A Novel Feature Based on Graph Signal Processing for Detection of Physical Access Attacks

Longting Xu¹, Mianxin Tian¹, Xing Guo¹, Zhiyong Shan¹, Jie Jia², Yiyuan Peng², Jichen Yang¹ and Rohan Kumar Das³

¹Donghua University Shanghai, College of Information Science and Technology, P. R. China
²Vivo AI Lab, P. R. China
³Fortemedia Singapore, Singapore

xlt@dhu.edu.cn

Abstract

Speaker verification systems face threat from various spoofing attacks and particularly, the physical access attacks or replay attacks that are most common show an imminent threat. Literature shows that graph signal processing (GSP) shows a better correlation between speech samples and explore more hidden information from speech than the traditional digital signal processing methods. With this motivation, we propose a novel feature based on GSP, namely, graph frequency cepstral coefficient (GFCC). We use the combined shift operator to construct the graph signal, and then carry out the graph Fourier analysis to extract GFCC features. It is observed that compared to fast Fourier transform, the GFT can more accurately represent the structural relationship of speech samples, which makes the real and replay speech very distinguishable in the frequency domain. We use the GFCC features with a light convolutional neural network system in our studies. The results on ASVspoof 2019 physical access corpus show that the proposed GFCC feature based system outperforms the challenge baselines by a large margin and emerge as one of the best performing state-of-the-art single systems.

1. Introduction

Speaker verification (SV) aims to authenticate the target person based on the voice samples [1]. The advancements in SV research has led to use of such systems in practical applications [2,3]. However, SV systems are very vulnerable to spoofing attacks [4,5]. Among these attacks, physical access attacks are the most common form of attacks as they can be performed without in-depth technical knowledge, which is required to generate logical access attacks [6]. If the recorded speech of the target user is somehow available, the attacker can simply use it for unauthorized access. Considering the extent of threat of physical access attacks in SV, we focus on detection of such attacks in this work.

Research on spoofing countermeasures witnessed several successful systems in the past decade. The features such as cochlear filter cepstral coefficient and instantaneous frequency [7], constant-Q cepstral coefficients (CQCC) [8], instantaneous phase [9], linear frequency cepstral coefficients (LFCC), subband spectral flux coefficients and spectral centroid frequency coefficients [10] are among some of the effective features for spoofing attack detection. The benchmark performance obtained using CQCC features also led to development of other features using constant-Q transform (CQT) [11–15]. Although Gaussian mixture model (GMM) was used as a classifier for most of the initial works for spoofing countermeasures, various advanced neural network based systems such as light convolutional neural networks (LCNN) [16,17], squeeze excitation residual networks [18,19], capsule networks [20], VGG, SincNet [21] and RawNet2 [22] are used in the recent years.

The graph is a form of representation, which helps to describe the geometric structure of the data field [23]. It contains information, which usually exists on the vertices and edges of the graph in the form of high-dimensional data (i.e., signals), and the weight related to each edge in the graph usually represents the similarity of the two vertices connected by it [24].

The graph signal processing (GSP) has been widely used in various research applications. Compared to traditional digital signal processing, the GSP can directly process non-linear and non-stationary signals. The authors of [25] proposed GSP based features of electroencephalogram signals to detect ictal class of epilepsy, which showed very effective results. The GSP was applied to New York City taxi data analysis from 2010 to 2013 as well [26]. The experimental results showed that graphic Fourier transform can extract the effective information hidden in the original data, and can be used as an effective tool to discover the potential behavior of New York city taxis. In [27], GSP based graph spectrum subtraction and iterative graph spectrum subtraction is proposed that resulted improved speech enhancement results showing scope of GSP in speech processing. The use of graph for attention models in anti-spoofing has been recently explored in [28]. However, the GSP based front-end for spoofing attack detection has not been explored before.

Most of the existing hand-crafted features used for replay attack detection is based on short-time analysis to solve the non-stationary problem of speech, and uses the approximate linear approach to deal with the non-linear problem. Hence, in order to obtain the characteristics of speech approximately, there is no doubt that some valuable information will be lost to some extent, which may affect in capturing the difference between replay and real speech. On the other hand, the GSP can process time-varying non-stationary signals, and it intuitively describes the potential relationship between the signals on the stored graph vertices by defining the edge and the weight of edge between the graph vertices.

