I. INTRODUCTION

THE rapid development of the Internet and multimedia technology has made the exchange of multimedia information reach an unprecedented depth. However, at the same time piracy and tampering have become rampant for it [1], [2]. Digital watermarking technology emerged as one of its solutions and has received widespread attention in the recent years [3], [4]. While various multimedia data (including audio, image, and video) require copyright protection, this paper focuses on protection of audio data.

Audio zero-watermarking technology is a promising technology for audio copyright protection due to its excellent imperceptibility, while there is potential to improve its robustness [5]. The zero-watermarking technology does not modify the data of the host audio signal, but constructs watermark information based on its content characteristics [6]. Based on different techniques for exploring stable content characteristics, audio zero-watermarks can be broadly grouped into two categories: time domain vector mapping and transform domain approaches. In general, the singular value decomposition (SVD) is used to seek stable characteristics based on the time domain. The largest singular value with stability is adopted to represent audio characteristics and the repetition of audio content makes the SVD based zero-watermark method to effectively resist serious synchronization attacks [7], [8].

Recently, researchers have been also studied watermarking technology based on the transform domain, including discrete cosine transform (DCT) [3], [4], [9], [10], discrete wavelet transform (DWT) [12], [13], Fourier transform (FT) [14], [15] and linear prediction cepstrum coefficients (LPCC) [16]. Specifically, DCT domain-based technologies [3], [4], [9], [11] apply the DCT to the host audio signal to obtain a set of audio segments. However, these technological innovations lie in the selection of audio segment characteristics. The DWT domain based technologies [12], [13] are similar to the DCT approach, which perform DWT on the host signal to select the stable characteristics of the audio segment. Furthermore, the phase information based on FT is also commonly used to characterize the characteristics of audio segments [14], [15].

The technologies based on amplitude information (such as DCT, DWT, or a combination of multiple transform domains [17], [19]) to characterize audio segment characteristics can effectively resist common attacks (such as noise, filtering and re-sampling attacks, etc.), whereas they generally do not have the ability to resist harsh synchronization attacks (such as time scale modify (TSM), cropping attacks). On the contrary, the technologies based on phase information of FT can resist synchronization attacks better than the technologies based on amplitude information. However, they are not very effective to resist common attacks.

To further enhance the robustness against various nature of attacks, a novel zero-watermarking technology for audio based on the graph Fourier transform (GFT) [20] is proposed in this work. The emerging graph signal processing (GSP) [21], [22] technology has been used in speech processing to express the structural relationship of speech sample data points. The graph topology constructed by the potential relationship between data points can determine the graph Fourier basis [23], and GFT can further analyze the characteristics of the graph signal in the graph frequency domain [20]. Progress has been made in using GFT to watermark unstructured data, such as point clouds [24], [25] and graphic data [27]. Here, we adopt GFT to transform the speech signal from the graph domain to the graph frequency domain for stabilizing characteristics of audio segments. We then encode all the selected graph Fourier coefficients that can characterize the audio segments to achieve zero-watermark embeddings. We evaluate our proposed method against various methods to show effectiveness against both common and synchronization attacks.
II. GRAPH SPECTRUM OF AUDIO SIGNAL

In this section, we describe the process to map the audio signal in the time domain to the graph domain and then the details to convert the graph signal to the graph frequency domain for further analyzing its characteristics.

A. Graph domain mapping of audio signal

1) Basic concept of graph signal: In order to map the audio signal \(x\) in the time domain to the graph audio signal \(y\) in the graph domain, it is necessary to exploit the graph in the GSP. The graph is composed of vertices, edges connecting the vertices, and edge weights. Mathematically, the graph can be expressed as, \(G = (V, E, W)\), where \(V\) represents the vertex set, \(E\) represents the edge set, and \(W\) represents the weight set \([22]\).

