Channel-wise Gated Res2Net: Towards Robust Detection of Synthetic Speech Attacks

Xu Li\textsuperscript{1}, Xixin Wu\textsuperscript{2}, Hui Lu\textsuperscript{1}, Xunying Liu\textsuperscript{1}, Helen Meng\textsuperscript{1}

\textsuperscript{1}Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, China
\textsuperscript{2}Department of Engineering, University of Cambridge, UK

\{xuli, luhui, xyliu, hmmeng\}@se.cuhk.edu.hk, xw369@cam.ac.uk

Abstract

Existing approaches for anti-spoofing in automatic speaker verification (ASV) still lack generalizability to unseen attacks. The Res2Net approach designs a residual-like connection between feature groups within one block, which increases the possible receptive fields and improves the system’s detection generalizability. However, such a residual-like connection is performed by a direct addition between feature groups without channel-wise priority. We argue that the information across channels may not contribute to spoofing cues equally, and the less relevant channels are expected to be suppressed before adding onto the next feature group, so that the system can generalize better to unseen attacks. This argument motivates the current work that presents a novel, channel-wise gated Res2Net (CG-Res2Net), which modifies Res2Net to enable a channel-wise gating mechanism in the connection between feature groups. This gating mechanism dynamically selects channel-wise features based on the input, to suppress the less relevant channels and enhance the detection generalizability. Three gating mechanisms with different structures are proposed and integrated into Res2Net. Experimental results conducted on ASVspoof 2019 logical access (LA) demonstrate that the proposed CG-Res2Net significantly outperforms Res2Net on both the overall LA evaluation set and individual difficult unseen attacks, which also outperforms other state-of-the-art single systems, depicting the effectiveness of our method.

Index Terms: channel-wise gated Res2Net, anti-spoofing, synthetic speech detection, automatic speaker verification

1. Introduction

Spoofing attacks on automatic speaker verification (ASV) have attracted ever-increasing security concerns in recent years, as they pose serious threats to essential applications of ASV, such as e-banking authentication, device activation, etc. These attacks can be categorized into human impersonation\cite{1,2}, audio replay\cite{3,4}, synthetic speech\cite{5,6} and the recently emerged adversarial attacks\cite{7-12}.

In the midst of the arms race between attack and defense for ASV, the speech community has held several ASVspoof Challenges\cite{13-15} to develop countermeasures mainly against audio replay, text-to-speech (TTS) and voice conversion (VC) attacks. ASVspoof 2019\cite{15} is the latest one that contains two sub-challenges: physical access (PA) and logical access (LA). PA considers spoofing attacks from replay while LA refers to attacks generated with TTS and VC techniques.

A model’s generalizability to unseen spoofing attacks is challenging but essential for developing reliable countermeasures\cite{16,17}. To tackle this issue, previous efforts dedicated to the design of powerful system architectures\cite{18-20} and generalized acoustic features\cite{18,21} have shown great enhancement of generalization to unseen PA attacks. However, the unseen nature of LA attacks has larger variations due to numerous available speech synthesis algorithms, and some of them are difficult to be detected, e.g. the A17 attack in the LA evaluation set\cite{15}. State-of-the-art (SOTA) countermeasures may easily overfit to the training and development sets, and lack good generalizability to unseen LA attacks\cite{17}. Hence, this work focuses on enhancing generalized detection of LA attacks.

Much promising effort has been dedicated to designing countermeasures against LA attacks\cite{20,22-25}. Das et al.\cite{26} augments the training data based on signal companding methods to enhance generalization. The RawNet2 architecture\cite{27} is applied to detect synthetic speech directly upon the raw speech waveform. Our earlier work\cite{19} leverages the Res2Net architecture to improve the model’s generalizability and demonstrates its superior detection accuracy on unseen LA attacks.

According to\cite{19}, Res2Net designs a residual-like connection between different feature groups within one block, which increases the possible receptive fields and helps the system generalize better to unseen attacks. However, such a residual-like connection is performed by a direct addition between feature groups without channel-wise priority. We argue that the information across channels within a group may not contribute to spoofing cues equally, and the less relevant channels are expected to be suppressed before adding to another feature group, so that the system can generalize better to unseen attacks.

From the above motivation, this work extends\cite{19} and proposes a novel network architecture, i.e. channel-wise gated Res2Net (CG-Res2Net). CG-Res2Net modifies the Res2Net block and enables a channel-wise gating mechanism in the residual-like connection between feature groups. This gating mechanism selects the more relevant channels while suppressing the less relevant ones to dynamically adapt to unseen attacks, hence enhances the model’s detection generalization.

