ONE-CLASS NEURAL NETWORK FOR ANTI-SPOOFING IN SPEAKER VERIFICATION

You Zhang
University of Rochester
you.zhang@rochester.edu

ABSTRACT

Human voices can be used to authenticate the identity of the speaker to ensure the security of certain processes, such as financial transactions. The current automatic speaker verification (ASV) systems are vulnerable to many kinds of spoofing attacks, especially text-to-speech (TTS) and voice conversion (VC). Recently, machine learning techniques have been applied to improve the reliability of ASV systems to deal with fake speech. However, most methods have limitations to detect fake speech that is generated by more advanced algorithms, since those algorithms have not been used to generate spoofing attacks in the training set. We proposed an anti-spoofing system based on one-class neural network to enhance the robustness of model against unknown spoofing attacks. Our results show that it outperforms many previous systems for unknown spoofing attacks when tested on ASVspoof 2019 dataset.

1. INTRODUCTION

Speaker verification plays an essential role in biometric authentication since it uses acoustics features to verify whether the given utterance is from a target person. The given utterances are usually expected to be genuine speech. However, speaker verification can be susceptible to spoofing attacks, such as impersonation, replay, text-to-speech, or voice conversion, that can fake speech to fool the verification system [16].

Before the microphone processes, Impersonation and replay attacks can be performed to spoofing the verification. Impersonation means human professional mimics or twins make very similar sounds that are almost the same as the target person's voice. Currently, there are no databases available for impersonation, so the research community does not take care of this attack. Replay spoofing attacks are to use pre-recorded audio as the input to the speaker verification system, which are the most accessible attacks since recording devices are nowadays easy to get. Text-to-speech (TTS) means to convert written text into spoken words and voice conversion (VC) is to convert speech from source speaker to target speaker's voice. These two spoofing attacks are both computer-generated which are posting more significant threats with the progress of developing algorithms.

Researchers have made some efforts to develop methods to detect whether the input speech is from a real person rather than spoofing attacks. Spoofing attacks detection is called anti-spoofing as a research topic. The ASVspoof challenge series [5, 12, 17] have been providing datasets and metrics to investigate countermeasures to defend against spoofing speaker verification. Similar to speaker verification, anti-spoofing uses some acoustic features to verify the genuineness of the given utterances. A complete speaker verification system should consist of front-end feature extraction and back-end ASV system with anti-spoofing module. The illustration for the speaker verification with standalone anti-spoofing system is shown in Figure 1.

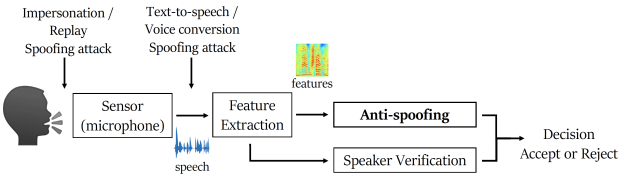


Figure 1. Overview of spoofing attacks and anti-spoofing in the speaker verification system.

Some combinations of magnitude-based and phase-based features had been proven to be efficient [7, 19, 20] in front-end feature extraction. Some useful features are based on wavelet transform [3, 11] and temporal or frequency modulation [4, 18]. There are also some traditional features, such as Cochlear Filter Cepstral Coefficients (CFCC) [7] and Constant Q Cepstral Coefficients (CQCC) [11], that reflect closely human perception system. As for the back-end classifier, previous works [7, 8, 14] showed that using Gaussian Mixture Model (GMM) works well for anti-spoofing. Still, GMM does not model well for clustering high dimensional features [10]. With the prevalent of deep learning, some basic deep learning models have also been proven to have good performance in spoofing detection [21]. In ASVspoof 2017 challenge, which focused on replay spoofing attacks, more neural network-based models are well-designed for replay detection [2, 6].

With the aforementioned features and models, most methods show good results in known spoofing attacks but do not generalize well on unknown attacks [17]. In real cases, however, usually the spoofing attacks are unknown and keep up-to-date, more generalized anti-spoofing methods are needed. This paper shows our proposed method for improving the generalization ability of anti-spoofing models. Our system based on One-class Neural Network (OCNN) is described in Section 2. The experiments are

conducted with ASVspoof 2019 dataset, and the results show that the proposed method outperforms baseline methods for dealing with unknown attacks, described in Section 3. In Section 4, some conclusions and future work are presented for discussion.