In order to utilize GSP for processing audio signals, the time domain signal \(x\) is first divided into frames, and then each frame is mapped. Assuming that \(x\) is divided into \(M\) frames with \(N\) sampling points, one of the frames can be expressed as \(x_m = [x_{m1}, x_{m2}, ..., x_{mN}]^T\) and \(m = 1, 2, ..., M\). Given a graph, \(y_m\) can be expressed as a graph signal, which is defined as a mapping as follows.

\[
S : x_m \rightarrow y_m,
\]

where, \(y_m = [y_{m1}, y_{m2}, ..., y_{mN}]^T\) indexed by \(G = (V, E, W)\) is a one-to-one mapping value of \(x_m\). Each element of \(y_m\) represents the intensity at a vertex in the corresponding graph and each vertex corresponds to a sampling point in the time domain. The corresponding graph \(G\) describes the relationship among the vertices and can be written in detail as given in Equation (2).

\[
V = [v_1, v_2, ..., v_N]^T,
E = \{e_{ij} \in \{0, 1\} | i=1,2, ..., N, j=1,2, ..., N \in \mathbb{R}^{N \times N},
W = \{w_{ij} \} \in \mathbb{R}^{N \times N}.
\]

Here, \(e_{ij}\) indicates that there is no edge connection between vertex \(v_i\) and \(v_j\), otherwise \(e_{ij} = 1\). The \(w_{ij}\) represents the weight of the edge between \(v_i\) and \(v_j\). The general weight matrix can be represented by the graph Laplacian matrix \(L\) or the graph adjacency matrix \(A\) \([28, 29]\). Among them, \(L\) is only applicable to undirected graphs, while the \(A\) is not \([21]\). Considering speech signal is a time series with obvious temporal relevance, directional weights can exactly represent the relationship between speech time sampling points. Therefore, this work adopts \(A\) as \(W\) and the value of \(A\)’s elements are 0 or 1 to achieve the purpose of focusing only on whether there is a connection between the vertices.

2) Construction of graph audio signal: In this work, the combined graph \(k\)-shift operator \(\Gamma_k\) is used to construct \(A\) to obtain the graph speech signal. According to the above analysis \(A\), \(A\) is equivalent to \(W\) and \(E\) as a binary matrix, the graph can be redefined as \(G_{\Gamma_k} = (V, \Gamma_k, \Gamma_k)\). \(\Gamma_k\) is defined as

\[
\Gamma_k = \sum_{t=0}^{k-1} \gamma_t, k = 1, 2, ..., \]

where \(\gamma_t \in \mathbb{R}^{N \times N}(t = 0, 1, ...)\) is a binary matrix and which denotes a \(t\)-shift operator. The element \(\gamma_{ij}\) of \(\Gamma_k\) satisfies the condition

\[
\gamma_{ij} = \begin{cases} 1, & \text{if } (j - i) \mod N = 0, ..., k - 1 \ \\
0, & \text{else} \end{cases}
\]

Obviously, when \(k = 1\), \(\Gamma_1 = \gamma_0\) is a unit matrix which implies that the signal was not shifted. The graph signal \(y_o\) obtained after implementing \(\Gamma_k\) on the time domain signal \(y_i\) can be expressed as \(y_o = \Gamma_k \cdot y_i\).

B. Spectrum of audio signal in graph frequency domain

With the aid of the adjacency matrix \(A\), the graph domain signal can be converted to the graph frequency domain. The specific method performs SVD on \(A\) to obtain the singular value decomposition of \(A\). We have \(A = Q \Sigma Q^{-1}\), where \(Q = [\epsilon_1, \epsilon_2, ..., \epsilon_N] \in \mathbb{R}^{N \times N}\) is formed by \(N\) eigenvectors of \(A\) and \(\Sigma = [\zeta_1, \zeta_2, ..., \zeta_N] \in \mathbb{R}^{N \times N}\) with \(N\) eigenvectors as the main diagonal of \(A\). The column \(\epsilon_k\) of \(Q\) represents the spectral components at the corresponding graph frequency \(\zeta_k\).

