



A Deep One-Class Learning Method for Replay Attack Detection

Yijie Lou, Shiliang Pu, Jianfeng Zhou, Xin Qi, Qinbo Dong, Hongwei Zhou

Hikvision Research Institute, Hangzhou, China

louyijie@hikvision.com, pushiliang.hri@hikvision.com

Abstract

Replay-attack is a serious issue for automatic speaker verification (ASV) and recently lots of countermeasures have been proposed to protect ASV from spoofing attacks. Traditional countermeasures are a binary-classification system which was observed to have limited generalization on unseen attacks. One-class learning methods which have been widely used in anomaly detection is a promising method to enhance the robustness of replay detection system. In this paper, we propose a deep one-class learning scheme to model the genuine speeches in a compact embedding space. To reduce the variance of genuine embedding space, we design an architecture unit, called residual variability block, which can be flexibly integrated into usual convolutional neural networks. Comprehensive experiments show that the proposed deep one-class learning scheme is effective for replay attack detection under cross-database scenarios. Besides, in our internal collected dataset, such a scheme shows better robustness under mismatched conditions between the enrollment and test phase.

Index Terms: one-class learning, replay attack detection, automatic speaker verification

1. Introduction

In the last few years, automatic speaker verification (ASV) has made significant progress and has been widely used in biometric authentication. However, ASV suffers a great threat due to varieties of spoofing attacks. Generally, there are four kinds of spoofing attacks including impersonation, text-to-speech (TTS), voice conversion (VC), and replay attacks [1]. Attacks based on TTS or VC technology require attackers have professional skills in speech synthesis. Replay-attack, however, is a relatively low-technology attack. Nowadays, with smartphones becoming increasingly popular, individuals can easily record and play back speech samples to attack ASV systems. Besides, high-quality recording and playback devices make the spoofing speech much closer to genuine speech, which poses a high threat to ASV systems [2, 3].

To protect ASV system from spoofing attacks, automatic speaker verification and spoofing countermeasures (ASVspoo) was first organized in 2015 [4] and ASVspoo 2021 was the 4th biannual competitive challenge [5]. The replay attack detection task was first introduced in ASVspoo 2017 [3], where the spoofing data were collected from real replayed attacks. In ASVspoo 2019, a controlled setup where replay attacks simulated using a range of real replay devices and carefully controlled acoustic conditions was adopted with the aim of bringing new insights into the replay spoofing problem [6]. In ASVspoo 2021, participants were required to use ASVspoo 2019 training data for model training, and evaluate on more realistic conditions [5]. The mismatch between training and test database requires the model to have generalization capability for unseen data. Results showed that in such a mismatched scenario,

replay-attack task becomes more challenging as compared to the one in previous ASVspoo challenges.

The poor generalization capability for replay-attack systems has been aroused increasing attention from earlier years [7–9]. This serious issue limits the technology to practical use. The generalization problem is mainly attributed to two aspects. The first one is pointed out by the work [10], where authors observed that spoofing audio tends to have longer silences at the end than genuine ones in ASVspoo 2019 dataset. Such an uneven distribution of silence duration may lead models to learn this irrelevant cue. This issue can be solved by using voice activity detection (VAD) to cut the silence or rearranging the silence to be evenly distributed for genuine and spoofing audio. The second one is that spoofing data is distributed in a large scope of variability space due to complicated attack configurations (various types of recording and playback devices) and authentication configurations (various types of environments and microphones). In reality, we cannot collect all kinds of data to train the model. Even so, the model may not be able to suppress the unrelated information focusing only on the replayed information. To make the model more generalizable to unseen replay attacks, some works have shown that one-class learning is a promising method where only genuine embedding space is modeled and various spoofing data are naturally distributed apart from genuine ones. [9, 11–13].