2. METHODS

Inspired by [13], one-class classification methods can set a tight boundary for genuine speech in high dimensional feature space. As a result, most spoofing attacks are considered as outliers which maps outside of genuine speech in feature space. In our proposed system, we first train a CNN as described in subsection 2.2 with some traditional features to learn the latent high dimensional features representation that can discriminate spoofing attacks and genuine speech. Then the extracted features are used to train One-class Neural Network (OCNN) to model the distribution of genuine speech. Finally, the proposed system outputs a score to indicate the probability that the input utterance belongs to real speech. The structure of the proposed system is shown in Figure 2.

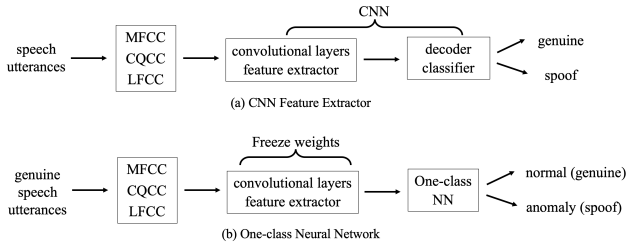


Figure 2. Model structure for the proposed anti-spoofing system.

2.1 Front-end

With a given utterance, we extract mel-frequency cepstral coefficients (MFCC), constant Q cepstral coefficients (CQCC) and linear frequency cepstral coefficients (LFCC) as front-end features to feed into CNN and train a classifier. MFCC is a common feature that is widely used in many speech processing tasks, so we assume it may contain some helpful information for detect spoofing. CQCC and LFCC are features used in baseline systems for ASVspoof 2019 challenge [15].

To form batches of data to feed into the network, the input features need to have same size. Since different utterances have different duration, padding or chopping are operated on the cepstral coefficients to make them the same fixed length. For example, MFCC has 50 frequency bins but the number of time frames are different. We took 220 time frames as the fixed length, which is a little above the average length. If the duration is less than 220 frames, we repeat pieces of data and concatenate them to the left or right. If the duration is longer than expected, we randomly select a consecutive piece of frames with the fixed length and chop the others. The same procedure is followed to form CQCC and LFCC batch.

2.2 CNN Feature Extractor

A convolutional neural network (CNN) is first trained to classify the spoofing attacks and genuine speech. Our CNN consists of 5 layers including three convolutional layers with $16 \times 7 \times 7$, $32 \times 5 \times 5$ and $32 \times 3 \times 3$ filters and two fully-connected layers with 32 nodes and two outputs. We divide the CNN into two parts: the convolutional layers are used as feature extractor, and the fully connected layers are decoder classifiers. After training, the feature extractor output high dimensional features for each utterances, and the decoder output the classification result of the given utterances. The CNN architectures for extracting features from MFCC, CQCC and LFCC are similar. The one for CQCC is shown in Figure 3 as an example.

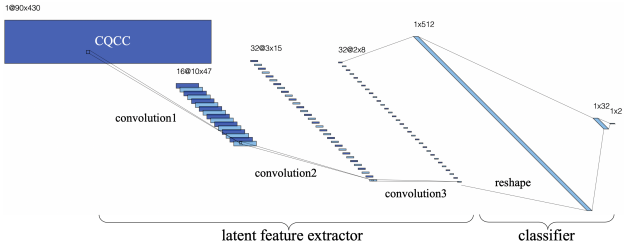


Figure 3. CNN architecture for extracting features from CQCC.

2.3 One-class Neural Network (OCNN)

One-class Neural Network (OCNN) is the state-of-the-art anomaly detection method [1]. In the original paper, the authors first trained an autoencoder to learn the latent representation of features that can reconstruct the original input. They then adopted the output of the pre-trained encoder to feed into the proposed OCNN. The idea behind this is using transfer learning to obtain features that best discriminate anomalies. The objective for one-class classification methods is to find a hyper-plane in feature space that sets a decision boundary for normal data.

