Robust Cross-SubBand Countermeasure Against Replay Attacks

Jingze Lu¹², Yuxiang Zhang¹², Wenchao Wang¹², Pengyuan Zhang¹²

¹Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China
²University of Chinese Academy of Sciences, China
{lujingze, zhangyuxiang, wangwenchao, zhangpenguay}@hccl.ioa.ac.cn

Abstract

In the ASVspoof2021 physical access (PA) task, due to the mismatch between the simulated training data and the evaluation data from the real scenario, performance of previous top-performing countermeasure systems had a significant degradation. The main reason for this phenomenon can be attributed to a simulation-to-real gap. In this work, the effect of sim-to-real gap is investigated on different datasets for replay attacks. Differences in the frequency domain between simulated and real datasets are investigated to cross the sim-to-real gap. On the basis of our previous work, different sub-band acoustic features have different capabilities in distinguishing spoof utterrances from bonafide ones. To decrease the effect of sim-to-real gap and build a robust anti-spoofing system against the replay attacks, a cross-subband countermeasure is proposed in this work. Furthermore, we use visualized heatmap to explore the artefacts captured by model trained with cross-subband method. To verify the generalization capability of the cross-subband method on different datasets, several sets of comparative experiments were also done. The results show that our cross-subband countermeasure is robust to sim-to-real gap in the PA task, and the fusion model based on it is regarded as one of the top-performing anti-spoofing systems in the ASVspoof2021 Challenge.

Index Terms— anti-spoofing, replay attack, simulation-to-real gap, speaker verification, ASVspoof 2021 Challenge

1. Introduction

Automatic speaker verification (ASV) [1], as one type of biometric system, plays an important role in data security and passing certification in recent years. Similar to other biometric systems, it is possible to illegally verify through the ASV system by imitating or tampering acoustic characteristics. Common spoofing techniques such as Text-To-Speech synthesis (TTS) [2] and Voice Conversion (VC) [3] are effective in deceiving ASV systems. In this paper, however, more attention is given to another effective yet more feasible attacking method, the replay attacks. Requiring no expertise and only a high-quality recording device such as a smart phone, replay attacks are the most accessible approach to spoofing the ASV system for almost everyone [4]. Experimental results from the ASVspoof2019 [5] and ASVspoof2021 Challenge [6] also show that compared to TTS and VC attacks, current methods are less effective for replay attacks.

In response to the threat of a variety of audio spoofing attacks, the Automatic Speaker Verification Spoofing and Countermeasures (ASVspoof) Challenges were successfully held every two years since 2015. The ASVspoof2021 Challenge is the fourth competitive challenge of the series, including logical access (LA), physical access (PA), and speech deepfake (DF) three different tasks. Among them, the PA task focuses on the replay attacks. The 2021 PA task re-collected the evaluation dataset, recording utterances with different equipments in different sized rooms, while inherited the training dataset of 2019 PA task, which is built based on simulation [6]. During the challenge, we noticed that our previous high accuracy systems performed mediocre on the ASVspoof2021 PA evaluation dataset, even though training on exactly the same training dataset. The difference in the data distribution across datasets is considered to be the main cause of the performance degradation of systems. This kind of difference can be also named as a simulation-to-real gap, which is a common challenge in the field of deep reinforcement learning for robotics [7]. However, based on simulation, infinite amount of data with complex controls could be easily obtained [8]. For the PA task, a simulated training dataset consists of utterances with different acoustic and different replay configurations could be created at low cost. Therefore, it is necessary to develop a sim-to-real transfer method for the task of replay attacks detection.

Our previous work verified that features in different sub-bands have different capabilities in distinguishing spoof utterrances from bonafide ones in the LA task [9]. In this work, we explored the robustness of sub-band features against the sim-to-real gap and proposed an effective yet counterintuitive cross-subband countermeasure (CM) for detecting replay attacks. In general, same features are used in the training and evaluating phases of deep learning. The novel cross-subband method train data on the high-frequency part of the spectrograms and evaluate on the low-frequency part, or vice versa. The cross-subband method utilizes information carried by different subbands and is capable of weakening the effect of sim-to-real gap. In order to investigate why the cross-subband countermeasure works, we did experiments on datasets from the ASVspoof 2017 V2.0 [10], 2019 [5] and 2021 [6], as well as REMASC [11].