Specifically, we propose and compare three possible gating mechanisms: single-group channel-wise gate (SCG), multi-group channel-wise gate (MCG) and multi-group latent-space channel-wise gate (MLCG). SCG automatically learns a channel-wise gate by a fully-connected layer, based on the current feature group. MCG differs from SCG by additionally considering the next feature group information as reference to compute the gate. Finally, MLCG modifies MCG to firstly project the features of two groups into separate latent spaces, then compute the gate based on the two latent spaces. Three gating mechanisms are integrated with Res2Net to form SCG-Res2Net, MCG-Res2Net and MLCG-Res2Net, respectively. The proposed systems are evaluated on the ASVspoof 2019 LA par-

http://dx.doi.org/10.21437/Interspeech.2021-2125
tion in terms of the performance on the overall LA evaluation set as well as individual difficult unseen attacks. Experimental results demonstrate the effectiveness of the proposed gating mechanisms. All three proposed CG-Res2Net models outperform other single, SOTA systems on ASVspoof 2019 LA evaluation, depicting the promising performance of the CG-Res2Net models.

The contributions of this work include: 1) Proposing a novel CG-Res2Net architecture which can integrate one of three different channel-wise gating mechanisms into the Res2Net block; 2) Demonstrating that three proposed CG-Res2Net models outperform Res2Net on the overall LA evaluation set as well as individual difficult unseen attacks; 3) The proposed CG-Res2Net models outperform other single, SOTA systems on the ASVspoof 2019 LA evaluation set.

The rest of this paper is organized as follows: Section 2 introduces the network structure of proposed CG-Res2Net. Section 3 and Section 4, respectively. Finally, Section 5 concludes this work.

2. Approach

2.1. Channel-wise gated Res2Net

This section introduces the network structure of proposed CG-Res2Net. CG-Res2Net modifies the Res2Net block to enable a channel-wise gating mechanism in the residual-like connection between feature groups. The comparison between the structures of Res2Net and CG-Res2Net blocks is illustrated in Fig. 1 (a) and (b). After a $1 \times 1$ convolution, both models evenly split the input feature map $X$ by the channel dimension into $s$ subsets, denoted by $x_i$, where $i \in \{1, 2, ..., s\}$. We assume that $X \in \mathbb{R}^{C \times D \times T}$ and each $x_i \in \mathbb{R}^{C_i \times D_i \times T_i}$, where $C, D$ and $T$ denote the dimensions of channel, spectrum and time, respectively. Res2Net enables a direct addition between feature groups before a $3 \times 3$ convolution. Each $y_i$ is derived as Eq. 1:

$$y_i = \begin{cases} x_i, & i = 1 \\ K_i(x_i), & i = 2 \\ K_i(x_i + y_{i-1}), & 2 < i \leq s \end{cases}$$

where each $K_i(\cdot)$ denotes a convolutional function with a parameter size of $3 \times 3$. CG-Res2Net adopts a gating mechanism in the residual-like connection. Each $y_i$ is derived as follows:

$$y_i = \begin{cases} x_i, & i = 1 \\ K_i(x_i), & i = 2 \\ K_i(x_i + z_{i-1}), & 2 < i \leq s \end{cases}$$

$$z_{i-1} = a_{i-1} \otimes y_{i-1}$$

where $z_i$ scales $y_i$ by a channel-wise gate $a_i \in \mathbb{R}^C$, and $\otimes$ denotes a channel-wise multiplication operation. We expect that such an gating mechanism gives priority to channels that contain most spoofing cues and suppresses the less relevant channels, then enhances the model’s generalizability to unseen attacks.

This work proposes three novel channel-wise gating modules to be integrated with the Res2Net block, as shown in Fig. 1 (c)-(e). The detailed functionality of each module is demonstrated in Section 2.2. Our codes have been made open-source.

2.2. Channel-wise gating mechanism

2.2.1. Single-group channel-wise gate

As shown in Fig. 1 (c), the single-group channel-wise gate (SCG) automatically learns a gate $a_i$ given the current feature group $y_i$. The mapping is achieved by a fully-connected layer $y_i$ is firstly squeezed to the channel dimension by averaging over the spectrum and time dimensions (Eq. 4), and then transformed by a fully-connected layer $W_{fc} \in \mathbb{R}^{C \times C}$ with a sigmoid activation function $\sigma$ to derive the gate $a_i$ (Eq. 5).

