Graph Fourier Transform based for Spoofing-Aware Speaker Verification

Daiyu Huang[1], *Xing Guo*[1], *Shanmin Liu*[1], *Mianxin Tian*[1], *Hongtao Zhang*[1], *Longting Xu*[1]

[1] Department of Information Science and Technology, Donghua University, Shanghai 200000, China
hdy_hdy0715@126.com, xlt@dhu.edu.cn

1. System descriptions

Proposed spoofing-aware speaker verification (SASV) system as shown in Fig. 1, which includes the automatic speaker verification (ASV) subsystem and countermeasure (CM) subsystem. The ASV subsystem and CM subsystem will produce different embeddings/scores. Then, a simple sum of the scores produced by the ASV and CM subsystems.

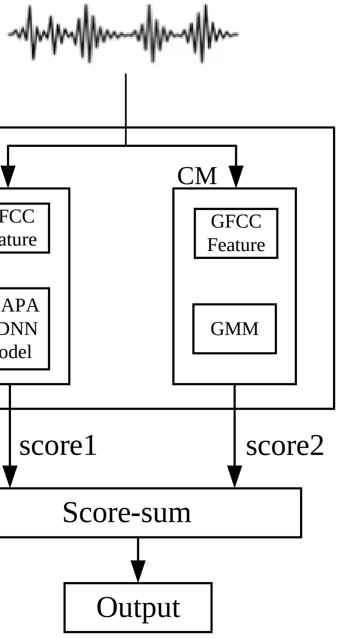

Figure 1: *The proposed SASV system flow chart.*

1.1. ASV subsystem

The same ASV subsystem as in the baseline was taken in our SASV system. The frequency MeI-Frequency Ceptral Coefficients (MFCC) features and the ECAPA-TDNN [1] were adopted for pre-trained ASV subsystem.

1.2. CM subsystem

The proposed CM subsystem use a novel feature based on graph signal processing, namely graph frequency cepstral coefficient (GFCC). Gaussian mixture model (GMM) is a clustering algorithm widely used in speaker verification in recent years, and we select it as suitable backend classifier of GFCC, the GMM structure is the same as the ASVspoof 2019 baseline. ASVspoof2019 LA train partition is used for training the system.

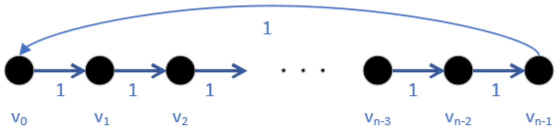

Figure 2: *Graph topology of finite time domain signals.*

1.3. SASV solutions

Score-sum ensembles using cosine similarity scores generated from speaker embeddings produced by a pre-trained ASV subsystem and the scores produced by the CM subsystem;

2. Proposed GFCC feature

In this section, we introduce the process of GFCC feature extraction, as shown in Fig. 3. To put it simply, six steps are required to obtain this feature, namely pre-emphasis, framing and construct graph signal, Graph Fourier transformation (GFT), log power spectrum and discrete cosine transform (DCT).

- Pre-emphasis: In order to enhance the amplitude of the high-frequency band of speech intentionally pre-emphasis is used in the first step. It can effectively compensate the loss of high-frequency components in the sound transmission, so as to balance the spectrum skew in speech.
- Framing: 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.
- Construct graph signal: The graph definition in graph signal processing 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 $\mathcal{G}$ is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{W})$, where $\mathcal{V}$ represents the vertex set, $\mathcal{E}$ represents edge set, and $\mathcal{W}$ represents the weight set, respectively. Generally, $\mathcal{W}$ can be represented by graph adjacency matrix $\mathcal{A}$. We construct a graph k-shift operator ψ_k to represent $\mathcal{A}$. $\mathcal{A}$ is equivalent to $\mathcal{W}$ and $\mathcal{E}$ as a 0-1 matrix, $\mathcal{G}$ can be redefined as $\mathcal{G}_{\psi_k} = (\mathcal{V}, \psi_k, \psi_k)$ [2].The graph signal y_{out} obtained after implementing ψ_k on the time domain signal y_{in} can be expressed as $\mathrm{y}_{\text{out}} = \psi_k \cdot \mathrm{y}_{\text{in}}$.

