Data augmentation and post selection for improved replay attack detection

Yuanjun Zhao, Roberto Togneri and Victor Sreeram
School of Electrical, Electronic and Computer Engineering, The University of Western Australia (UWA), Australia
E-mail: yuanjun.zhao@research.uwa.edu.au, (roberto.togneri, victor.sreeram)@uwa.edu.au

***Abstract*—Vulnerabilities of the Automatic Speaker Verification (ASV) technology have been recognized and have generated much interest to design anti-spoofing detectors. Replay attacks pose a severe threat due to the relative difficulty for detection and the ease in mounting spoofing attacks. In this paper, a high performing spoofing detection countermeasure is presented. Deep Learning (DL) based speech embedding extractors and a novel data augmentation approach are combined to improve the detection performance. To select augmented samples with high quality and diversity and avoid the bias caused by human subjective perception, we propose the use of a Support Vector Machine (SVM) based post-filter. With the generated extra informative training data, problems of over-fitting and lack of generalization can be significantly alleviated. Experimental results measured by equal error rates (EERs) indicate a relative improvement of 30% on the development and evaluation subsets. This provides the motivation for the proposed audio data augmentation and also promotes the future research on generated samples selection in the application of speaker spoofing detection.**

I. Introduction

The Automatic Speaker Verification (ASV) system has been widely embedded in consumer electronics and security check scenarios as a reliable solution to person authentication [1]. Recently, the ASV technology itself is encountering intractable security problems in regards to spoofed speech attacks [2]. Among all the spoofing approaches, replay attacks can be performed with accessible devices like mobile phones and this ease poses a significant threat to ASV systems [3]. Due to the success of deep learning (DL) technology in classification and recognition tasks, it is a powerful motivation to apply Deep Neural Networks (DNNs) for ASV anti-spoofing tasks [4]. DL based architectures are usually adopted for detecting spoofing attacks. In the example works of [5], [6] the best single system performance was achieved by employing DNNs to extract speech embeddings for back-end classification tasks.

However, for acoustic models trained by deep learning based systems, performance degradation is often observed when the training data are insufficient or imbalanced [7]. With insufficient training data, the training process will be trapped with the issue of over-fitting to the seen data, which leads to a model of lower generalization ability to unseen data [8]. Table I shows the statistics of the ASVspoof 2017 2.0 corpus [9]. A total of 13306 speech samples were collected in the evaluation subset, which contains data with a wider range of variations than present in the training subset. This condition is not friendly to deep learning based systems, although the purpose is to encourage research in more generalized countermeasures. Integrating audio augmentation with speaker anti-spoofing systems is still an under-explored direction in the community. In addition, it is also not clear whether every data point generated would have equal impact in classifier performance [10]. For replay spoofing detection the unknown variations in the quality of the replay attack recording and playback make it more difficult to determine this [3]. As a consequence there is a pressing need for a post selection process to only keep generated data of high quality. A related work in [11] showed the benefit of the data augmentation strategy on the first version of the challenge corpus. However, generated samples were only augmented for the replay subset, which potentially causes an imbalanced training subset.

TABLE I
Statistics of the ASVspoof 2017 Version 2.0 database

Subset	# Spk	# Replay sessions	# Replay Config	# Utterances	
				Bona fide	Replay
Training	10	6	3	1507	1507
Dev.	8	10	10	760	950
Eval.	24	161	57	1298	12008
Total	42	177	61	3565	14465

In this paper, we propose a replay detection countermeasure based on data augmentation and post selection. The whole training subset of the ASVspoof 2017 Version 2.0 corpus is used to generate more speech samples with preserved labels. Moreover, an SVM based post selection process [10], [12] is applied to select generated samples. The contributions of this work are threefold. Firstly, this work proposes a solution to the more challenging problem of data augmentation of both the Bona fide and replay data for speaker anti-spoofing. Then, for the first time, the state-of-the-art SVM based post selection process is combined with traditional data augmentation approaches to select generated speech samples of high quality. Consequently, by integrating the proposed countermeasure, the performance of two efficient deep learning based spoofing detection benchmark systems are significantly improved. Experimental results confirm the advantage of data augmentation and post selection.