Making a compact genuine feature space is essential for one-class learning. For shallow models, substantial feature engineering is commonly required [9, 11, 13]. For example, works [9] utilized enrollment data to suppress the data variability of unrelated factors such as speaker traits and authentication recording devices. However, the underlying assumption is that enrollment and test speech would be under the same authentication environment and microphone. In reality, the enrollment and test speech exists mismatched situation, which may degrade the performance of shallow one-class model such as GMM, OC-SVM [14], KDE [15]. Recently, some one-class learning methods based on neural networks have been proposed, such as OC-Softmax loss [12], Deep-VGG [16], VAE [17]. These deep learning models present a way to automatically learn deep feature representation and shows a powerful ability to deal with complicated data distribution. In this work, we propose a deep one-class learning scheme aiming to enhance the model generalizability under cross-database replay detection scenario and meanwhile improve the performance on mismatched conditions. Specifically, we first use OC-Softmax loss to build a one-class learning setup, which makes genuine speech embeddings have a compact boundary while spoofing data are kept away from the genuine data by a certain margin. To reduce the genuine feature variability, we design a Residual-Variability (RV) block which computes the residual between enrollment and test deep features. We expect that such residual features can suppress the unrelated variability and highlight the replay information.

The rest of the paper is organized as follows. In section

2, we review the long-term averaged spectrum based residual variability (LTAS-RV) feature which was proposed in work [9]. Such an important concept will be later extended to build the RV block. Section 3 describes the proposed deep one-class learning scheme. Detailed implementations are provided in section 4. Section 5 describes and discusses the results based on our proposed method. Conclusions are provided in section 6.

2. Related Work

To enhance the model generalizability for cross-database replay attack detection, authors used GMM to perform one-class classification [9]. To effectively build the one-class model, the authors proposed the feature named LTAS-RV to reduce the variability of genuine speeches. The feature extraction can be schematically shown in Fig. 1(a). Specifically, long-term averaged spectrum (LTAS) is first estimated by averaging the log-magnitude spectrogram along the time axis to reduce the short-term variability. After that, the LTAS-RV is generated by making the spectrum difference between the enrollment and test LTAS, which aims to reduce the long-term variability.

3. Deep one-class learning scheme

The traditional network based systems for replay detection execute binary classification. Specifically, network first receives a single speech feature and calculate the corresponding embedding. Then a binary classification loss function makes the embedding discriminative for genuine and spoofing speeches. These methods may lead the model to overfit to known replay attacks. To tackle this problem, we build a one-class learning scheme. Firstly, OC-Softmax loss function is used to optimize the embedding in a compact space for the genuine speech and keep spoofing embedding away from genuine one. Secondly, we consider that speech contains much non-replay information such as speaker traits, speech contents, etc. These unrelated information increases the variability in embedding space, which is a serious problem for one-class modeling. To suppress the embedding variability for genuine speech, the Residual-Variability (RV) block is designed which encodes the residual information between enrollment and test utterance. Here, the test utterance is to be detected whether it belongs to genuine speech or not, while the corresponding enrollment utterance is a reference to the test one, and is always assumed as genuine speech. We note that the proposed RV block can be flexibly integrated into standard deep convolutional neural network (CNN) systems such as VGGNet [18], ResNet [19], etc.

In this section, we first briefly review the OC-Softmax loss function, then the architecture of RV block is introduced. Finally, we discuss the proposed network structure called ResNet-RV to fulfill the residual embedding extraction.

3.1. One-class Softmax loss

When dealing with logical spoofing attack detection task, Zhang et al. found that the model may overfit to known attacks if we train a compact embedding space for spoofing speech just as the one for genuine speech [12]. From this point of view, they designed a loss function called one-class Softmax (OC-Softmax) loss to learn a feature space in which the genuine speech embeddings have a compact boundary while spoofing data are kept away from genuine data by a certain margin. Such a one-class learning method behaves robust when the data distribution is mismatched between training and test for spoofing speeches.