The rest of the paper is organized as follows: Section 2 describes the sim-to-real gap cross datasets and the proposed cross-subband countermeasure against it. Section 3 introduces the front-end features and network architectures of all single systems, as well as experimental parameter settings. While in section 4, the results and analysis of all systems are presented. The paper ends in section 5 with conclusions.

2. Cross-Subband Countermeasure

2.1. Cross-datasets Sim-to-Real Mismatch

In the PA task, spoof utterances are high-quality replays of bonafide utterances, and most of the acoustic characteristics have subtle difference between bonafide utterances and the spoof ones. Only some artefacts are introduced in the process of recording and replaying, such as additive noise and reverber-
In order to verify the effect of simulation-to-real gap on the replay attacks, experiments are operated on three commonly used anti-spoofing datasets, the PA datasets of ASVspoof 2017 V2.0, 2019 and 2021, as well as the REMASC dataset. Models are trained on the training dataset of the ASVspoof 2019 PA task, and evaluated on all four evaluation datasets. Among them, the evaluation dataset of ASVspoof 2017 V2.0, the ASVspoof 2021 PA task and the REMASC dataset contain utterances collected in real scenarios. In contrast, data in the ASVspoof 2019 PA dataset is merely generated based on simulation.

Figure 2 shows the difference between spectrogram of utterances from the PA datasets of ASVspoof 2019 and 2021. It can be clearly observed that the speech in the ASVspoof 2021 PA dataset has additive noise in all frequency bands, while speech from the ASVspoof 2019 PA dataset is much cleaner. In Figure 3, we compare the average SNR of speech across PA datasets of ASVspoof 2017 V2.0, 2019 and 2021. The figure shows that simulated data from the ASVspoof 2019 dataset has much higher SNR than data in other datasets. It is also revealed that, in the real environment, utterances have lower SNR in high-frequency sub-band.

The most obvious difference between utterances in these datasets is the energy distribution in the frequency domain. Figure 4 shows the energy distribution of bonafide and spoofing utterances in the ASVspoof 2019 training dataset and 2021 evaluation datasets. In both mentioned datasets, bonafide and spoofing utterances come in pairs. Since every spoofing replay attack has a corresponding bonafide utterance, in the frequency domain, the effect of the replay attacks could be visualized as the difference value between the energy of spoofing and bonafide utterances, named as the artefact energy in this paper. As shown in Figure 4(a), in the ASVspoof 2019 training dataset, artefacts are evenly distributed over the entire frequency band, which is obviously different from dataset collected in real scenarios, as Figure 4(b) shows. In the simulated dataset, the artefacts exhibit consistent patterns in low and high frequency bins. In the real scenarios, spoof utterances have higher energy than bonafide ones in low frequency bins, while lower in high frequency bins. The reason for this phenomenon is that the speech tend to lose energy in high frequency bins during the process of recording.
and replaying. However, the simulated methods would amplify energy in different subbands. Simulated dataset has different pattern of the artefact energy in frequency domain compared to dataset collected in real scenarios. Based on the above observations, we assume that the key to cross the sim-to-real gap is to utilize the implicit information in frequency domain. Our previous work demonstrated that features in different frequency sub-bands have different capability in distinguishing spoof utterances from bonafide ones in the LA task [9]. In this work, a counterintuitive method was found to be valid for crossing the sim-to-real gap in PA task, which is the cross-subband countermeasure. In section 2.2 we describe this method in detail and try to come up with the theory that supports it.

2.2. Cross-Subband Method

Cepstral coefficients and spectrograms can characterize the change of the speech frequency spectrum over time and are widely used as features in building CM systems. There are many traditional CM systems that use cepstral coefficients as features and Gaussian Mixture Models (GMM) as classifiers that demonstrate that the performance of cepstral features varies greatly across frequency bands for both LA and PA tasks [12]. It is also found that for anti-spoofing systems based on spectrograms and DNN, not all frequency bands of spectrograms are useful for spoofing detection in PA tasks. And the effect of different frequency bands of spectrograms varies for different training sets. [13].

The full frequency band is more suitable when the training set and evaluation set are equally distributed. In this work, to decrease the effect of sim-to-real gap, we propose a cross-subband countermeasure. By convention, models with sub-band features should use spectrograms from the same sub-band during training and evaluating. However, for the cases where sim-to-real gap exists between training data and evaluation data in the ASVspoof2021 PA task, we found that the cross-utilization of high-frequency (HF) part and low-frequency (LF) part of the original spectrograms has achieved better results. In our CM system, we trained data on the HF part of the spectrograms and evaluated on the LF part, or vice versa. Figure 1 shows the framework of our proposed cross-subband countermeasure.