In this work, the graph topology constructed by the potential relationship between speech samples is used to determine the graph Fourier basis similar to that in [27]. The GFT is then used to transform speech signal from graph domain to graph frequency domain for reflecting the structural relationship between the speech samples. We believe as a result of this the distinction between the real and replayed speech will become more significant in the graph frequency domain as GFT can discover more hidden properties of graph signals in the graph fre-
frequency domain [29]. A novel feature namely, graph frequency cepstral coefficient (GFCC) is proposed based on the GSP. We consider a light convolutional neural network (LCNN) classifier for developing the spoofing countermeasure with the GFCC features for our studies on ASVspoof 2019 physical access corpus. The primary contribution of this work is the use of GSP in anti-spoofing and the proposal of GFCC features for effective detection of physical access attacks.

The remainder of the paper is organized as follows. Section 2 describes representing a speech signal in graph domain. Section 3 and Section 4 discuss the details of GFT and the proposed GFCC feature derived using GFT, respectively. The experiments and their results are reported in Section 5. Finally, Section 6 concludes the work.

2. Representing Speech Signal in Graph Domain

The graph definition in GSP plays a key role in speech transformation from time domain to graph domain, where the graph consists of vertexes, edges connecting vertexes and edge’s weights. Mathematically, a graph \( G \) is defined as \( G = (V, E, \mathcal{W}) \), where \( V \) represents the vertex set, \( E \) represents edge set, and \( \mathcal{W} \) represents the weight set, respectively.

In order to apply GSP to speech signal, it is necessary to frame the speech signal in time domain. Consider the speech signal \( s \) is divided into \( M \) frames and each frame has \( N \) sample points, one of the frames can be represented as \( s_m = \{s_{m1}, s_{m2}, \ldots, s_{mN}\}^T \) and \( m = 0, 1, \ldots, M \). Then based on the directionality of the speech sample sequence after framing, any frame of speech sample can be regarded as a limited time sequence signal with directivity. In principle, the graph topology of speech signal can be constructed by finite time sequence signal, then, the speech signal \( s \) is transformed from time domain to graph domain, each sampling point is mapped to \( y_m \). Here, \( y_m \) can be expressed as a graphic speech signal, which is defined as a mapping as follows.

\[
s_m \rightarrow y_m \tag{1}
\]

where \( y_m = [y_{m1}, y_{m2}, \ldots, y_{mN}]^T \) indexed by \( G = (V, E, \mathcal{W}) \), \( V \) is a one-to-one mapping value of \( s_m \). Each element of \( y_m \) indicates the intensity of corresponding vertex in the graph field and each vertex corresponds to a sampling point in the time domain. The corresponding graph \( G \) describes the relationship between the vertexes, the set of all vertexes in graph domain is expressed as \( V = \{v_0, v_1, \ldots, v_{N-1}\} \), \( E \in \mathbb{R}^{N \times N} \) is the edge set. The \( e_{ij} \in \{0, 1\} \) indicates that there is an edge between \( v_i \) and \( v_j \). While \( \mathcal{W} = \{\omega_{ij}\}_{(i,j)\in E} \in \mathbb{R}^{N \times N} \) is the graph weighting matrix, \( \omega_{ij} \) represents the weight of the edge between \( v_i \) and \( v_j \) [27].

Generally, \( V \) can be represented by graph Laplacian matrix \( \mathcal{L} \) or graph adjacency matrix \( A \). The \( \mathcal{L} \) is only applicable to undirected graph. The vertices in graph domain are directional, so graph adjacency matrix \( A \) is used as \( W \). This work mainly focuses on whether there is connection between vertexes, not the strength between them. Hence, \( A \) is a 0-1 matrix, where \( a_{ij} = 1 \) indicates there is connection from \( v_i \) to \( v_j \), otherwise \( a_{ij} = 0 \).

