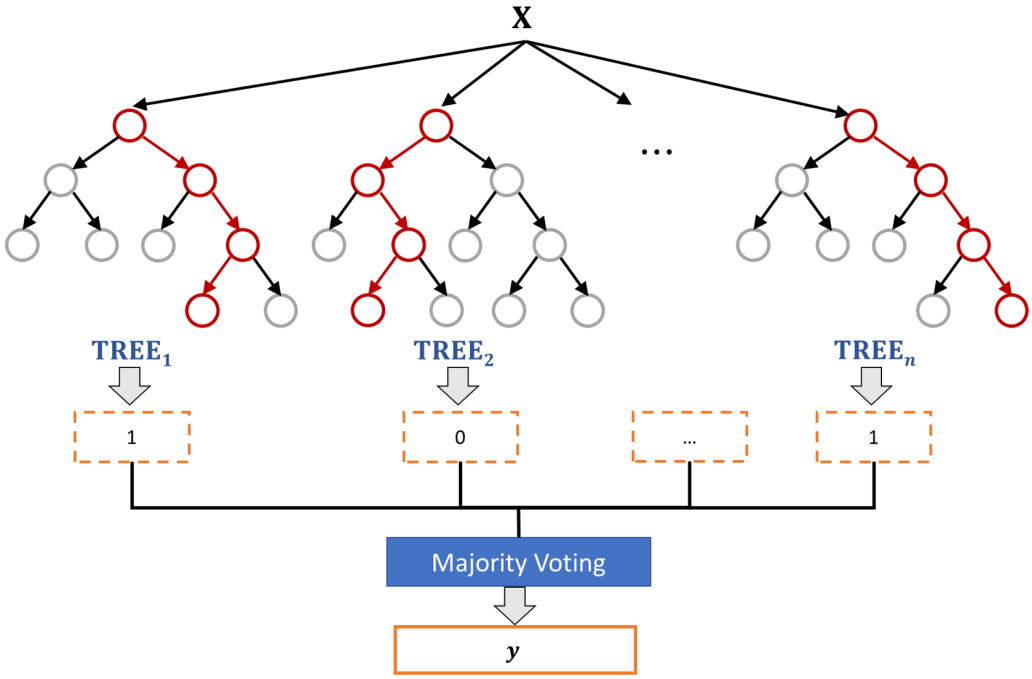


Figure 4.13: Random Forest classifier for data-detection. $\mathbf{X}$ is set of features, y is classification label ("0" or "1"), n is number of trees.

Before deciding on a random forest algorithm for generating our data-detection model, we carried out a preliminary experiment. We compared three watermark detection methods: (1) using rule-based with thresholds on power spectral density and pole location; (2) using a decision tree model; (3) using a random forest model. The features for constructing the decision tree and random forest models are power spectral density, pole locations, and statistical features (minimum, maximum, mean, standard deviation, skewness, and kurtosis) of the frequency-response envelope of the watermarked speech frames (without any pre-processing). We evaluated these three methods using a dataset consisting of 100 utterances randomly selected from the LibriSpeech[89] and VCTK[120] corpora. These 100 utterances were selected from one female speaker (LibriSpeech) and one male speaker (VCTK). We set the watermarking payload to 4 bps, a fixed set of McAdams coefficients for watermarking ($\{\alpha_0, \alpha_1\} = \{0.9, 1\}$), and a fixed frame size (20 ms) without a sliding window. The average classification errors of methods using rule-based, decision tree model, and random forest model in a 10-fold cross-validation evaluation were 32.12%, 26.25%, and 16.42%, respectively. On the basis of these results, we chose the random forest algorithm because it is the most stable and robust against outliers than the others used in our preliminary

experiment.

To improve our random forest classifier model for the blind-detection process, we use a short-term analysis frame with a fixed length (default frame size = 20 ms) with a sliding window. The features for constructing this data-detection model are power spectral density, complex pole locations, and statistical features of line spectral frequencies (LSFs) pairs on the frequency-response envelope. The statistical features consist of mean, standard deviation, skewness, and kurtosis. The statistical features of LSFs are used because they are more relevant than the global statistical features of the frequency-response envelope to represent the McAdams coefficient manipulation. Figure 4.14 the LSF positions on the frequency-response envelope of modified speech signals when $\alpha = \{1, 0.95, 0.9, 0.85\}$. We generate two modified speech signals through McAdams coefficient modification ($a_0(n)$ and $a_1(n)$) following the process explained in Subsection 4.2.1 for the training process. The label corresponds to the binary bit represented by the McAdams coefficient ("0" or "1").

Figure 4.15 shows the block diagram of the data-detection process. The details of this process is as follows:

Step 1: The watermarked signal ($y(n)$) is passed through a pre-emphasis filter. This filter is used to compensate for the average spectral shape that emphasizes the higher frequency components. A finite impulse response (FIR) filter is used as the pre-emphasis filter ($P(z)$), which is expressed as

$$P(z) = 1 - 0.95z^{-1}. \tag{4.4}$$

Step 2: Since the constructed random forest classifier works on a short-time frame basis, we used the sliding window technique to obtain more analysis speech frames for optimizing the data-detection process. For example, if the sampling frequency (Fs) is 16 kHz and default frame size is set to 20 ms, we have almost double the number of speech frames when the shift length is set to half the frame size. Figure 4.16 illustrates the data-detection process using a sliding window.

Step 3: From each watermarked speech frame obtained from the sliding window, we conduct feature extraction, i.e., complex pole positions, statistical features of LSF pairs from the frequency-response envelope, and power spectral density. The complex pole positions and statistical features of LSF pairs from the frequency-response envelope are derived from LP analysis. The power spectral density is obtained using the Fourier transform.

Step 4: Finally, our random forest classifier is used to generate the detected watermarked bit-stream ($w'(k)$) on the basis of the majority voting of the detected bit in all sliding window sub-frames in the corresponding

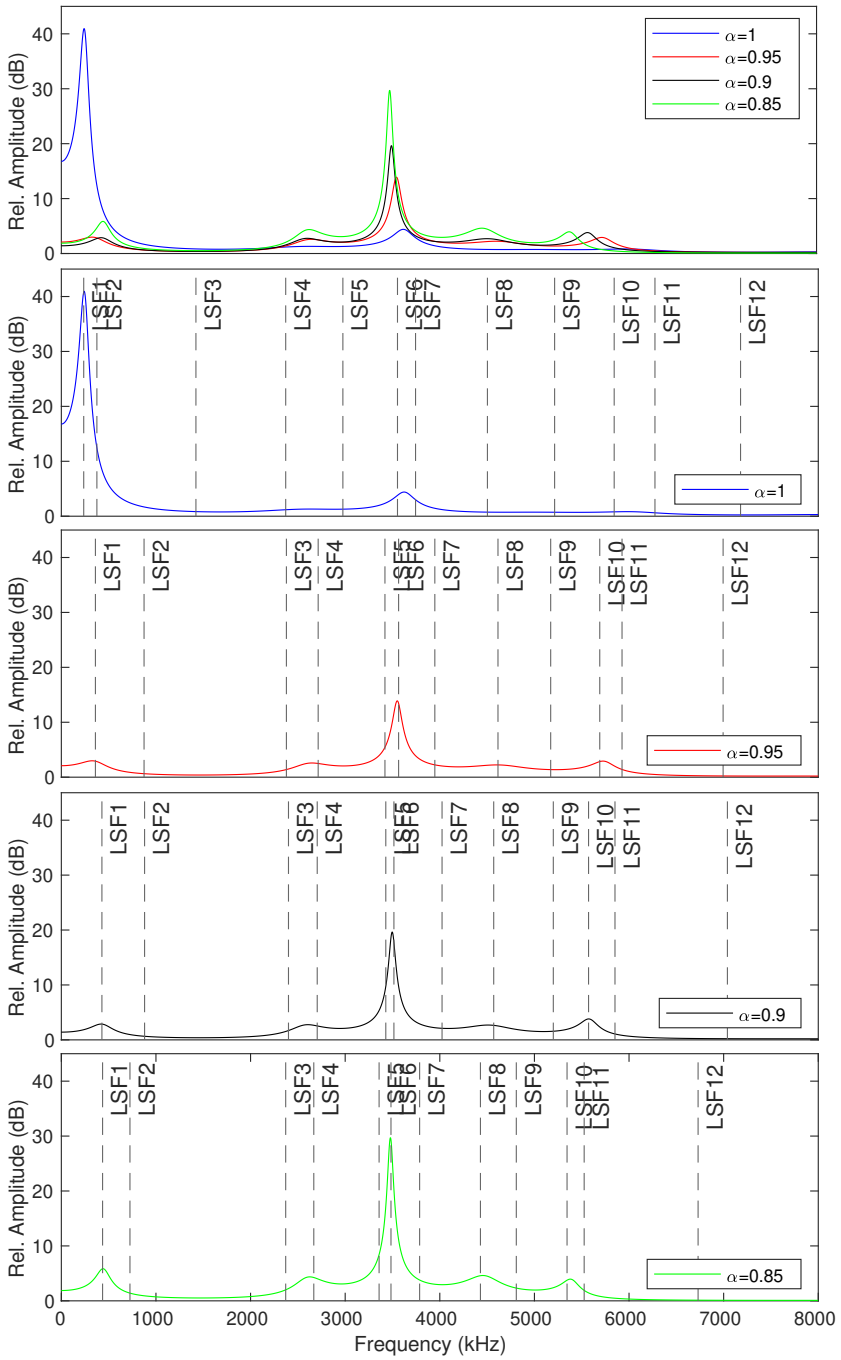


Figure 4.14: LSF positions on frequency-response envelopes obtained from various McAdams coefficients ($\alpha = \{1, 0.95, 0.9, 0.85\}$).

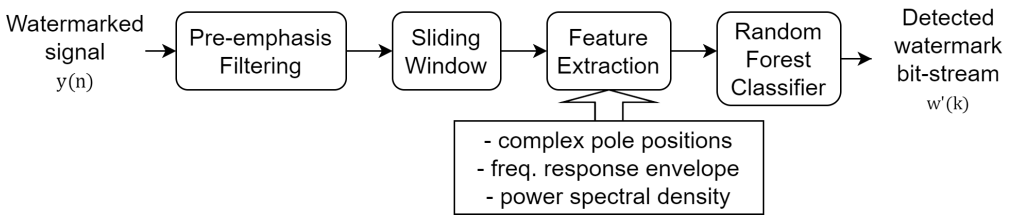


Figure 4.15: Block diagram of blind-detection process. $w'(k)$ is the detected watermark bit-stream of k-th frame.

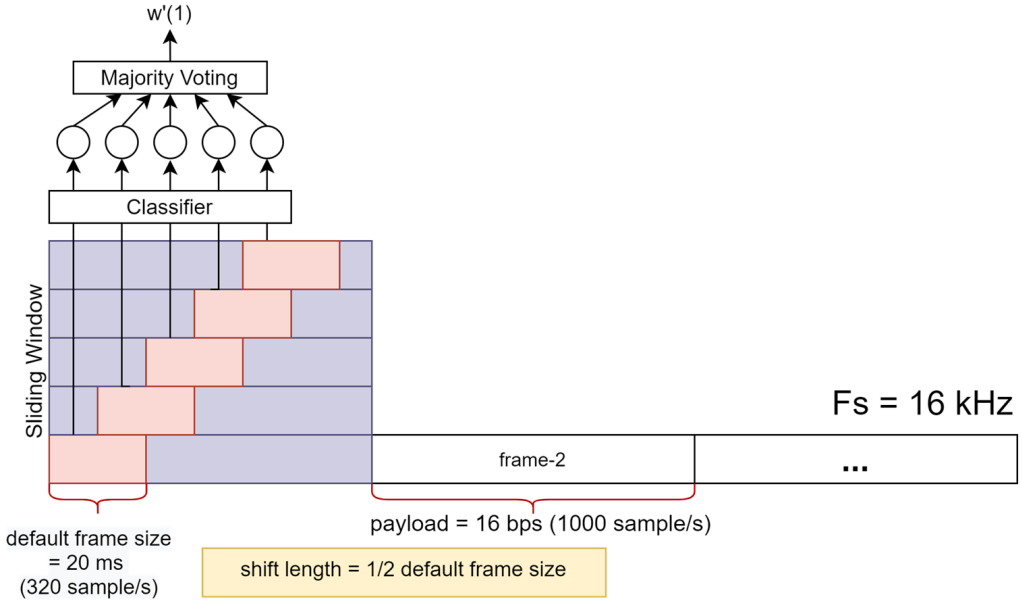


Figure 4.16: Illustration of watermark detection using sliding window. Sampling frequency (Fs) is 16 kHz, payload is 16 bps, and shift length was set to half default short-time frame size (10 ms).

watermarked frame. The most common bit information shown in Fig. 4.16 in five detected bits from the classification task of five sub-frames determine the detected watermark bit of the first frame.

4.2.4 Experimental Setup

This section describes the dataset, random forest classifier for the data-detection process and evaluation setting to analyze the performance of our proposed method.

Table 4.5: Statistics of dataset

Subset	Number of Speakers			Number of Utterances
	Male	Female	Total	
LibriSpeech (train)	4	4	8	225
LibriSpeech (test)	2	2	4	25
VCTK (train)	1	1	2	225
VCTK (test)	1	1	2	25
Total	8	8	8	500

Dataset

We semi-randomly selected 250 utterances from LibriSpeech [89], and 250 utterances from VCTK [120]. Semi-randomly means that we selected utterances from a particular number of speakers (with balance gender distribution). LibriSpeech is sampled at 16 kHz and designed for automatic speech recognition research, and VCTK is sampled at 48 kHz and designed for text-to-speech research. We unified the sampling rate of both corpora to 16 kHz. The selected utterances varied depending on the speaker and speech content.

Table 4.5 shows the distribution of the dataset. LibriSpeech has less utterances but relatively long duration, and VCTK has more utterances (almost 10 times that of LibriSpeech) but relatively shorter in duration. Due to these differences, we used a different number of speakers from each corpus. A total of 500 utterances were then split into 90% for the training set and 10% for the testing set. The training set was used for constructing the random forest classifier for blind detection. The testing set was then used for evaluating speech watermarking performance.

Evaluation setting

The testing set consisted of 50 utterances. The objective evaluation of our proposed speech watermarking method was based on the information-hiding criteria suggested in [51]. There were two main goals for this evaluation, i.e., (1) to investigate the trade-off between inaudibility and detection rate of our method using various gaps between McAdams coefficients; and (2) investigate the robustness of our method against various speech processing operations.

To reach the first goal, we considered five different McAdams coefficients ($\alpha_0 = \{0.95,0.925,0.9,0.875,0.85\}$) as representations of bit-"0", where we kept $\alpha_1 = 1$ as representation of bit-"1". These values were chosen to analyze the optimal gap to balance the inaudibility and robustness requirements. We thus constructed five random forest classifiers for blindly detecting the watermarks

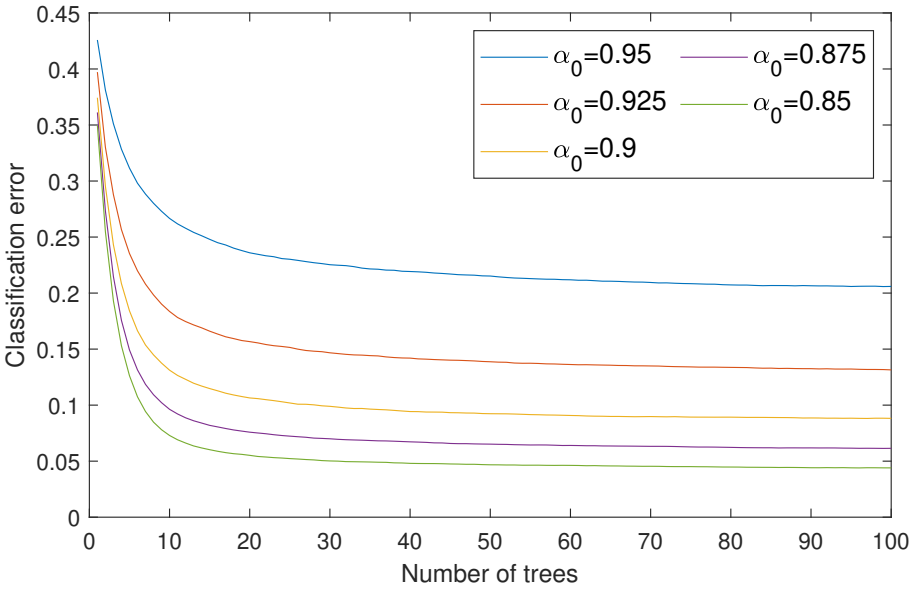


Figure 4.17: Classification errors of constructed random forest classifiers using several McAdams coefficients for representing bit-"0" ($\alpha_0 = \{0.95, 0.925, 0.9, 0.875, 0.85\}$). Maximum number of trees was set to 100.

with regards to the McAdams coefficient. The classification errors of all random forest classifiers are shown in Fig. 4.17. The metrics for evaluating the inaudibility requirement are log spectral distance (LSD) [42] and perceptual evaluation of speech quality (PESQ) [97] ITU-T P.862 (see the description in Tab. 2.1). LSD is used to measure the spectral distortion of watermarked signal ($y(n)$) in comparison with the original signal ($x(n)$) in decibels (dB) (as shown in Eq. 2.9).

For evaluating watermark detection accuracy and security level, we used the bit error rate (BER), false acceptance rate (FAR), false rejection rate (FRR), and F1-score. The threshold set for an acceptable BER is 10% [51]. In the evaluation, we defined the watermarked bit-stream ($w(k)$) as a random binary stream with the length depending on the payload. We investigated the payloads of 2, 4, 8, 16, and 32 bps. For robustness evaluation, we considered eight cases of non-malicious signal processing operations, i.e., normal (no attack), down-sampling to 12 kHz (resample-12), up-sampling to 24 kHz (resample-24), bit compression to 8 bits (requant-8), bit extension to 24 bits (requant-24), conversion to Ogg format (Ogg), conversion to MPEG-4 Part 14 or MP4 format (MP4), and conversion to G723.1 codec (G723). The bitrate of G723.1 codec is 5.3 kbps with algebraic code-excited linear prediction

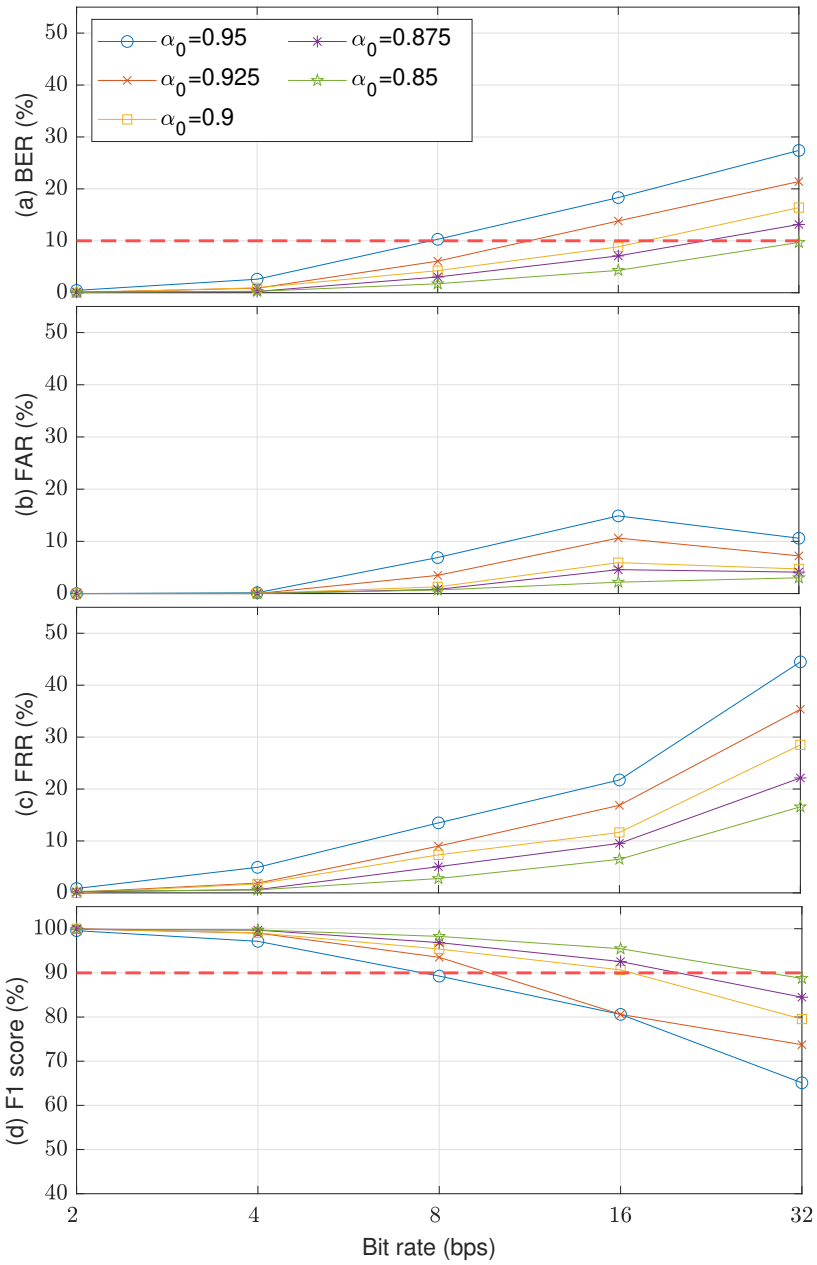


Figure 4.18: Watermark detection accuracy results using several McAdams coefficients for representing bit-"0" ($\alpha_0 = \{0.95, 0.925, 0.9, 0.875, 0.85\}$) in terms of: (a) BER, (b) FAR, (c) FRR, and (d) F1-score.

(ACELP) algorithm.

We also carried out a comparison analysis among our proposed method using McAdams coefficients $(\alpha_0, \alpha_1) = (0.9, 1.0)$ (Proposed) and two other well-known speech watermarking methods, i.e., LSB and DSS. These two methods were chosen because they can clearly represent the inaudibility and robustness trade-off. LSB works by modifying the most insignificant bits of the speech signal with watermarks, thus achieving high performance in inaudibility requirements but very fragile against any signal processing operation. In contrast, DSS works by spreading the watermarks over the whole frequency band. Therefore, it is preferred due to its robustness, but it causes significant distortion throughout the speech (lack of inaudibility). We conducted the comparative analysis using payloads of 4, 8, 16, and 32 bps.

4.2.5 Results

Figure 4.18 shows the watermark detection accuracy and security level results in terms of BER, FAR, FRR, and F1-score. Five McAdams coefficients were used to represent bit-"0" ($\alpha_0 = \{0.95,0.925,0.9,0.875,0.85\}$),whereas the McAdams coefficient for representing bit-"1" was set to 1 ($\alpha_1 = 1$). The results indicate a similar tendency for all these metrics when using a larger gap between α_0 and α_1, i.e., better detectability, except a slight anomaly in FAR for payloads 16 and 32 bps. Considering the detectability threshold (BER = 10%), only when $\alpha_0 = 0.85$, the embedding payload was up to 32 bps. With $\alpha_0 = \{0.875, 0.9\}$, the payload was 16 bps. For other observed α_0 the payload was less than 16 bps. A similar error rate for FAR and FRR was also found when we considered the observed payloads. When considering the overall security level in F1-score with a threshold of 90%, the proposed methods with $\alpha_0 \leq 0.9$ reached a payload of 16 bps.

The results of the inaudibility test are shown in Fig. 4.19. On the basis of the inaudibility threshold, the evaluation results indicate that with $\alpha_0 \leq 0.9$, both PESQ and LSD scores satisfied the requirement of up to 32 bps. The inaudibility requirement could be satisfied by watermarked signals with $\alpha_0 = 0.875$ up to 16 bps and $\alpha_0 = 0.85$ up to 8 bps. We will thus consider using $\alpha_0 = 0.9$ for further analysis of robustness. As a reference, we provide demo speech outputs from our proposed method that can be accessed publicly[1].