The rest of the paper is organized as follows. The detailed concepts and the proposed framework are introduced in section II. In section III, the experimental results and relevant discussion are given. Finally, a conclusion of this paper is detailed in section IV.

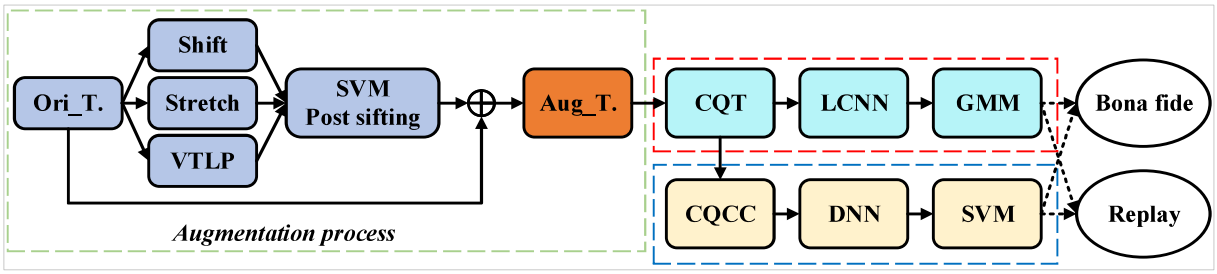

Fig. 1. Block diagram of the replay anti-spoofing countermeasure proposed in this work. The system in red dashed box is the CQT-LCNN-GMM system [5] and the one in blue dashed box is the CQCC-DNN-SVM system [6].

II. Data Augmentation and SVM Post Selection

A. Data augmentation

In this work, we apply the time-frequency representation (TFR) of the constant Q transformation (CQT) and the Constant Q cepstral coefficient (CQCC) as acoustic features. CQT based TFRs and CQCCs give a higher frequency resolution for lower frequencies and a higher temporal resolution for higher frequencies [13]. These constant Q transformation based features have been widely adopted to detect spoofing attacks and have demonstrated excellent accuracy over other traditional acoustic features based approaches such as the Mel-frequency cepstral coefficients (MFCCs) [14], [15].

The scheme of the data augmentation process is shown as a part of the entire countermeasure as in Fig.1. Several audio augmentation methods are used to generate new speech samples. In the simplest cases, the original audio files are loaded into the computer and transformed with straightforward variations such as shifting and stretching. The ratio parameters of stretching are set to 1.2 and 0.8, which provides extensional and compressed speech signals. For the shifting transformation, the number of the sampling points shifted are set randomly to maximize the diversity of produced speech files. Transformed audio files are sifted and assembled with the original training subset to build the augmented training dataset.

Another straightforward but effective way for audio data augmentation is the vocal tract length perturbation (VTLP), which is derived from the vocal tract length normalization (VTLN) [16]. For VTLP, a random warp factor α ranging between 0.9 and 1.1 is selected for each utterance. Then the frequency axis of this utterance is warped such that a frequency f is mapped to a new frequency f' using the following approach:

$$f' = \begin{cases} f\alpha & f \le F_{hi}\frac{\min(\alpha,1)}{\alpha} \\ S/2 - \frac{S/2 - F_{hi}\min(\alpha,1)}{S/2 - F_{hi}\frac{\min(\alpha,1)}{\alpha}}(S/2 - f) & \text{otherwise} \end{cases}$$

where S denotes the sampling frequency and F_{hi} is a boundary frequency chosen to cover the formant regions of interest. In this work, the sampling frequency is 16kHz and the F_{hi} is set to 4800.

B. SVM post selection

An SVM based post-filter is applied in this work to sift away generated samples with low impact for the training process. The generated samples with the same class label are collected together as the generation pool, which is a mixture of various audio files. However, not all of the generated samples are highly correlative with the original ones. To solve this problem of sample selection after data augmentation, a sifting process was introduced and achieved promising results in the application of acoustic scene classification [10]. We employ the similar SVM based iterative sifting process to refine the augmented samples and the process is depicted in Fig.2.