We construct a graph \( k \)-shift operator \( \psi_k \) to represent \( A \). After the above analysis, \( A \) is equivalent to \( W \) and \( \mathcal{L} \) as a 0-1 matrix, \( G \) can be redefined as \( G_{\psi_k} = (V, \psi_k, \psi_k) \) [27], where

\[
\psi_k = \sum_{t=0}^{k-1} \gamma_t, k = 1, 2, \ldots, N
\tag{2}
\]

where \( \gamma_t \in \mathbb{R}^{N \times N} (t = 0, 1, \ldots, k - 1) \) is a 0-1 matrix. The element \( \gamma_{ij} \) of \( \gamma_t \) satisfies the following condition

\[
\gamma_{ij} = \begin{cases} 1, &t, f(j - i) \ mod \ N = t \\ 0, &\text{else} \end{cases}
\tag{3}
\]

The graph signal \( y_{\text{out}} \) obtained after implementing \( \psi_k \) on the time domain signal \( y_m \) can be expressed as \( y_{\text{out}} = \psi_k \cdot y_m \). We consider an example to illustrate when \( (y_1, \ldots, y_{N-1}, y_0)^T = \psi_1 \cdot (y_0, y_1, \ldots, y_{N-1})^T \), the transformation of \( \psi_1 \) into a matrix is expressed as follows

\[
\psi_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}_{N \times N}
\tag{4}
\]

According to the Theorem 2 in [30], the visualized topological structure can be obtained, as shown in Figure 1.

![Figure 1: Graph topology of finite time domain signals.](image)

How to map the signal without obvious graph topology to graph signal is the key of GSP technology from theory to practical application. For the topology without a priori, the graph learning algorithm can be used to estimate the topology according to the smoothness, sparsity and other attributes of the graph signal.

3. Graph Fourier Transformation

In order to analyze the characteristics of the graph signal, it is usually transformed into the graph frequency domain, and the graph adjacency matrix is eigen-decomposed. The specific method is singular value decomposition (SVD) of adjacency matrix \( A \).

\[
A = \zeta \Lambda u^{-1}
\tag{5}
\]

where \( \Lambda = \text{diag} \{\lambda_0, \lambda_1, \ldots, \lambda_{N-1}\} \), \( \lambda_k \) represents different eigenvalues of \( \Lambda \), \( \lambda_k \) indicates different frequencies in the graph frequency domain. \( \zeta = [\eta_0, \eta_1, \ldots, \eta_{N-1}] \) is a matrix, where column \( \eta_k \) represents the spectral components at the graph frequency of \( \lambda_k \). Since \( A \) is a full rank matrix, \( A \) has \( n \) linearly independent eigenvectors, so we can conclude that \( \zeta \) is invertible. The Fourier matrix of graph is defined as \( \mathcal{F} \)

\[
\mathcal{F} = \zeta^{-1} = [\eta_0, \eta_1, \ldots, \eta_{N-1}]^{-1}
\tag{6}
\]

By graph Fourier transformation, we can obtain the spectrum \( \hat{y}_m \) of graph signal \( y_m \).

\[
\hat{y}_m = \mathcal{F} \cdot y_m = [f_0 y_0, f_1 y_1, \ldots, f_{N-1} y_{N-1}]^T
\]

\[
= [\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{N-1}]^T.
\tag{7}
\]

\( \hat{y}_k \) is the graph Fourier coefficient corresponding to each graph frequency \( \lambda_k \), where \( k = 0, 1, \ldots, N - 1 \). \( y_m \) is given by \( y_m = \psi_k \cdot s_m \).
4. Proposed GFCC feature

In this section, we introduce the process of GFCC\(^1\) feature extraction, as shown in Figure 2. The first step of GFCC extraction algorithm is pre-emphasis, which is used to enhance the amplitude of the high-frequency band of speech intentionally that can effectively compensate the loss of high-frequency components in the sound transmission, so as to balance the spectrum skew in speech. The pre-emphasis transfer function is given by the following equation

\[ H(z) = 1 - b z^{-1} \]  

where \( b \) is a constant and its typical value is 0.97.

Owing to the characteristic parameters of speech signal are time-varying, speech is regarded as non-stationary signal. Hence, a short-term processing is performed that divides a given speech into smaller blocks called as frames, under which they are assumed as stationary signals. Considering the speech signal \( s \) is evenly divided into \( M \) non-overlapping and consecutive frames, the length of each frame is represented by \( N \), where the value of \( N \) can be expressed by the following

\[ N = 1/M \]  

where \( l \) represents the length of speech signal \( s \).