The OCNN is designed as a feedforward neural network with one hidden layer and one output node. The training objective for OCNN is formulated to be equivalent to the loss function of One-class Support Vector Machine (OC-SVM):

$$\min_{w,V,r} \frac{1}{2}\|w\|_2^2 + \frac{1}{2}\|V\|_F^2 + \frac{1}{\nu} \cdot \frac{1}{N} \sum_{n=1}^{N} \max\left(0, r - \langle w, g\left(V\mathbf{X}_{n:}\right)\rangle\right) - r \tag{1}$$

where w is the weight matrix from hidden layers to output, V is the weight matrix from input to hidden layers. $g(\cdot)$ is the activation function for the hidden layer. Given training data $\mathbf{X}$, $g\left(V\mathbf{X}_{n:}\right)$ is the mapped vectors that separate most normal data points as far as possible from origin. The non-linearity of the mapping function makes it outperform OC-SVM [9] which uses $\Phi\left(\mathbf{X}_{n:}\right)$ for this term.

$\langle w, g(V\mathbf{X}_n :)\rangle$ is the scalar output of the OCNN and r is the bias of the hyper-plane. $\nu \in (0, 1)$ is a parameter to control the significance of the regulation term in order to prevent overfitting.

During training our proposed system, the OCNN takes the latent features extracted from pre-trained CNN and concatenate them together as input. The illustration is shown in Figure 4. Only genuine speech utterances is used as the front-end input since we formulate anti-spoofing as anomaly detection problem in this step. To optimize (1), r is fixed first to find optimal w and V with backpropagation. After each epoch, r is solved using the νth-quantile of the scalar output of OCNN. During the inference stage, the decision score y_n is the output of OCNN and r is the threshold. If the decision score y_n is larger than r, the sample is considered as normal data, otherwise anomalies.

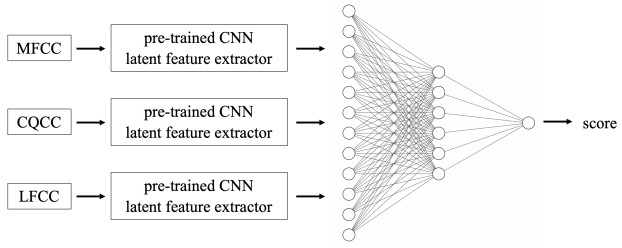


Figure 4. One-class Neural Network architecture.

3. EXPERIMENTS

3.1 ASVspoof 2019 dataset

The ASVspoof 2019 challenge [12] provides a standard database to and evaluation metric for anti-spoofing. The logical access part of provided dataset includes genuine speech and different kinds of TTS and VC spoofing attacks. In the training and validation set, there are 6 known attacks consisting of 4 TTS and 2 VC algorithms. In the test set, there are 11 unknown attacks including combinations of different TTS and VC attacks. The test set also includes two similar attacks which use same algorithms as two of the attacks in training set but trained with different data. The TTS and VC algorithms are mostly based on neural networks which is commonly used in real cases.

3.2 Evaluation metrics

To evaluate the performance of the anti-spoofing system, we take care of the output score of the anti-spoofing system and the speaker verification system. The output of the anti-spoofing system is called countermeasure (CM) score which indicates the similarity of the given utterance with genuine speech. The output of the speaker verification system is ASV score which indicates the probability that the given utterance belongs to the target speaker.

3.2.1 Equal Error Rate (EER)

Equal Error Rate is a threshold for the countermeasure decision score where false alarm rate is equal to the miss rate. If two probability of (2) are equal, the θ is called EER. The lower the EER is, the better the anti-spoofing system can detect spoofing attacks.

$$\begin{aligned} P_{fa}(\theta) &= \frac{\#\{\text{ spoof trials with score } > \theta\}}{\#\{\text{ total spoof trials }\}} \\ P_{\text{miss}}(\theta) &= \frac{\#\{\text{ human trials with score } \leq \theta\}}{\#\{\text{ total human trials }\}} \end{aligned} \tag{2}$$

3.2.2 tandem detection cost function (t-DCF)