Figure 5 shows the difference between bonafide utterances and spoof ones in the ASVspoof2019 simulated dataset in spectrogram. Many horizontal stripes could be observed in the spoof utterance. We believe that these stripes reflect artefacts stemming from recording and presentation devices, mainly the amplitude-frequency characteristics of the loudspeakers added in the simulation process. The ASVspoof2019 PA dataset modelled both the linear and non-linear characteristics of lots of loudspeakers [14]. Different kinds of amplitude-frequency characteristics caused the artefacts of the simulated data fluctuate widely at both high and low frequencies, as shown in Figure 4(a). However, the simulated data amplifies the influence of the loudspeaker’s amplitude-frequency characteristics on the frequency spectrum. As shown in Figure 4(b), the characteristics only have a slight effect on the low frequency bins (below 4kHz) in real scenarios, including amplify the lowest frequency band (below 1kHz) and two valley points (1kHz and 3kHz). In the high frequency subband (higher than 4kHz), the effect of amplitude-frequency characteristics on the spectrum is not obvious. Figure 6 shows the heatmaps of CNN based classifier in the high and low frequency subbands. We use heatmaps based on GradCAM [15], a visualization method, on the magnitude spectrogram to show where the CNN based model really focus on. It could be observed from the heatmaps that the CM system trained on the high frequency subband captures the horizontal stripes artefacts well. However, CM system trained on the low frequency subband could not capture such artefacts effectively.

In conclusion, the reason for the excellent performance of cross-subband method against sim-to-real gap could be explained as follows:

- During the training period, model trained with high frequency feature has stronger ability to capture the horizontal stripes artefacts.
- Such horizontal stripes artefacts, which reflect the trough of amplitude-frequency characteristic of loudspeakers, maintain the same pattern at low and high frequency bins.
- In real scenarios, horizontal stripes artefacts mainly exist in the low frequency subband.

For these reasons, the cross-subband method works well in the ASVspoof2021 PA task. Several sets of comparative experiments were also done to check the generalization capability of the proposed method, as described in section 3.
3. Experimental Setup

3.1. Dataset and Metrics

To investigate the capability of cross-subband countermeasure against sim-to-real gap across datasets, CM systems were trained on the ASVspoof 2019 training dataset and evaluated under the evaluation datasets of the ASVspoof 2021 PA task. As comparative tests, we also evaluated the CM systems under the evaluation datasets of the ASVspoof 2017 V2.0 and 2019 PA task. Meanwhile, we use the REMASC dataset to verify the robustness of cross-subband method. REMASC dataset is a realistic replay attack corpus for voice controlled systems, which collected data from different environments. In our experiments, we eliminated the data recorded outdoor or in vehicles, only retained the data recorded indoor. The reason for filtering the REMASC dataset is that all training data is simulated indoor. We want to eliminate the effect of external factors on the experiment.

Two kinds of performance metrics are used in this work. One of them is the equal error rate (EER), which is defined as the point where the false acceptance rate (FAR) and the false rejection rate (FRR) are equal. The other metric is minimum normalized tandem detection cost function (min-tDCF) [16].

3.2. Front-End and Model Architecture

The main acoustic feature used in our model is log magnitude spectrogram. The short-time Fourier Transform (STFT) spectrogram was extracted with window length 1728, hop length 130 and a 1728-point FFT. The selection of 1728 point FFT as the feature follows [17], which describe the system performs well in both LA task and PA task in the ASVspoof2019 Challenge. To maintain a consistent dimension in the time domain in each batch, we truncated or concatenated the length of each spectrogram into 600 frames. For the speech whose number of frames is lower than 600, spectrograms are flipped horizontally before concatenating to prevent interference from discontinuities. For the speech whose number of frames is higher than 600, only the first 600 frames are preserved and fed into the model.

For the cross-subband method, the spectrogram is divided into high-frequency and low-frequency, two parts of the same size. The sampling rate of all of the utterances provided in every datasets we used is 16kHz. According to the Nyquist Sampling theorem, for a given sample rate \( f_s \), perfect reconstruction is guaranteed possible for a bandlimit \( B < f_s/2 \), which is 8kHz. To keep the same size of dimensions between the training and evaluating stages, we select 4kHz as low/high frequency boundary. The shape of the final high-frequency and low-frequency features is 433×600.