The robustness results in eight cases are shown in Fig. 4.20. The watermarked signal was generated with $\alpha_0 = 0.9$ and $\alpha_1 = 1$. By only considering detectability and security level threshold, the results indicate that our pro-

[1] `http://www.jaist.ac.jp/~s1920436/Entropy2021/demo/demo.html`

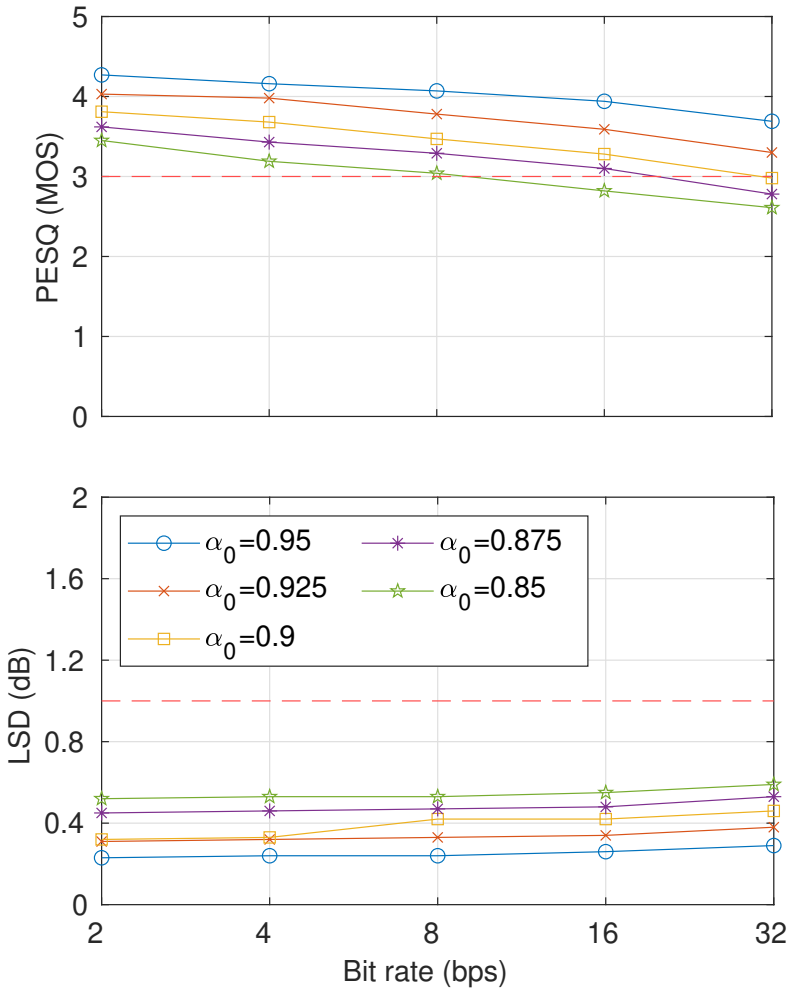


Figure 4.19: Sound-quality results using several McAdams coefficients for representing bit-"0" ($\alpha_0 = \{0.95, 0.925, 0.9, 0.875, 0.85\}$) in terms of PESQ (top) and LSD (bottom).

posed method had similar robustness with the normal case when dealing with up-sampling (resample-24), bit extension (requant-24), Ogg, and MP4 processing operations. Robustness degraded when down-sampling (resample-12), bit compression (requant-8), and G723.1 codec (G723) were applied. For resample-12 and requant-8, we can say that our proposed method is robust when the payload is 4 bps (BER $< 10\%$). Unfortunately, our proposed method is not robust when the G723.1 codec is applied (BER $> 10\%$).

The results on the comparative analysis among our proposed method with

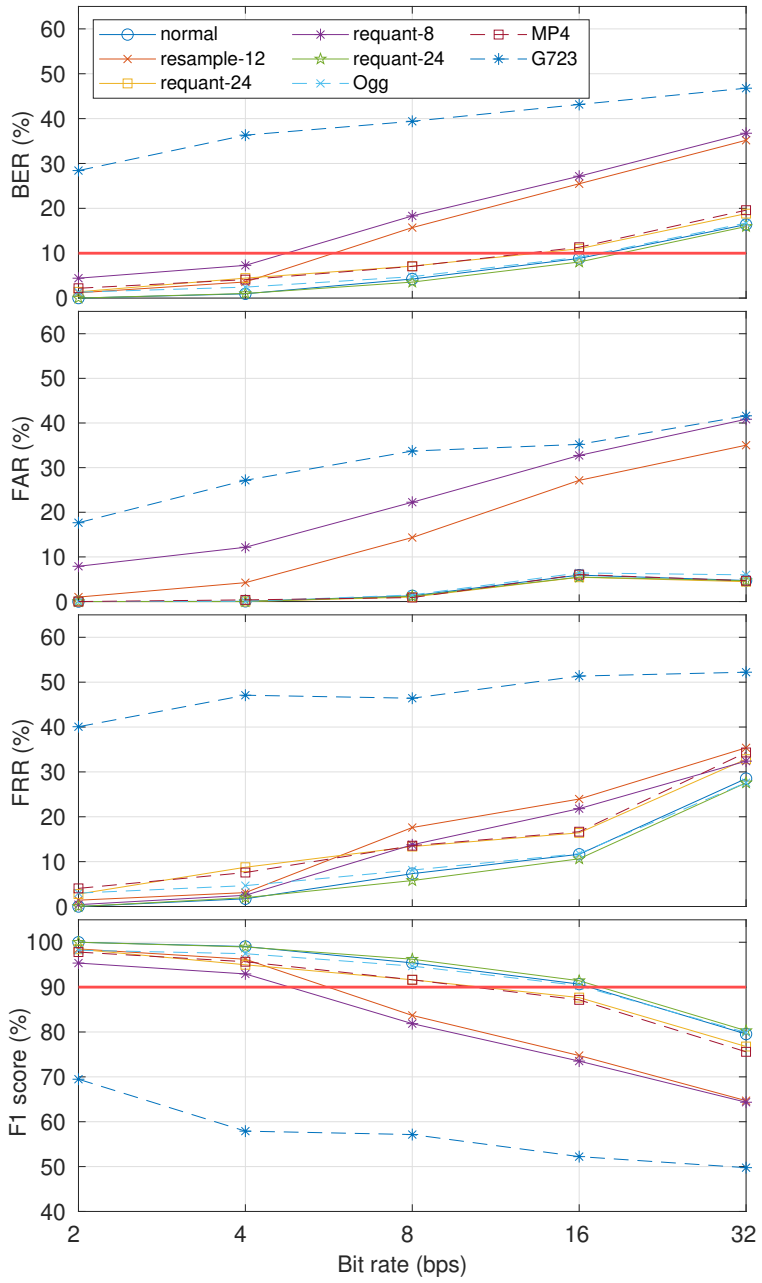


Figure 4.20: Robustness results in terms of BER, FAR, FRR, and F1-score in eight cases: normal, resample-12, resample-24, requant-8, requant-24, Ogg, G723, and MP4. The McAdams coefficient for representing bit-"0" was 0.9 ($\alpha_0 = 0.9$).

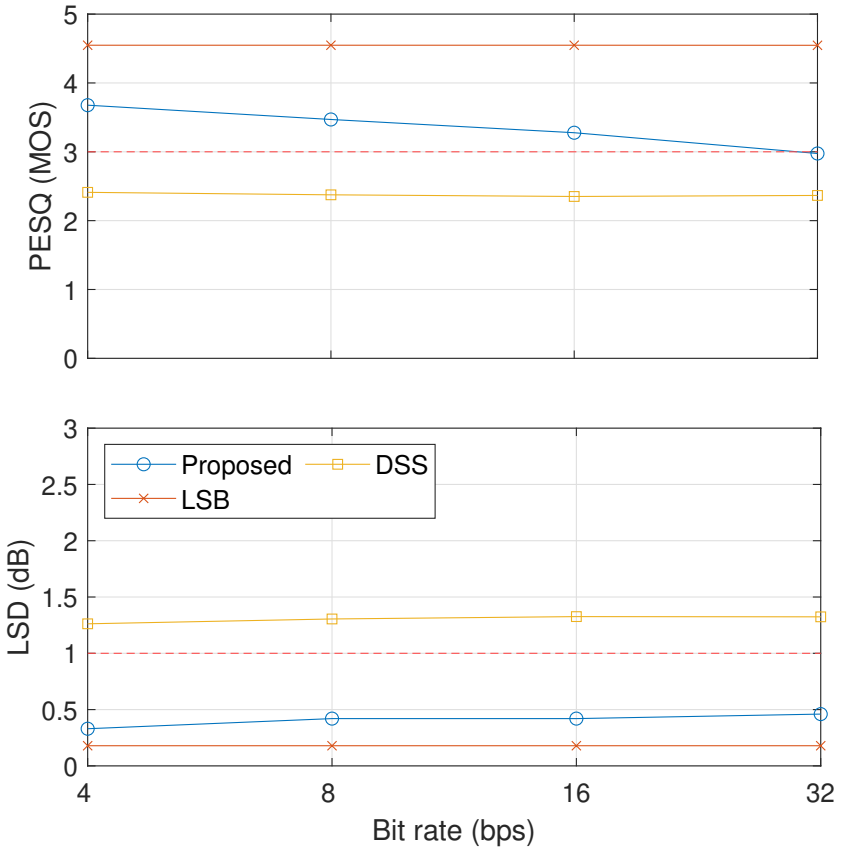


Figure 4.21: Evaluation of inaudibility results of three compared methods (Proposed, LSB, and DSS).

$(\alpha_0, \alpha_1) = (0.9, 1.0)$ (Proposed), LSB, and DSS are shown in Figs. 4.21 and 4.22. Figure 4.21 shows the inaudibility comparison results in terms of PESQ and LSD. These results indicate that LSB and our proposed method could pass the threshold of inaudibility but not DSS.

Figure 4.22 shows the robustness comparison results in terms of BER. In contrast to the inaudibility results, the robustness results indicate that LSB was fragile in dealing with almost all observed signal processing operations, except with the up-sampling to 24 kHz. However, DSS was very robust even in a higher payload, except with the G723.1 speech codec. Although not as robust as DSS, Our proposed method had better robustness against most of the observed signal processing operations (down-sampling, re-quantization, Ogg format, and MP4 format) than LSB.

To better represent the application of speech watermarking, we embedded an image as a watermark to a speech signal. The watermark detection

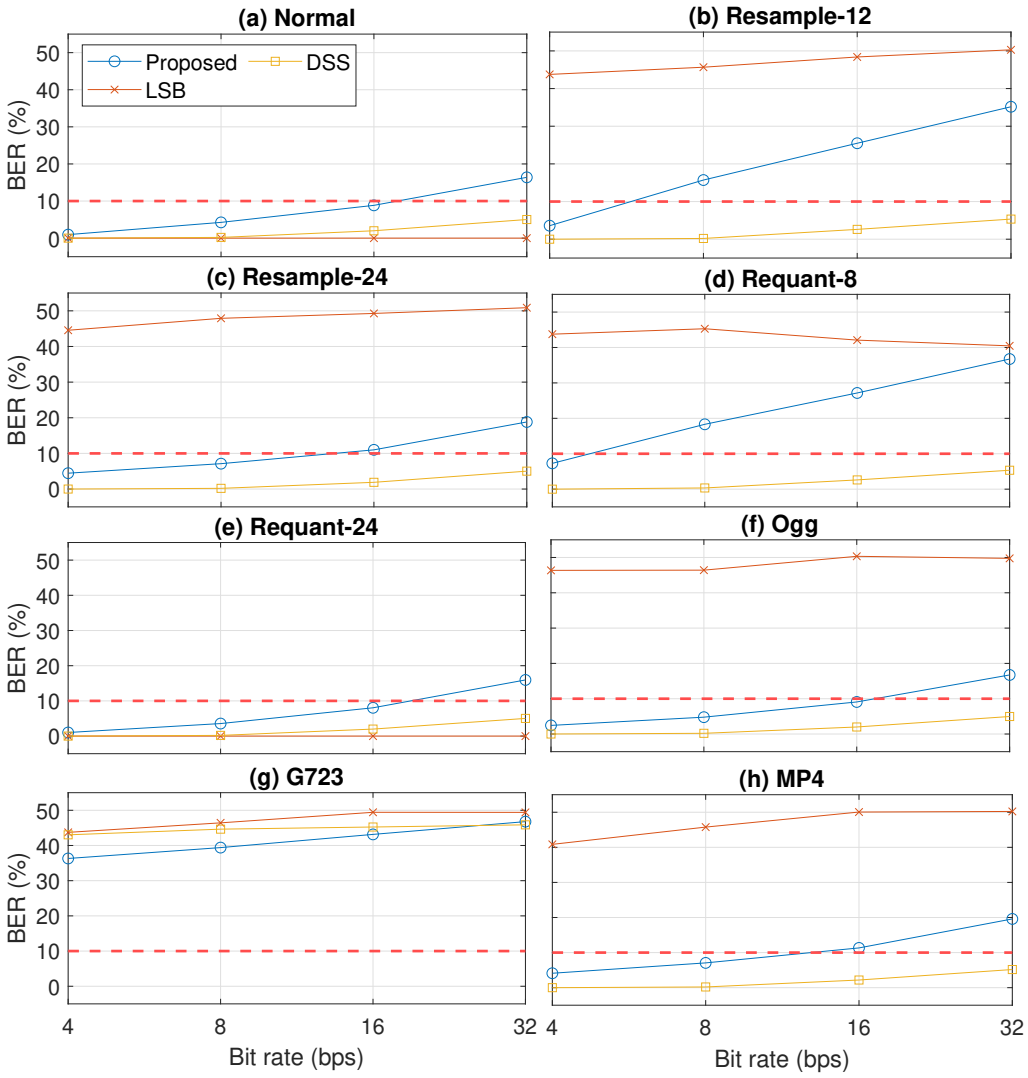


Figure 4.22: Robustness results of three compared methods (Proposed, LSB, and DSS) in terms of BER in eight cases: normal, resample-12, resample-24, requant-8, requant-24, Ogg, G723, and MP4.

results are shown in Fig. 4.23. The size of the image in the binary bit-stream was 80×192. The watermarked signal was generated using $\alpha_0 = 0.9$ and $\alpha_1 = 1$ with 4-bps payload. Although not perfectly accurate, we could observe the reflection of embedded image information even after certain operations, including re-sampling, re-quantization, and conversion to Ogg and MP4 formats.

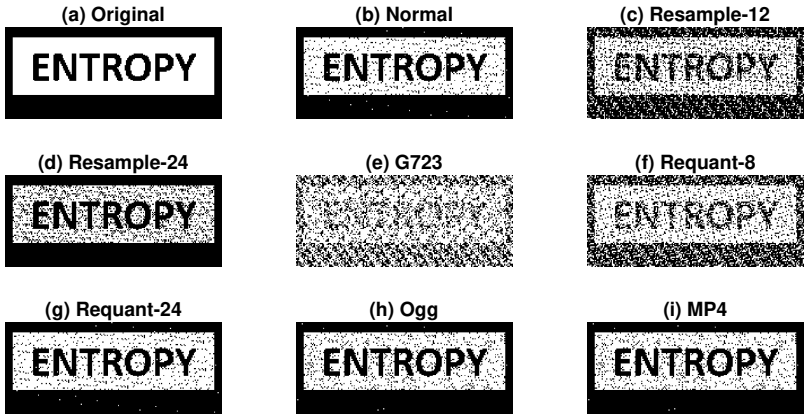


Figure 4.23: Application of embedding image information using proposed method with 4-bps payload after several non-malicious signal processing operations, i.e., (a) original watermark, (b) normal, (c) resample-12, (d) resample-24, (e) G723, (f) requant-8, (g) requant-24, (h) Ogg, and (i) MP4. McAdams coefficients for representing bit-"0" and bit-"1" were 0.9 and 1, respectively ($(\alpha_0, \alpha_1) = (0.9, 1.0)$).

Chapter 5

Voice Privacy Protection Based on Speaker Anonymization

Speaker anonymization (also known as de-identification) is a method of protecting voice privacy. It works by concealing the personally identifiable information of uttered speech without degrading the linguistic information [35]. Previously, voice transformation was utilized to suppress speaker identity for anonymization purposes [53, 54]. Kaldi phone, a diphone-based syntactic source speech, was transformed to attack the speaker identification system, successfully fooling the Gaussian mixture model (GMM)-based speaker identification system [53]. Subsequently, de-identification of online speakers was feasible with a voice transformation method that de-identifies speaker using GMM mapping and harmonic-stochastic models [93]. Next, a voice transformation technique that uses a target's natural speech instead of a synthetic voice was developed to conceal speaker identity [1]. Cepstral frequency warping is another alternative approach implemented with an amplitude scaling technique to transform speech and hide identity [72].

Recently, an anonymization method based on a neural source-filter (NSF) model was proposed by Fang et al. [35]. This method separates the speaker identity and the linguistic content from the input speech. An x-vector that refers to the speaker identity was modified to hide personal information before resynthesizing the speech data. In the Voice Privacy Challenge 2020 [112], this method was introduced as the primary baseline system because the x-vector could effectively encode speaker identity as a feature in a speaker verification system [107]. Another baseline introduced in the Voice Privacy Challenge 2020 used the McAdams coefficient [77] to transform the spectral envelope of speech signals to achieve speaker de-identification [90]. The objective evaluation results showed the primary baseline based on the NSF model hid speaker information better than the second baseline based on the McAdams

coefficient [112], confirming that the x-vector is a practical feature to represent speaker identity information.

In this study, we improve the primary baseline system by modifying the x-vector singular value for speaker anonymization. Our preceding experimental results in [74] showed that our method improves the anonymization rate and is comparable with the baseline system. There is also room for improvement that we intend to present in this study. First, we thoroughly analyze the effectiveness of x-vector modification with singular value decomposition (SVD) by considering various singular value thresholds. We predict that modifying the significant elements represented in an x-vector singular value (SV) can fulfill the speaker-to-speaker correspondence requirement in an anonymization system. Second, despite using a regression model as in our prior work [74], we construct a clustering model for selecting a set of x-vectors for generating the pseudo-target x-vector. Third, we modify acoustic features such as fundamental frequency (F_0) and speech duration to improve our method. The F_0 and speech duration are strongly related to the perception of speaker individuality [3, 27], so modifying these features should de-identify speaker individuality. To evaluate the performance and effectiveness of our method, we conduct an objective evaluation that follows the Voice Privacy Challenge 2020 [112], and we propose a more reliable subjective evaluation for assessing the privacy and utility-related metrics in a speaker anonymization system.

5.1 Speaker Anonymization Based on X-Vector Singular Value Modification

This section demonstrates our contributions based on the B1 system. Our hypothesis is that modifying the F_0 and the x-vector anonymization model by using SVD could improve the verifiability performance of a speaker anonymization system. We apply the modification to components 4 and 5 in Fig. 2.7. Figure 5.1 shows the schematic diagram of our proposed method.

5.1.1 Pseudo-target Generation

In contrast to a voice conversion system, the speaker target is unknown in an anonymization system, so a target anonymized speaker (pseudo-speaker) must be determined. The x-vector of an input speaker in the B1 system was modified as a speaker individuality feature using a selection algorithm on a pool of x-vectors (as explained in Subsection 2.3.3) [112]. This selection algorithm was utilized by randomly choosing 100 x-vectors from a set of 200

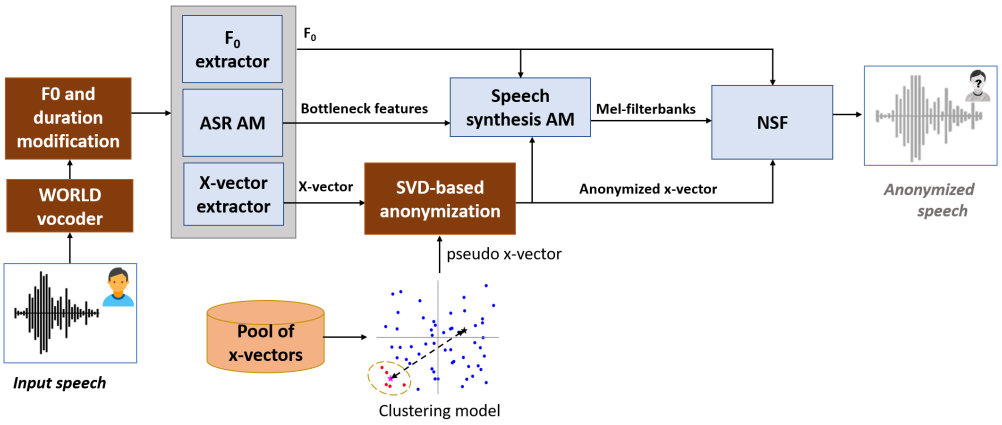


Figure 5.1: Schematic diagram of proposed speaker anonymization system.

x-vectors obtained from speakers who were the furthest distance from the input speaker. The distance was determined using the PLDA.

As Fig. 5.2 (a) shows, the average x-vector from the randomly selected subset from the furthest x-vectors set can cause the input speaker's x-vector to be given nearby. To reduce occurrences of this problem, we constructed a gender-dependent clustering model based on k-means as the selection algorithm for a set of the furthest x-vectors of the same-gender utterances. K-means clustering is commonly used because of its simple implementation and because it scales to large datasets and guarantees convergence [11]. The pseudo x-vector was then determined using the mean of the set x-vectors. Figure 5.2 (b) illustrates the selection algorithm of our method.

5.1.2 SVD-based X-vector Anonymization

For the x-vector anonymization technique, we applied one of the matrix factorization concepts in linear algebra, namely, SVD [40]. SVD is widely used for dimension reduction applications (e.g., data compression and denoising) because it provides a more stable matrix decomposition than the other methods [41, 99]. The SVD technique decomposes a given input matrix into its constituent elements based on the polar decomposition. Mathematically, the SVD is expressed by the following equation.

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}, \tag{5.1}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the orthonormal eigenvectors of $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, respectively, and $\mathbf{\Sigma}$ consists of the square roots of the eigenvalues of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$.

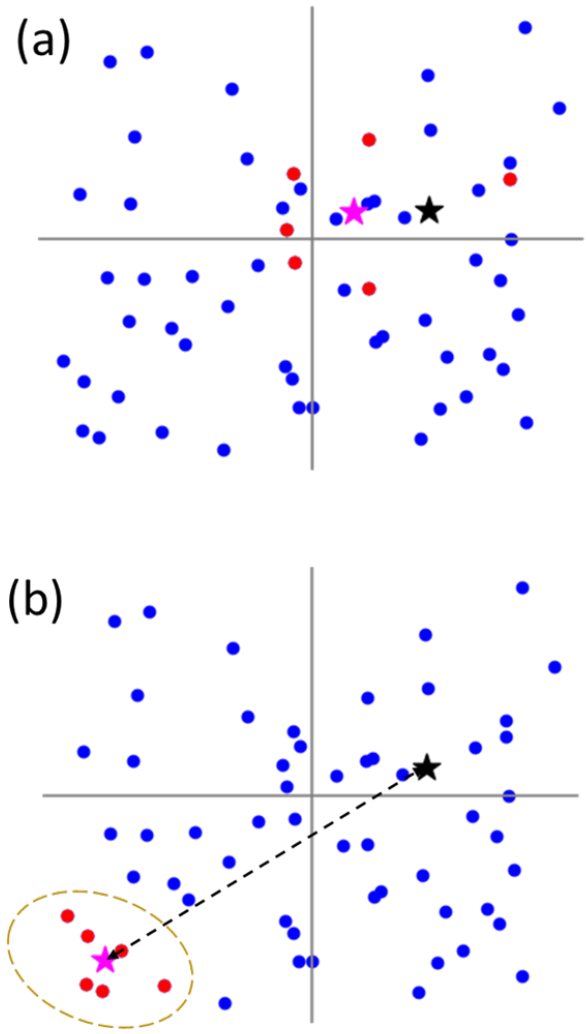


Figure 5.2: Illustration of x-vector selection algorithm using: (a) random selection and (b) clustering-based selection. Round blue markers indicate set of x-vector candidates, round red markers indicate chosen x-vector candidates, black star markers indicate given input x-vectors, and magenta star markers indicate chosen pseudo-target x-vectors.

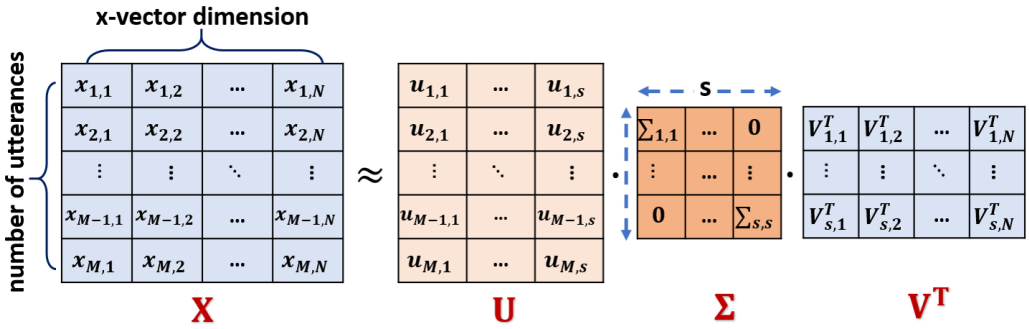


Figure 5.3: Modification of x-vector SVs [74]. The $x_{i,j}$ refers to the element of matrix $\mathbf{X}$ in row i and column j. Similarly, $u_{i,j}$, $\Sigma_{i,j}$, and $V^T_{i,j}$ are the elements of matrix $\mathbf{U}$, $\mathbf{\Sigma}$, and $\mathbf{V^T}$ in row i and column j, respectively. The $\mathbf{V^T}$ is the transpose matrix of $\mathbf{V}$. The $\mathbf{s}$ determines the number of singular values.