The formulation of OC-Softmax loss function is denoted as follows.

$$L_{OCS} = \frac{1}{N} \sum_{i=1}^N \log \left(1 + e^{\alpha(m_{y_i} - \mathbf{w}_0 \mathbf{x}_i)(-1)^{y_i}} \right) \quad (1)$$

Where α is a scale factor and $y_i = \{0, 1\}$ represents the genuine and spoofing classes. The vector \mathbf{w}_0 represents the direction for target class embeddings. The loss function aims to minimize the angle θ between embedding \mathbf{x}_i and \mathbf{w}_0 for genuine speech ($y_i = 0$) meanwhile maximize the angle θ for spoofing speech ($y_i = 1$). Two margins ($m_0, m_1 \in [-1, 1], m_0 > m_1$) are hyper-parameters which can be tuned to constrain the embedding space for the two classes.

3.2. RV block

We extend the procedure of LTAS-RV feature extraction to design the RV block structure which is depicted in Fig. 1(b). As we can see, the RV block supports multiple feature map pairs as input. These feature map pairs can be extracted from different layers of neural networks. For each pair, the block accepts the two feature maps from enrollment and test utterances, which we denote as $X_{in} \in \mathbb{R}^{C \times T_X \times D}$ and $Y_{in} \in \mathbb{R}^{C \times T_Y \times D}$ respectively. Here the C represents the number of channels, $T_{X(Y)}$ and D denote the dimensions along the time and frequency axis respectively. The RV block then computes the average pooling for the two feature maps through spatial dimensions (time and frequency axis). Formally, the output vectors $X_{out}, Y_{out} \in \mathbb{R}^C$ are generated by shrinking X_{in} and Y_{in} to the same dimension, which can be mathematically expressed as follows.

$$X_{out}(i) = \frac{1}{T_X D} \sum_{m=1}^{T_X} \sum_{n=1}^D X_{in}(i, m, n) \quad (2)$$

$$Y_{out}(i) = \frac{1}{T_Y D} \sum_{m=1}^{T_Y} \sum_{n=1}^D Y_{in}(i, m, n) \quad (3)$$

After executing the average pooling process for all pairs, the RV block concatenates all these output vectors to form the feature representations for enrollment and test utterances respectively. Finally, the difference between the two feature representations is calculated to generate the residual embedding. To stabilize the training process, the two normalization operations called L_2 normalization and batch normalization are applied before and after the subtraction process respectively.

3.3. ResNet-RV architecture

The RV block can be integrated into many state-of-the-art CNN architecture to suppress the embedding variability for genuine speech. In this work, we use ResNet as backbone and build the ResNet-RV architecture to evaluate the effectiveness of the proposed RV block. The detailed configuration of ResNet can be referred to Table 1. The ResNet-RV architecture receives the pair of enrollment and test feature maps as input, and extract residual embedding where the information unrelated to replay-attack is suppressed. A straightforward way to construct ResNet-RV architecture is to integrate RV block at the last layer of ResNet. Considering the work [9], researches used handcrafted features such as CQCC [20], LFCC for residual variability extraction, which shows discriminative for genuine and spoofing speeches. We believe that more shallow