In this work, SENet is used as classifier. The SENet is integration of the ResNet with the squeeze-and-excitation (SE) block, which can effectively distinguish the bonaﬁde speech from the spoof attacks. A Global Average Pooling layer is connected after the SENet block, which calculate the mean value of each channels. Due to the reason that the angular margin based softmax loss (A-softmax) [19] is used as loss function, an Angular Linear layer replace the common fully-connected layer connected after the pooling layer. The classiﬁer used in this work based on SENet34 model architecture with a (16, 32, 64, 128) number of channels.

3.3. Details of Systems Implementation

Based on the features and model architecture above, several single CM systems are prepared for the experiment:

- **low-frequency-cross-subband (LFCs):** The FFT spectrogram used for training is the lower frequency half FFT bins (0-4kHz) extracted as described in 3.2, while the FFT spectrogram for evaluating is the higher frequency half FFT bins (4-8kHz).
- **high-frequency-cross-subband (HFCs):** The FFT spectrogram used for training and evaluating is the opposite of the previous.
- **low-frequency-subband (LFS):** The FFT spectrogram used for training and evaluating are both taken from the 0-4kHz.
- **high-frequency-subband (HFS):** The FFT spectrogram used for training and evaluating are both taken from the 4-8kHz.
- **full-band (FB):** The FFT spectrogram used for training and evaluating are complete spectrogram.

In addition to the above models, we also trained some single CM models with different features and neural networks to achieve an effective fusion system:

- **FFT-SENet-DA:** The FFT spectrograms used for training and evaluating are complete spectrograms. To develop a more robust countermeasure, a data augmentation (DA) strategy is added to the training and development sets. We recorded part of the bonaﬁde utterances in the ASVspoof 2019 training and development sets, and added them to both sets.
- **MGD-SENet:** The MG gram is extracted using the same method as [20], with 800 window length and 400 hop length. Hamming window, 1024 FFT point, \( \alpha = 0.6 \), and \( \gamma = 0.3 \).
- **LFCC-SENet:** LFCC follows exactly the oﬃcial baseline provided in ASVspoof 2019 [5] to produce 60-dimensional feature vectors using a 20ms window length, 512 FFT points and 20 ﬁlters with their delta and double delta coefficients extracted.
- **CQCC-GMM-Baseline:** The baseline system of the ASVspoof 2019 [5].
- **CQCC-GMM-DA:** Model and feature is identical to the baseline. The training data is a combination of the entire ASVspoof 2019 training set and our own recorded data on the basis of the ASVspoof 2019 training set.

Loss function is minimized using Adam [21] optimizer with \( \beta_1=0.9, \beta_2=0.98, \epsilon=10^{-9} \) and weight decay \( 10^{-4} \). Warm-up steps are set to 1000, where the learning rate increases linearly. After warm-up steps, the learning rate decreases proportionally to the inverse square root of the step number. All models were trained with 36 epochs, in which the model with the lowest loss on the development dataset was selected as the final model.

4. Result and Discussion

4.1. Cross-dataset Results

As shown in Table 1, for the same CM system, the performance on the ASVspoof2019 PA dataset is quite better than that on the
Table 1: Performance of sub-band and cross-subband systems in ASVspoof 2017 V2.0, 2019, 2021 PA task and REMASC. All systems are trained with ASVspoof 2019 simulated dataset. Systems are described in detail in section 3.3. The performance metrics used are EER and min-tDCF.

<table>
<thead>
<tr>
<th>Evaluation Dataset</th>
<th>System</th>
<th>Progress Phase</th>
<th>Evaluation Phase</th>
<th>Progress Phase</th>
<th>Evaluation Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min-tDCF EER/%</td>
<td>min-tDCF EER/%</td>
<td></td>
<td>min-tDCF EER/%</td>
<td></td>
</tr>
<tr>
<td>ASVspoof 2021 PA</td>
<td>FB</td>
<td>0.8769 37.91 0.9194 40.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFS</td>
<td>0.9735 44.50 0.9631 43.19</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>0.9079 35.44 0.9093 35.77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFCS</td>
<td>0.9030 36.86 0.9276 39.04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>0.8035 31.33 0.8320 32.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFCS</td>
<td>0.1318 5.02 0.1094 4.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>0.0160 0.59 0.0190 0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFCS</td>
<td>0.0162 0.74 0.0257 0.95</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>0.0596 2.41 0.0742 3.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFCS</td>
<td>0.2829 10.65 0.2626 10.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>0.0131 5.00 0.0194 4.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASVspoof 2017 V2.0</td>
<td>FB</td>
<td>- 50.00 - 51.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFS</td>
<td>- 43.45 - 56.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>- 53.68 - 57.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFCS</td>
<td>- 44.45 - 52.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>- 46.96 - 43.76</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>REMASC</td>
<td>FB</td>
<td>- - - 51.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INDOOR</td>
<td>LFS</td>
<td>- - - 47.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>- - - 54.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LFCS</td>
<td>- - - 62.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HFS</td>
<td>- - - 44.00</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2017 V2.0 and 2021 dataset. One of the most important reasons is the sim-to-real gap between datasets. ASVspoof2019 dataset is created based on simulation methods, so the good performance on it can not demonstrate the robustness of the CM system.