The x-vectors are extracted from a variety of utterances spoke by a speaker which is not equivalent to each other. However, a PLDA classifier distinguishes which speaker the x-vectors originated from [107]. Principal component analysis shows that the distribution of a single speaker's x-vectors are clearly clustered close together (as shown in Fig. 5.4). Considering those preliminary studies, x-vector anonymization by SVD could capture the eigenstructure and result in a better representation of intra-speaker information. Thus, modifying the SV of the x-vectors matrix could satisfy the speaker-to-speaker correspondence requirement (the x-vectors of a given speaker should not be similar to the other speakers).

Our SVD-based x-vector technique [74] is conducted in the following three steps (shown in Fig. 5.5):

i. *Pseudo-target x-vector matrix formation*

   After the pseudo-target x-vectors of a given speaker from all available train utterances were chosen using the clustering model, we concatenated those x-vectors into a matrix ($\mathbf{X}$) that had an $M \times N$ dimension, where M is the total available utterances and N is the dimension of the x-vector (512).

ii. *SV decomposition and modification*

   The pseudo-target x-vector matrix was decomposed by SVD (as shown in Eq. 5.1) into two singular matrices ($\mathbf{U}$ and $\mathbf{V}$) and a diagonal singular values matrix ($\mathbf{\Sigma}$). In this approach, $\mathbf{U}$ could be interpreted as the utterance-to-concept similarity matrix and $\mathbf{V}$ as the x-vector-to-concept similarity matrix. $\mathbf{\Sigma}$ represents the strength of each concept involved.

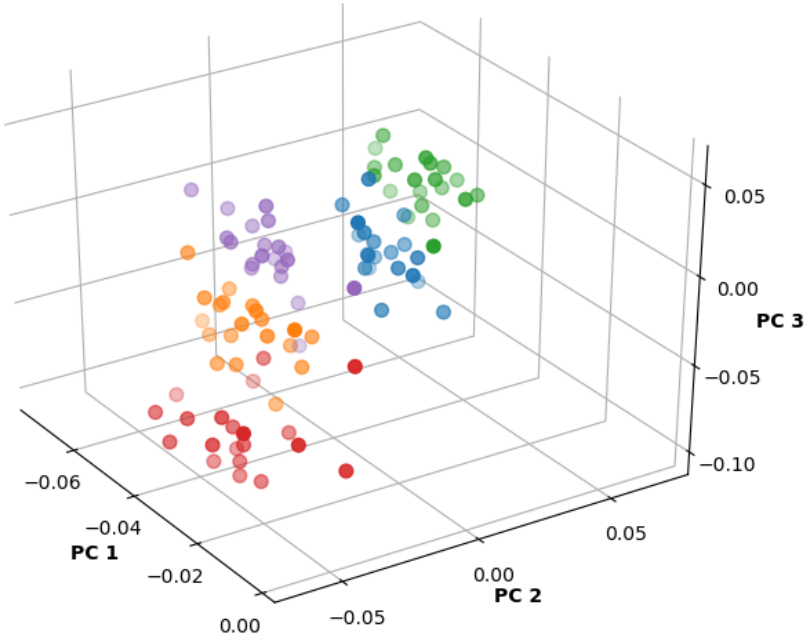


Figure 5.4: Principal components (PCs) of x-vectors from five speakers in VCTK development dataset for enrollment in 3D space. Colors represent speaker labels (e.g., round orange markers represent class of x-vectors of speaker with ID label "p234").

The anonymization was conducted by controlling the dimension of $\mathbf{\Sigma}$ using a threshold parameter ($\mathbf{s}$) to obtain more general constituent elements of the x-vector. Figure 5.3 shows the x-vector anonymization by SV modification.

iii. *Anonymized x-vector reconstruction*

After the SV modification, we reconstructed the modified matrix using $\mathbf{U}$, $\mathbf{V}$, and the modified $\mathbf{\Sigma}$. The anonymized x-vector of the given utterance was then extracted accordingly.

5.2 Development of Speaker Anonymization by Modification of Speech Prosody

F_0 contours and their dynamics are strongly related to speaker individuality [3, 27] because F_0 is an important physical factor that affects pitch perception.

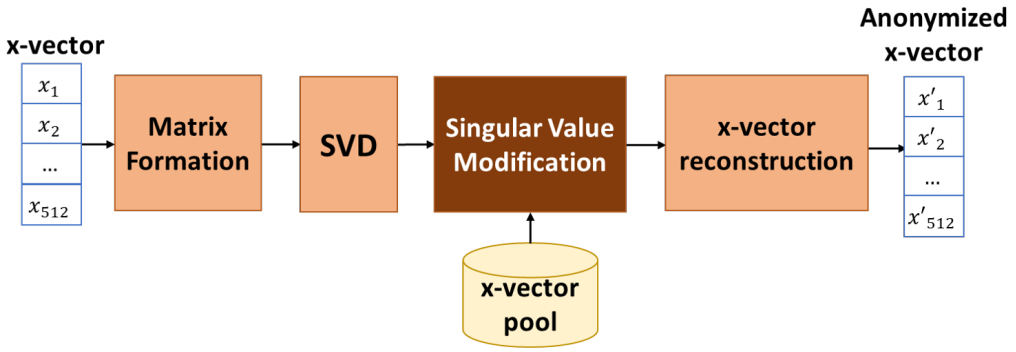


Figure 5.5: Schematic diagram of x-vector modification by SVD [74]. x_i and the x'_i are the i-th element of input x-vector and anonymized x-vector, respectively

Furthermore, it accommodates the perception of several kinds of paralinguistic and prosodic information [44, 47]. For instance, we could classify a speaker's gender solely by using the F_0 as the feature because the F_0 of female speech is generally higher than that of male speech.

In this study, we modify the F_0 using the mean F_0 information of adult female and male speakers from a previous study [114] by using WORLD vocoder [79]. We classify the speech into a high F_0 and a low F_0 by comparing the mean F_0 of the input utterance and the mean F_0 regarding gender. Accordingly, we convert the low F_0 into a high F_0 by 1.5 times and vice versa (as shown in Eq. 5.2). Factor 1.5 was chosen because our preliminary experiment showed that it is the largest factor that outcome the possible F_0 range [114] in the dataset.

$$f'_0(n) = \begin{cases} 1.5 \times f_0(n), & \overline{f_0}(n) \leq \overline{F_0} \\ f_0(n)/1.5, & \overline{f_0}(n) > \overline{F_0} \end{cases}, \tag{5.2}$$

where m indicates the time frame, $f_0(n)$ and $f'_0(n)$ are the original F_0 and the modified F_0 in the time domain, respectively. The $\overline{f_0}(n)$ is the mean F_0 value of original F_0, whereas $\overline{F_0}$ is the mean female or male F_0 value based on [114].

In addition to F_0 modification, we carry out speech duration modification. Duration is a speech property relevant to expressing "stress" in speaking. Consequently, the speaking rate varies from speaker to speaker [27]. Speaking rates have been reported to significantly affect speaker verification system performance [23]. In this study, we lengthen speech by increasing the frame duration by 1.2 times because the mismatched speech tempo could be minimized by this factor (minimizing the possible distortion caused by this modification)

[23].

5.3 Experiments using SVD-based X-vector Speaker Anonymization

The experiments were entirely based on the protocols and datasets provided in the Voice Privacy Challenge 2020[1] [112]. In this Section, we provide the specific description of our method, including the datasets we used, our experiments, and our evaluation settings.

5.3.1 Datasets

We conducted our experiments using four publicly open-source corpora as described in Section 4 of the Voice Privacy Challenge 2020's evaluation plan [112]: LibriSpeech (libri) [89], LibriTTS [133], the voice cloning toolkit (VCTK) [120], and VoxCeleb-1,2 [19, 80]. Each corpus was split into training, development, and testing data. Additionally, "common part" and "different part" subsets of trial utterances were constructed specifically for the VCTK dataset to evaluate speaker verifiability regardless of text-dependency. The common part consisted of the utterances that were identical for all the speakers, and the different part consisted of the distinct utterances for all the speakers.

We utilized the available training subsets from LibriTTS (train-other-500 and train-clean-100) [133], comprised of approximately 1,400 speakers and 240,000 total utterances. Table 5.1 shows the statistics of these datasets. We evaluated our method with both development and test data of libri [89] and VCTK [120] in ASVeval and ASReval.

5.3.2 Experimental Setting

The experiments were conducted using the Kaldi toolkit [94] for the main anonymization framework, WORLD for F_0 modification [79], and the scikit-learn software [92] for the k-means clustering model. First, we analyzed the input signal using WORLD to obtain the F_0, the aperiodicity, and the spectral envelope. Subsequently, we modified the F_0 based on Eq. 5.2 and the frame duration before re-synthesizing the speech.

The output of the resynthesized speech was given input to the NSF-based anonymization system. In the x-vector anonymization block (sub-element

[1]https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

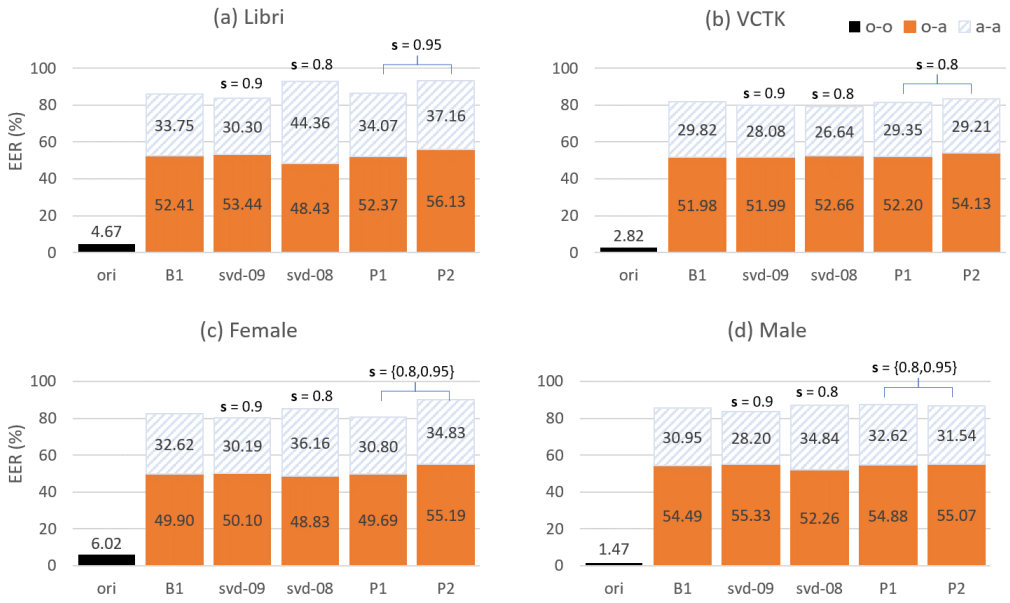


Figure 5.6: Average ASVeval results from controlling SV threshold using k-means clustering and modification of F_0 and duration. Original speech as "ori" denotes ASVeval results using both original enrollment and trials (o-o). "B1" denotes results of ASVeval using primary baseline model [112]. "svd-09" and "svd-08" denote ASVeval results by x-vector SV modification with thresholds (**s**) 0.9 and 0.8, respectively. "P1" denotes ASVeval results obtained by x-vector SV modification with k-means clustering, whereas "P2" denotes results with additional F_0 and speech duration modification. Orange bars represent results in pairs of original enrollment and anonymized trials (o-a). Gray bars represent results in pairs of anonymized enrollment and anonymized trials (a-a).

Table 5.1: Training data for pool of x-vectors.

Subset	Female	Male	Total	#Utter
train-clean-100	123	124	247	33,236
train-other-500	560	600	1160	205,044

4 in Fig. 2.7), we utilized the pseudo-target generation and x-vector SV modification as explained in Section 5.1. We conducted our method's entire process separately for each gender.

In pseudo-target generation, we chose the 200 furthest x-vectors using PLDA and clustered those x-vectors into 50 groups by the k-means algorithm. Finally, the pseudo x-vector of a given speaker was determined by the centroid furthest from the corresponding x-vector. In these experiments, we also analyzed the effect of controlling the SV threshold parameter. The threshold values are 0.9 and 0.8.

5.3.3 Evaluation

We evaluated our method with both objective and subjective tests. The evaluation was conducted based on a comparative study of the B1 method and our method.

Objective Test

The general procedure of the objective test was based on the Voice Privacy Challenge 2020 [112] and investigated three points. First, we investigated how effectively we could control the SV threshold. Second, we investigated how effectively we could select pseudo x-vectors using k-means clustering. Third, we investigated how the anonymization system performs by modifying the F_0 and speech duration.

Figure 5.6 compares the average results of the corresponding development and test datasets of the speaker verifiability assessment using ASVeval with the B1 system with those of our method. These results were obtained by averaging the EER results of development and test datasets. The subset of the common part is not available in the LibriSpeech dataset; therefore, the results in Fig. 5.6 were derived from all the results, excluding the subset of the common part of the VCTK dataset. To determine the effectiveness of the x-vector SV modification, we conducted the experiments without using the random selection provided in the baseline (only the mean value of the 200 furthest x-vectors). The results shown in Fig. 5.6 indicate results comparable

Table 5.2: Detailed ASVeval results using only x-vector SV modification with 0.95 threshold for LibriSpeech and 0.8 threshold for VCTK (SV Modif), our P1 method, and our P2 method. "Gen" stands for gender (F: female and M: male). "=" stands for the equivalent results to the left columns.

Dataset	**Gen**	**Anonymization**		**SV Modif**			**SV Modif + k-means (P1)**			**SV Modif + k-means + F_0 (P2)**		
		Enroll	**Trial**	**EER (%)**	C_{llr}^{min}	C_{llr}	**EER (%)**	C_{llr}^{min}	C_{llr}	**EER (%)**	C_{llr}^{min}	C_{llr}
Libri (dev)	F	ori	ori	8.67	0.30	42.86	=					
			anon	51.99	1.00	147.21	50.57	0.998	145.131	**55.82**	1.00	**156.61**
		anon		32.95	0.86	14.25	35.37	0.88	14.694	**38.35**	0.92	**27.00**
	M	ori	ori	1.24	0.03	14.25	=					
			anon	58.70	1.00	170.42	57.76	0.999	169.887	**60.4**	1.00	**174.72**
		anon		28.88	0.78	18.43	34.01	0.861	24.696	**34.32**	0.87	**29.11**
Libri (test)	F	ori	ori	7.66	0.18	26.79	=					
			anon	48.72	1.00	151.98	48.36	0.996	152.426	**55.29**	1.00	**153.69**
		anon		28.65	0.78	12.73	31.02	0.819	15.449	**39.23**	0.91	**36.93**
	M	ori	ori	1.11	0.04	15.30	=					
			anon	54.34	1.00	168.93	52.78	**0.999**	169.064	**53.01**	1.00	**168.19**
		anon		30.73	0.81	24.20	35.86	0.903	34.784	**36.75**	**0.90**	**39.89**
VCTK common (dev)	F	ori	ori	2.62	0.09	0.87	=					
			anon	50.87	1.00	167.48	49.71	1.00	175.25	**54.07**	1.00	**187.25**
		anon		24.42	0.70	7.12	**26.16**	0.71	6.66	24.71	0.72	**21.19**
	M	ori	ori	1.43	0.05	1.56	=					
			anon	**57.26**	1.00	191.60	55.27	1.00	194.49	56.13	1.00	**207.22**
		anon		25.93	0.71	18.20	**32.76**	0.84	23.69	26.21	0.72	**23.89**
VCTK diff (dev)	F	ori	ori	2.86	0.10	1.13	=					
			anon	50.14	0.99	165.94	51.04	0.99	168.53	**54.86**	1.00	**188.41**
		anon		26.78	0.77	8.72	25.32	0.74	8.13	**26.67**	0.73	**12.88**
	M	ori	ori	1.44	0.05	1.16	=					
			anon	**55.98**	1.00	166.42	54.39	1.00	168.77	52.06	1.00	**175.87**
		anon		25.31	0.74	18.28	**29.73**	0.82	**22.89**	25.46	0.74	17.49
VCTK common (test)	F	ori	ori	2.89	0.09	0.87	=					
			anon	50.00	1.00	156.09	50.00	1.00	156.09	**56.36**	1.00	**178.95**
		anon		28.61	0.80	8.81	**30.92**	0.83	9.12	30.64	0.83	**21.41**
	M	ori	ori	1.13	0.04	1.04	=					
			anon	**55.65**	1.00	186.48	54.80	1.00	191.83	52.54	1.00	**204.88**
		anon		20.34	0.62	9.79	**30.51**	0.82	20.60	25.42	0.72	**23.73**
VCTK diff (test)	F	ori	ori	4.89	0.17	1.50	=					
			anon	49.64	1.00	142.88	48.77	1.00	148.15	**54.78**	1.00	**161.07**
		anon		32.66	0.87	11.36	31.48	0.84	11.50	**35.08**	0.87	**18.15**
	M	ori	ori	2.07	0.07	1.82	=					
			anon	54.31	1.00	164.68	54.59	1.00	168.63	**54.82**	1.00	**179.63**
		anon		21.81	0.67	13.26	**30.88**	0.84	**22.90**	29.62	0.83	18.26

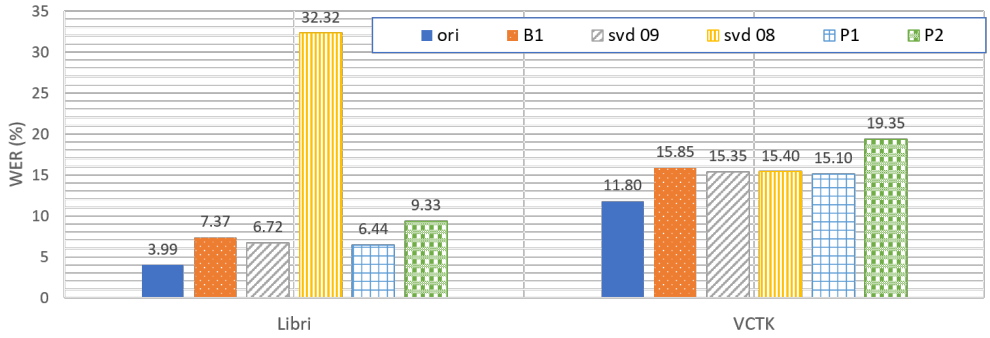


Figure 5.7: ASReval results of ori, anonymized speech by B1, controlling SV threshold (svd-09, svd-08), modifying x-vectors SV ($\mathbf{s}$={0.8,0.95}) with k-means clustering (P1), and modifying F_0, speech duration, and x-vector SV modification ($\mathbf{s}$={0.8,0.95}) with k-means clustering (P2).

to the B1. By controlling the threshold, we could slightly improve speaker verifiability.

The corresponding results of ASReval are provided in Fig. 5.7. The ASReval performance with x-vector SV modification was slightly better than the baseline, except for the LibriSpeech dataset with a 0.8 threshold (svd-08). Although the ASVeval results from using svd-08 in the a-a case were improved for the LibriSpeech dataset, the intelligibility in terms of ASReval degraded significantly. We predict that this occurred because the LibriSpeech dataset is a clean dataset. The dimension reduction modification on the SV could distort the constituent elements of the x-vectors. To compensate for this degradation, we used the 0.95 threshold parameter for the LibriSpeech dataset and 0.8 for the VCTK dataset in the following experiments. In practical use, when we have information about the characteristics of the dataset, we could follow the threshold parameters of proposed methods (e.g., $\mathbf{s}$=0.8 when the dataset has high variabilities as the VCTK dataset). We suggest the value of 0.9 as the threshold parameter for a completely unknown dataset to maintain a general anonymization performance based on x-vectors SV modification.

The two right bars in Fig. 5.6 show the comparative average ASVeval results of our method using x-vector SV modification with k-means clustering (P1) and additional F_0 and speech duration modification (P2). Overall results from this figure show no significant difference between the B1 method and the P1 method. However, the additional F_0 and speech duration modification in the P2 method could effectively improve the performance of the anonymization approximately 5% of the EER score in the o-a case and up to approximately 3% in the a-a case. The P2 method clearly yields better results than the B1 method, especially with the LibriSpeech dataset and the female utterances.

In addition to the EER score, we calculated the log-likelihood-ratio cost function for evaluating the speaker verifiability of the methods using x-vector SV modification with 0.95 threshold for LibriSpeech and 0.8 threshold for VCTK (SV Modif) only, SV Modif with k-means clustering (P1), and SV Modif with k-means and F_0 modification. The details of ASVeval are provided in Table 5.2. The increasing trend in the EER and C_{llr} scores indicates improvement in the privacy metric of a speaker anonymization system. In terms of the utility metric, the ASReval results of the P1 and P2 systems shown in Fig. 5.7 are almost similar to the B1 system.

Subjective Test

Compared with the B1 method, our P2 method showed distinguishable results for the speaker verifiability assessment using ASVeval and slightly inferior assessment results using ASReval. However, these results cannot sufficiently determine the effectiveness of an anonymization system. For instance, Table 5.2 shows that the P2 results are not always better than the P1 results. Therefore, we propose a subjective test that considers human hearing perception in the speaker anonymization system assessment.

In the initial state, we focused on the main purpose of an anonymization system, which is to conceal as much personally identifiable information as possible while maintaining the naturalness and intelligibility of the speech. The attack model has not yet been considered in this test. Three metrics were used for the subjective evaluation: speech intelligibility, speech naturalness, and speaker dissimilarity. Since the listeners did not know the context of the spoken utterances, we define "intelligible speech" as speech that contains words that can clearly be heard in the corresponding language. In this experiment, the words are in English. Meanwhile, "natural speech" is the speech most closely perceived as a human voice.

We conducted our subjective evaluation with a listening test divided into two main parts. The first part was measuring the intelligibility and naturalness metrics. A 5-point scale was used for both intelligibility (1-mostly unintelligible, 2-somewhat unintelligible, 3-cannot decide, 4-somewhat intelligible, 5-mostly intelligible) and naturalness (1-mostly unnatural, 2-somewhat unnatural, 3-cannot decide, 4-somewhat natural, 5-mostly natural). The second part was measuring the verifiability metric. We provided paired stimuli (original and anonymized utterances) and asked the participants to determine whether the speakers of those two stimuli were the same. The similarity metric was also a 5-point scale (1-completely similar, 2-mostly similar, 3-somewhat similar, 4-mostly different, and 5-completely different).

Twenty-four participants (thirteen men and eleven women, aged 20–35)

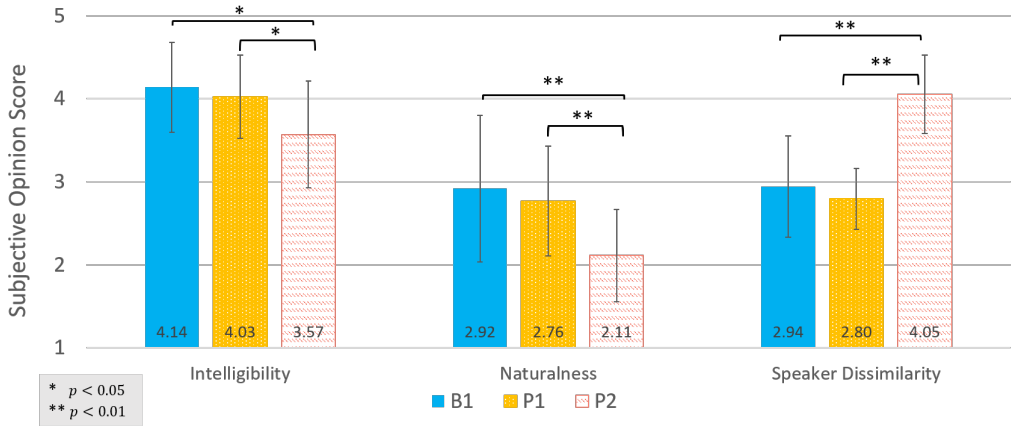


Figure 5.8: Overall subjective evaluation results in terms of intelligibility, naturalness, and speaker dissimilarity.

were employed in our subjective test. Each participant had a normal hearing ability and was a non-native speaker with a B2 English proficiency level. We conducted a paired-comparison test (with the original utterance provided as reference) to compensate for any bias from the participants as non-native speakers.

The subjective evaluation compared three methods: B1, P1, and P2. Nine stimuli containing both female and male utterances were randomly chosen from both LibriSpeech and VCTK datasets to evaluate speech intelligibility and naturalness (three stimuli from each method). The two stimuli used to compare speaker dissimilarity consisted of 36 pairs (twelve of the same stimuli from each method). These twelve pairs were randomly chosen from the development and test data of the LibriSpeech and VCTK datasets. There was an equal distribution of female and male utterances.

This test was conducted in a standard soundproof room equipped with a computer, an audio interface (Roland OCTA-CAPTURE), and headphones (SENNHEISER HDA 200) to avoid environmental bias. The sound pressure level of the background noise in the room was lower than 28 dB. We also randomized the order of the stimuli and normalized all the sound data of the listening test at the same sound pressure level of -20 decibels relative to full scale (dBFS) and sampled at 16 kHz. Before the experiment, we explained the test to each participant and instructed them to ensure they understood the test and the metrics. During the test, each stimulus was played only once to prevent bias.