After the speech signal \( s \) is divided into frames, the \( m \) frame \( s_m \) of signal in time domain can be transformed into signal \( y_m \) in graph domain by graph \( k \)-shift operator \( \psi_k \).

\[ y_m = \psi_k \cdot s_m \]  

As shown in Eq. (5), the SVD decomposition of \( \psi_k \) is used to get \( c \), and the graph Fourier transformation of \( y_m \) is used to get the spectral coefficient \( \hat{y}_m \) as

\[ \hat{y}_m = \varsigma^{-1} \cdot y_m \]  

The logarithm is then taken to compress the dynamic range of the power spectrum and improve its display effect. After the graph speech signal \( y_m \) is processed by GFT, we can obtain \( \hat{y}_m = [\hat{y}_0, \hat{y}_1, \ldots, \hat{y}_{N-1}]^T \). Further, we can obtain log power spectrum \( \hat{Y}_m(i) \) as

\[ \hat{Y}_m(i) = [\log|\hat{y}_0|^2, \log|\hat{y}_1|^2, \ldots, \log|\hat{y}_i|^2]^T, i = N - 1 \]  

Finally, we apply the discrete cosine transform (DCT) to derive the cepstral coefficients from the log power spectrum as follows

\[ F(z) = c(z) \sum_{i=0}^{N-1} \log|\hat{y}_i|^2 \cos \left[ \frac{(i + 0.5)\pi}{N} \right] \]  

where \( \log|\hat{y}_i|^2 \) is the log power spectrum of speech signal, \( N \) is the number of points of the speech signal, \( F(z) \) is the extracted GFCC after DCT transformation, \( z = 0, \ldots, N - 1 \), \( c(z) \) can be considered as a compensation coefficient, where

\[ c(z) = \sqrt{\frac{1}{N}}, z = 0, \text{where } c(z) = \sqrt{\frac{2}{N}}, z \neq 0. \]

\(^1\)https://github.com/DHUspeech/GFCC-feature

5. Experiments and Results

5.1. Experimental setup

The studies in this work are conducted on ASVspoof 2019 physical access database [31]. The replay attacks of the database are created under a simulated setup. It consists of three subsets, which are train, development and evaluation set. The performance metric for evaluation are tandem detection cost function (t-DCF) and equal error rate (EER) [32].

The LCNN architecture used in our studies follows that reported in [33], which is one of the state-of-the-art classifiers for anti-spoofing. It considers Adam optimizer with \( \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8} \) [34]. The initial learning rate of \( 3 \times 10^{-4} \) is multiplied by 0.5 for every ten epochs, and batch size is 8.

5.2. Analysis of results

We now focus on the results of the proposed system developed using GFCC features. As the GFCC is proposed as a new feature, we evaluate it for various feature dimensions (static) ranging from 20 to 60 in steps of 10. Table 1 shows the performance of our proposed GFCC feature with LCNN classifier (GFCC-LCNN) for various feature dimensions. We observe that the best performance is achieved when feature dimension is equal to 60. However, we mention that the performances of 40 and 60 feature dimensions are very close. The optimal feature dimension for GFCC is considered as 60 from this analysis.

Table 1: Performance in t-DCF and EER(%) on ASVspoof 2019 physical access evaluation set for different feature dimensions (FD) of GFCC features.

<table>
<thead>
<tr>
<th>FD</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-DCF</td>
<td>0.0556</td>
<td>0.0496</td>
<td>0.0450</td>
<td>0.0535</td>
<td>0.0429</td>
</tr>
<tr>
<td>EER(%)</td>
<td>1.85</td>
<td>1.71</td>
<td>1.52</td>
<td>1.83</td>
<td>1.51</td>
</tr>
</tbody>
</table>

We are now interested to compare the performance of our GFCC-LCNN system to the ASVspoof 2019 challenge baselines as well as some of the other existing single systems evaluated on ASVspoof 2019 physical access corpus. The ASVspoof 2019 considers CQCC and LFCC feature with GMM based classifier as two baselines of the challenge [31]. Table 2 first shows the comparison of our proposed system using GFCC features with the ASVspoof 2019 baselines. It is observed that the GFCC-LCNN system outperforms both the challenge baselines CQCC-GMM and LFCC-GMM by a large margin. Then we consider some of the well performing single systems on ASVspoof 2019 physical access corpus for comparing to our proposed system.