On the datasets consist of real replay attacks (ASVspoof 2017 V2.0 and 2021 PA dataset), our cross-subband countermeasure did promote the robustness of CM system. On the ASVspoof2021 PA dataset, the result shows that using only sub-band features did not achieve better results than using the full spectrogram. However, by crossing the sub-bands, performance gains are obtained significantly. The best performing single model is cross-subband countermeasure that evaluated on low-frequency features and trained on high-frequency features. The low-frequency cross-subband system also shows a huge performance improvement compared to the system using the same low-frequency model but evaluated with low-frequency features.

On the ASVspoof2017 V2.0 dataset, our cross-subband countermeasure (trained on the high-frequency sub-band and evaluated on the low-frequency sub-band) also shows a great progress (7.32% EER decrease and 14.33% performance promotion) compared to the CM system using only full spectrogram. The result supports the conclusion that the cross-subband countermeasure is effective to the sim-to-real gap between datasets and can promote the robustness of CM systems. High frequency cross-subband method also shows superiority on the REMASC dataset.

Table 2 shows the results for single models and the final fusion system submitted to the PA task of the ASVspoof21 Challenge.

<table>
<thead>
<tr>
<th>System</th>
<th>Evaluation Phase</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min-tDCF</td>
<td>EER/%</td>
</tr>
<tr>
<td>FFTlow-SENet-CS</td>
<td>0.9276 39.04</td>
<td>0.1</td>
</tr>
<tr>
<td>FFTHigh-SENet-CS</td>
<td>0.8320 32.90</td>
<td>0.2</td>
</tr>
<tr>
<td>FFTHigh-SENet</td>
<td>0.9093 35.77</td>
<td>0.1</td>
</tr>
<tr>
<td>FFT-SENet</td>
<td>0.9194 40.25</td>
<td>0.1</td>
</tr>
<tr>
<td>FFT-SENet-DA</td>
<td>0.9342 40.23</td>
<td>0.1</td>
</tr>
<tr>
<td>MGD-SENet</td>
<td>0.8994 38.22</td>
<td>0.1</td>
</tr>
<tr>
<td>LFCC-SENet</td>
<td>0.9394 39.19</td>
<td>0.1</td>
</tr>
<tr>
<td>CQCC-GMM</td>
<td>0.9436 38.15</td>
<td>0.1</td>
</tr>
<tr>
<td>CQCC-GMM-DA</td>
<td>0.9735 41.08</td>
<td>0.1</td>
</tr>
</tbody>
</table>

| fusion | 0.7122 | 27.59 |

4.2. Performance Comparison and Discussion

Table 2 shows the results of four baseline models provided by the organizer of the ASVspoof2021 Challenge, three top-performing CM systems without data augmentation submitted into the ASVspoof 2021 Workshop as well as the cross-subband countermeasure.
it cause performance degradation. Among all the single model
we trained, the high frequency cross-subband single model is
the best one, and has the greatest contribution in the process of
results fusion. The fusion system reduces the minimum t-DCF
and EER by 14.40% and 16.14% respectively compared to the
best cross-subband based single system in the evaluation phase.

Table 4 shows part of the ASVspoof 2021 evaluation results
for the PA conditions [6]. The EER of our method is lower than
[24] and [23]. The reason is that compared to their models, we
neither did too much design on features and models, nor per-
formed data augmentation on the proposed method. The basi-
line model we used in the ASVspoof21 PA task in a quite simple
model, only consists of a STFT feature frontend and a ResNet34
model with SE-block. Our proposed cross-subband counter-
measure is simple but effective. The fusion system based on it
reached the top-performing level in the ASVspoof2021 Chal-
lenge.

Table 4: Part of the ASVspoof 2021 evaluation results for the
PA conditions.