The overall results of our subjective evaluation are shown in Fig. 5.8. The figure shows that the P1 method performed similarly to the B1 method with

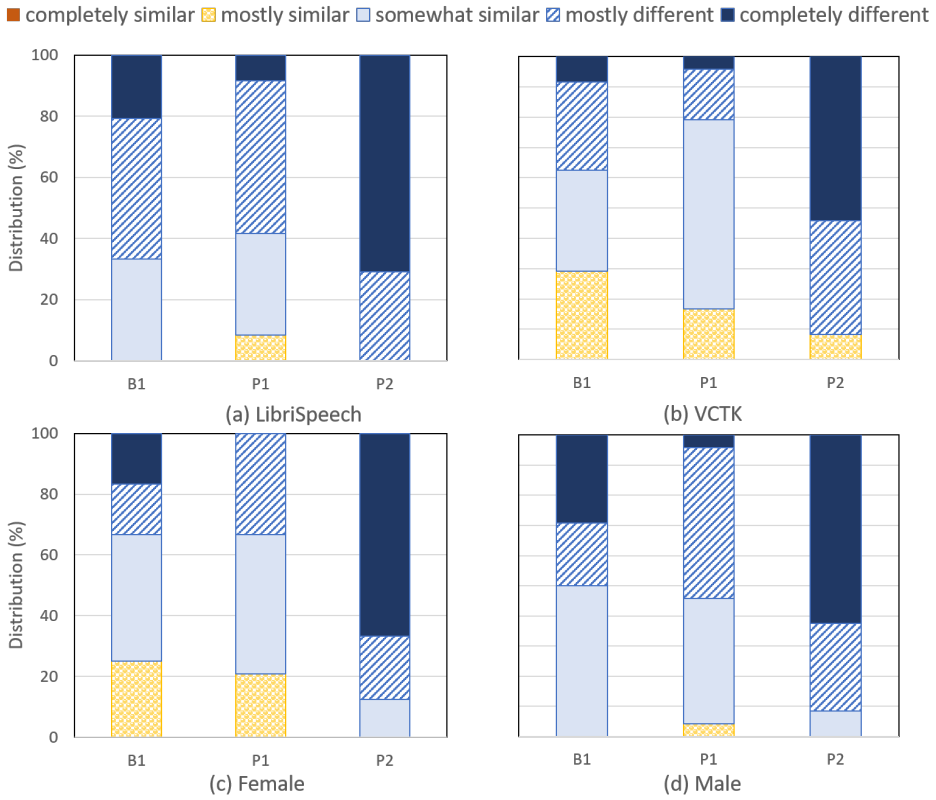


Figure 5.9: Subjective evaluation results of speaker dissimilarity in utterances from (a) LibriSpeech dataset, (b) VCTK dataset, (c) female speakers, and (d) male speakers.

all the metrics, while the P2 method performed significantly better than the B1 and P1 methods regarding dissimilarity. An apparent limitation of the P2 method is the slight reduction in the intelligibility and naturalness metrics. To verify this result, we conducted a single-factor analysis of variance (ANOVA) test using the mean of the total stimuli per method of the subject's rating score, i.e., the mean rating scores from three stimuli for the intelligibility & naturalness and twelve stimuli for the speaker dissimilarity.

The results of the ANOVA test showed significant differences between the three methods (B1, P1, and P2) in speech intelligibility ($F(2, 69) = 3.90, p < 0.05$), naturalness ($F(2, 69) = 6.28, p < 0.01$), and speaker dissimilarity ($F(2, 69) = 23.47, p < 0.01$) between the three compared methods (B1, P1, and P2). Subsequently, we conducted a post-hoc Tukey honestly significant difference test to determine the differences between the two methods for each metric. The results indicated that there is a statistically significant difference between the B1 and the P2 ($p < 0.05$ for speech intelligibility, $p < 0.01$ for naturalness, and $p < 0.01$ for speaker dissimilarity). Similarly, the difference between the P1 and the P2 is also significant ($p < 0.05$ for speaker intelligibility, $p < 0.01$ for naturalness, and $p < 0.01$ for speaker dissimilarity). Meanwhile, there is no significant difference between the B1 and the P1 ($p > 0.05$ for speaker intelligibility, naturalness, and speaker dissimilarity).

Figure 5.9 shows the speaker dissimilarity test distribution regarding datasets and gender. The top two figures show the speaker dissimilarity results of the three methods from the LibriSpeech dataset (left) and the VCTK dataset (right). The bottom two figures show the speaker dissimilarity results of the three methods from female utterances (left) and male utterances (right). These figures are consistent with the overall results in Fig. 5.6 that denote how the results of the P1 are relatively similar to the B1, whereas speaker dissimilarity significantly improved using the P2 method.

A demo page of the output anonymized speech of this system is available publicly as a reference[2].

5.4 Comparison Analysis on Speaker Anonymization Approaches

We proposed techniques to improve the primary baseline system introduced in the Voice Privacy Challenge 2020. An ablation test was conducted to determine the effectiveness of each method and its combinations. The de-

[2]http://www.jaist.ac.jp/~s1920436/anon/demo.html

Table 5.3: System description of related system anonymization methods.

System	Description
B1 [112]	primary baseline (x-vector anonymization)
B2 [112]	secondary baseline (McAdams coefficient modification)
P1	proposed method 1 (B1 using SVD and k-means clustering)
P2	proposed method 2 (P1 + modification on F_0 and speech duration)
O1 [115]	B1 using cosine distances, GMM for sampling vectors in a PCA-reduced space
S2 [34]	B1 using doman-adversarial training autoencoders
K2 [45]	anonymization using x-vectors and SS models, voice-indistinguishability metric, Griffin-Lim algorithm based waveform vocoder
I1 [31]	modification on formants, F_0, and speaking rate
C1 [18]	B1 + F_0 modification

tailed results were excluded to condense the excess results obtained from our experiments. Based on our experimental results shown in Subsection 5.3.3, in this section, we discuss the effectiveness of each technique in our methods, evaluation design & metrics for speaker anonymization, and we discuss the limitations in the current methods and evaluation protocols.

Four key questions about our current study are the following:

1. **How effective is modifying x-vector SV for speaker anonymization with a k-means clustering model?**

   In this study, we conducted experiments using several SV thresholds for anonymizing x-vectors. The average results are shown in Fig. 5.6 for speaker verifiability and in Fig. 5.7 for speech intelligibility. These results suggest that anonymization by modifying x-vector SV could achieve a performance comparable to the primary baseline. Speaker verifiability slightly improved when the threshold of the SV was reduced to 0.8, especially in the a-a scenario. Unfortunately, this improvement significantly distorted speech intelligibility for the LibriSpeech dataset.

   The trade-off between verifiability and intelligibility occurred from using SVD. The SVD technique is used to capture the intra-speaker characteristics. Consequently, the optimal SV threshold for the LibriSpeech dataset is higher than for the VCTK dataset because the VCTK dataset contains more variation (in accents, etc.) than the LibriSpeech dataset. To control the trade-off between speaker verifiability and speech intelligibility, we selected the best threshold for each dataset.

   In our methods, we used the k-means clustering model to choose the pseudo-target x-vector. This differs from our prior work [74] (labeled as "A2" in Figs. 5.10 and 5.12) in which we constructed a regression

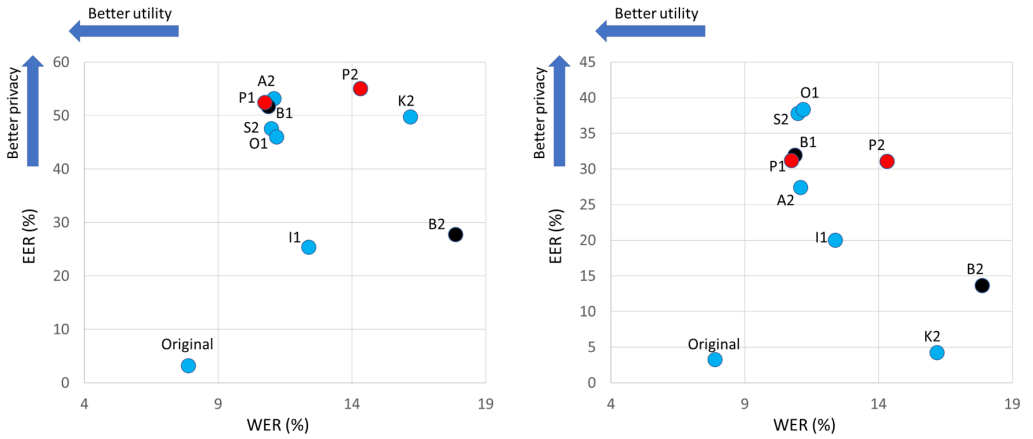


Figure 5.10: Mean WER versus mean EER over all LibriSpeech and VCTK datasets in (o-a) and (a-a) scenarios obtained from various systems proposed in Voice Privacy Challenge 2020. Black dot refers to results obtained by baseline system. Red dot refers to results obtained by our proposed system. Blue dot refers to results obtained by other systems proposed in Voice Privacy Challenge 2020. Table 5.3 describes each system.

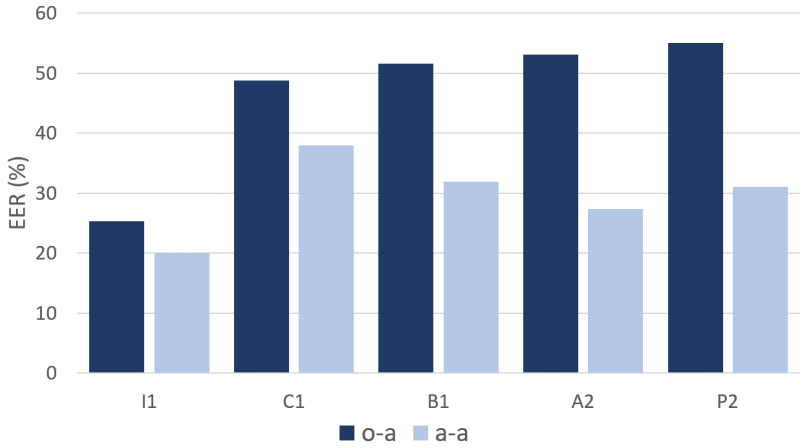


Figure 5.11: Mean EER values over LibriSpeech (test set) in (o-a) and (a-a) scenarios obtained by systems related to modifying speech prosody. Table 5.3 describes each method.

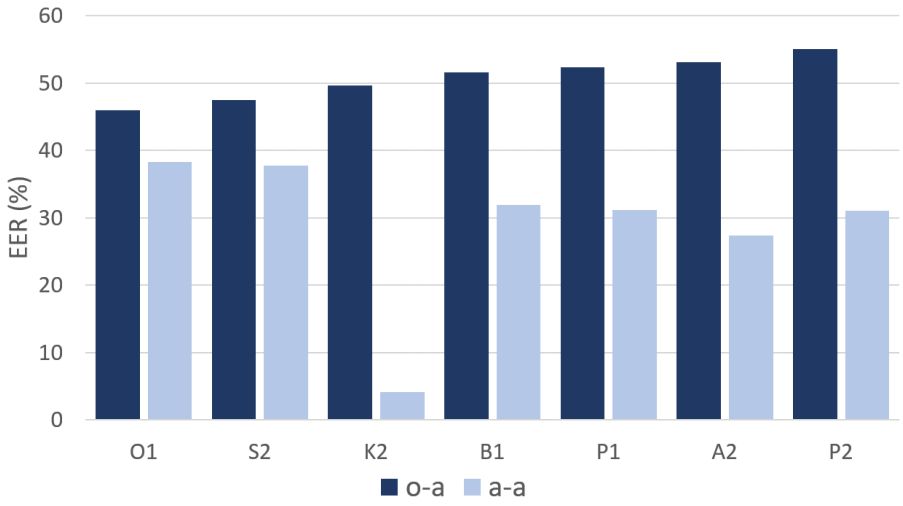


Figure 5.12: Mean EER values over all LibriSpeech and VCTK datasets in (o-a) and (a-a) scenarios obtained by systems related to x-vector anonymization. Table 5.3 describes each method.

model that chose the pseudo-target x-vector. The contribution of this clustering model is that it improves the speaker verifiability in the a-a scenario without degrading the intelligibility achieved by our prior work (compare A2 and P1 in Fig. 5.10).

2. **How effective is modifying speech prosody, including the F_0 and speech duration?**

   The evaluation results show the effectiveness of modifying the F_0 and speech duration for speaker anonymization. The objective evaluation results of P1 and P2 (shown in Fig. 5.6) show a slightly better performance of P2 in speaker verifiability, especially with the LibriSpeech dataset and female utterances. Contrary to the results of ASVeval, P2 caused slightly reduced speech intelligibility. Additionally, the subjective evaluation results of P1 and P2 showed more significant differences in speaker verifiability. Unfortunately, P2 caused slightly more perceivable distortion than P1 in terms of utility (intelligibility and naturalness).

   To further investigate the effect of speech prosody modification, we conducted comparative analysis of related speaker anonymization systems. Figure 5.11 shows the ASVeval results of five systems labeled I1, C1, B1, A2, and P2 (detailed in Table 5.3). We could not obtain full results for all the systems; therefore, the results shown are only for LibriSpeech (test data). The results of I1 [31] in Figs. 5.11 and 5.12

show that modifying formants, the F_0, and the speaking rate alone improves the anonymization performance (compared with the original in Fig. 5.12). It supported the previous studies [3, 27] that described the strong relationship between speaker individuality and the F_0 & speech duration. Unfortunately, modifying the speech prosody degraded the intelligibility.

Combining the speech prosody modification with the main framework in the primary baseline (systems C1, A2, and P2) improves speaker anonymization. A study by Champion et al. [18] investigated the effect of F_0 modification in the primary baseline system across gender. Our current work did not focus on cross-gender modification, so we only compared their results obtained from same-gender modification labeled C1. Although the C1 reduced the speaker verifiability performance in the o-a scenario compared with B1, it significantly improved the performance in the a-a scenario. The ASReval results showed a similar tendency to other speech prosody modification systems in the slight reduction in WER that occurred [18]. Due to incomplete results, the mapping of WER and EER results across all datasets for the C1 could not be included in Fig. 5.10.

In contrast to the C1, we carried out experiments using WORLD for the F_0 modification (systems A2 and P2). In A2 (our prior work [74]), we used the estimated F_0 obtained from the SWIPE algorithm in WORLD in the primary baseline framework with SVD-based modification. Overall, the A2 performed slightly better than the B1 in the o-a scenario but not as well in the a-a scenario. A more significant improvement could be achieved by modifying the estimated F_0 and speech duration (method P2), especially in the o-a scenario (as shown in Fig. 5.11). Despite improving privacy, we could see drawbacks similar to those that occurred due to the speech prosody modification in Fig. 5.10.

In summary, all systems related to modifying speech prosody can significantly improve privacy, but they also slightly degrade utility. In addition, the P2 performed better in ASVeval for female utterances than for male utterances. We predict that these results occur due to the different F_0 range between female and male speakers [114] and the linear transformation utilized. An effective transformation of the F_0 for anonymization is gender-dependent because female speakers have a wider F_0 range than male speakers.

3. **How does the performance of the proposed method compare with existing speaker anonymization systems?**

Figure 5.10 shows the scatter plot of the mean WER and mean EER results of several existing systems in both scenarios (o-a and a-a). The mean WER and mean EER were calculated by averaging the results obtained from all the LibriSpeech and VCTK datasets (development and test sets). In the Voice Privacy Challenge 2020 results, one criteria for determining better privacy is a higher EER. Subsequently, better utility is determined by having a lower WER. Although solely using these two metrics might be considered an oversimplified evaluation, the results in Fig. 5.10 gives insight into comparing the existing systems. For example, one of the overall conclusions provided in the Voice Privacy Challenge 2020 results indicated that the systems based on x-vector anonymization (B1, P1, P2, O1, S2, and K2) could perform better than the ones based on signal-processing methods (B2 and I1).

The overall results show that some systems (A2, P1, O1, and S2) have nearly similar results as the primary baseline system (B1), especially in the o-a scenario. Our prior work (A2) was less effective than B1 in the a-a scenario. However, the results of other participants that also used the primary baseline framework (O1 and S2) showed improved privacy in the a-a scenario (mean EER increased around 5%), but the mean EER reduced around 5% to 10% for the o-a scenario. Subsequently, our P2 method, slightly improved in the o-a scenario but had a result similar to B1 regarding privacy. Unfortunately, it increased the mean WER by approximately 3.5% in comparison with B1. The B2, I1, and K2 systems not based on B1 were less effective than B1 in both privacy and utility. Figure 5.12 compares ASVeval results of anonymization systems that use x-vectors. Although there is a relatively slight degradation in intelligibility, our P2 method achieved the highest mean EER in the o-a scenario.

4. **How reliable and significant is the subjective evaluation for the speaker anonymization system?**

   There are limitations to the existing objective evaluation. One such limitation is that speaker verifiability is evaluated using x-vector embedding [107] in the ASVeval. Although this system performed the best in terms of the EER for the VoxCeleb dataset, Hautamäki and Kinnunen's study [46] indicates that an x-vector embedding system that only uses the Mel-spectrogram as its input is not robust with intra-speaker variations. Reportedly, this primarily results from the mismatch between the F_0 mean and the associated formant frequencies. Consequently, a better objective evaluation for assessing the verifiability of a speaker

anonymization system must be considered because it should distinguish the uniqueness of anonymized speech for each speaker. Another limitation is that it is quite difficult to decide which system is the most effective in all cases regardless of the datasets and genders using only the objective evaluation results (such as the results shown in Figs. 5.6 and 5.7).

Accordingly, we proposed a subjective evaluation that can more reliably determine the effectiveness of different speaker anonymization systems. This subjective evaluation differs from the one introduced in the Voice Privacy Challenge 2020 [112]. Even though native speakers have a better understanding of their mother language, we considered the difficulty in gathering an adequate number of native speakers as suggested for the challenge's subjective evaluation. Collecting evaluation results via the internet could be an alternative solution to deal with this problem; however, there will be biases due to different environments, equipment, etc. Furthermore, instead of using a 10-point scale opinion score, we used a 5-point scale based on psychological studies that suggested higher reliability regarding the response rate and quality [15, 32]. The 10-point scale opinion score is too difficult and could increase the "frustration level" even for a native speaker [15]. Furthermore, Eli P. Cox's 1980 study on the optimal number of alternatives for a scale suggested that an odd number of alternatives is preferable to enable a neutral response [32].

To improve the reliability of our subjective evaluation, we anticipated bias based on environment, equipment, understanding, and/or human perceptual phenomena in the hearing system. We also compensated for any potential misunderstandings from non-native speakers by using pair-comparison even though we verified the participants' English skills (detailed in Subsection 5.3.3). We also conducted ANOVA tests to analyze significant differences in the systems.

Figures 5.8 and 5.9 show our subjective evaluation results, comparing B1 with our P1 and P2 systems. Unfortunately, the results obtained using the P1 method are not significantly different from the results obtained using the B1 method. However, combining all the techniques proposed in this study (P2) could improve the speaker dissimilarity significantly compared with the B1 method regardless of the dataset and gender.

Chapter 6

Evaluation and Discussion

This chapter provides an explanation of the extensive evaluation of the proposed framework. We take a SIH method that utilized the McAdams coefficient as a study case for evaluating both watermarking and anonymization performance.

6.1 Evaluation

The evaluation is comprised of three sections, i.e., (1) assessment of the analysis and synthesis process, (2) watermarking method in terms of inaudibility and robustness, and (3) anonymization performance in terms of speaker verifiability and speech intelligibility. The comparison analysis for each evaluation is also conducted to compare the performance of our proposed method and baseline systems. For instance, we analyze the performance of speaker anonymization by comparing our proposed method with the primary baseline anonymization system based on neural source-filter (NSF) model & x-vector speaker embedding and secondary baseline anonymization system based on McAdams coefficient.

6.1.1 Analysis and Synthesis Assessment

This assessment aims to investigate the reliability of the analysis and synthesis methods utilized for SIH. Here, we assume that the hidden message ($s(m)$) is an empty set. By investigating the analysis and synthesis methods, we expect to infer the effectiveness of the SIH method regardless of the analysis and synthesis methods.

We semi-randomly selected a total of 100 utterances from LibriSpeech [89] and utterances from VCTK [120] (similar subset with the dataset for

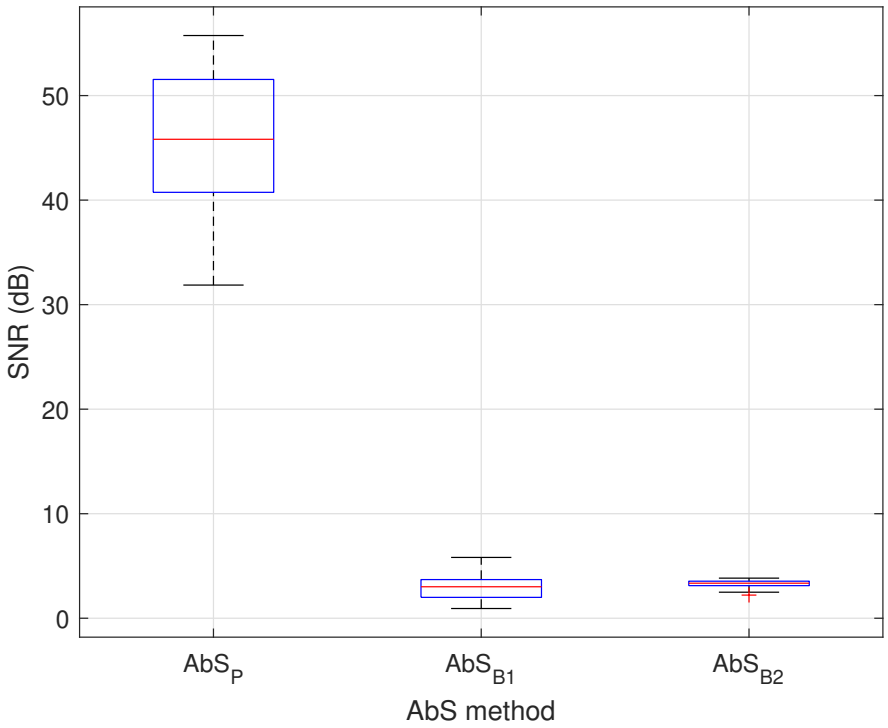


Figure 6.1: Analysis and Synthesis Assessment using SNR.

speech watermarking evaluation in [75]). Semi-randomly means that we selected utterances from a particular number of speakers (with balance gender distribution). LibriSpeech is sampled at 16 kHz and designed for automatic speech recognition research, and VCTK is sampled at 48 kHz and designed for text-to-speech research. We unified the sampling rate of both corpora to 16 kHz. The selected utterances varied depending on the speaker and speech content. For the assessment metrics, we used signal-to-noise ratio (SNR), log-spectral density (LSD), and perceptual evaluation of speech quality (PESQ).

The SNR is widely used to compare the noise level in comparison to the signal for transmission. A higher SNR value indicates that the signal strength is stronger than the noise levels. In SIH, noise is often referred to as the subtraction of watermarked signal with the original signal. The SNR value is expressed in dB. Although it is arguable to interpret the value of SNR for speech quality assessment, the value greater than 40 dB is considered excellent in several studies. Figure 6.1 shows the assessment results using (SNR) metric. These results indicate that the output signal of the AbS

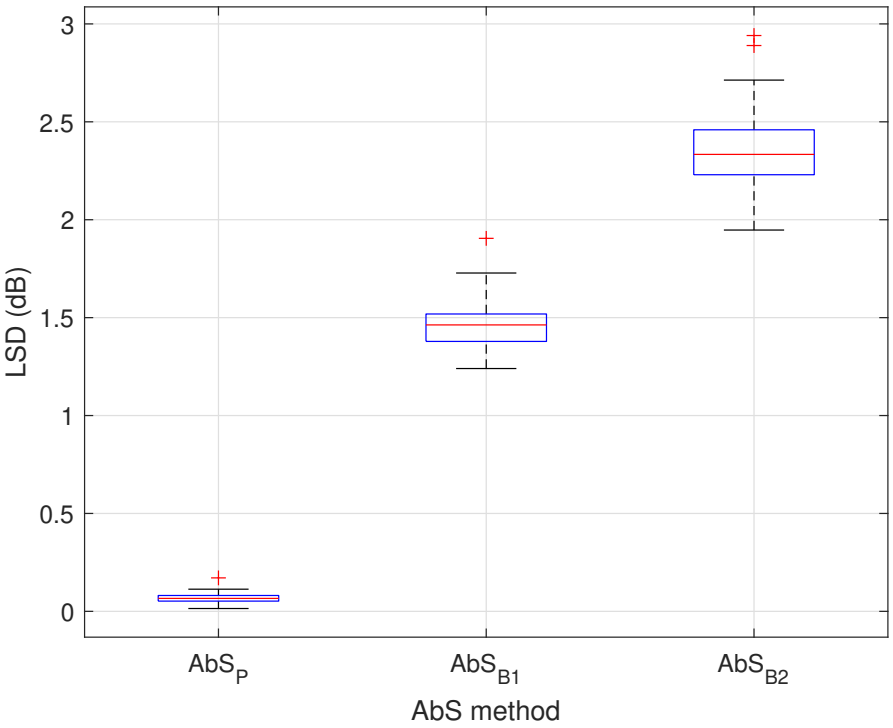


Figure 6.2: Analysis and Synthesis Assessment using LSD.

method of our proposed method (AbS_P) could achieve higher SNR than the AbS method of baseline systems (AbS_{B1} and AbS_{B2}). The mean SNR for AbS_P is approximately 45 dB, while the AbS_{B1} and AbS_{B2} are both less than 10 dB.