Table 2 reports various well performing single system results from their respective published works. These systems use various front-ends such as short-time Fourier transform (STFT), group delay gram (GD-gran), spectrum from various transforms, and modified magnitude phase spectrum (MMPS), apart from the CQCC and LFCC based front-ends of the ASVspoof 2019 baselines. In addition, different deep learning framework are used in those systems, which are convolutional neural network (CNN), gated recurrent unit (GRU), residual network
Table 2: Comparison of proposed GFCC-LCNN system to ASVspoof 2019 baselines and a some single systems on ASVspoof 2019 physical access evaluation set.

<table>
<thead>
<tr>
<th>System</th>
<th>t-DCF</th>
<th>EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed: GFCC-LCNN</td>
<td>0.0429</td>
<td>1.31</td>
</tr>
<tr>
<td>ASVspoof 2019 Baseline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CQCC-GMM [31]</td>
<td>0.2454</td>
<td>11.04</td>
</tr>
<tr>
<td>LFCC-GMM [31]</td>
<td>0.3017</td>
<td>13.54</td>
</tr>
<tr>
<td>Comparison to Other Single Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNN [35]</td>
<td>0.1577</td>
<td>5.75</td>
</tr>
<tr>
<td>STFT-GRU [30]</td>
<td>0.1235</td>
<td>4.79</td>
</tr>
<tr>
<td>CQT-LCNN [16]</td>
<td>0.0295</td>
<td>1.23</td>
</tr>
<tr>
<td>LFCC-LCNN [16]</td>
<td>0.1053</td>
<td>4.60</td>
</tr>
<tr>
<td>DCFT-LCNN [16]</td>
<td>0.5600</td>
<td>2.06</td>
</tr>
<tr>
<td>GD gram-ResNet [37]</td>
<td>0.0439</td>
<td>1.79</td>
</tr>
<tr>
<td>SincNet-Softmax [21]</td>
<td>0.0527</td>
<td>2.11</td>
</tr>
<tr>
<td>Spec-ResNet [38]</td>
<td>0.0994</td>
<td>3.81</td>
</tr>
<tr>
<td>CQCC-ResNet [38]</td>
<td>0.1070</td>
<td>4.43</td>
</tr>
<tr>
<td>STFT-CapsNetFC [20]</td>
<td>0.1198</td>
<td>4.93</td>
</tr>
<tr>
<td>(CQT-MMPS)-LCNN [39]</td>
<td>0.0240</td>
<td>0.90</td>
</tr>
</tbody>
</table>

(ResNet), capsule network (CapsNetFC) and SincNet.

On comparing the performances of various single systems with our proposed GFCC-LCNN system in Table 2, we observe that the GFCC based front-end feature performs better than other cepstral features such as CQCC and LFCC when they are used with similar deep learning systems. In addition, our GFCC-LCNN emerges as one of the top performing single systems on the ASVspoof 2019 physical access evaluation set. Among all these single systems, the (CQT-MMPS)-LCNN shows the best performance, which is relatively much better than other systems. However, we note that it considers a hybrid front-end information in terms of magnitude and phase spectrum unlike other systems. In addition, it is noted that we do not claim that the GFCC based proposed GFCC feature is the best front-end for spoofing attack detection, rather we are interested in showing that it is one of the strongest front-ends for the detection of physical access attacks.

6. Conclusion

In this work, we studied the application of GSP in anti-spoofing by proposal of a new feature, namely, GFCC for detection of physical access attacks. The studies conducted on ASVspoof 2019 physical access database using LCNN classifier project the effectiveness of the proposed GFCC based front-end for identifying the replay attacks. In addition, it proves to be one of the strongest front-ends when compared to other existing state-of-the-art systems on the ASVspoof 2019 physical access corpus. The future work will focus on extending the studies to the detection of realistic replay attacks and other categories of spoofing attacks in ASVspoof 2021.

7. Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grants 62071242 and 62001100, in part by Shanghai Sailing Program under Grant 19YF1402000.

8. References