  For example, when $(y_{N-1}, y_0, y_1, \cdots, y_{N-2})^{\mathrm{T}} = \psi_1 (y_0, y_1, \cdots, y_{N-1})^{\mathrm{T}}$, the transformation of ψ_1 into

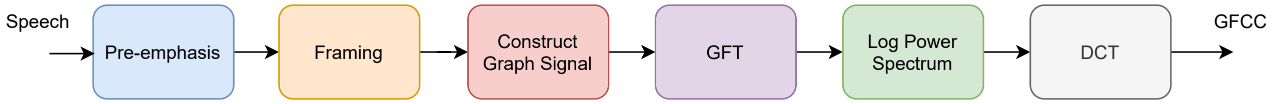

Figure 3: *Schematic diagram of proposed GFCC feature extraction.*

a matrix is expressed as follows

$$\psi_1 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & \cdots & 1 & 0 \end{bmatrix}_{N\times N} \tag{1}$$

According to the Theorem 2 in [3], the visualized topological structure can be obtained, as shown in Fig. 2.

- GFT: After knowing the graph topology, we can further study the graph frequency domain characteristics of graph signal, GFT provides the possibility for the representation of graph signals in the ”spectrum domain”. In GSP theory, different graph topologies make different graph Fourier transform bases, which will map signals to different graph Fourier domains. Therefore, graph Fourier transform is a mapping technology that can map speech to different ”frequency” domains, it focuses on the interaction between the internal structure of graph and the corresponding graph signal, and provides a perspective for the analysis of GSP eigenvalue spectrum.
- Log power spectrum: The logarithm is then taken to compress the dynamic range of the power spectrum and improve its display effect.
- DCT: Compared with discrete Fourier transform, DCT has better frequency-domain energy aggregation, and DCT can well describe the relevant characteristics of human speech signals and obtain more discrimination information. We apply the DCT to derive the cepstral coefficients from the log power spectrum.

3. Experimental results and analysis

Table 1: *The SV-EER, SPF-EER and SASV-EER for both development and evaluation protocols based on proposed SASV system and baselines.*

	SV-EER		**SPF-EER**		**SASV-EER**	
	Dev	Eval	Dev	Eval	Dev	Eval
Baseline1	14.89	35.1	6.94	0.50	2.09	19.15
Baseline2	14.38	16.01	0.01	1.23	5.41	8.75
Proposed	12.33	19.14	0.00	7.00	5.18	12.48

According to the data provided in the table above, several conclusions can be drawn as follows:

- From the performance of the development set, our proposed model outperforms the three models on both SV-EER and SPF-EER. The Proposed SV-EER of 12.33% correpondes to a relative reduction of 14.26% compared to the Baseline2 solution (14.38%). At the same time, the SPF-EER of our proposed solution achieves at 0 compared to the Baseline2's SPF-EER of 0.01%. However the Proposed performs not as well as the Baseline1 in SASV-EER which increases from 2.09% to 5.18%.
- In term of evaluation set, our proposed model performs slightly worse than on the development set. In general, the Proposed performs not as well as the better performer between Baseline1 and Baseline2 in all three assessments. Details are as follows: the Proposed SV-EER of 19.14% corresponds to a relative increment of 19.55% compared to the Baseline2 solution (19.55%). The SPF-EER and SASV-EER of the Proposed increased from 0.5% to 7% and 8.75% to 12.48% respectively compared to the top performers.
- Since our proposed solution is based on Baseline1, it is necessary to compare the performances of the two solutions. In term of SV-EER, the Proposed achieves makes progress in both Dev and Eval compared to the Baseline1, the SV-EERs of Dev and Eval dorp at 17.19% and 45.47% respectively compared to those of the Baseline1.When referring to SPF-EER, the Proposed performs better than the Baseline1 on Dev but not Eval. The SPF-EER of Dev and Eval belonging to the Proposed reduces from 6.94% to 0 and increases from 0.5% to 7.00% respectively compared to the Baseline1.Unlike performance in SPF-EER, the Proposed performs better on Eval but worse on Dev than the Baseline1. In Dev, the SASV-EER of the Proposed increases from 2.09% to 5.18% compared to the Baseline1. While, the Proposed SASV-EER of 12.48% corresponds to a relative reduction of 34.83% compared to the Baseline1 solution (19.15%).

4. Score files description

The dev_scores.txt and dev_scores.txt submitted in our email are for development and the evaluation protocols respectively.

5. References

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” 2020.

[2] X. Yan, Z. Yang, T. Wang, and H. Guo, “An iterative graph spectral subtraction method for speech enhancement,” *Speech Communication*, vol. 123, pp. 35–42, 2020.

[3] A. Gavili and X.-P. Zhang, “On the shift operator, graph frequency, and optimal filtering in graph signal processing,” *IEEE Transactions on Signal Processing*, vol. 65, no. 23, pp. 6303–6318, 2017.