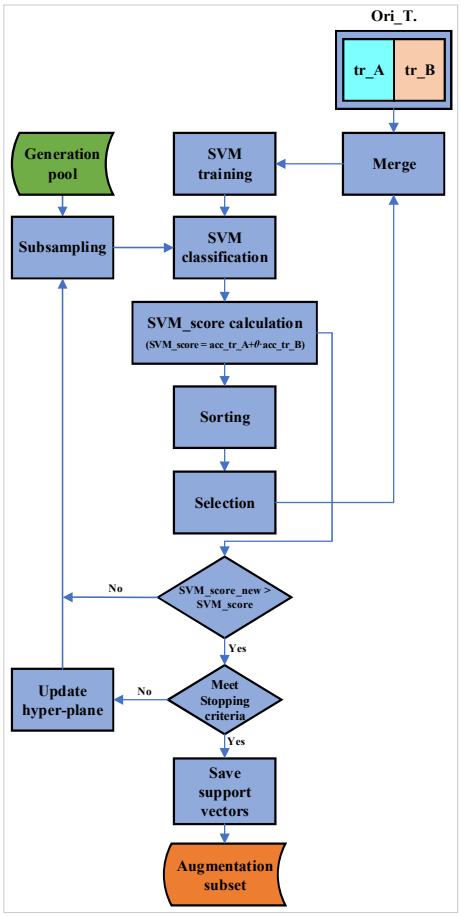

Fig. 2. Flowchart of the SVM based sifting process.

The generated samples for each class, i.e. Bona fide and replay, are organized to build two generation pools. Note that only half of the original training data (namely tr_A) is used for training the SVM hyper-plane while the other half (namely tr_B) is kept for validating. An SVM hyper-plane for each class is derived from the tr_A as a baseline performance. The trained hyper-plane is applied on relevant generated samples and the classification performance of the SVM is measured by the sum of the training and validation set accuracy (namely acc_tr_A and acc_tr_B). The measurement equation is given as below:

$$\text{SVM\_score} = \text{acc\_tr\_A} + \theta \cdot \text{acc\_tr\_B} \quad (1)$$

where a linear weighting factor θ is added to the validation set score considering the iterative update.

We take the spoof class as an example to describe the detailed processing steps of the SVM based post-filter. The main steps are the same as in [10] except that adding random perturbations on the generated samples is removed to make the sifting system applicable for samples produced by the augmentation approaches in this work. The generated spoofed speech pool is randomly subsampled. The distances between these samples and the SVM hyper-plane trained from the tr_A set are calculated. After sorting the subsampled generated spoofed speech by the distance order, only a preset number of the nearest ones are retained. We then merge these preserved spoof samples with the original samples from the tr_A set and use this merged set to train a new SVM hyper-plane. The classification performance of the new SVM is obtained by equation (1). The weighting factor θ is set to 1.5 as suggested in [10]. If the accuracy score of the new SVM outperforms the previous SVM score, the previous SVM hyper-plane is replaced with the new one and this iteration continues again with re-subsampled generated spoofed speech. All the steps of subsampling, sorting, selecting, merging and performance checking are repeated until the sifting process is completed. The associated support vectors of spoofed speech are employed for the augmented dataset once the SVM cannot be optimized anymore. The whole process is repeated with the tr_B as the training set for SVM and the tr_A as the validation set. As well as the spoof class, the entire process is repeated for the Bona fide class.

The statistics of the augmented training subset after the SVM based post selection is shown in Table II. The first two rows are the original and a partial segment of the training subset. This segment is called Mini_T. and is built by randomly selecting speech samples from the original training subset. The Mini_T. is used to investigate the impact of data quantity on the model training. The next four rows give the amount of the generated samples by different audio augmentation methods. Considering the balance between the original data and the augmented data, we limit the number of augmented samples to 1200 for each class by tuning the parameters of the SVM post selection process. The size of the final augmented training subset is provided in the last row, which is an assembly of the Ori_T. and four augmentation subsets.

