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 automatic speaker verification (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.

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