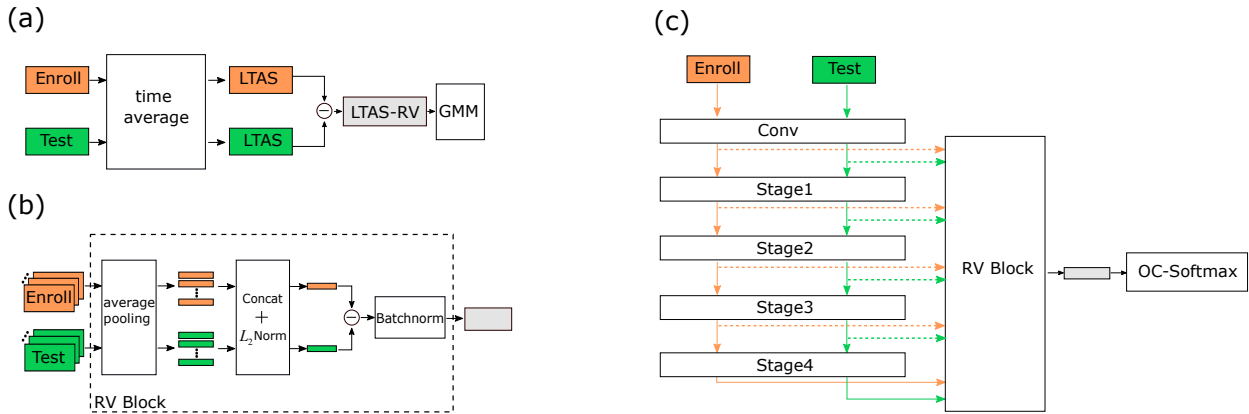


Figure 1: (a) Schematic diagram for residual variability feature extraction and one-class GMM modeling. (b) The structure of RV block. (c) Illustration of the proposed ResNet-RV architecture.

feature maps contain discriminative information and may contribute towards more robust residual embedding. In the field of ASV, varieties of multi-layer feature aggregation structures were proposed [21–23] and shown effectiveness for ASV systems. Hence, we borrow such an idea to design the ResNet-RV architecture which is shown in Fig. 1(c). As we can see, the paired feature maps with different resolutions are extracted from the convolution layer and the following 4 stages of ResNet. Then the RV block processes these feature maps to generate the final residual embedding. Under the multiple enrollment utterances scenario, the test utterance is paired to each enrollment one to generate the multiple residual embeddings. Then we just average them as the final embedding.

4. Experiment

4.1. Datasets

All experiments in this work were conducted based on ASVspooft2017 V2.0 (AS17) database [2], ASVspooft2019 physical access (AS19-PA) database [6], and Hikspeech-PA dataset. The details of the three datasets are illustrated as follows.

- **AS17 and AS19-PA:** The open-sourced datasets were released in the ASVspooft2017 and ASVspooft2019 challenges. For AS17 dataset, all genuine utterances come from a subset of the original RedDots corpus, while the spoofed utterances are the result of replaying and recording the RedDots utterances with heterogeneous devices and acoustic environments. For AS19-PA dataset, both genuine and spoofing utterances are generated according to a simulation of their presentation to the microphone of an ASV system within a reverberant acoustic environment.
- **Hikspeech-PA:** The AS17 and AS19-PA datasets are not suitable for studying the mismatched situation where authentication conditions (recording devices and environments) are different in the enrollment and test phase. Such mismatched scenarios may be harmful to the system where enrollment data are utilized for replay detection. To evaluate the performance in such a mismatched condition, we constructed the Hikspeech-PA dataset where data were collected from the real world. First, the enrollment data were collected from 20 speak-

ers. We used Huawei smart phone as the recording device and chose office as the acoustic environment. Then, we prepared test data which contain genuine and spoofing utterances. The genuine utterances were collected from the same 20 speakers with two recording devices (Huawei and Xiaomi smart phones) and three types of environments (office, outdoor and meeting room). The spoofing utterances collecting process was similar to the one used in ASVspooft 2017. Specifically, we used a laptop to play back these recorded test data and recorded them again with different configurations including recording devices (Nova3i, Oppo, Iphone), recording distances (10cm, 100cm), and room sizes (small and large meeting room). For each speaker, the number of both enrollment and test (genuine) utterances are 5, and each utterance contains a random string of 8 digits.

4.2. Baseline system

In this work, we opted for the standard ResNet with Softmax loss as the baseline. ResNet is similar to a standard multi-layer CNN, but with added skip connections such that the layers add residuals to an identity mapping on the channel outputs. In this paper, we experiment with ResNet-18 architecture, and modify the layers to adapt the feature input. The detailed architectures for ResNet-18 are shown in Table 1.