TABLE II
STATISTICS OF THE AUGMENTED TRAINING SUBSET

Subsets	# Utterances	
	Bona fide	Replay
Ori_T.	1507	1507
Mini_T.	380	380
Shift	1200	1200
Stretch(0.8)	1200	1200
Stretch(1.2)	1200	1200
VTLP	1200	1200
Aug_T.	6307	6307

C. Replay spoofing detection framework

The block diagram of two replay detection systems is given in Fig.1. These high performing benchmark systems are based on deep learning architectures to obtain speech embeddings. The features extracted are all derived from the CQT transformation, as used in the baseline system of the challenge [3]. With the CQT-LCNN-GMM system [5] when the augmented training subset is ready, the constant Q transformation is adopted to extract the time-frequency representation. A light CNN (LCNN) is used to gain the embeddings from the CQT spectral features and a Gaussian Mixture Model (GMM) classifier is adopted as the back-end.

On the other hand, with the CQCC-DNN-SVM system [6] the CQCC features are extracted by applying a uniform resampling on the logarithmic power spectrum of the CQT spectrograms. A DNN based architecture is built to extract high level representations from the CQCC features and an SVM based classifier is used as the back-end. The output of these two classifiers is classification as the speech as either Bona fide or replay.

III. EXPERIMENTAL RESULTS

A. Experiments settings

The CQT spectral features and CQCC features are extracted by an open-source MATLAB toolkit[1]. The maximum and the minimum frequency in the constant Q transform are set as $F_{\max} = F_{sample}/2$ and $F_{\min} = F_{\max}/2^9$ respectively. The Nyquist frequency of the database is $F_{sample} = 16\text{kHz}$. The number of octaves is 9 and the number of bins per octave B is set to 96, which results in a time shift of 8 ms. The parameter γ is set to $\gamma = \Gamma = 228.7 * \left(2^{(1/B)} - 2^{(-1/B)}\right)$. In the process of CQCC features extraction, the resampling period is $d = 16$. By truncating the spectrum along the time axis with a fixed size of 863×400, CQT spectral features with unified time-frequency shape are obtained. The dimension of the CQCC static coefficients is set to 20 and appended with their first and second derivatives.

The architecture of the LCNN consists of 5 convolution layers, 4 Network in Network layers, 10 Max-Feature-Map layers, 4 max-pooling layers and 2 fully connected layers. The network of the DNN contains 3 convolution layers, 1 max-pooling layer and 3 fully connected layers. More detailed

[1] http://audio.eurecom.fr/content/software

settings for LCNN and DNN networks can be found in [5], [6]. The GMM back-end classifier utilizes 512-component models which are trained using the EM algorithm, on Bona fide and replay speech, respectively. The SVM back-end classifier is trained using a linear kernel to discriminate two classes.

All the scores from classifiers are represented by the Log-likelihood ratio (LLR). The EER is defined as the operating point on the Detection error tradeoff (DET) curve, where the false acceptance rate (FAR) is equal to the false rejection rate (FRR). A lower EER(%) indicates a better detection performance.

B. Results and analysis

To assess the proposed detection method, we compare with two benchmark anti-spoofing systems including the CQT-LCNN-GMM system [5] and the CQCC-DNN-SVM system [6]. The results are demonstrated as in Table III. The baseline CQCC-GMM system released in the ASVspoof 2017 challenge [3] is also presented. Over both the Dev. and Eval. subsets, the two deep learning based systems can outperform the baseline with decreased EERs. A serious degradation of detection performance can be seen when the systems are trained with the Mini_T.. This is a further evidence to confirm the importance of augmented data when using deep learning systems.

TABLE III
THE EERS(%) OF THE DL BASED SYSTEMS.