The LSD is often used to measure the log-spectral distortion in speech coding. It has become one of the standards for measuring the performance of quantization or interpolation. The threshold value for LSD to be considered as good quality is less than 1 dB. Figure 6.2 shows the assessment results using LSD metric. These results indicate that only AbS_P could pass the requirement of LSD (mean LSD value is approximately 0.05 dB). Meanwhile, the analysis and synthesis method of both AbS_{B1} and AbS_{B2} caused high spectral distortion with mean LSD values are around 1.45 dB and 2.35 dB, respectively.

The PESQ is also one of the standards for automatic speech quality assessment, which was originally addressed for a telephony system. The PESQ represents the perceptual speech quality of $y(n)$ with $x(n)$ as the reference in mean opinion scores (MOS). The MOS varies from a scale of 1

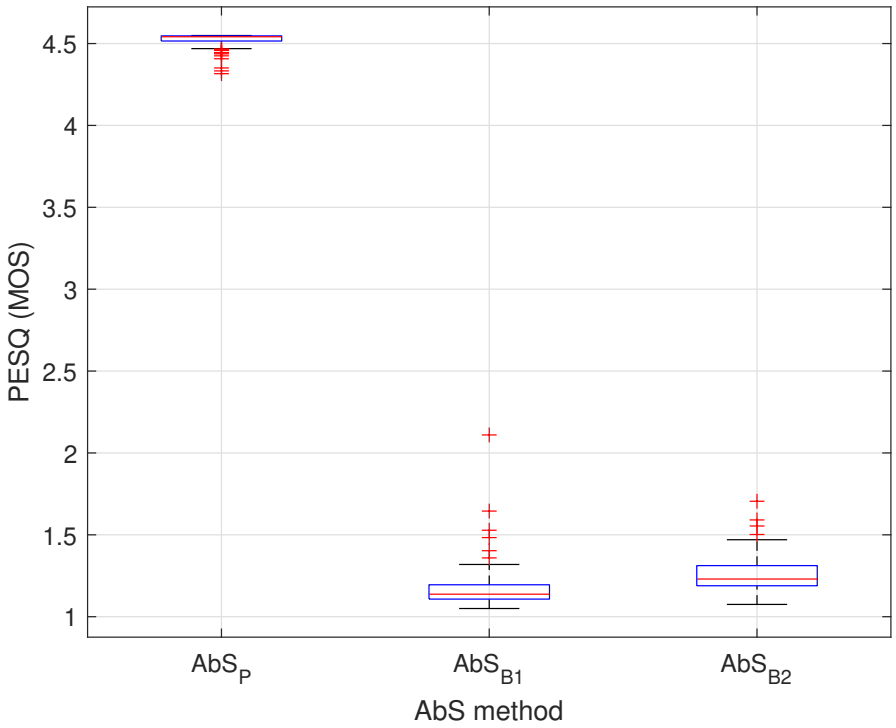


Figure 6.3: Analysis and Synthesis Assessment using PESQ.

(bad) to 5 (excellent). Typically, the threshold of PESQ value is 3. Figure 6.3 shows the assessment results using PESQ metric. These results indicate that the quality of output resynthesized speech using analysis and synthesis method P is almost excellent (mean PESQ value is approximately 4.5). On the other hand, the quality of output resynthesized signals of AbS_{B1} and AbS_{B2} are not good (mean PESQ value is less than 3).

6.1.2 Watermarking Assessment

The watermarking assessment aims to investigate the reliability of our proposed method in terms of watermarking requirements based on IHC [51]. Most of the evaluation setting follows the evaluation explained in Section 4.2.

We semi-randomly selected 250 utterances from LibriSpeech [89], and 250 utterances from VCTK [120]. We unified the sampling rate of both corpora to 16 kHz. The selected utterances varied depending on the speaker and speech content. A total of 500 utterances were then split into 90% for the training set and 10% for the testing set. The training set was used for constructing

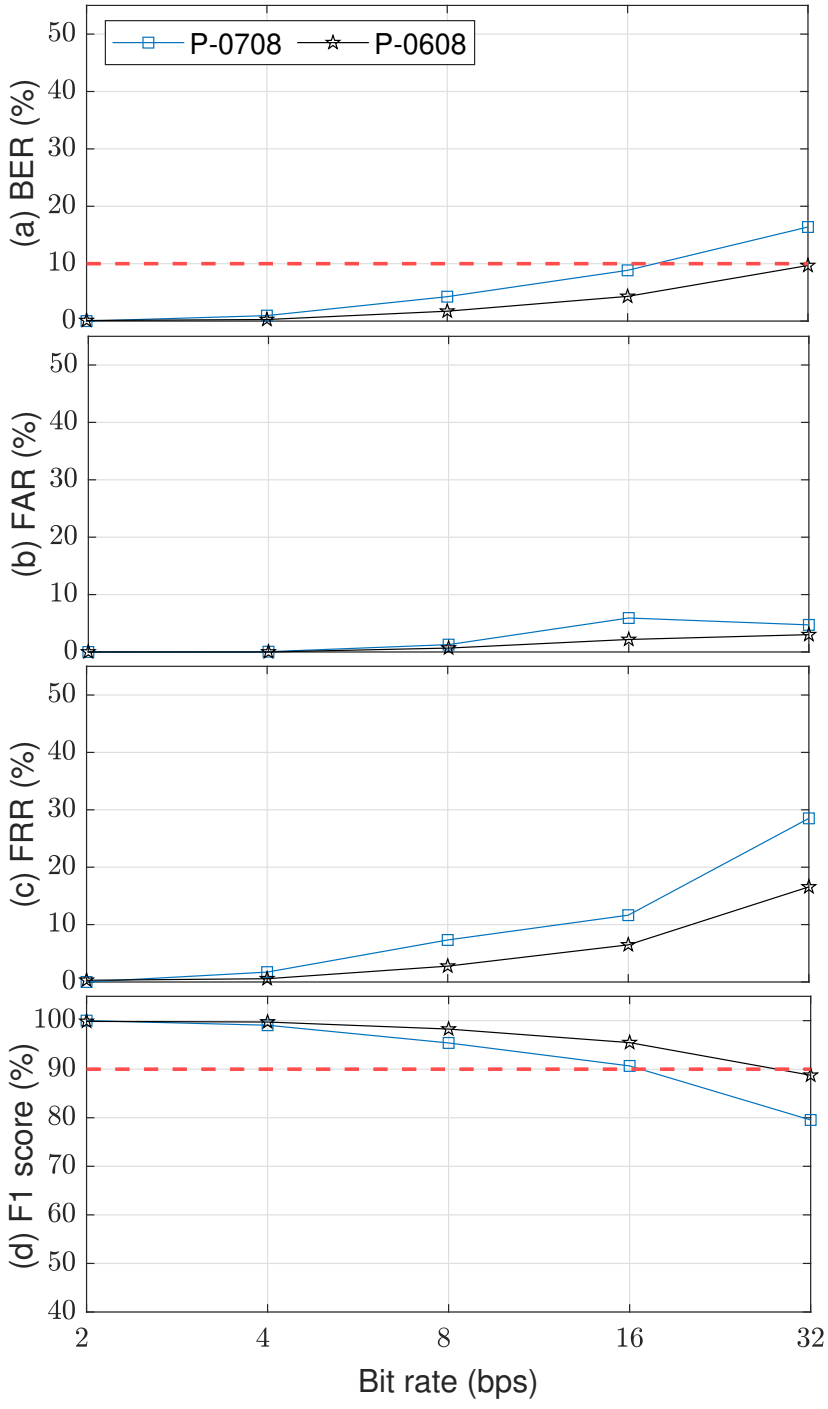


Figure 6.4: Watermarking detection accuracy evaluation in terms of: (a) BER, (b) FAR, (c) FRR, and (d) F1-score.

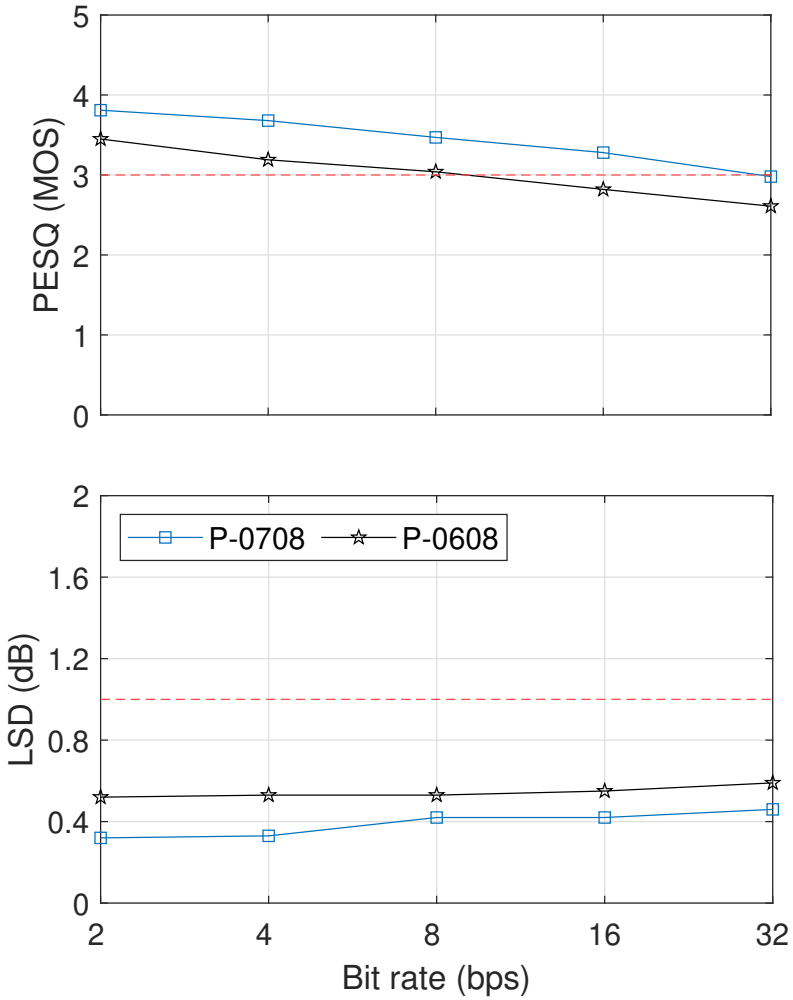


Figure 6.5: Sound-quality evaluation results in terms of PESQ (top) and LSD (bottom).

the random forest classifier for blind detection. The testing set consisted of 50 utterances. The testing set was then used for evaluating speech watermarking performance. The metrics for evaluating the inaudibility requirement are log spectral distance (LSD) [42] and perceptual evaluation of speech quality (PESQ) [97] ITU-T P.862.

For evaluating watermark detection accuracy and security level, we used the bit error rate (BER), false acceptance rate (FAR), false rejection rate (FRR), and F1-score. The threshold set for an acceptable BER is 10% [51]. In the evaluation, we defined the watermarked bit-stream ($w(k)$) as a random

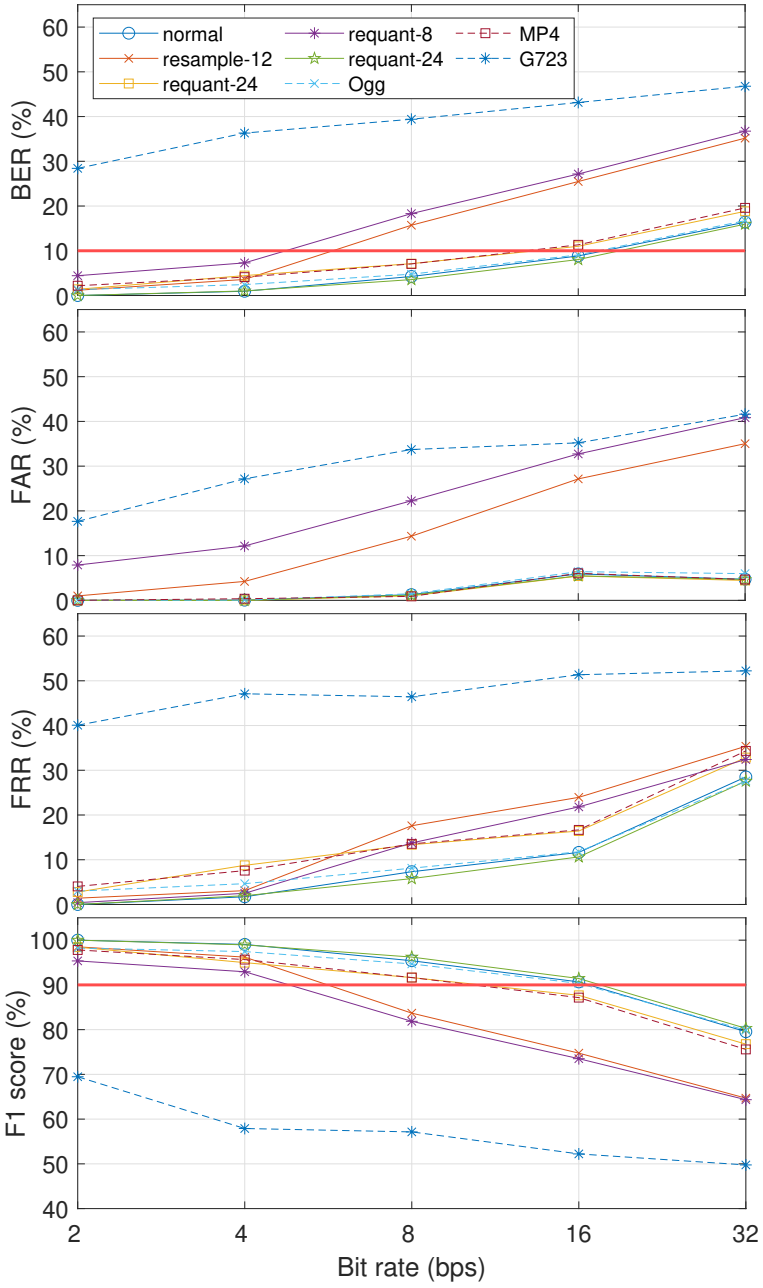


Figure 6.6: Robustness results in terms of BER, FAR, FRR, and F1-score in eight cases: normal, resample-12, resample-24, requant-8, requant-24, Ogg, G723, and MP4.

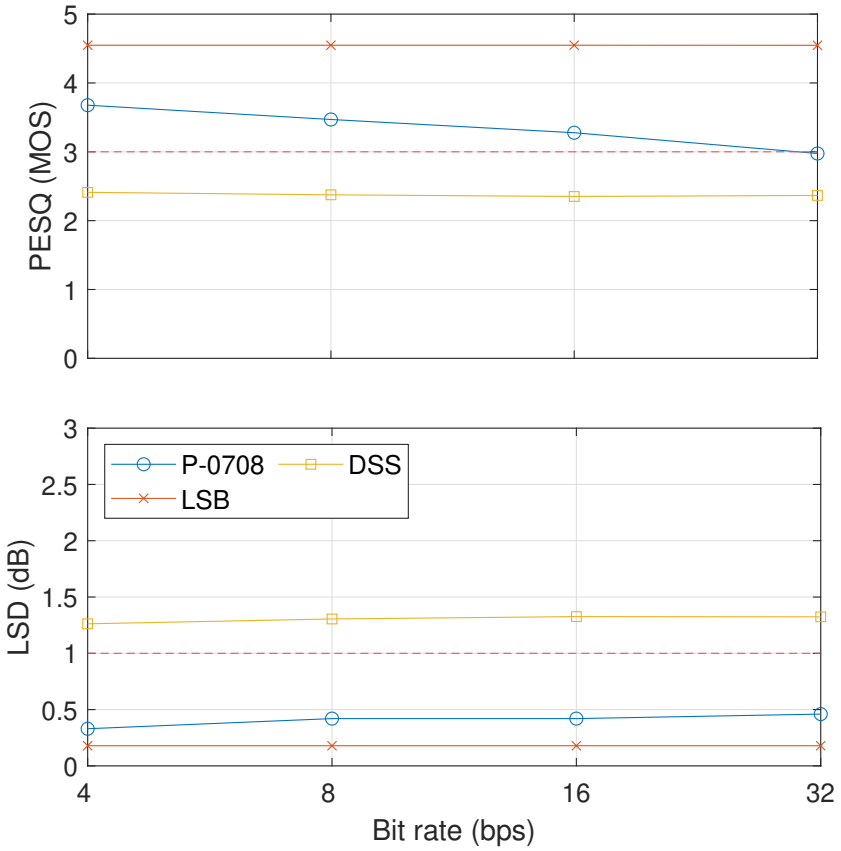


Figure 6.7: Evaluation of inaudibility results of three compared methods (P-0708, LSB, and DSS).

binary stream with the length depending on the payload. We investigated the payloads of 2, 4, 8, 16, and 32 bps. For robustness evaluation, we considered eight cases of non-malicious signal processing operations, i.e., normal (no attack), down-sampling to 12 kHz (resample-12), up-sampling to 24 kHz (resample-24), bit compression to 8 bits (requant-8), bit extension to 24 bits (requant-24), conversion to Ogg format (Ogg), conversion to MPEG-4 Part 14 or MP4 format (MP4), and conversion to G723.1 codec (G723). The bitrate of G723.1 codec is 5.3 kbps with algebraic code-excited linear prediction (ACELP) algorithm.

We also carried out a comparison analysis among our proposed method using McAdams coefficients $(\alpha_0, \alpha_1) = \{(0.6, 0.8), (0.7, 0.8)\}$ (P-0608, P-0708) and two other well-known speech watermarking methods, i.e., LSB and DSS. We conducted the comparative analysis using payloads of 4, 8, 16, and 32 bps.

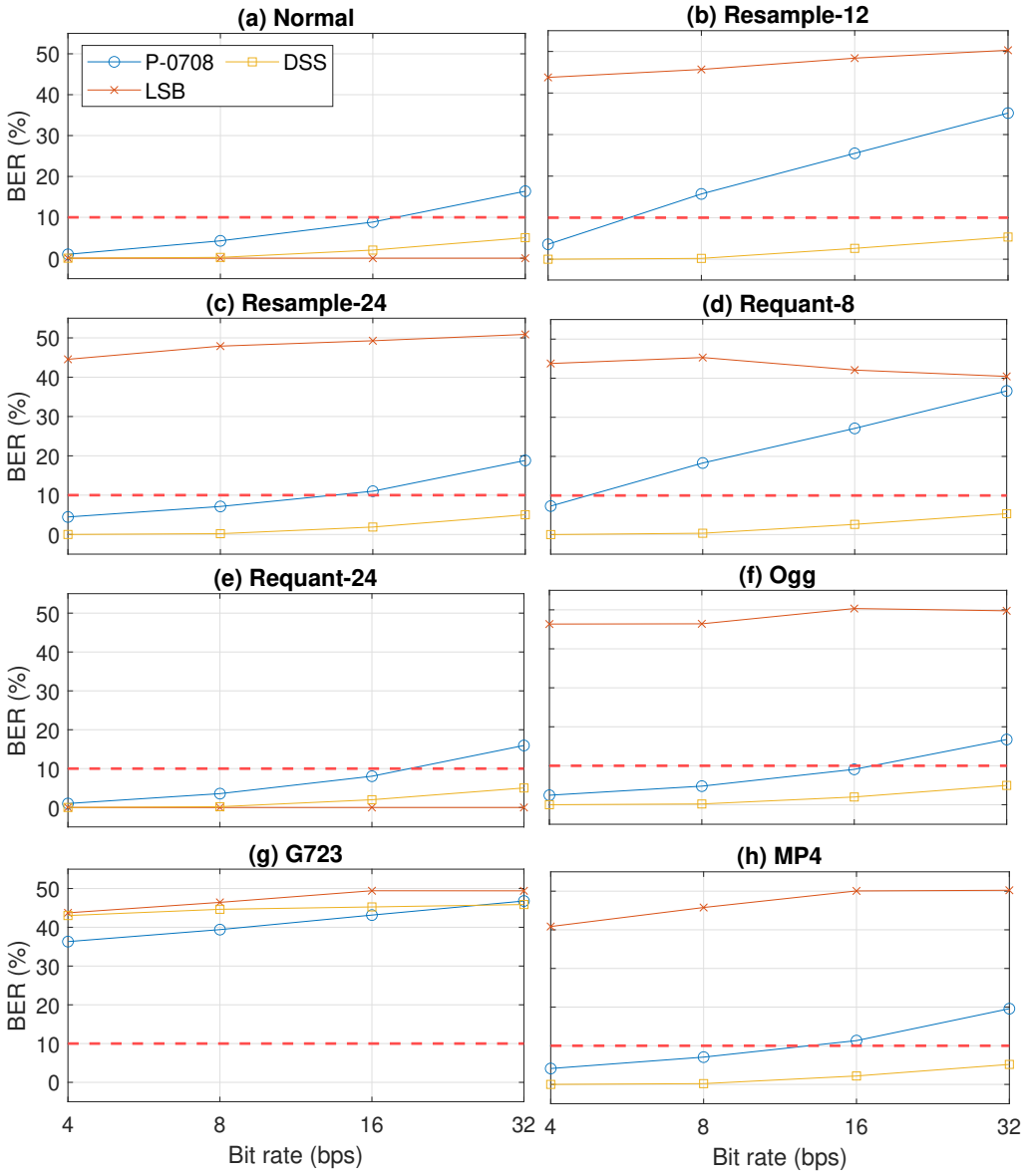


Figure 6.8: Robustness results of three compared methods (P-0708, LSB, and DSS) in terms of BER in eight cases: normal, resample-12, resample-24, requant-8, requant-24, Ogg, G723, and MP4.

Figure 6.4 shows the watermark detection accuracy and security level results in terms of BER, FAR, FRR, and F1-score. Considering the detectability threshold (BER = 10%), the embedding payload for P-0608 was up to

32 bps, where as for P-0708 was up to 16 bps. A similar error rate for FAR and FRR was also found when we considered the observed payloads. When considering the overall security level in F1-score with a threshold of 90%, both of the proposed methods could pass the threshold level.

The results of the inaudibility test are shown in Fig. 6.5. On the basis of the inaudibility threshold, the evaluation results indicate that both PESQ and LSD scores for P-0708 satisfied the requirement of up to 32 bps. Meanwhile, for P-0608, although the LSD score is acceptable even with embedding payload 32 bps, the PESQ score could be satisfied by embedding up to 8 bps watermark signal. We will thus consider P-0708 for further analysis of robustness.

The robustness results in eight cases are shown in Fig. 6.6. The watermarked signal was generated with $\alpha_0 = 0.7$ and $\alpha_1 = 0.8$. By only considering detectability and security level threshold, the results indicate that our proposed method had similar robustness with the normal case when dealing with up-sampling (resample-24), bit extension (requant-24), Ogg, and MP4 processing operations. Robustness degraded when down-sampling (resample-12), bit compression (requant-8), and G723.1 codec (G723) were applied. For resample-12 and requant-8, we can say that our proposed method is robust when the payload is 4 bps (BER $< 10\%$). Unfortunately, our proposed method is not robust when the G723.1 codec is applied (BER $> 10\%$).

The results on the comparative analysis among our proposed method with $(\alpha_0, \alpha_1) = (0.7, 0.8)$ (P-0708), LSB, and DSS are shown in Figs. 6.7 and 6.8. Figure 6.7 shows the inaudibility comparison results in terms of PESQ and LSD. These results indicate that LSB and our proposed method could pass the threshold of inaudibility but not DSS.

Figure 6.8 shows the robustness comparison results in terms of BER. In contrast to the inaudibility results, the robustness results indicate that LSB was fragile in dealing with almost all observed signal processing operations, except with the up-sampling to 24 kHz. However, DSS was very robust even in a higher payload, except with the G723.1 speech codec. Although not as robust as DSS, Our proposed method had better robustness against most of the observed signal processing operations (down-sampling, re-quantization, Ogg format, and MP4 format) than LSB.