Tandem detection cost function (t-DCF) is a new evaluation metric adopted in ASVspoof 2019 challenge. While EER only evaluates the performance of anti-spoofing system, The tandem detection cost function assess the influence of countermeasure system on the reliability of ASV system. The formulation of t-DCF is as follows:

$$\begin{aligned} \text{t}-\text{DCF}(s,t) = &C_{\text{miss}}^{\text{asv}} \cdot \pi_{\text{tar}} \cdot P_{\text{a}}(s,t) \\ &+ C_{\text{fa}}^{\text{asv}} \cdot \pi_{\text{non}} \cdot P_{\text{b}}(s,t) \\ &+ C_{\text{fa}}^{\text{cm}} \cdot \pi_{\text{spoof}} \cdot P_{\text{c}}(s,t) \\ &+ C_{\text{miss}}^{\text{cm}} \cdot \pi_{\text{tar}} \cdot P_{\text{d}}(s) \end{aligned} \tag{3}$$

where s and t are EER for anti-spoofing and speaker verification system. C are the cost for the four situations and π are the corresponding parameters. T-DCF is a function of thresholds for the CM score and ASV score, so usually the minimum t-DCF is used for evaluation. The lower the min t-DCF is, the better the anti-spoofing system impacts on a more reliable speaker verification system.

3.3 Results

3.3.1 Baseline

The baseline in ASVspoof 2019 challenge is adopted here. Baseline 1 uses CQCC as feature and GMM as back-end model for anti-spoofing, whereas baseline 2 uses LFCC as feature and GMM as back-end model. The ASV module uses x-vector as front-end feature and PLDA as back-end classifier and the ASV scores are already provided by the organisers. We also constructed a deep learning baseline with CNN for anti-spoofing which has the same structure as described in section 2.2. Table 1 shows the comparison of baseline methods.

Methods	EER(%)	min t-DCF
LFCC + GMM	8.09	0.2116
CQCC + GMM	9.57	0.2366
LFCC + CNN	6.45	0.1633
CQCC + CNN	13.26	0.3578

Table 1. Comparison of baseline methods

The LFCC+CNN method outperforms the others with the lowest EER and the lowest min t-DCF. The distribution of the output score is shown in Figure 5. From the figure, LFCC+CNN successfully separate spoofing attacks from genuine speech with small overlapped distribution.

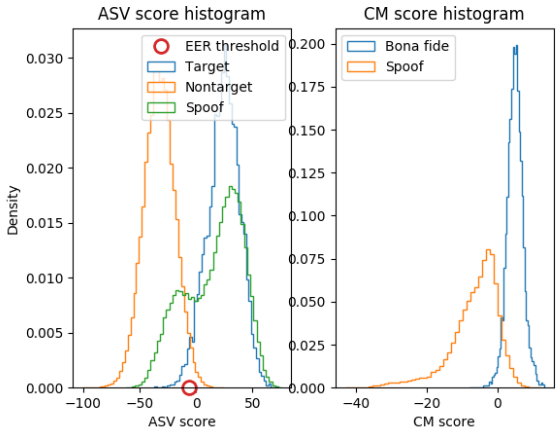


Figure 5. Distribution of ASV score and CM score for LFCC+CNN.

3.3.2 Proposed method

With the proposed method, the EER is 42.14% and min t-DCF is 0.6212. The current result is much worse than expected. The distribution of the ASV and CM score is shown in Figure 6. It seems that the one-class method did push the boundary of the genuine speech much further from the origin, but it did not benefit separate spoofing attacks.

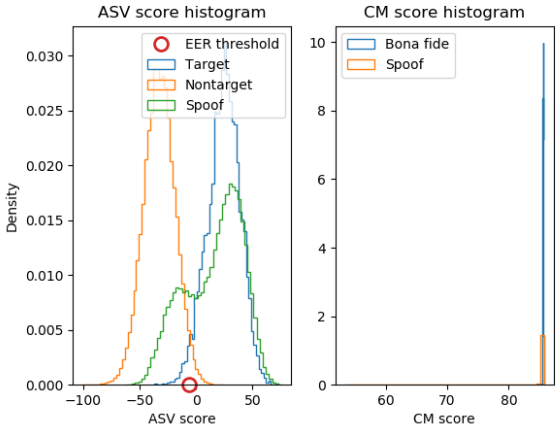


Figure 6. Distribution of ASV score and CM score for proposed OCNN.