Table 1: The detailed configuration for ResNet-18 architecture

Layer	Out Channel	Blocks	Stride	Out shape
Conv	16	-	1	$16 \times T \times 257$
Stage1	16	2	1	$16 \times T \times 257$
Stage2	32	2	2	$32 \times T/2 \times 129$
Stage3	64	2	2	$64 \times T/4 \times 65$
Stage4	128	2	2	$128 \times T/8 \times 33$

4.3. Feature engineering

Before executing the feature extraction process, we applied VAD operation to all the raw waveform data. One reason is that the VAD operation is commonly applied in front-end of real ASV system. VAD operation thus should be introduced into the replay detection system to better work in conjunction with ASV system [24]. The second reason is that uneven distribution

Table 2: The in-domain (AS17 and AS19-PA) and out-of-domain (Hikspeech-PA) performances were evaluated for models trained on the combination of AS17 and AS19-PA training set. The enrollment data of Hikspeech-PA dataset were recorded by Huawei phone in the office environment, and test data were collected under various configurations to form mismatched conditions.

Model	Loss	AS17	AS19-PA	Hikspeech-PA					
				Huawei			Xiaomi		
				Office	Outdoor	Meeting	Office	Outdoor	Meeting
ResNet	Softmax	12.63	6.30	8.85	8.76	19.00	8.89	6.21	12.21
ResNet-RV	OC-Softmax	7.40	5.90	5.00	13.01	12.00	3.02	2.99	7.02

of silence duration is observed in work [10], which may lead the model to learn this unrelated information. Therefore, we want to remove this effect by cutting all silence segments.

For feature extraction, the spectrogram were generated in a sliding window fashion using a povey window of width 25ms and step 10ms. The dimension of the spectrogram is 257.

4.4. Data preparation

- **Training set:** The original training and development of AS17 and AS19-PA datasets were used for model training. We first extended all spectrogram to 400 frames by repeating their contents. Then the extended feature maps were broken down into segments with length 200 frames and overlap 20 frames. For each training batch, we randomly sampled 32 speakers from training data. For each speaker, we randomly selected two genuine segments as a positive pair. Besides we also randomly sampled one genuine and one spoofing segment as a negative pair.
- **Test set:** The original evaluation set of AS17 and AS19-PA as well as Hikspeech-PA were used to evaluate the model performance. For AS17 and AS19-PA evaluation sets, we partitioned them into enrollment and test sets. The enrollment sets were constructed by randomly selecting one genuine speech for each speaker. Meanwhile, we excluded these enrollment utterances from evaluation sets to form the test sets. For Hikspeech-PA dataset, we used 5 utterances as enrollment data for each speaker.

4.5. Details of Systems Implementation

We used PyTorch framework to train the ResNet-RV network. Training used SGD optimizer with a momentum of 0.9 and weight decay of $1e-2$. We adopted ReduceLROnPlateau scheduler with a frequency of validating every 500 iterations, the patience is 4, and the decay factor is 0.5. The learning rate is initially set to $1e-3$ and the minimum learning rate is $9e-7$. During training, SpecAugment method [25] including time masking and frequency masking was used to enhance the model robustness. For hyper-parameters in OC-Softmax loss function, we set $\alpha = 20$, $m_0 = 0.9$ and $m_1 = 0.2$, based on the observations made in [12].

5. Results and Analysis

The performances under various systems are shown in Table 3. Firstly, we conducted several ablation experiments under publicly available AS17 and AS19-PA datasets to demonstrate the effectiveness for RV block and OC-Softmax loss under cross-dataset scenarios. When the baseline system (ResNet with Softmax loss) trained on AS17 dataset, the performance shows reasonable on AS17 test dataset, but degrades drastically on AS19-PA test dataset. When OC-Softmax loss is implemented, we observed that the performances were improved both on the two

datasets. The performance is further enhanced if the proposed RV block is used. When the above systems were trained on AS19-PA dataset, we observed that these systems obtain similar performances on the in-domain dataset. While, for out-of-domain dataset, these systems exhibit consistent performance gain as compared to the ones trained on AS17 dataset. Such a phenomenon demonstrates that the binary classification models have a good discrimination ability to detect known attacks but cannot work well on unknown attacks. As a comparison, the proposed one-class learning scheme can effectively enhance the generalization capability under the cross-database scenarios.

Table 3: Ablation studies for the effectiveness of the RV block and OC-Softmax loss on in-domain and out-of-domain test datasets. The performance is evaluated using the equal error rate (EER).