6.1.3 Speaker Anonymization Assessment

We evaluated the speaker verifiability of our proposed method by using a pretrained automatic speaker verification system (ASVeval) and the intelligibility by using a pretrained automatic speech recognition system (ASReval), similar to the protocol in VP2020 (as explained in Section 2.3). Our intent with this evaluation is mainly to investigate the effectiveness and reliability

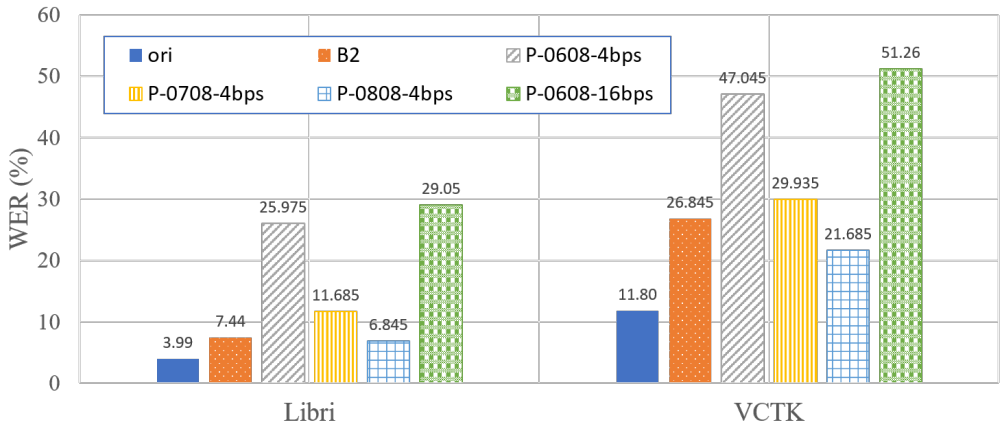


Figure 6.9: ASReval results of ori, anonymized speech by B2, proposed methods with several McAdams coefficient pairs and embedding payloads.

of the proposed method in anonymizing the PII of the speaker individuality information.

For the speaker anonymization, we conducted our evaluation using ASVeval to check the speaker verifiability performance in three cases: original enrolls and original trials case (o-o), original enrolls and anonymized trials case (o-a), and anonymized enrolls and anonymized trials case (a-a). The metric used in ASVeval is EER, where higher EER is regarded as higher anonymity.

Figure 6.10 compares the average results of the corresponding development and test datasets of the speaker verifiability assessment using ASVeval with the B1 system with those of our method. These results were obtained by averaging the EER results of development and test datasets. These results are categorized into female and male genders as the subset of the datasets, i.e., (a) Libri female, (b) Libri male, (c) VCTK female, and (d) VCTK male. The overall results show the same tendency that the proposed method with McAdams coefficients pairs 0.6 and 0.8 could surpass the performance of the B2 system. When the payload is set higher, the speaker verifiability also improves (compare P-0608-4bps and P-0608-16bps). The results in Fig. 6.11 also show similar trend to Fig. 6.10.

The corresponding results of ASReval are provided in Fig. 6.9. The ASReval results showed that the P-0808-4bps could slightly improve the speech intelligibility than the B2 in terms of WER. Meanwhile, embedding watermarks with McAdams coefficients pair 0.7 and 0.8 (P-0708-4bps) caused slightly more distortion than the B2. Although the speaker verifiability in the ASVeval surpassed the B2 system, embedding watermarks with McAdams coefficients pair 0.6 and 0.8 (P-0608-4bps and P-0608-16bps) caused around

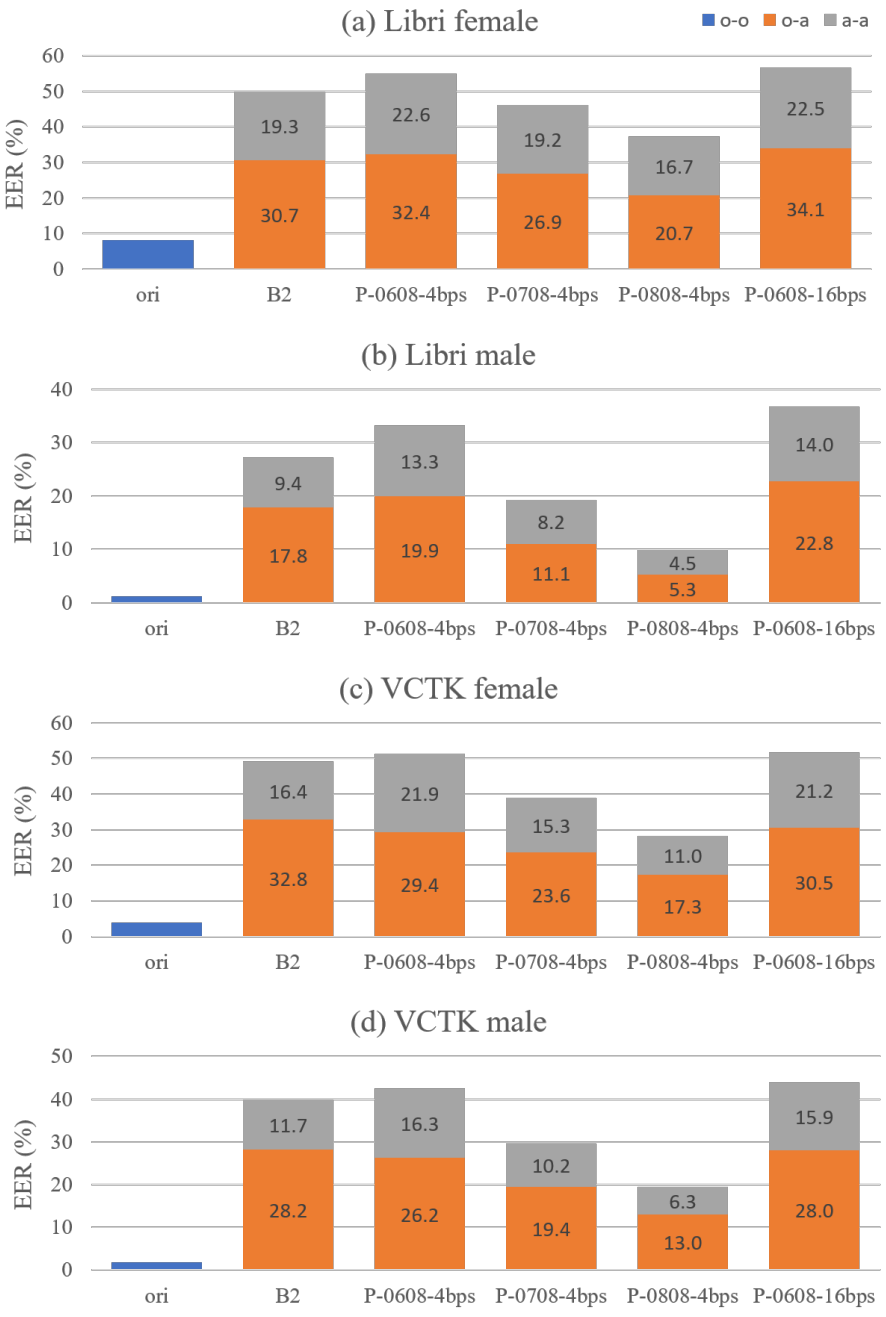


Figure 6.10: Average ASVeval results of ori, anonymized speech by B2, proposed methods with several McAdams coefficient pairs and embedding payloads. Orange bars represent results in pairs of original enrollment and anonymized trials (o-a). Gray bars represent results in pairs of anonymized enrollment and anonymized trials (a-a).

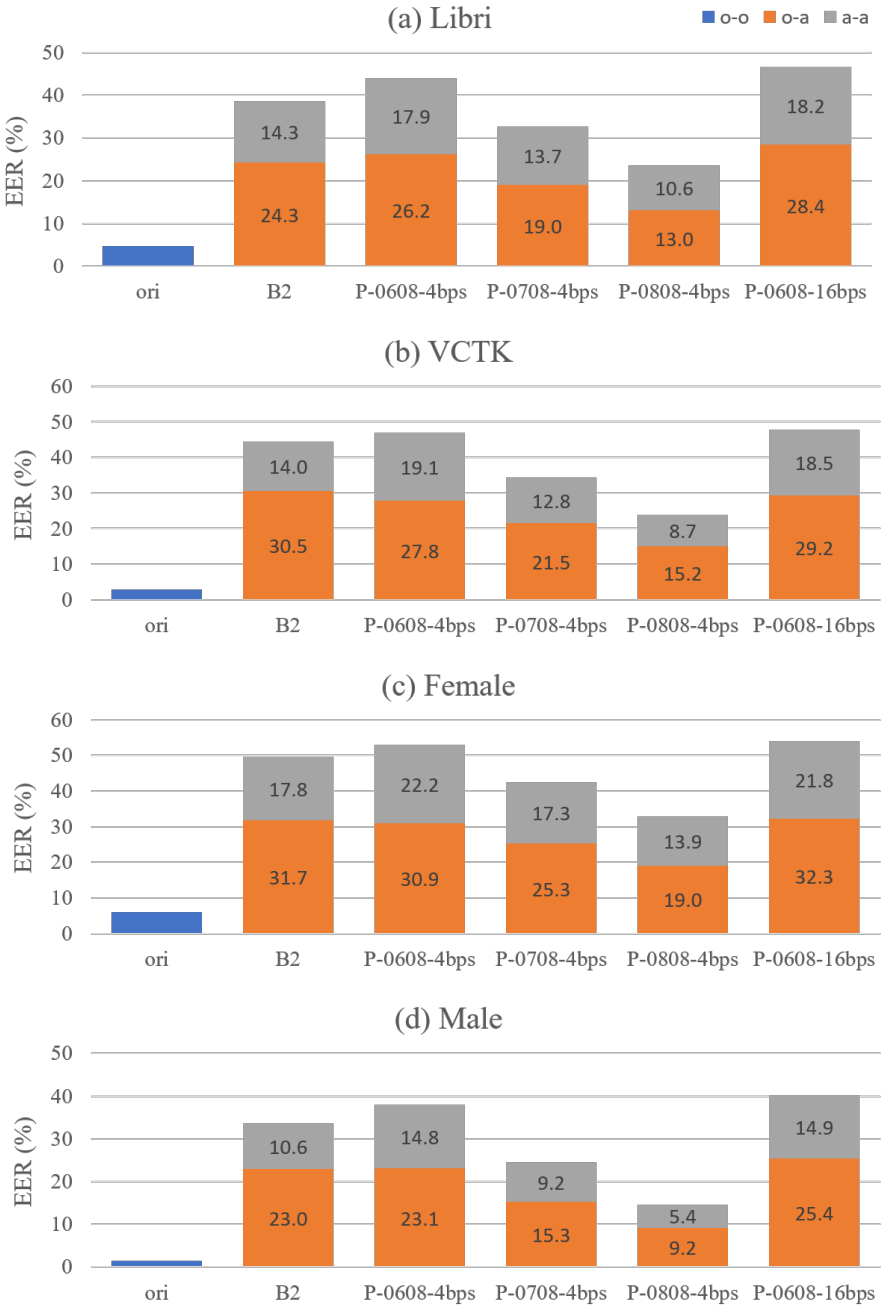


Figure 6.11: Average ASVeval results based on dataset and gender of ori, anonymized speech by B2, proposed methods with several McAdams coefficient pairs and embedding payloads. Orange bars represent results in pairs of original enrollment and anonymized trials (o-a). Gray bars represent results in pairs of anonymized enrollment and anonymized trials (a-a).

20% more WER than the B2 system, which apparently remains as the limitation of this work.

In addition to the EER score, we calculated the log-likelihood-ratio cost function for evaluating the speaker verifiability of the baseline systems (B1 and B2) and proposed methods with several McAdams coefficient pairs and embedding payloads. As to compare the overall results of privacy metrics versus utility metrics, Fig. 6.12, Fig. 6.13, Fig. 6.14, Fig. 6.15 show the performance of several methods of speaker anonymization. The details of ASVeval are provided in Appendix A. The increasing trend in the EER and C_{llr} scores indicates improvement in the privacy metric of a speaker anonymization system. In terms of the utility metric, the corresponding ASReval results of compared methods are also available in Appendix A.

6.2 Discussion

In this chapter, we have conducted a thorough evaluation of our proposed SIH framework for the purpose of watermarking and anonymization. We took methods based on McAdams coefficient manipulation as our study case. The evaluation was conducted into three main parts: analysis and synthesis assessment, watermarking evaluation, and speaker anonymization evaluation. The key information that we could draw from the evaluation results are as follows:

1. The analysis and synthesis assessment results showed that the output speech signals from our analysis and synthesis method have significantly better quality than those from baseline systems. These results indicate that the analysis and synthesis of baseline systems itself are causing distortion and might cause a certain level of anonymity error.

2. The watermarking evaluation results showed that the proposed methods with McAdams coefficients pairs $(\alpha_0, \alpha_1) = \{(0.6, 0.8), (0.7, 0.8)\}$ (P-0608, P-0708) are reliable (accomplished the requirements of SIH). The embedding payload that satisfies both robustness and inaudibility is 8 bps and 16 bps for P-0608 and P-0708, respectively.

3. Finally, the speaker anonymization evaluation results showed that the P-0608 method could improve the performance of speaker verifiability of the B2 but not the speech intelligibility. Increasing the embedding payload could also improve the speaker verifiability performance.

There are several limitations related to existing techniques and evaluations. Per the summary in the Voice Privacy Challenge 2020, x-vector-anonymization

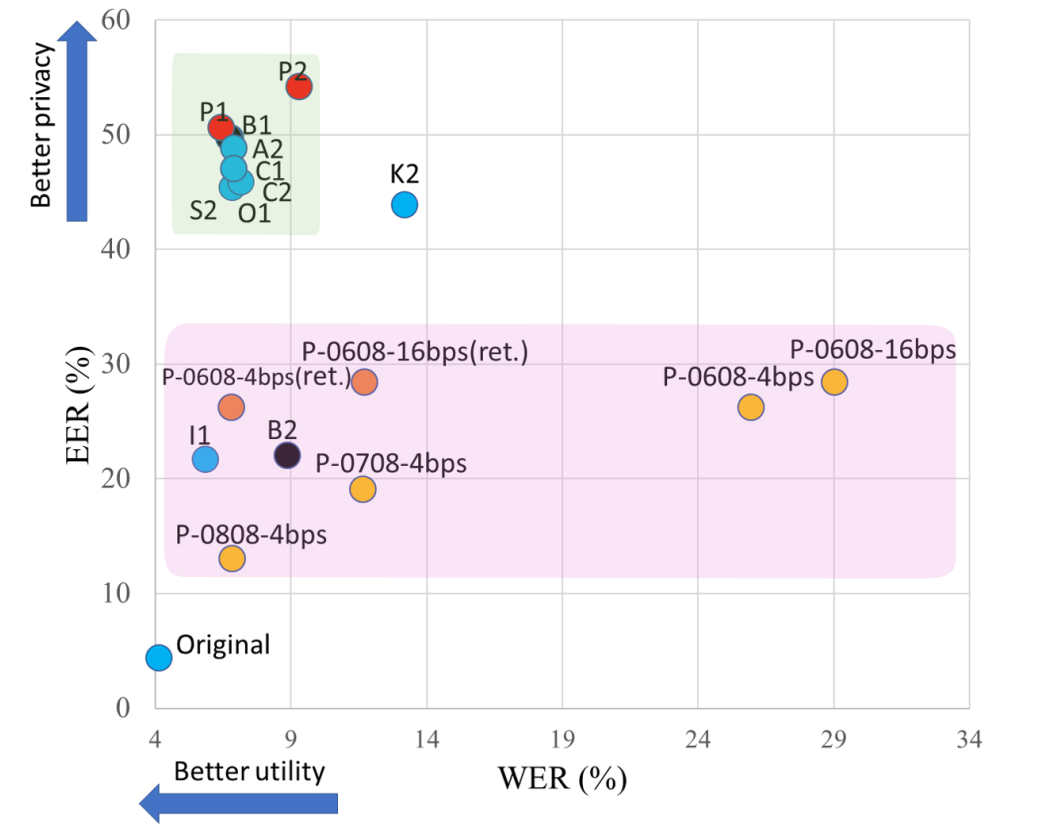


Figure 6.12: Mean WER versus mean EER over all **VCTK datasets in (a-a)** scenario obtained from various systems proposed in Voice Privacy Challenge 2020. Black dot refers to results obtained by the baseline system. Red dot refers to results obtained by methods in [76]. Yellow and orange dots refer to results obtained by proposed methods. Blue dot refers to results obtained by other systems proposed in Voice Privacy Challenge 2020. The dots in the green shaded area are methods based on a neural vocoder (mainly based on the primary baseline framework). The dots in the pink shaded area are methods based on the LP vocoder.

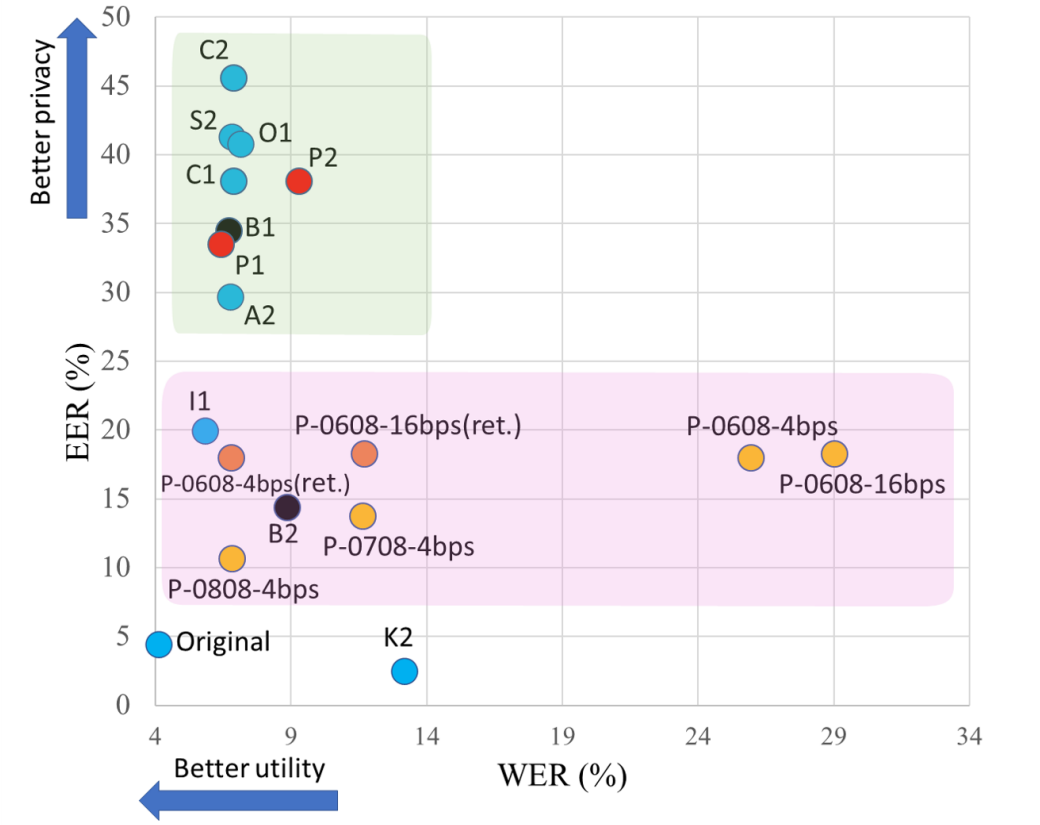


Figure 6.13: Mean WER versus mean EER over **LibriSpeech dataset in (a-a)** scenario obtained from various systems proposed in Voice Privacy Challenge 2020. Black dot refers to results obtained by the baseline system. Red dot refers to results obtained by methods in [76]. Yellow and orange dots refer to results obtained by proposed methods. Blue dot refers to results obtained by other systems proposed in Voice Privacy Challenge 2020. The dots in the green shaded area are methods based on a neural vocoder (mainly based on the primary baseline framework). The dots in the pink shaded area are methods based on the LP vocoder.

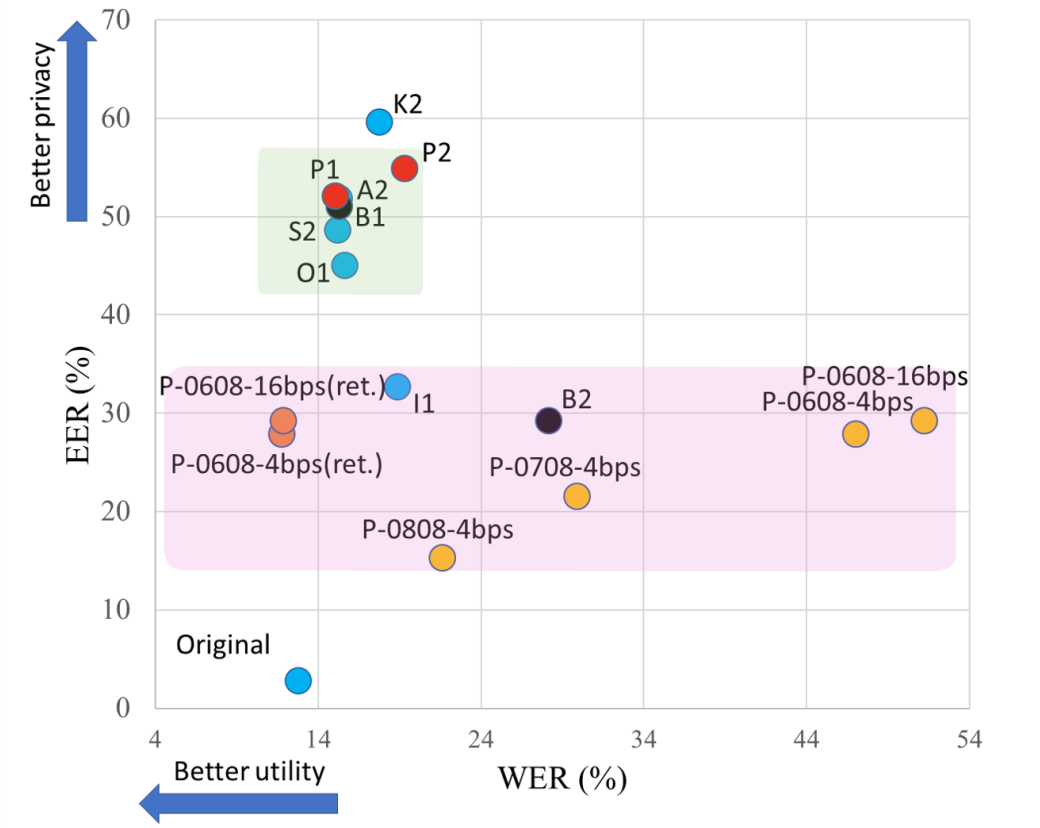


Figure 6.14: Mean WER versus mean EER over **VCTK dataset in (o-a)** scenario obtained from various systems proposed in Voice Privacy Challenge 2020. Black dot refers to results obtained by the baseline system. Red dot refers to results obtained by methods in [76]. Yellow and orange dots refer to results obtained by proposed methods. Blue dot refers to results obtained by other systems proposed in Voice Privacy Challenge 2020. The dots in the green shaded area are methods based on a neural vocoder (mainly based on the primary baseline framework). The dots in the pink shaded area are methods based on the LP vocoder.