4. DISCUSSIONS

4.1 Conclusions

A novel framework based on one-class neural network is proposed. The proposed method aims to set a tight boundary for genuine speech and everything outside the boundary is considered as spoofing attacks. Our proposed method is expected to show better performance on unknown spoofing attacks when tested on ASVspoof 2019 dataset, but the current results are still promising.

4.2 Future works

Reproduction of [1, 13] need to be done to double check the pipeline and implementation of the proposed models.

Feature selection should be taken into more consideration. Some magnitude-based perceptual features like CFCC are also in use to develop TTS and VC algorithms to generate speech that sounds more real. Some artifacts that may not be sensitive to humans in spoofed speech could be the key to solve anti-spoofing task. Cepstral coefficients computed from several kinds of spectrograms may show some efficiency in some speech processing tasks, but there are still lots of features they may not contain. Raw time-domain signals may be a better input for latent feature extraction.

Model selection may be another issue. Current CNN classifier may capture features that are useful for classify but may lose some information that best represent genuine speech. Autoencoder has also been experimented as a trial. Some more powerful models could be explored to extract latent features.

The training procedure and loss functions could be improved in training OCNN to involve spoofing data. Currently we only make use of genuine speech when training OCNN since it is originally an anomaly detection method. However, in our dataset, there are much more spoofing attacks that may also be useful in training OCNN. Some terms in loss function could be designed to take advantage of those fake speech data.

5. REFERENCES

[1] Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. *arXiv preprint arXiv:1802.06360*, 2018.

[2] Bhusan Chettri, Saumitra Mishra, Bob L Sturm, and Emmanouil Benetos. Analysing the predictions of a cnn-based replay spoofing detection system. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 92–97. IEEE.

[3] Tharshini Gunendradasan, Eliathamby Ambikairajah, Julien Epps, and Haizhou Li. An adaptive-q cochlear model for replay spoofing detection.

[4] Tharshini Gunendradasan, Buddhi Wickramasinghe, Phu Ngoc Le, Eliathamby Ambikairajah, and Julien Epps. Detection of replay-spoofing attacks using frequency modulation features. In *Interspeech*, pages 636–640.

[5] Tomi Kinnunen, Md Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017.

[6] Cheng-I Lai, Alberto Abad, Korin Richmond, Junichi Yamagishi, Najim Dehak, and Simon King. Attentive

filtering networks for audio replay attack detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6316–6320. IEEE.

[7] Tanvina B Patel and Hemant A Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[8] Jon Sanchez, Ibon Saratxaga, Inma Hernaez, Eva Navas, Daniel Erro, and Tuomo Raitio. Toward a universal synthetic speech spoofing detection using phase information. *IEEE Transactions on Information Forensics and Security*, 10(4):810–820, 2015.

[9] Hyun Joon Shin, Dong-Hwan Eom, and Sung-Shick Kim. One-class support vector machines—an application in machine fault detection and classification. *Computers Industrial Engineering*, 48(2):395–408, 2005.

[10] Xiaohai Tian, Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Spoofing detection from a feature representation perspective. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2119–2123. IEEE.

[11] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Odyssey*, volume 45, pages 283–290.

[12] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Hector Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[13] Jesus Villalba, Antonio Miguel, Alfonso Ortega, and Eduardo Lleida. Spoofing detection with dnn and one-class svm for the asvspoof 2015 challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[14] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa. Relative phase information for detecting human speech and spoofed speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[15] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Hector Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, and Kong Aik Lee. The asvspoof 2019 database. *arXiv preprint arXiv:1911.01601*, 2019.

[16] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, 66:130–153, 2015.

[17] Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li. Synthetic speech detection using temporal modulation feature. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7234–7238. IEEE.

[19] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Eng Siong Chng, and Haizhou Li. Spoofing speech detection using high dimensional magnitude and phase features: The ntu approach for asvspoof 2015 challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*.

[20] Hong Yu, Achintya Sarkar, Dennis Alexander Lehmann Thomsen, Zheng-Hua Tan, Zhanyu Ma, and Jun Guo. Effect of multi-condition training and speech enhancement methods on spoofing detection. In *2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, pages 1–5. IEEE.

[21] Chunlei Zhang, Chengzhu Yu, and John HL Hansen. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):684–694, 2017.