<table>
<thead>
<tr>
<th>System</th>
<th>Evaluation Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min-tDCF</td>
</tr>
<tr>
<td>T07 [24]</td>
<td>0.6853</td>
</tr>
<tr>
<td>T16 (ours)</td>
<td>0.7122</td>
</tr>
<tr>
<td>T23 [23]</td>
<td>0.7296</td>
</tr>
<tr>
<td>T01</td>
<td>0.7446</td>
</tr>
<tr>
<td>T04</td>
<td>0.7462</td>
</tr>
<tr>
<td>T33</td>
<td>0.7648</td>
</tr>
<tr>
<td>T08</td>
<td>0.7670</td>
</tr>
<tr>
<td>T37</td>
<td>0.8216</td>
</tr>
<tr>
<td>T27</td>
<td>0.8307</td>
</tr>
<tr>
<td>B01</td>
<td>0.9434</td>
</tr>
</tbody>
</table>

5. Conclusion

In this work, to deal with the sim-to-real gap between the sim-
ulated datasets and the ones collected in real scenarios for the
replay attacks, a cross-subband countermeasure is proposed.
Low-frequency and high-frequency sub-bands have different
ability of distinguishing genuine utterances from spoof ones,
which provide the basis of cross-subband method. After cross-
ing features from different frequency sub-bands in training and
evaluating, single countermeasure systems achieve better per-
formance. Fusion model based on our cross-subband CM sys-
tem is regarded as one of the top-performing systems for the
ASVspoof2021 Challenge.

6. References

[1] Douglas A Reynolds, “Speaker identification and verifi-
cation using gaussian mixture speaker models,” Speech
inig vulnerability of voice verification systems to spoof-
ing attacks by means of a tts system,” in International
132–137.
lak, Eng Siong Chng, and Haizhou Li, “Vulnerability
of speaker verification systems against voice conversion
spoofing attacks: The case of telephone speech,” in 2012
IEEE International Conference on Acoustics, Speech and
[4] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi
Yamagishi, Federico Alegre, and Haizhou Li, “Spooing
and countermeasures for speaker verification: A survey,”
Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi
Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and
Kong Aik Lee, “ASVspoof 2019: Future Horizons in
Spoofed and Fake Audio Detection,” in Proc. Interspeech
2019, 2019, pp. 1008–1012.
ing progress in spoofed and deepfake speech detection,” in
Proc. 2021 Edition of the Automatic Speaker Verification
and Spoofing Countermeasures Challenge, 2021, pp. 47–
54.
[7] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Wester-
lund, “Sim-to-real transfer in deep reinforcement learn-
ing for robotics: a survey,” in 2020 IEEE Symposium Se-
ries on Computational Intelligence (SSCI). IEEE, 2020,
pp. 737–744.
[8] Mengyuan Yan, Qingyun Sun, Iuri Frosio, Stephen Tyree,
and Jan Kautz, “How to close sim-real gap? transfer with
[9] Yuxiang Zhang, Wenchao Wang, and Pengyuan Zhang,
“The Effect of Silence and Dual-Band Fusion in Anti-
4279–4283.
[10] Héctor Delgado, Massimiliano Todisco, Md Sahidullah,
Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, and Ju-
nichi Yamagishi, “Asvspoof 2017 version 2.0: meta-data
analysis and baseline enhancements,” in Odyssey 2018-
The Speaker and Language Recognition Workshop, 2018.
[11] Yuan Gong, Jian Yang, Jacob Huber, Mitchell MacKnight,
and Christian Poellabauer, “Remasc: Realistic replay at-
tack corpus for voice controlled systems,” Proc. Inter-
[12] Hemlata Tak, Jose Patino, Andreas Nautsch, Nicholas
Evans, and Massimiliano Todisco, “Spoofing Attack De-
tection Using the Non-Linear Fusion of Sub-Band Classi-
[13] Bhusan Chettingi, Tomi Kinnunen, and Emmanouil Benetos,
“Subband modeling for spoofing detection in automatic
speaker verification,” in Proceedings of Odyssey 2020:
The Speaker and Language Recognition Workshop. ISCA,
2020.
[14] Xin Wang, Junichi Yamagishi, Massimiliano Todisco,
Héctor Delgado, Andreas Nautsch, Nicholas Evans,
Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik
of synthesized, converted and replayed speech,” Com-
[15] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek
Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Ba-
tra, “Grad-cam: Visual explanations from deep networks
via gradient-based localization,” in Proceedings of the
IEEE international conference on computer vision, 2017,
pp. 618–626.