Since \(A\) here is a row trapezoidal matrix with full row rank, this will result in \(N\) linearly independent eigenvectors. Correspondingly, \(Q\) is invertible, the graph Fourier matrix \(F\) can be defined as follows \([20]\).

\[
F = Q^{-1} = [\epsilon_1, \epsilon_2, ..., \epsilon_N]^{-1} = [f_1, f_2, ..., f_N].
\]

The graph spectrum \(\tilde{y}\) obtained after performing the GFT on the graph signal \(y\) can be expressed as

\[
\tilde{y} = F \cdot y = Q^{-1} \cdot y = [f_1y_1, f_2y_2, ..., f_Ny_N]^T = [\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_N]^T,
\]

where \(\tilde{y}_n\) represents the graph Fourier coefficients at the corresponding graph frequency \(\zeta_n\). In addition, combining the equations (3) and (6) to note that GFT is essentially a simple matrix operation process, and it is a method without latency.

III. PROPOSED GFT BASED AUDIO ZERO-WATERMARKING

The proposed framework performs zero-watermark processing in the GFT domain. Fig. 1 shows the flow diagram of the zero-watermarking generation and extraction process. The zero-watermarking embedding and extraction have some common processes that include framing, constructing graph signal, GFT and encoding. We note that their XOR process is slightly different.

A. The common processes

1) Framing: According to the length \(M\) of the watermark sequence obtained by watermark image dimensionality reduction, the audio signal \(x\) is evenly divided into \(M\) non-overlapping frames. The length of each frame is represented by \(N\). Hence, we have \(N = \text{floor}(\text{x_len}/M)\), where \(x\_len\) denotes the length of the audio signal \(x\).
2) Constructing graph signal: Once the time domain signal is framed, one of the frames can be expressed as \( x^{(m)} = [x_1^{(m)}, x_2^{(m)}, \ldots, x_N^{(m)}]^T \) and \( m = 1, 2, \ldots, M \). By performing the combined graph \( k \)-shift operator \( \Gamma_k \) on \( x^{(m)} \), we can obtain the graph signal \( y^{(m)} \).

\[
y^{(m)} = \Gamma_k \cdot x^{(m)}. \tag{7}
\]

3) GFT: Based on applying SVD on \( \Gamma_k \), the GFT base \( \mathcal{F}_{\Gamma_k} \) can be obtained. The graph spectrum coefficients \( \tilde{y}_{\Gamma_k} \) of the graph signal \( y^{(m)} \) can be then obtained by the formula (8).

\[
\tilde{y}_{\Gamma_k} = \mathcal{F}_{\Gamma_k} y^{(m)}. \tag{8}
\]

Considering graph spectrum is mainly concentrated at lower frequencies, and when \( k \) is small, the spectrum is relatively stable [23]. In addition, a larger \( k \) would cause a higher amount of calculation. We mainly discuss the case when \( k = 3 \) in this work.

4) Encoding: For obtaining the stability characteristic sequence of the audio segment, we analyzed absolute value of the maximum spectral coefficient of the segments in audio with a duration of 22 seconds. Fig. 2 shows the maximum absolute graph spectral coefficient values under different attacks of the first 256 frames in 1024 frames. From Fig. 2, it can be clearly observed that these values have undergone some changes after these attacks, but the trends are relatively stable. Therefore, these values can be used to represent the feature sequence \( F \) of each frame to resist attacks.

In order to obtain the feature binary sequence \( B \) of each audio segment, the K-means clustering algorithm is used to divide the feature sequence \( F \) into two categories, which are coded as 0 and 1, respectively.

B. The different processes

1) XOR of zero-watermarking embedding: The obtained signal feature binary sequence is XOR-ed with the watermark sequence to obtain the watermark key \( K \) as follows.

\[
K = \{ k(m) = B(m) \oplus W(m) \mid 1 \leq m \leq M \}, \tag{9}
\]

where \( W(m) \) is the pixel point value of the binary image.