Model	Loss	Training Set	Test Set	
			AS17	AS19-PA
ResNet	Softmax	AS17	14.99	39.01
ResNet	OC-Softmax		14.45	37.66
ResNet-RV	OC-Softmax		10.38	32.58
ResNet	Softmax	AS19-PA	43.56	5.51
ResNet	OC-Softmax		41.92	5.83
ResNet-RV	OC-Softmax		35.40	5.60

Secondly, we also tested the proposed system on Hikspeech-PA dataset to evaluate the performance on enrollment and test mismatched scenarios (Table 2). The enrollment data were recorded by Huawei phone in office environment. The test data were collected under various configurations. Here we combined the AS17 and AS19-PA training sets to train the models. We observed that the proposed scheme outperforms the baseline under in-domain datasets (AS17 and AS19-PA) and almost all cases under the mismatched conditions. Such results indicate that the ResNet-RV architecture can be learned to suppress the unrelated information such as recording devices and environments.

6. Conclusion

In this paper, we propose a deep one-class learning scheme and verify the effectiveness of enhancing the robustness on the out-of-domain replay attack scenarios. Besides, the proposed system also shows superior performance under various mismatched scenarios. Although acquiring performance gain, experimental results show that the out-of-domain replay detection is still a challenging task. The proposed deep one-class learning approach suggests a promising framework for further investigations. One avenue for our future research is to design a more effective RV block which is crucial for one-class learning. Besides, more powerful one-class learning loss functions are worth designing to make the system better represent genuine embedding space.

7. References

- [1] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.
- [2] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. A. Lee, and J. Yamagishi, "Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *Odyssey 2018-The Speaker and Language Recognition Workshop*, 2018.
- [3] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.
- [4] Z. Wu, T. Kinnunen, N. Evans, J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Sixteenth annual conference of the international speech communication association*, 2015.
- [5] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A. Nautsch, J. Patino, M. Sahidullah, M. Todisco, X. Wang *et al.*, "Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *arXiv preprint arXiv:2109.00535*, 2021.
- [6] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:1904.05441*, 2019.
- [7] P. Korshunov and S. Marcel, "Cross-database evaluation of audio-based spoofing detection systems," in *Interspeech*, no. CONF, 2016.
- [8] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Cross-domain replay spoofing attack detection using domain adversarial training," in *Interspeech*, 2019, pp. 2938–2942.
- [9] X. Cheng, M. Xu, and T. F. Zheng, "Cross-database replay detection in terminal-dependent speaker verification," *Proc. Interspeech 2021*, pp. 4274–4278, 2021.
- [10] B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramírez, E. Benetos, and B. L. Sturm, "Ensemble models for spoofing detection in automatic speaker verification," *arXiv preprint arXiv:1904.04589*, 2019.
- [11] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The dku-cmri system for the asvspoof 2021 challenge: Vocoder based replay channel response estimation," *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pp. 16–21, 2021.
- [12] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.
- [13] F. Alegre, A. Amehraye, and N. Evans, "A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. IEEE, 2013, pp. 1–8.
- [14] B. Schölkopf, R. C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," *Advances in neural information processing systems*, vol. 12, 1999.
- [15] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.
- [16] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.
- [17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [20] M. Todisco, H. Delgado, and N. W. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Odyssey*, vol. 2016, 2016, pp. 283–290.
- [21] Z. Gao, Y. Song, I. V. McLoughlin, P. Li, Y. Jiang, and L.-R. Dai, "Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system," in *INTERSPEECH*, 2019, pp. 361–365.
- [22] B. Desplanques, J. Thienpondt, and K. Demuyne, "Ecapattmn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [23] Y.-Q. Yu and W.-J. Li, "Densely connected time delay neural network for speaker verification," in *INTERSPEECH*, 2020, pp. 921–925.
- [24] Y. Zhang¹², W. Wang¹², and P. Zhang¹², "The effect of silence and dual-band fusion in anti-spoofing system," 2021.
- [25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.