Systems	Subsets	EER%	
		Dev.	Eval.
CQCC-GMM(baseline)	Ori_T.	11.46	29.71
CQT-LCNN-GMM	Mini_T.	18.56	27.37
	Ori_T.	15.37	20.35
	Aug_T.(no sifting)	10.25	14.88
	Aug_T.(with sifting)	8.02	12.14
CQCC-DNN-SVM	Mini_T.	17.43	26.11
	Ori_T.	6.90	16.57
	Aug_T.(no sifting)	6.30	15.62
	Aug_T.(with sifting)	**4.80**	**11.30**

To investigate whether the SVM based post-filter can contribute to a lower EER performance we included both data augmentation without sifting and with our novel use of SVM based sifting. From Table III it is evident that when the post-filter is integrated the detection performance can be significantly improved. The best performance is obtained with the CQCC-DNN-SVM system, which gets very good results with EERs of 4.80% and 11.30% on the development and evaluation sets, respectively. These promising results indicate the significance of and verify the effectiveness of the post selection process when using data augmentation. With sufficient training data of informative variants generated using data augmentation with our proposed use of post selection, the detection performance can be improved when using deep learning systems for spoofing detection.

IV. CONCLUSIONS

In this paper, a replay anti-spoofing countermeasure based on data augmentation and features post selection was proposed. A novel audio augmentation method was adopted to generate speech samples. An SVM based post-filter was applied for sifting out samples with low relevance and diversity. With the expanded training subset, the issues of over-fitting and deficient generalization were alleviated. The effectiveness of the post selection process was revealed and confirmed for spoofing detection. Anti-spoofing systems were assessed on the ASVspoof 2017 Version 2.0 corpus and promising results were achieved on the development and evaluation subsets.

REFERENCES

[1] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine,* vol. 11, no. 2, pp. 23-61, Secondquarter 2011.

[2] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanilçi, M. Sahidullah, A. Sizov, et al, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 4, pp. 588-604, June 2017.

[3] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, et al, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," *Proc. Interspeech 2017,* 2-6, DOI: 10.21437/Interspeech.2017-1111, 2017.

[4] C. Zhang, C. Yu, and J. H. Hansen, "An investigation of deep-learning frameworks for speaker verification antispoofing," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 4, pp. 684-694, 2017.

[5] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks." *Proc. Interspeech 2017,* pp. 82-86. 2017.

[6] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, Replay attack detection using dnn for channel discrimination. *Proc. Interspeech 2017,* pp. 97-101. 2017.

[7] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 23, no. 9, pp. 1469-1477, 2015.

[8] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," *in Sixteenth Annual Conference of the International Speech Communication Association,* 2015.

[9] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, et al, "Asvspoof 2017 version 2.0: metadata analysis and baseline enhancements," *in Odyssey 2018 The Speaker and Language Recognition Workshop,* 2018.

[10] S. Mun, S. Park, D. K. Han, and H. Ko, "Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane," *Proc. DCASE,* pp. 93-97, 2017.

[11] W. Cai, D. Cai, W. Liu, G. Li, and M. Li, "Countermeasures for automatic speaker verification replay spoofing attack: On data augmentation, feature representation, classification and fusion." *Proc. Interspeech 2017,* pp. 17-21. 2017.

[12] M. Li, D. Tang, J. Zeng, T. Zhou, H. Zhu, B. Chen, and X. Zou. "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder." *Computer Speech & Language* vol. 56, pp. 80-94, 2019

[13] M. Todisco, H. Delgado, and N. Evans, "Constant q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Computer Speech & Language,* vol. 45, pp. 516-535, 2017.

[14] Y. Zhao, R. Togneri, and V. Sreeram, "Spoofing detection using adaptive weighting framework and clustering analysis," *in Proc. Interspeech 2018,* pp. 626-630. 2018.

[15] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, et al, *Introduction to Voice Presentation Attack Detection and Recent Advances.* Cham: Springer International Publishing, pp. 321-361. 2019.

[16] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," *in Proc. ICML Workshop on Deep Learning for Audio, Speech and Language,* vol. 117, 2013.