2) XOR of zero-watermarking extraction: The watermarked signal feature binary sequence \( B' \) is XOR-ed with the watermark key \( K \) to obtain the watermark sequence \( W' \) as follows.

\[
W' = \{ w(m) = B'(m) \oplus K(m) \mid 1 \leq m \leq M \}. \tag{10}
\]

Finally, the obtained watermark binary sequence can be restored to a watermark image by increasing the dimension, as shown in Fig. 3.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental setup

1) Database: In order to verify the effectiveness of the proposed zero-watermarking scheme, various styles of audio clips are randomly selected from the DSD100 database [30], including rock, classical, jazz, country and pop music. There are a total of forty pieces of music and each audio clip with

| Schemes | Attacks | AWGN (10dB) | AWGN (20dB) | LPH | Re-sampling | Re-quantization | MP3 | Amplitude (1.5) | Amplitude (2) |
|---------|---------|-------------|-------------|-----|-------------|----------------|-----|----------------|---------------|
| Our     |         | 0.0308/0.9786 | 0.0083/0.9942 | 0.0130/0.9910 | 0.0017/0.9988 | 0.0275/0.9910 | 0.0055/0.9961 | 0.0003/0.9996 | 0.0004/0.9995 |
| DWT     | 12      | 0.0764/0.9467 | 0.0350/0.9758 | 0.0097/0.9932 | 0.0032/0.9977 | 0.0729/0.9492 | 0.0164/0.9837 | 0/1            | 0/1           |
| STFT    | 15      | 0.2542/0.8141 | 0.2432/0.8227 | 0.2181/0.8419 | 0.1521/0.8916 | 0.2326/0.8307 | 0.2091/0.8490 | 0/0.9999       | 0/1           |
| DWT-DCT | 18      | 0.1501/0.8934 | 0.0680/0.9527 | 0.0376/0.9740 | 0.0155/0.9893 | 0.0794/0.9445 | 0.0240/0.9834 | 0/1            | 0/1           |
| DWT-DCT-SVD | 19 | 0.0326/0.9774 | 0.0093/0.9935 | 0.0146/0.9899 | 0.0009/0.9993 | 0.1002/0.9295 | 0.0036/0.9975 | 0.5333/0.5669 | 0.5255/0.5776 |
We observe that both the proposed scheme and the first four schemes can effectively extract the watermark bits under the amplitude attacks, while the scheme presented in [19] is unable to resist the amplitude attacks. In addition, the robustness of the proposed scheme under the first six attacks is as superior as the scheme [19], yet it is significantly better than the schemes based on the other three transform domains. The reason behind this result may be attributed to the robust graph Fourier coefficients and use of K-means to cluster the feature sequence to obtain the feature binary sequence. Overall, the proposed scheme has an outstanding performance in resisting common attacks and outperforms the baselines. In addition, the watermark images located in the first row of Fig. 3 are extracted after the host signal is subjected to common attacks. The extracted images appear almost the same as the original, which depicts that the robustness of the proposed method.

### V. Conclusion

In this work, we propose a novel zero-watermarking technique based on the GFT. We note that the combined shift operator is used to construct the graph signal, and then the stable graph Fourier coefficients are selected for encoding to obtain zero-watermark embedding. Our experimental results on DSD100 database show that the proposed scheme is more robust than the traditional transform domains such as DCT, DWT, FT against common and synchronization attacks.

---

TABLE II: Robustness comparison of the proposed scheme and the baselines under synchronization attacks, where ***/*** indicates average metrics BER/NC and the bold mark indicates the best metric across all the schemes under each attack.

| Schemes       | Attacks                          | Metrics (BER/NC) |
|---------------|----------------------------------|------------------|
|               | TSM+1%  | TSM+10% | TSM-1% | TSM-10% | Crop (15 front) | Crop (15 back) | Crop (20 front) | Crop (20 back) |
| DWT           | 0.1720  | 0.1767  | 0.1162  | 0.1328  | 0.1830  | 0.2122  | 0.2427  | 0.2180  |
| STFT          | 0.3343  | 0.3347  | 0.2920  | 0.3005  | 0.3422  | 0.3599  | 0.3750  | 0.3200  |
| DW/DCT        | 0.4700  | 0.4760  | 0.4200  | 0.4300  | 0.4820  | 0.4920  | 0.4920  | 0.4850  |
| DW/DCT/SVD    | 0.3340  | 0.3370  | 0.2920  | 0.3000  | 0.3420  | 0.3599  | 0.3750  | 0.3200  |

Fig. 3: Extracted watermark images of the proposed scheme under attacks: (a) Original image; (b) AWGN (10dB); (c) LPF; (d) Re-sampling; (e) MP3; (f) Re-quantization; (g) Amplitude (2 times); (h) TSM+1%; (i) TSM+10%; (j) TSM-1%; (k) TSM-10%; (l) Cropping 5 frames (front); (m) Cropping 10 frames (front); (n) Cropping 20 frames (front)
REFERENCES

[1] T. Zong, Y. Xiang, I. Natgunanathan, L. Gao, G. Hua, and W. Zhou, “Non-linear-echo based anti-collusion mechanism for audio signals,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 969–984, 2021.