$$F_{ap}(y_i) = \frac{1}{D \times T} \sum_{d=1}^{D} \sum_{t=1}^{T} y_i(:,d,t)$$

$$a_i = \sigma(W_{fc}F_{ap}(y_i))$$

2.2.2. Multi-group channel-wise gate

Since the residual-like connection is operated between $y_i$ and $x_{i+1}$, it may be helpful to consider $x_{i+1}$ as reference when ap-

---

1https://github.com/lixucuhk/Channel-wise-Gated-Res2Net
by the tandem detection cost function (t-DCF) [15] and equal error rate (EER) [15]. The log-probability of the bonafide class is adopted as the score for t-DCF and EER computation. As indicated in Fig. 1 (a) and (b), all ResNet, Res2Net and CG-Res2Net models in this work integrate the SE block without explicit denotation. The hyper-parameter $s$ in the Res2Net block and $r$ in the MLCG mechanisms are both experimentally set as 4. For acoustic features, this work adopts the constant-Q transform (CQT), which achieves the best results when incorporating with SERes2Net50 [19]. The CQT is extracted with 16ms step size, Hamming window, 9 octaves with 48 bins per octave. All samples are truncated along the time axis to reserve exactly 400 frames. The samples with less than 400 frames would be extended by repeating their contents.

Training strategy: The training strategy is identical with [19], where binary cross entropy is used to train all models. Adam [29] is adopted as the optimizer with $\beta_1=0.9$, $\beta_2=0.98$ and initial learning rate being $3 \times 10^{-4}$. All models are trained for 20 epochs, and the model with lowest EER on development set is chosen to be evaluated.

4. Results

4.1. Effectiveness of channel-wise gated Res2Net

This section evaluates the effectiveness of proposed CG-Res2Net models for enhancing the generalizability to unseen attacks. Table 2 shows the EER and t-DCF performance of different systems. Notice that the robustness of a spoofing countermeasure depends on its effectiveness on detecting unseen attacks in the evaluation set. Consistent with [19], we observe that Res2Net50 performs much better than ResNet50 due to its efficient residual-like connection between feature groups. Compared with ResNet50, all three CG-Res2Net models demonstrate a superior detection performance on unseen attacks in the evaluation set. Specifically, SGC-Res2Net50 performs slightly better than ResNet50, while MCG-Res2Net50 and MLCG-Res2Net50 both show a significant improvements over Res2Net50. MCG-Res2Net50 achieves the most promising performance and outperforms Res2Net50 by a relative EER reduction of 28.8% and a relative t-DCF reduction of 29.7%. Such results verify the effectiveness of the proposed gating mechanisms, and the necessity of considering the reference information $(x_{i+1}$ in Fig. 1d and Fig. 1e) in the gating module. MLCG-Res2Net50 did not outperform MCG-Res2Net50 on the overall attacks in the evaluation set, but it has better generalizability to the most difficult unseen attack (A17) in the evaluation set, as will be discussed in Section 4.2. Finally, it is also observed that the model complexity of CG-Res2Net models is comparable to that of Res2Net50 and smaller than that of ResNet50, which verifies the efficiency of the proposed gating mechanisms.

4.2. Detection performance on each unseen attack

As mentioned in Section 3, the LA evaluation set has 11 unseen attacks (A07-A15, A17 and A18) and two attacks (A16 and A19) from the training set but trained with different data. As reported in the official ASVspoof 2019 summary [15], A17 is the most difficult, such that most submitted systems failed
to detect it. To perform a detailed system evaluation on each unseen attack, this section reports the detection accuracy on data of each attack, given the system’s operating point in terms of EER, as shown in Fig. 2. It has been observed that A17 and A18 are two difficult unseen attacks. In particular, for the most difficult A17 attack, ResNet50 has a detection accuracy below 50% and Res2Net50 only achieves an accuracy of 81.48%. All the three proposed CG-Res2Net50 models outperform Res2Net50 by a large margin. MLCG-Res2Net50 achieves the highest detection accuracy of 87.63%, which outperforms Res2Net50 by 6.15% in absolute accuracy. For the A18 attack, ResNet50 and Res2Net50 achieve a detection accuracy of 79.41% and 92.55%, respectively. The proposed MCG-Res2Net50 and MLCG-Res2Net50 outperform them with a detection accuracy of 99.21% and 98.27%, respectively. These observations verify the effectiveness of MCG-Res2Net50 and MLCG-Res2Net50 on generalization to difficult unseen attacks. For other easily detectable attacks, CG-Res2Net50 models perform comparably well with Res2Net50.