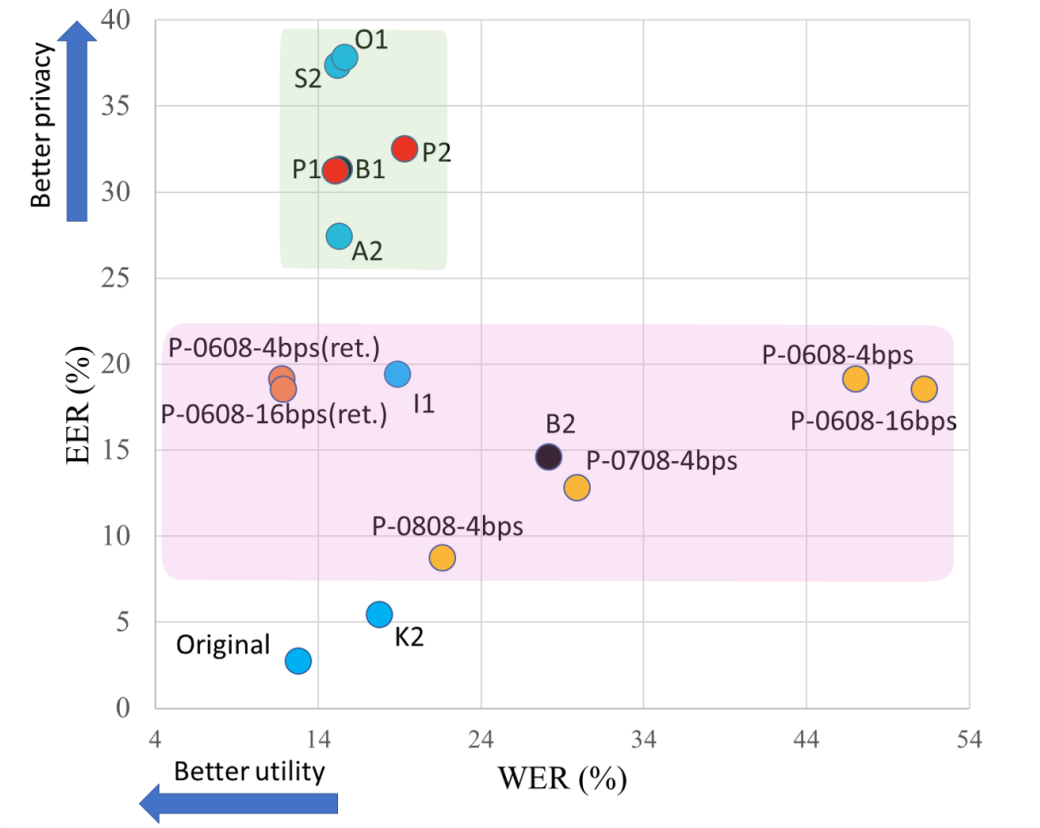


Figure 6.15: Mean WER versus mean EER over all **VCTK dataset in (a-a)** scenario obtained from various systems proposed in Voice Privacy Challenge 2020. Black dot refers to results obtained by the baseline system. Red dot refers to results obtained by methods in [76]. Yellow and orange dots refer to results obtained by proposed methods. Blue dot refers to results obtained by other systems proposed in Voice Privacy Challenge 2020. The dots in the green shaded area are methods based on a neural vocoder (mainly based on the primary baseline framework). The dots in the pink shaded area are methods based on the LP vocoder.

methods could be more effective than signal-processing methods proposed by other participants. However, we predicted a potential bias from the use of x-vectors in the objective evaluation by the ASV system. Even if there is a slightly better performance using any x-vector modification technique, the output is greatly affected by the NSF model. We determined this through our subjective evaluation, which showed very similar results between the B1 and the P1 (as shown in Chapter 5).

Regarding evaluation limitations, it could be argued that the current evaluations remain insufficient for assessing a speaker anonymization system. Although the metrics used in the current evaluations could give useful information, there are many critical points that are not captured by those metrics. Thus, it is quite difficult to conduct comparative studies solely using those metrics because the results are inconsistent. For instance, the objective evaluation results obtained by a system proposed in [45] are the opposite of a subjective evaluation in terms of privacy metrics. Furthermore, the difficulty in assessing the quality of a speaker anonymization system should be considered. For example, the degradation of anonymized speech quality could cause improve the EER despite the poor performance of the anonymization task. Therefore, we attempted to provide a considerably more consistent and reliable subjective evaluation. However, there are limitations, especially regarding the attack models.

Chapter 7

Conclusion

This chapter contains the summary and highlights of this research contributions. Finally, the remaining works and ideas for improving this research in the future are also described.

7.1 Summary

This study addressed an information hiding approach to preserve both security and privacy simultaneously in speech communication. We proposed a framework to secure speech communication. The proposed framework integrates the speech information hiding approach to secure the speaker anonymization, which mainly consists of two main parts, i.e., encoder and decoder. The encoder is mainly aimed to protect the speaker's identity by using an anonymization approach. In contrast to the other works, the anonymization is conducted with a parameter that will be used to represent watermarks. The result of anonymized speech should be able to conceal the sensitive personal information in speech while maintaining the naturalness and intelligibility of the speech. Meanwhile, the decoder is aimed to protect the authentication of the speech by accurately detecting the embedded watermarks.

This study investigated the analysis and synthesis model from both the conventional approach and the one using neural networks to propose robust methods for protecting speech content and privacy. As a study case for the conventional approach, we utilized the conventional CELP codecs based on linear predictive coding (LPC). Meanwhile, as a study case for the analysis and synthesis based on neural networks, we utilized the neural vocoder, which is based on a neural source-filter (NSF) model.

The first proposed method utilized features related to formant frequencies in conventional speech codecs. From the LPC analysis, the direct-form of linear

predictor coefficients (LPCs) and the residual signal were obtained. LPCs are often derived into line spectrum pairs (LSPs) or line spectral frequencies (LSFs) for robust representation in the quantization of the excitation codebook. LSFs are highly related to formant frequencies. Consequently, the modification of LSFs is promising to simultaneously affects the content-related and speaker-related information. We proposed SIH methods by direct modification on LSFs quantization bits and by modifying the McAdams coefficient.

The second proposed method utilizes the state-of-the-art speaker individuality feature in speaker recognition systems, namely, x-vector [107]. X-vector is derived from an identity vector (i-vector) modeling approach with speaker embedding. We modify the x-vector using singular vector decomposition (SVD) to anonymize speaker identity. Besides, we also improve the proposed method by modifying speech prosodies, such as fundamental frequency and speech duration.

Finally, the proposed framework is evaluated to ensure reliability and robustness by general datasets and protocols established in the Information Hiding Criteria (IHC) [51] and the Voice Privacy Challenge 2020 [112]. In these challenges, the full universal dataset from many speakers, various sources, and recording environments were used to verify the performance of the proposed framework.

7.2 Contributions

The main contribution of this study is to establish a framework for protecting speech security and voice privacy in digital speech communication systems. Privacy-preserving technology is required to protect the speech content and personal profiles of the speaker. In this study, the framework for preserving speech security and privacy using an information hiding approach has been proposed. Besides, speech perception and production systems are considered to solve the essential problems in existing methods, especially in promoting naturalness and intelligibility. Besides, this study also contributes as an alternative for detecting speech tampering and spoofing countermeasure.

7.3 Future Work

The SIH framework for protecting speech content and privacy in digital speech communication has several rooms for improvement, as follows:

1. The proposed framework mainly applied to the conventional and neural vocoders has shown a promising improvement in security and robustness

compared to other existing methods. However, there is still a limitation, especially when using a neural vocoder, that the results of anonymization performance (as shown in Chapter 5) are not much different than the baseline proposed in Voice Privacy Challenge 2020. This is mainly due to the main scheme of NSF that is highly dependent on the pre-trained models in each step. We would like to investigate speech analysis and synthesis methods that better suit x-vector singular value modification in future work.

2. Secondly, the method of speaker anonymization based on x-vector singular value modification is possible to be improved by the SIH method that is also based on singular value decomposition (SVD). SVD could capture the eigenstructure and result in a better representation of intra-speaker information. By controlling the less significant eigenstructure, we could provide better protection to speech signals with additional watermarks.

3. Besides features related to formant frequencies, we will investigate other prominent features that are often utilized in speech codecs to improve the performance of SIH and anonymization, including the robustness, capacity, and intelligibility. Subsequently, we will develop a method based on the proposed framework to deal with tampering and spoofing countermeasure problems.

4. Furthermore, we plan to improve our evaluation methods, especially the subjective evaluation, considering the attack models described in the Voice Privacy Challenge 2020.

Bibliography

[1] Mohamed Abou-Zleikha, Zheng-Hua Tan, Mads Græsbøll Christensen, and Søren Holdt Jensen. A discriminative approach for speaker selection in speaker de-identification systems. In *23rd European Signal Processing Conference, EUSIPCO 2015, Nice, France*, pages 2102–2106. IEEE, 2015.

[2] Manu Airaksinen, Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi, and Paavo Alku. A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1658–1670, 2018.

[3] Masato Akagi and Taro Ienaga. Speaker individuality in fundamental frequency contours and its control. *Journal of the Acoustical Society of Japan (E)*, 18(2):73–80, 1997.

[4] G. Akers and Matthew Lennig. Intonation in text-to-speech synthesis: Evaluation of algorithms. *Journal of the Acoustical Society of America*, 77:2157–2165, 1985.

[5] Sabur Alim and Nahrul Khair Alang Md Rashid. Some commonly used speech feature extraction algorithms, 12 2018.

[6] Sercan Ömer Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. *CoRR*, abs/1802.06006, 2018.

[7] Michael Arnold, Martin Schmucker, and Stephen D. Wolthusen. *Techniques and Applications of Digital Watermarking and Content Protection*. Artech House, Inc., USA, 2003.

[8] Bishnu S. Atal and Joel R. Remde. A new model of LPC excitation for producing natural-sounding speech at low bit rates. In *IEEE Interna-*

tional Conference on Acoustics, Speech, and Signal Processing, ICASSP '82, Paris, France, May 3-5, 1982, pages 614–617. IEEE, 1982.

[9] Walter Bender, Daniel F. Gruhl, Norishige Morimoto, and Anthony Lu. Techniques for data hiding. *IBM Syst. J.*, 35:313–336, 1996.

[10] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10):763–786, 2007. Intrinsic Speech Variations.

[11] Léon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7, [NIPS Conference, Denver, Colorado, USA, 1994]*, pages 585–592. MIT Press, 1994.

[12] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[13] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[14] Niko Brümmer and Johan A. du Preez. Application-independent evaluation of speaker detection. *Comput. Speech Lang.*, 20(2-3):230–275, 2006.

[15] Francis Buttle. Servqual: review, critique, research agenda. *European Journal of Marketing*, 30:8–32, 01 1996.

[16] João P. Cabral, Korin Richmond, Junichi Yamagishi, and Steve Renals. Glottal spectral separation for speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):195–208, 2014.

[17] Joseph P. Campbell, Thomas E. Tremain, and Vanoy C. Welch. The federal standard 1016 4800 bps celp voice coder. *Digit. Signal Process.*, 1:145–155, 1991.

[18] Pierre Champion, Denis Jouvet, and Anthony Larcher. A Study of F0 Modification for X-Vector Based Speech Pseudonymization Across Gender. In *PPAI 2021 - The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence*, Virtual, China, February 2021.

[19] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 1086–1090. ISCA, 2018.

[20] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. *Digital Watermarking and Steganography*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2 edition, 2007.

[21] Nedeljko Cvejic, Nedeljko Cvejic, and Tapio Seppanen. *Digital Audio Watermarking Techniques and Technologies: Applications and Benchmarks*. IGI Global, USA, 2007.

[22] Nedeljko Cvejic and Tapio Seppänen. Increasing robustness of lsb audio steganography by reduced distortion lsb coding. *J. Univers. Comput. Sci.*, 11:56–65, 2005.

[23] Rohan Kumar Das, Bidisha Sharma, and S. R. Mahadeva Prasanna. Significance of duration modification for speaker verification under mismatch speech tempo condition. *International Journal of Speech Technology*, 21(3):401–408, 2018.

[24] Gilles Degottex and Daniel Erro. A uniform phase representation for the harmonic model in speech synthesis applications. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014:1–16, 01 2014.

[25] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[26] Najim Dehak, Pedro Torres-Carrasquillo, Douglas Reynolds, and R. Dehak. Language recognition via i-vectors and dimensionality reduction. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 857–860, 01 2011.

[27] Volker Dellwo, Mark A. Huckvale, and Michael Ashby. How is individuality expressed in voice? an introduction to speech production and description for speaker classification. In *Speaker Classification I: Fundamentals, Features, and Methods*, volume 4343 of *Lecture Notes in Computer Science*, pages 1–20. Springer, 2007.

[28] Li Deng, Dong Yu, and A. Acero. Structured speech modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1492–1504, 2006.

[29] Charles Dodge and Thomas A. Jerse. Computer music: Synthesis, composition, and performance, 1997.

[30] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis, 2019.

[31] S. Pavankumar Dubagunta, Rob J. J. H. van Son, and Mathew Magimai. Adjustable Deterministic Pseudonymisation of Speech: Idiap-NKI's submission to VoicePrivacy 2020 Challenge, 2020.

[32] III Eli P. Cox. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research*, 17(4):407–422, 1980.

[33] Daniel Erro, Iñaki Sainz, Eva Navas, and Inma Hernaez. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 8(2):184–194, 2014.

[34] Fernando M. Espinoza-Cuadros, Juan M. Perero-Codosero, Javier Antón-Martín, and Luis A. Hernández Gómez. Speaker de-identification system using autoencoders and adversarial training. *CoRR*, abs/2011.04696, 2020.

[35] Fuming Fang, Xin Wang, Junichi Yamagishi, Isao Echizen, Massimiliano Todisco, Nicholas W. D. Evans, and Jean-François Bonastre. Speaker anonymization using x-vector and neural waveform models. *CoRR*, abs/1905.13561, 2019.

[36] Gunnar Fant. *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*. De Gruyter Mouton, 2012.

[37] Marcos Faundez-Zanuy, Martin Hagmüller, and Gernot Kubin. Speaker verification security improvement by means of speech watermarking. *Speech Commun.*, 48(12):1608–1619, December 2006.

[38] Hiroya Fujisaki. Prosody, models, and spontaneous speech. In *Computing Prosody*, pages 27–42, 1997.

[39] Hiroya Fujisaki. Information, prosody, and modeling - with emphasis on tonal features of speech -. In *Proc. Speech Prosody 2004*, pages 1–10, 2004.

[40] G. H. Golub and C. Reinsch. Singular Value Decomposition and Least Squares Solutions. *Numer. Math.*, 14(5):403–420, April 1970.

[41] Ian J. Goodfellow, Yoshua Bengio, and Aaron C. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016.

[42] A. Gray and J. Markel. Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(5):380–391, 1976.

[43] Daniel Gruhl, Anthony Lu, and Walter Bender. Echo hiding. In Ross Anderson, editor, *Information Hiding*, pages 295–315, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.

[44] Carlos Gussenhoven, B Repp, A Rietveld, H Rump, and Jacques Terken. The perceptual prominence of fundamental frequency peaks. *The Journal of the Acoustical Society of America*, 102:3009–22, 12 1997.

[45] Yaowei Han, Yang Cao, Sheng Li, Qiang Ma, and Masatoshi Yoshikawa. Voice-indistinguishability - protecting voiceprint with differential privacy under an untrusted server. In *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA*, pages 2125–2127. ACM, 2020.

[46] Rosa González Hautamäki and Tomi Kinnunen. Why did the x-vector system miss a target speaker? impact of acoustic mismatch upon target score on voxceleb data. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China*, pages 4313–4317. ISCA, 2020.

[47] Keikichi Hirose and Hiromichi Kawanami. Temporal rate change of dialogue speech in prosodic units as compared to read speech. *Speech Communication*, 36(1-2):97–111, 2002.

[48] Ching-Tang Hsieh and Pei-Ying Sou. Blind cepstrum domain audio watermarking based on time energy features. *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No.02TH8628)*, 2:705–708 vol.2, 2002.

[49] Guang Hua, Jiwu Huang, Yun Q. Shi, Jonathan Goh, and Vrizlynn L. L. Thing. Twenty years of digital audio watermarking - a comprehensive review. *Signal Processing*, 128:222–242, 2016.

[50] Won Hwang, Hwan Kang, Seung-Soo Han, Kab Kim, and Hwan Kang. Robust audio watermarking using both dwt and masking effect. In *IWDW*, volume 2939, pages 382–389, 10 2003.

[51] IHC Committee. IHC evaluation criteria and competition. URL: `https://www.ieice.org/iss/emm/ihc/IHC_criteriaVer6.pdf`, visited on 2021-06-08.

[52] N.S. Jayant. Analog scramblers for speech privacy. *Computers & Security*, 1(3):275–289, 1982.

[53] Qin Jin, Arthur R Toth, Alan W Black, and Tanja Schultz. Is voice transformation a threat to speaker identification? In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4845–4848. IEEE, 2008.

[54] Qin Jin, Arthur R Toth, Tanja Schultz, and Alan W Black. Speaker de-identification via voice transformation. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 529–533. IEEE, 2009.

[55] P. Kabal. Ill-conditioning and bandwidth expansion in linear prediction of speech. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I, 2003.

[56] Jessada Karnjana, Masashi Unoki, Pakinee Aimmanee, and Chai Wutiwiwatchai. Tampering detection in speech signals by semi-fragile watermarking based on singular-spectrum analysis. In *Advances in Intelligent Information Hiding and Multimedia Signal Processing*, pages 131–140, Cham, 2017. Springer International Publishing.

[57] Hideki Kawahara, Jo Estill, and Osamu Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *MAVEBA*, 2001.

[58] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible

role of a repetitive structure in sounds. *Speech Commun.*, 27:187–207, 1999.

[59] Angelos D. Keromytis. A comprehensive survey of voice over ip security research. *IEEE Communications Surveys & Tutorials*, 14:514–537, 2012.

[60] Hyoung Kim and Choi yong hee. A novel echo-hiding scheme with backward and forward kernels. *Circuits and Systems for Video Technology, IEEE Transactions on*, 13:885 – 889, 09 2003.

[61] Darko Kirovski and Henrique S. Malvar. Robust spread-spectrum audio watermarking. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*, pages 1345–1348. IEEE, 2001.

[62] W. B. Kleijn and K. K. Paliwal. *Speech Coding and Synthesis*. Elsevier Science Inc., USA, 1995.

[63] Chin-Su Ko, Ki-Young Kim, Rim-Wo Hwang, Youngseop Kim, and Sang-Burm Rhee. Robust audio watermarking in wavelet domain using pseudorandom sequences. In *4th Annual ACIS International Conference on Computer and Information Science (ICIS 2005), 14-16 July 2005, Jeju Island, South Korea*, pages 397 – 401, 02 2005.

[64] P. Kumsawat, K. Attakitmongcol, and A. Srikaew. Digital audio watermarking for copyright protection based on multiwavelet transform. In *EuroISI*, 2008.

[65] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. Atr japanese speech database as a tool of speech recognition and synthesis, 1990.

[66] Lin-Shan Lee and Ger-Chih Chou. A general theory for asynchronous speech encryption techniques. *IEEE Journal on Selected Areas in Communications*, 4(2):280–287, 1986.

[67] K. Li, Y.C. Soh, and Z.G. Li. Chaotic cryptosystem with high sensitivity to parameter mismatch. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 50(4):579–583, 2003.

[68] Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, and Thomas Fang Zheng. Deep factorization for speech signal. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing,*

ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018, pages 5094–5098. IEEE, 2018.

[69] Xin Li and Hong Heather Yu. Transparent and robust audio data hiding in cepstrum domain. In *2000 IEEE International Conference on Multimedia and Expo, ICME 2000, New York, NY, USA, July 30 - August 2, 2000*, page 397. IEEE Computer Society, 2000.

[70] Yiqing Lin and W. Abdulla. *Audio Watermark: A Comprehensive Foundation Using MATLAB*. Springer, 2014.

[71] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. MOSNet: Deep learning based objective assessment for voice conversion, 2019.

[72] Carmen Magariños, Paula Lopez-Otero, Laura Docio-Fernandez, Eduardo Rodriguez-Banga, Daniel Erro, and Carmen Garcia-Mateo. Reversible speaker de-identification using pre-trained transformation functions. *Computer Speech & Language*, 46:36–52, 2017.

[73] Hafiz Malik, Ashfaq A. Khokhar, and Rashid Ansari. Robust audio watermarking using frequency selective spread spectrum theory. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2004, Montreal, Quebec, Canada, May 17-21, 2004*, pages 385–388. IEEE, 2004.

[74] Candy Olivia Mawalim, Kasorn Galajit, Jessada Karnjana, and Masashi Unoki. X-vector singular value modification and statistical-based decomposition with ensemble regression modeling for speaker anonymization system. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China*, pages 1703–1707. ISCA, 2020.

[75] Candy Olivia Mawalim and Masashi Unoki. Improving security in mcadams coefficient-based speaker anonymization by watermarking method. *CoRR*, abs/2107.07223, 2021.

[76] Candy Olivia Mawalim, Shengbei Wang, and Masashi Unoki. Speech information hiding by modification of LSF quantization index in CELP codec. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2020, Auckland, New Zealand, December 7-10, 2020*, pages 1321–1330. IEEE, 2020.

[77] Stephen Mcadams. Spectral fusion, spectral parsing and the formation of auditory images. *Ph. D. Thesis, Stanford*, 1984.

[78] Ian Vince McLoughlin. Review: Line spectral pairs. *Signal Process.*, 88(3):448–467, March 2008.

[79] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99.D:1877–1884, 07 2016.

[80] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. VoxCeleb: A Large-Scale Speaker Identification Dataset. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 2616–2620. ISCA, 2017.

[81] Yuta Nakashima, Ryuki Tachibana, and Noboru Babaguchi. Watermarked movie soundtrack finds the position of the camcorder in a theater. *Trans. Multi.*, 11(3):443–454, April 2009.

[82] Ryouichi Nishimura. Audio watermarking using spatial masking and ambisonics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:2461–2469, 2012.

[83] Ryouichi Nishimura and Yoiti Suzuki. Audio watermark based on periodical phase shift. *The Journal of the Acoustical Society of Japan*, 60:268–272, 05 2004.

[84] Ryouichi Nishimura, Yôiti Suzuki, and B.-S Ko. *Advanced Audio Watermarking Based on Echo Hiding: Time-Spread Echo Hiding*, pages 123–151. IGI Global, 01 2007.

[85] Jonas Obleser and Frank Eisner. Pre-lexical abstraction of speech in the auditory cortex. *Trends in Cognitive Sciences*, 13(1):14–19, 2009.

[86] Tokunbo Ogunfunmi and Madihally J. Narasimha. Speech over voip networks: Advanced signal processing and system implementation. *IEEE Circuits and Systems Magazine*, 12(2):35–55, 2012.

[87] Hyen-O Oh, Jong Won Seok, Jin Woo Hong, and Dae Hee Youn. New echo embedding technique for robust and imperceptible audio watermarking. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace*

Convention Center, Salt Lake City, Utah, USA, Proceedings, pages 1341–1344. IEEE, 2001.

[88] Scott Wilkinson Ozge Koymen, John Morton. Digital speech encryption, compression, and transmission, May 1996.

[89] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia*, pages 5206–5210. IEEE, 2015.

[90] Jose Patino, Natalia A. Tomashenko, Massimiliano Todisco, Andreas Nautsch, and Nicholas W. D. Evans. Speaker anonymisation using the McAdams coefficient. *CoRR*, abs/2011.01130, 2020.

[91] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. A time delay neural network architecture for efficient modeling of long temporal contexts. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany*, pages 3214–3218. ISCA, 2015.

[92] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[93] Miran Pobar and Ivo Ipsic. Online speaker de-identification using voice transformation. In *37th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2014, Opatija, Croatia*, pages 1264–1267. IEEE, 2014.

[94] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hawaii, US*. IEEE Signal Processing Society, December 2011.

[95] Tuomo Raitio, Antti Suni, Junichi Yamagishi, Hannu Pulakka, Jani Nurminen, Martti Vainio, and Paavo Alku. Hmm-based speech syn-

thesis utilizing glottal inverse filtering. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19:153 – 165, 02 2011.

[96] Karthikeyan Natesan Ramamurthy and Andreas Spanias. *MATLAB Software for the Code Excited Linear Prediction Algorithm: The Federal Standard-1016*. Synthesis Lectures on Algorithms and Software in Engineering. Morgan & Claypool Publishers, 2010.

[97] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*, pages 749–752. IEEE, 2001.

[98] Eng. Sattar B. Sadkhan and Nidaa Abbas. Speech scrambling based on wavelet transform, 04 2012.

[99] Md. Sahidullah and Tomi Kinnunen. Local spectral variability features for speaker verification. *Digit. Signal Process.*, 50:1–11, 2016.

[100] S. Saito, T. Furukawa, and K. Konishi. A digital watermarking for audio data using band division based on qmf bank. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (CASSP'02)*, volume 4, pages IV–3473–IV–3476, Los Alamitos, CA, USA, may 2002. IEEE Computer Society.

[101] Redwan Salami, Claude Laflamme, Jean-Pierre Adoul, Akitoshi Kataoka, Shinji Hayashi, Takehiro Moriya, Claude Lamblin, Dominique Massaloux, Stéphane Proust, Peter Kroon, and Yair Shoham. Design and description of CS-ACELP: a toll quality 8 kb/s speech coder. *IEEE Trans. Speech Audio Process.*, 6(2):116–130, 1998.

[102] Manfred R. Schroeder and Bishnu S. Atal. Code-excited linear prediction(celp): High-quality speech at very low bit rates. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '85, Tampa, Florida, USA, March 26-29, 1985*, pages 937–940. IEEE, 1985.

[103] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:132–157, January 2021.

[104] D. Snyder, D. Garcia-Romero, and D. Povey. Time delay deep neural network-based universal background models for speaker recognition. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 92–97, 2015.

[105] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018.

[106] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification, 08 2017.

[107] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-Vectors: Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada*, pages 5329–5333. IEEE, 2018.

[108] F. Soong and B. Juang. Line spectrum pair (lsp) and speech data compression. In *ICASSP '84. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 9, pages 37–40, 1984.

[109] N. Sugamura and F. Itakura. Speech data compression by lsp analysis/synthesis technique. *Trans. IEICE.*, J64, 1981.

[110] Haoran Sun, Lantian Li, Yunqi Cai, Yang Zhang, Thomas Fang Zheng, and Dong Wang. Deep generative factorization for speech signal. *CoRR*, abs/2010.14242, 2020.