[2] P. Dhar and T. Shimamura, “Blind audio watermarking in transform domain based on singular value decomposition and exponential-log operations,” Radioengineering, vol. 26, pp. 552–561, 06 2017.

[3] Y. Xiang, I. Natgunanathan, Y. Rong, and S. Guo, “Spread spectrum-based high embedding capacity watermarking method for audio signals,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 12, pp. 2228–2237, 2015.

[4] Y. Xiang, I. Natgunanathan, D. Peng, G. Hua, and B. Liu, “Spread spectrum audio watermarking using multiple orthogonal pn sequences and variable embedding strengths and polarities,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 3, pp. 529–539, 2018.

[5] K. Yang, W. Wang, Z. Yuan, and W. Zhao, “Strong robust zero watermarking algorithm based on stft transform and image normalization,” in 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2018, pp. 236–240.

[6] Y. Peng and M. Yue, “A zero-watermarking scheme for vector map based on feature vertex distance ratio,” Journal of Electrical and Computer Engineering, vol. 2015, 01 2015.

[7] M. Cao, C. Li, and L. Tian, “Content-based audio zero-watermarking algorithm against tsn,” in 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016, pp. 297–301.

[8] Y. Sun, L. Tian, and C. Li, “Robust zero-watermarking algorithm based on audio beats,” in 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), 2021, pp. 308–312.

[9] A. Kanhe and A. Ganasekaran, “A blind audio watermarking scheme employing dct–ht–sd technique,” Circuits, Systems, and Signal Processing, vol. 38, no. 8, pp. 3607–3714, 2019.

[10] H. Al-khafaji and C. Abhayaratne, “Graph spectral domain blind watermarking for unstructured data from sensor networks,” in Information and Schur decomposition,” in Proceedings of the IEEE, vol. 106, no. 5, pp. 808–828, 2018.

[11] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, “The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains,” IEEE Signal Processing Magazine, vol. 30, no. 3, pp. 83–98, 2013.

[12] Y. Xiang, Y. Wang, Z. Yang, and H. Guo, “An iterative graph spectral subtraction method for speech enhancement,” Speech Communication, vol. 123, pp. 35–42, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639320302405

[13] E. E. Abdallah, A. B. Hamza, and P. Bhattacharya, “Spectral graph-theoretic approach to 3d mesh watermarking,” in Proceedings of Graphics Interface 2007, ser. GI ’07. New York, NY, USA: Association for Computing Machinery, 2007, pp. 327–334. [Online]. Available: https://doi.org/10.1145/1268517.1268570

[14] H. Al-Khafaji and C. Abhayaratne, “Graph spectral domain blind watermarking,” in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2492–2496.

[15] P. K. Dhar and T. Shimamura, “Blind audio watermarking in transform domain based on singular value decomposition and exponential-log operations,” Radioengineering, vol. 26, no. 2, pp. 552–561, 2017.

[16] J. Panda, S. Choudhary, K. Nath, and S. Kumar, “Audio zero watermarking scheme based on sub band mean energy comparison using dwt-dct,” in 2016 International Conference on Signal Processing and Communication (ICSPC), IEEE, 2016, pp. 352–357.

[17] A. E. A. Jayarani, M. R. Bhatt, and D. Geetha, “Zero watermarking on audio based on stft,” in 2018 International Conference on Computing, Electronics & Communications Engineering (ICCCECE), IEEE, 2018, pp. 253–256.

[18] S.-M. Tsai, “An efficient and robust zero-watermarking scheme for digital audio,” in 2013 IEEE International Conference on Circuits and Systems (ICCCAS). IEEE, 2013, pp. 51–54.