4.3. Comparison with the state-of-the-art single systems

This section compares the proposed CG-Res2Net models with some reported single, SOTA systems evaluated on the ASVspoof 2019 LA partition, including systems that are submitted to the ASVspoof 2019 competition and those reported in works afterwards (according to our best knowledge). The EER and t-DCF performance are shown in Table 3. The systems are denoted by a name that encodes the input features, system architecture and loss criteria.

We observe that existing efforts dedicated into acoustic features and data augmentation [23, 26, 30], system architecture [19, 27, 31] and loss criteria [31] have achieved very promising performance. As an extension of [19], the proposed CG-Res2Net models outperform other SOTA systems, depicting the effectiveness of the gating mechanisms within the Res2Net block. Moreover, the proposed CG-Res2Net models can be utilized as a backbone network, to be integrated with other effective strategies, e.g., loss criteria, for stronger generalization to unseen attacks.

5. Conclusion

This work proposes a novel network architecture, i.e., CG-Res2Net, to enhance the model’s generalization to unseen attacks. It modifies the Res2Net block to enable a channel-wise gating mechanism in the residual-like connection between feature groups. Such a gating mechanism dynamically selects channel-wise features based on the input, to suppress the less relevant channels and enhance the detection generalizability. Three gating mechanisms are proposed and verified to be effective in enhancing generalization. In terms of overall performance on the LA evaluation set, MCG-Res2Net achieves the best performance and outperforms the Res2Net by a relative EER reduction of 28.8% and a relative t-DCF reduction of 29.7%. On the most difficult unseen attack (A17), MLCG-Res2Net achieves the best performance, which outperforms Res2Net by 6.15% absolute detection accuracy. The proposed CG-Res2Net models outperform other single, SOTA systems on the ASVspoof 2019 LA evaluation, depicting the effectiveness of our method. Given the superior generalizability of CG-Res2Net, future work will investigate its applicability to other speech applications.

6. Acknowledgement

We sincerely thank Mr. Zhiyuan Peng from the Chinese University of Hong Kong for some meaningful discussions. This work is supported by HKSAR Government’s Research Grants Council General Research Fund (Project No. 14208718).

Figure 2: The detection accuracy on each attack, given different network architectures with EER operating points. A16 and A19 are two attacks from the training set but trained with different training data. A17 is the most difficult unseen attack [15].

Table 3: Performance comparison of CG-Res2Net models to some known state-of-the-art single systems on the ASVspoof 2019 LA evaluation set.

<table>
<thead>
<tr>
<th>System</th>
<th>EER (%)</th>
<th>t-DCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spec+ResNet+CE [24]</td>
<td>11.75</td>
<td>0.216</td>
</tr>
<tr>
<td>Spec+ResNet+CE [23]</td>
<td>9.68</td>
<td>0.274</td>
</tr>
<tr>
<td>MFCC+ResNet+CE [23]</td>
<td>9.33</td>
<td>0.204</td>
</tr>
<tr>
<td>CQCC+ResNet+CE [23]</td>
<td>7.69</td>
<td>0.217</td>
</tr>
<tr>
<td>LFCN+LCNN+A-softmax [20]</td>
<td>5.06</td>
<td>0.100</td>
</tr>
<tr>
<td>FFT+LCNN+A-softmax [20]</td>
<td>4.53</td>
<td>0.103</td>
</tr>
<tr>
<td>Raw-Audio+RawNet2+CE [27]</td>
<td>4.66</td>
<td>0.129</td>
</tr>
<tr>
<td>FG-CQTN+LCNN+CE [30]</td>
<td>4.07</td>
<td>0.102</td>
</tr>
<tr>
<td>Spec+LCGRNN+GKDE-Softmax [31]</td>
<td>3.77</td>
<td>0.084</td>
</tr>
<tr>
<td>Spec+LCGRNN+GKDE-Triplet [31]</td>
<td>3.03</td>
<td>0.078</td>
</tr>
<tr>
<td>DASC-CQTN+LCNN+CE [26]</td>
<td>3.13</td>
<td>0.094</td>
</tr>
<tr>
<td>CQTN+SERes2Net50+CE [19]</td>
<td>2.50</td>
<td>0.074</td>
</tr>
<tr>
<td>Ours: CQTN+SCG-Res2Net50+CE</td>
<td>2.43</td>
<td>0.076</td>
</tr>
<tr>
<td>Ours: CQTN+MCG-Res2Net50+CE</td>
<td>1.78</td>
<td>0.052</td>
</tr>
<tr>
<td>Ours: CQTN+MLCG-Res2Net50+CE</td>
<td>2.15</td>
<td>0.069</td>
</tr>
</tbody>
</table>
7. References


