Abstract

In recent years, all kinds of information have been digitized, collected, analyzed, and used. In such cyber-physical space, a system to ensure the authenticity of information is required for appropriate information utilization. In particular, audio zero-watermarking technology has been proposed as a mechanism to ensure the authenticity of audio signals. Audio zero-watermark technique creates a detection key from watermark and binary pattern generated from features of the audio signal instead of embedding watermark directly into the signal. This technique apparently embeds the watermark into the signal and uses this key to detect the watermark from the audio signal.

However, this technique has the problem of incorrectly detecting the watermark when the audio signal is subjected to sound information processing such as information compression and speech coding. These sound information processing are unavoidable in the process of communication and storage of digital audio signals. Therefore, a high robustness against these processing is essential. In information compression and speech coding, information is thinned out to the extent that it does not affect human hearing. Therefore, the digital signal itself will be very different, even though the human ear does not notice the change. Audio zero-watermarking generates a binary pattern from the features of the audio signal to create a detection key. This is then used to detect the watermark from the audio signal. If audio zero-watermarking method can generate a binary pattern from an audio signal that has undergone sound information processing using the same signal features, it is expected to be robust to various types of sound information processing. In order to be unaffected by changes in the signal due to sound information processing, it is desirable to use as signal features only those essential elements of the signal that are not subject to information compression (important for human hearing).

Conventional methods extract features unique to a signal from the time and frequency information of an audio signal and generate a binary pattern using these features. However, these methods still have issues, especially in terms of robustness against speech coding. Therefore, this paper decided to use Spikegram, an auditory spectral representation, for feature representation of audio signals. Spikegram represents information on human auditory nerve firings and is derived by sparse coding. Spikegram is a sparse representation of audio signals with a small number of firing points (spikes). The horizontal axis of each spike represents the timing of firing, the vertical axis represents the location of firing, and the shading represents the intensity of the firing. The small number of spikes in the Spikegram are not only important for the signal, but are also essential for human hearing. In particular, this paper investigated the derivation process of each Spike

in the computation of Spikegram with iterative processing, and found that the Spikes are derived in the order of the Spikes with the highest intensity. Therefore, this paper decided to use the Spikegram consisting of Spikes in the initial iteration (up to 100 iterations) as the signal feature. Spikes obtained in the early iterations are the most important of all the spikes and are the essential elements of the signal. Therefore, these spikes are considered to be robust features that cannot be subtracted by information compression or other processes.

The proposed method in this study generates a binary pattern using a Spikegram (consisting of 100 Spikes) derived from an audio signal. The conversion from Spikegram to binary pattern is realized by a binarization process. If the binary pattern has the same size as the watermark, the element is set to 1 if there is even a single Spike in the Spikegram corresponding to the element, and is set to 0 if there is no Spike. The watermark embedding process creates a detection key by calculating the exclusive or of the binary pattern generated by the binarization process and the watermark. In the process of detecting watermark, the exclusive or of the binary pattern generated from the signal to be detected in the same way as in the embedding process and the detection key.

Simulations were conducted to investigate the robustness against the proposed method to sound information processing. In the simulation, a detection key is created from the watermark (W) and the audio signal (original signal) and, and the watermark (W') is detected using the detection key from the audio signal (target signal) in which sound information processing is applied to the original signal. BER (Bit Error Rate) is calculated from the embedded watermark (W) and the detected watermark (W'), and robustness is evaluated using BER as an indicator. First, this simulation used 1-10 seconds of Japanese speech as the audio signal. The watermark information is a random binary matrix of various sizes. Low-pass filtering, re-quantization, re-sampling, information compression (MP3 and AAC), and speech coding (G.711 and G.729) were used for sound information processing. Simulation results showed that the proposed method can detect watermark information without error in the absence of sound information processing. The proposed method also achieves high robustness with BER less than 1% for all sound information processing except for some speech coding. In general, the proposed method achieves the same high robustness against sound information processing as the conventional method, and also achieves sufficient robustness against re-quantization and speech coding, which the conventional method is not good at. In addition, the relationship between the payload and the robustness of the proposed method was investigated. The results showed that the robustness of the proposed method tended to decrease slightly as the payload was increased, but the increase was small, indicating that the method remained robust enough even when the payload was increased. In addition, to confirm the effectiveness of the proposed method for audio signals other than speech signals, simulations were conducted using music signals. The results showed a similar trend to the results for speech signals, suggesting that the proposed method can be used for all audio signals, both speech and music.

Finally, an application of the audio zero watermarking method based on auditory spectral representation is examined for detection of speech tampering. The determination of tampering will be based on BER calculated by the proposed method. If there is a sufficient difference between BER of the audio information processing and BER of the tampering attack, a binary decision (tampered/no tampered) can be made using a certain BER as a threshold value. First, this paper investigated BER of the proposed method when a tampering attack was applied. The result showed that BER for tampering attacks such as adding noise that drowns out the original speech signal, silencing of the speech signal, or changing the pitch of the speech signal was approximately 2.4%. The BER for sound information processing was about 1.6% for G.729 speech coding, which is the highest BER, indicating that there is a sufficient difference between BER with tampering attacks and BER with sound information processing.

Next, to find the appropriate threshold (BER), this paper investigated the ROC curve and relationship between FRR (False Rejected Rate) and FAR (False Acceptance Rate) when the threshold BER was varied from 1.0% to 5.0% in 0.1% increments for the noise addition, zero interpolation (silence), and pitch shift tampering attacks. The ROC curve showed that the appropriate threshold was 1.2%, while the FRR/FAR relationship showed that the appropriate threshold was 1.1%. Therefore, in this paper, the threshold BER was set at 1.15%, and areas exceeding this threshold were judged to be tampered areas, while areas below this threshold were judged to be non-tampered areas.

The evaluation of speech tampering detection methods against tampering attacks showed that the method has a certain detection capability for zero interpolation, pitch shift, and sample replacement. On the other hand, when sound information processing was applied at the same time as the tampering attack, the detection capability was found to deteriorate.

As a result, it is showed that the audio zero-watermarking method based on auditory spectral representation proposed in this study can be used as a mechanism to guarantee the authenticity of audio signals. It is also suggested that this method can be applied to detect tampering with speech signals.