[111] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas W. D. Evans, Tomi H. Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 1008–1012. ISCA, 2019.

[112] Natalia Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. The VoicePrivacy 2020 Challenge evaluation plan, 2020. URL: `https://www.voiceprivacychallenge.org/docs/VoicePrivacy_2020_Eval_Plan_v1_3.pdf`, visited on 2021-06-10.

[113] Natalia A. Tomashenko, Brij Mohan Lal Srivastava, Xin Wang, Emmanuel Vincent, Andreas Nautsch, Junichi Yamagishi, Nicholas W. D. Evans, Jose Patino, Jean-François Bonastre, Paul-Gauthier Noé, and Massimiliano Todisco. Introducing the VoicePrivacy Initiative. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China*, pages 1693–1697. ISCA, 2020.

[114] Hartmut Traunmüller and Anders Eriksson. The frequency range of the voice fundamental in the speech of male and female adults, 01 1995.

[115] Henry Turner, Giulio Lovisotto, and Ivan Martinovic. Speaker Anonymization with Distribution-Preserving X-Vector Generation for the VoicePrivacy Challenge 2020. *CoRR*, abs/2010.13457, 2020.

[116] Masashi Unoki and Daiki Hamada. Method of digital-audio watermarking based on cochlear delay characteristics. *International Journal of Innovative Computing, Information and Control*, 6, 03 2010.

[117] Masashi Unoki and Ryota Miyauchi. Detection of tampering in speech signals with inaudible watermarking technique. In *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 118–121, 2012.

[118] Masashi Unoki and Ryota Miyauchi. Robust, blindly-detectable, and semi-reversible technique of audio watermarking based on cochlear delay characteristics. *IEICE Trans. Inf. Syst.*, 98-D:38–48, 2015.

[119] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *ArXiv*, abs/1609.03499, 2016.

[120] Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit, 2017.

[121] Katharina von Kriegstein, David R. R. Smith, Roy D. Patterson, Stefan J. Kiebel, and Timothy D. Griffiths. How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience*, 30(2):629–638, 2010.

[122] Ted S. Wada, Biing-Hwang Juang, and Rafid A. Sukkar. Measurement of the effects of nonlinearities on the network-based linear acoustic

echo cancellation. In *2006 14th European Signal Processing Conference*, pages 1–5, 2006.

[123] Shengbei Wang and Masashi Unoki. Watermarking method for speech signals based on modifications to lsfs. *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 283–286, 2013.

[124] Shengbei Wang and Masashi Unoki. Watermarking of speech signals based on formant enhancement. In *22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014*, pages 1257–1261. IEEE, 2014.

[125] Shengbei Wang, Weitao Yuan, Jianming Wang, and Masashi Unoki. Speech watermarking based on robust principal component analysis and formant manipulations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 2082–2086. IEEE, 2018.

[126] Shengbei Wang, Weitao Yuan, Jianming Wang, and Masashi Unoki. Detection of speech tampering using sparse representations and spectral manipulations based information hiding. *Speech Communication*, 112:1–14, 2019.

[127] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter-based waveform model for statistical parametric speech synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom*, pages 5916–5920. IEEE, 2019.

[128] Chung-Ping Wu and C.-C. Jay Kuo. Fragile speech watermarking based on exponential scale quantization for tamper detection. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages IV–3305–IV–3308, 2002.

[129] Zhijun Wu. *Information Hiding in Speech Signals for Secure Communication*. Syngress Publishing, 1st edition, 2014.

[130] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification. *Speech Commun.*, 66(C):130–153, February 2015.

[131] Zhizheng Wu and Haizhou Li. Voice conversion and spoofing attack on speaker verification systems. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9, 2013.

[132] Yoichi Yamashita. A review of paralinguistic information processing for natural speech communication. *Acoustical Science and Technology*, 34:73–79, 03 2013.

[133] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria*, pages 1526–1530. ISCA, 2019.

Appendix A – Speaker Anonymization Evaluation

A.1 – Primary Baseline (B1)

Table 1: ASV results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	EER, %	C_{llr}^{min}	C_{llr}	Enroll	Trial	Gen	Test set	EER, %	C_{llr}^{min}	C_{llr}
1	libri_dev	8.665	0.305	42.930	o	o	f	libri_test	7.664	0.184	26.799
2	libri_dev	50.140	0.996	144.312	o	a	f	libri_test	47.260	0.995	151.803
3	libri_dev	36.510	0.889	16.238	a	a	f	libri_test	31.390	0.832	16.157
4	libri_dev	1.242	0.035	14.275	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	57.760	0.999	168.576	o	a	m	libri_test	52.560	0.999	166.873
6	libri_dev	34.010	0.867	24.644	a	a	m	libri_test	35.410	0.898	33.938
7	vctk_dev_com	2.616	0.089	0.874	o	o	f	vctk_test_com	2.890	0.092	0.858
8	vctk_dev_com	49.420	0.995	171.902	o	a	f	vctk_test_com	47.980	0.994	161.972
9	vctk_dev_com	27.330	0.741	7.273	a	a	f	vctk_test_com	31.790	0.840	9.291
10	vctk_dev_com	1.425	0.051	1.565	o	o	m	vctk_test_com	1.130	0.035	1.029
11	vctk_dev_com	55.560	0.999	193.242	o	a	m	vctk_test_com	52.820	1.000	190.361
12	vctk_dev_com	32.480	0.843	23.793	a	a	m	vctk_test_com	31.640	0.833	21.642
13	vctk_dev_dif	2.920	0.102	1.152	o	o	f	vctk_test_dif	4.990	0.170	1.501
14	vctk_dev_dif	50.140	0.988	165.941	o	a	f	vctk_test_dif	48.510	0.998	146.746
15	vctk_dev_dif	26.450	0.760	8.480	a	a	f	vctk_test_dif	32.250	0.849	11.377
16	vctk_dev_dif	1.439	0.052	1.164	o	o	m	vctk_test_dif	2.067	0.071	1.819
17	vctk_dev_dif	53.950	1.000	167.862	o	a	m	vctk_test_dif	53.730	1.000	167.741
18	vctk_dev_dif	31.360	0.843	24.112	a	a	m	vctk_test_dif	31.230	0.842	23.781

Table 2: ASR results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	WER, %		Data	Test set	WER, %	
		LM_s	LM_l			LM_s	LM_l
1	libri_dev	5.25	3.82	o	libri_test	5.55	4.15
2	libri_dev	8.82	6.37	a	libri_test	9.12	6.65
3	vctk_dev	14.04	10.79	o	vctk_test	16.39	12.81
4	vctk_dev	18.89	15.31	a	vctk_test	18.90	15.24

A.2 – Secondary Baseline (B2)

Table 3: ASV results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	EER, %	C_{llr}^{min}	C_{llr}	Enroll	Trial	Gen	Test set	EER, %	C_{llr}^{min}	C_{llr}
1	libri_dev	8.665	0.305	42.930	o	o	f	libri_test	7.664	0.184	26.799
2	libri_dev	8.665	0.305	42.930	o	a	f	libri_test	26.090	0.686	115.617
3	libri_dev	32.670	0.840	108.906	a	a	f	libri_test	14.960	0.490	12.539
4	libri_dev	1.242	0.035	14.275	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	18.010	0.527	105.706	o	a	m	libri_test	17.820	0.501	106.594
6	libri_dev	10.560	0.359	11.920	a	a	m	libri_test	8.241	0.262	15.375
7	vctk_dev_com	2.616	0.089	0.874	o	o	f	vctk_test_com	2.890	0.092	0.858
8	vctk_dev_com	34.010	0.880	85.953	o	a	f	vctk_test_com	30.350	0.807	93.854
9	vctk_dev_com	11.630	0.367	43.531	a	a	f	vctk_test_com	14.160	0.464	42.744
10	vctk_dev_com	1.425	0.051	1.565	o	o	m	vctk_test_com	1.130	0.035	1.029
11	vctk_dev_com	23.930	0.671	90.882	o	a	m	vctk_test_com	24.580	0.716	99.328
12	vctk_dev_com	10.540	0.319	24.985	a	a	m	vctk_test_com	11.860	0.353	28.205
13	vctk_dev_dif	2.920	0.102	1.152	o	o	f	vctk_test_dif	4.990	0.170	1.501
14	vctk_dev_dif	35.540	0.908	90.635	o	a	f	vctk_test_dif	29.990	0.795	93.136
15	vctk_dev_dif	15.780	0.504	39.850	a	a	f	vctk_test_dif	16.920	0.546	41.371
16	vctk_dev_dif	1.439	0.052	1.164	o	o	m	vctk_test_dif	2.067	0.071	1.819
17	vctk_dev_dif	28.290	0.743	98.595	o	a	m	vctk_test_dif	28.190	0.721	101.692
18	vctk_dev_dif	11.170	0.383	23.060	a	a	m	vctk_test_dif	12.230	0.398	25.125

Table 4: ASR results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	WER, %		Data	Test set	WER, %	
		LM_s	LM_l			LM_s	LM_l
1	libri_dev	5.25	3.82	o	libri_test	5.55	4.15
2	libri_dev	8.22	5.95	a	libri_test	11.75	8.93
3	vctk_dev	14.04	10.79	o	vctk_test	16.39	12.81
4	vctk_dev	30.06	25.52	a	vctk_test	33.26	28.17

A.3 – Proposed Methods

A.3.1 P-0608 with 4-bps embedding payload

Table 5: ASV results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	EER, %	C_{llr}^{min}	C_{llr}	Enroll	Trial	Gen	Test set	EER, %	C_{llr}^{min}	C_{llr}
1	libri_dev	8.665	0.305	42.930	o	o	f	libri_test	7.664	0.184	26.799
2	libri_dev	38.640	0.835	117.674	o	a	f	libri_test	26.090	0.663	110.969
3	libri_dev	26.850	0.705	15.622	a	a	f	libri_test	18.250	0.534	14.741
4	libri_dev	1.242	0.035	14.275	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	20.960	0.575	103.135	o	a	m	libri_test	18.930	0.517	103.554
6	libri_dev	13.510	0.431	9.314	a	a	m	libri_test	13.140	0.412	13.054
7	vctk_dev_com	2.616	0.089	0.874	o	o	f	vctk_test_com	2.890	0.092	0.858
8	vctk_dev_com	30.520	0.816	84.261	o	a	f	vctk_test_com	29.190	0.802	88.586
9	vctk_dev_com	14.240	0.466	39.201	a	a	f	vctk_test_com	18.500	0.562	36.712
10	vctk_dev_com	1.425	0.051	1.565	o	o	m	vctk_test_com	1.130	0.035	1.029
11	vctk_dev_com	25.640	0.702	91.400	o	a	m	vctk_test_com	24.010	0.691	90.957
12	vctk_dev_com	11.970	0.413	21.099	a	a	m	vctk_test_com	10.450	0.374	21.652
13	vctk_dev_dif	2.920	0.102	1.152	o	o	f	vctk_test_dif	4.990	0.170	1.501
14	vctk_dev_dif	32.400	0.855	90.458	o	a	f	vctk_test_dif	26.340	0.753	81.154
15	vctk_dev_dif	21.170	0.646	26.852	a	a	f	vctk_test_dif	22.530	0.680	27.273
16	vctk_dev_dif	1.439	0.052	1.164	o	o	m	vctk_test_dif	2.067	0.071	1.819
17	vctk_dev_dif	27.590	0.754	95.502	o	a	m	vctk_test_dif	24.860	0.713	89.943
18	vctk_dev_dif	17.220	0.534	14.107	a	a	m	vctk_test_dif	15.440	0.501	13.896

Table 6: ASR results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	WER, %		Data	Test set	WER, %	
		LM_s	LM_l			LM_s	LM_l
1	libri_dev	5.25	3.82	o	libri_test	5.55	4.15
2	libri_dev	31.49	26.47	a	libri_test	29.98	25.48
3	vctk_dev	14.04	10.79	o	vctk_test	16.39	12.81
4	vctk_dev	49.44	45.33	a	vctk_test	52.85	48.76

A.3.2 P-0708 with 4-bps embedding payload

Table 7: ASR results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	WER, %		Data	Test set	WER, %	
		LM_s	LM_l			LM_s	LM_l
1	libri_dev	15.92	11.95	a	libri_test	14.89	11.42
2	vctk_dev	33.09	28.38	a	vctk_test	36.57	31.49

Table 8: ASV results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	EER, %	C_{llr}^{min}	C_{llr}	Enroll	Trial	Gen	Test set	EER, %	C_{llr}^{min}	C_{llr}
1	libri_dev	8.665	0.305	42.930	o	o	f	libri_test	7.664	0.184	26.799
2	libri_dev	32.670	0.778	107.673	o	a	f	libri_test	21.170	0.551	96.312
3	libri_dev	25.140	0.649	9.507	a	a	f	libri_test	13.320	0.435	6.493
4	libri_dev	1.242	0.035	14.275	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	11.490	0.363	83.463	o	a	m	libri_test	10.690	0.339	85.415
6	libri_dev	8.696	0.304	3.065	a	a	m	libri_test	7.795	0.263	2.736
7	vctk_dev_com	2.616	0.089	0.874	o	o	f	vctk_test_com	2.890	0.092	0.858
8	vctk_dev_com	25.870	0.724	62.693	o	a	f	vctk_test_com	23.700	0.704	65.093
9	vctk_dev_com	9.593	0.333	24.118	a	a	f	vctk_test_com	11.850	0.406	22.352
10	vctk_dev_com	1.425	0.051	1.565	o	o	m	vctk_test_com	1.130	0.035	1.029
11	vctk_dev_com	18.520	0.546	67.914	o	a	m	vctk_test_com	19.490	0.593	69.656
12	vctk_dev_com	7.407	0.274	8.183	a	a	m	vctk_test_com	6.780	0.264	9.816
13	vctk_dev_dif	2.920	0.102	1.152	o	o	f	vctk_test_dif	4.990	0.170	1.501
14	vctk_dev_dif	26.560	0.749	72.751	o	a	f	vctk_test_dif	20.630	0.635	56.653
15	vctk_dev_dif	14.090	0.459	15.077	a	a	f	vctk_test_dif	16.560	0.532	16.813
16	vctk_dev_dif	1.439	0.052	1.164	o	o	m	vctk_test_dif	2.067	0.071	1.819
17	vctk_dev_dif	22.330	0.623	74.767	o	a	m	vctk_test_dif	16.530	0.515	66.357
18	vctk_dev_dif	10.270	0.341	5.980	a	a	m	vctk_test_dif	10.160	0.343	6.750

A.3.3 P-0808 (enhanced version of B2) with 4-bps embedding payload

Table 9: ASV results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	EER, %	C_{llr}^{min}	C_{llr}	Enroll	Trial	Gen	Test set	EER, %	C_{llr}^{min}	C_{llr}
1	libri_dev	8.665	0.305	42.930	o	o	f	libri_test	7.664	0.184	26.799
2	libri_dev	26.140	0.690	96.768	o	a	f	libri_test	15.330	0.441	81.382
3	libri_dev	21.020	0.570	10.075	a	a	f	libri_test	12.410	0.385	5.936
4	libri_dev	1.242	0.035	14.275	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	6.211	0.200	60.645	o	a	m	libri_test	4.454	0.164	64.308
6	libri_dev	5.280	0.196	2.083	a	a	m	libri_test	3.786	0.124	0.932
7	vctk_dev_com	2.616	0.089	0.874	o	o	f	vctk_test_com	2.890	0.092	0.858
8	vctk_dev_com	20.930	0.638	44.948	o	a	f	vctk_test_com	17.920	0.532	44.433
9	vctk_dev_com	7.558	0.260	16.426	a	a	f	vctk_test_com	7.803	0.281	14.287
10	vctk_dev_com	1.425	0.051	1.565	o	o	m	vctk_test_com	1.130	0.035	1.029
11	vctk_dev_com	10.830	0.351	44.467	o	a	m	vctk_test_com	15.250	0.463	48.008
12	vctk_dev_com	4.843	0.175	3.181	a	a	m	vctk_test_com	4.237	0.169	3.270
13	vctk_dev_dif	2.920	0.102	1.152	o	o	f	vctk_test_dif	4.990	0.170	1.501
14	vctk_dev_dif	18.420	0.584	55.530	o	a	f	vctk_test_dif	16.100	0.520	32.814
15	vctk_dev_dif	9.882	0.326	9.350	a	a	f	vctk_test_dif	12.190	0.399	11.307
16	vctk_dev_dif	1.439	0.052	1.164	o	o	m	vctk_test_dif	2.067	0.071	1.819
17	vctk_dev_dif	15.580	0.465	52.854	o	a	m	vctk_test_dif	10.510	0.350	42.699
18	vctk_dev_dif	6.203	0.213	2.727	a	a	m	vctk_test_dif	6.487	0.223	2.794

Table 10: ASR results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	WER, %		Data	Test set	WER, %	
		$\mathbf{LM}_s$	$\mathbf{LM}_l$			$\mathbf{LM}_s$	$\mathbf{LM}_l$
1	libri_dev	9.43	6.69	a	libri_test	9.48	7.00
2	vctk_dev	24.67	20.42	a	vctk_test	27.58	22.95

A.3.4 P-0608 with 16-bps embedding payload

Table 11: ASV results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	EER, %	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$	Enroll	Trial	Gen	Test set	EER, %	$\mathbf{C}_{llr}^{min}$	$\mathbf{C}_{llr}$
1	libri_dev	8.665	0.305	42.930	o	o	f	libri_test	7.664	0.184	26.799
2	libri_dev	39.910	0.866	124.824	o	a	f	libri_test	28.280	0.740	118.686
3	libri_dev	25.850	0.702	23.618	a	a	f	libri_test	19.160	0.571	25.027
4	libri_dev	1.242	0.035	14.275	o	o	m	libri_test	1.114	0.041	15.342
5	libri_dev	25.310	0.664	112.781	o	a	m	libri_test	20.270	0.577	111.402
6	libri_dev	13.040	0.429	18.250	a	a	m	libri_test	14.920	0.438	23.244
7	vctk_dev_com	2.616	0.089	0.874	o	o	f	vctk_test_com	2.890	0.092	0.858
8	vctk_dev_com	32.560	0.857	87.035	o	a	f	vctk_test_com	29.480	0.801	91.132
9	vctk_dev_com	12.790	0.440	52.581	a	a	f	vctk_test_com	18.790	0.579	48.899
10	vctk_dev_com	1.425	0.051	1.565	o	o	m	vctk_test_com	1.130	0.035	1.029
11	vctk_dev_com	26.210	0.693	96.002	o	a	m	vctk_test_com	25.990	0.718	94.127
12	vctk_dev_com	10.830	0.380	36.521	a	a	m	vctk_test_com	10.730	0.366	34.889
13	vctk_dev_dif	2.920	0.102	1.152	o	o	f	vctk_test_dif	4.990	0.170	1.501
14	vctk_dev_dif	34.190	0.893	93.492	o	a	f	vctk_test_dif	26.800	0.773	86.021
15	vctk_dev_dif	20.550	0.616	38.573	a	a	f	vctk_test_dif	21.760	0.647	39.856
16	vctk_dev_dif	1.439	0.052	1.164	o	o	m	vctk_test_dif	2.067	0.071	1.819
17	vctk_dev_dif	28.680	0.762	98.909	o	a	m	vctk_test_dif	27.320	0.753	91.545
18	vctk_dev_dif	17.420	0.546	25.231	a	a	m	vctk_test_dif	14.290	0.470	24.479

Table 12: ASR results for both development and test partitions (o-original, a-anonymized speech).

#	Dev. set	WER, %		Data	Test set	WER, %	
		$\mathbf{LM}_s$	$\mathbf{LM}_l$			$\mathbf{LM}_s$	$\mathbf{LM}_l$
1	libri_dev	35.09	29.98	a	libri_test	32.95	28.12
2	vctk_dev	53.67	49.92	a	vctk_test	56.56	52.60

PUBLICATIONS

Main Publications

(International Journal)

1. Candy Olivia Mawalim and Masashi Unoki, "Feasibility of Audio Information Hiding Using Linear Time Variant IIR Filters Based on Cochlear Delay", *Journal of Signal Processing*, Research Institute of Signal Processing, vol. 23, no. 4, 2019.

2. Candy Olivia Mawalim and Masashi Unoki, "Speech Watermarking by McAdams Coefficient Scheme Based on Random Forest Learning", *Entropy*, MDPI, vol. 23, no. 10, 2021.

3. Candy Olivia Mawalim, Kasorn Galajit, Jessada Karnjana, Shunsuke Kidani, and Masashi Unoki, "Speaker Anonymization by Modifying Fundamental Frequency and X-Vectors Singular Value", *Computer Speech and Language*, Elsevier, vol. 73, 101326, 2022.

(Book Chapter)

1. Candy Olivia Mawalim and Masashi Unoki, "Audio Information Hiding based on Cochlear Delay Characteristics with Optimized Segment Selection", *Advances in Intelligent Systems and Computing*, Springer, vol. 1145, 2020.

(International Conference)

1. Candy Olivia Mawalim and Masashi Unoki, "Improving Security in McAdams Coefficient-Based Speaker Anonymization by Watermarking

Method", *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Tokyo, Japan, December 2021.

2. Candy Olivia Mawalim, Shengbei Wang, and Masashi Unoki, "Speech Information Hiding by Modification of LSF Quantization Index in CELP Codec", *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Virtual Event, Auckland, New Zealand, pp. 1321–1330, December 2020.

3. Candy Olivia Mawalim, Kasorn Galajit, Jessada Karnjana, and Masashi Unoki, "X-Vector Singular Value Modification and Statistical-Based Decomposition with Ensemble Regression Modeling for Speaker Anonymization System", *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association*, Virtual Event, Shanghai, China, pp. 1703–1707, October 2020.

4. Candy Olivia Mawalim and Masashi Unoki, "Audio Information Hiding based on Cochlear Delay Characteristics with Optimized Segment Selection", *Proc. 3rd International Conference on Security with Intelligent Computing and Bigdata Services (SICBS 2019)*, New Taipei City, Taiwan, December 2019.

5. Candy Olivia Mawalim and Masashi Unoki, "Feasibility of Audio Information Hiding using Linear-Time Variant IIR Filter based on Cochlear Delay", *NCSP'19 Proceedings*, Honolulu, Hawaii, USA, pp. 323–326, March 2019.

(Domestic Conference)

1. Candy Olivia Mawalim and Masashi Unoki, "Study on Audio Information Hiding using Linear-Time Variant IIR Filter based on Cochlear Delay Characteristics", in IEICE Technical Report, EMM2018–85(2019–01), pp: 19–24, Tohoku University, Sendai, January 2019. (non-referred, oral presentation)

2. Candy Olivia Mawalim and Masashi Unoki, "Study on inaudible audio information hiding using linear-time variant IIR filter based on cochlear delay characteristics", in IEICE Technical Report, EMM2018–108(2019–03), pp: 89–94, March 2019. (non-referred, poster presentation)

3. Candy Olivia Mawalim, Kasorn Galajit, Jessada Karnjana, and Masashi Unoki, "X-vector anonymization using regression modeling

with statistical and singular value decomposition", in IEICE Technical Report, EMM2021, Online, January 2021. (non-referred, oral presentation)

Other Publications

(International Journal)

1. Candy Olivia Mawalim, Shogo Okada, and Yukiko I. Nakano, "Task-Independent Recognition of Communication Skills in Group Interaction Using Time-Series Modeling", *ACM Transactions on Multimedia Computing Communications and Applications*, vol. 17, no. 4, pp. 1–27, 2021.

2. Candy Olivia Mawalim, Shogo Okada, Yukiko I. Nakano, and Masashi Unoki, "Personality Trait Estimation in Group Discussions using Multimodal Analysis and Speaker Embedding", *Journal on Multimodal User Interfaces*, Springer, Under review.

(Book Chapter)

1. Candy Olivia Mawalim, Shogo Okada, Yukiko Nagano, and Masashi Unoki, "Multimodal BigFive Personality Trait Analysis using Communication Skill Indices and Multiple Discussion Types Dataset", *Springer LNCS Social Computing and Social Media: Design, Human Behavior, and Analytics*, Springer, vol. 11578, 2019.

(International Conference)

1. Candy Olivia Mawalim, Shogo Okada, Yukiko Nagano, Masashi Unoki, "Multimodal BigFive Personality Trait Analysis using Communication Skill Indices and Multiple Discussion Types Dataset", *HCI International 2019 Proceedings*, Florida, USA, July 2019.

Other Achievements

(Awards)

1. Student Paper Award (2019 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing), March 2019.

2. Best Paper Award of the 11$^{\text{th}}$ International Conference on Social Computing and Social Media, July 2019.

(Grants)

1. Research Fellow DC1, Japan Society for the Promotion of Science (JSPS), Special Grants-in-Aid for Scientific Research (20J20580), April 2020 – March 2023.

2. Doctoral Research Fellowship (DRF), JAIST, October 2019 – March 2020.