One-class Neural Network for Anti-spoofing in Speaker Verification

You Zhang

University of Rochester

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 [6].

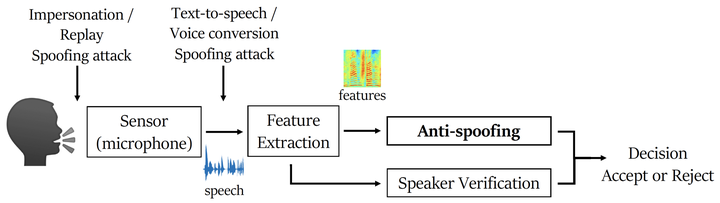
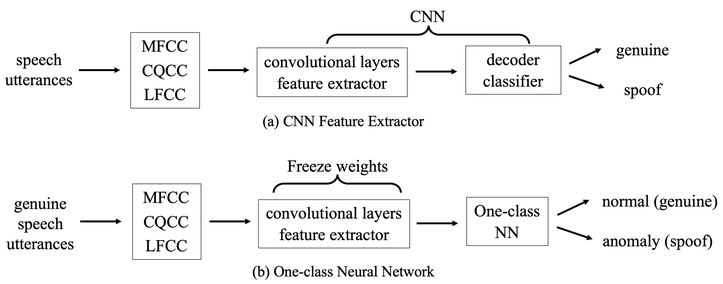


Fig. 1: Anti-spoofing and speaker verification systems.

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, 2, 3] have been providing datasets and metrics to investigate countermeasures to defend against spoofing speaker verification.

Methods

Inspired by [4], 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 classifier 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.



(a) CNN Feature Extractor

(b) One-class Neural Network

Fig. 2: Proposed model structure.

Model Architecture

A convolutional neural network (CNN) is first trained to classify the spoofing attacks and genuine speech. We divide the CNN into two parts: the convolutional layers are used as feature extractor, and the fully connected layers are decoder classifiers.

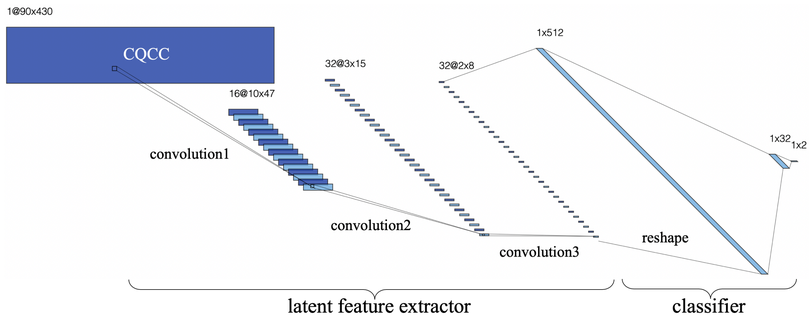


Fig. 3: Convolutional neural classifier.

One-class Neural Network (OCNN) is the state-of-the-art anomaly detection method [1]. 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 (OCSVM):

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

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 nonlinearity of the mapping function makes it outperform OCSVM which uses $\Phi\left(\mathbf{X}_{n:}\right)$ for this term. $\langle w, g\left(V\mathbf{X}_{n}:\right)\rangle$ 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.

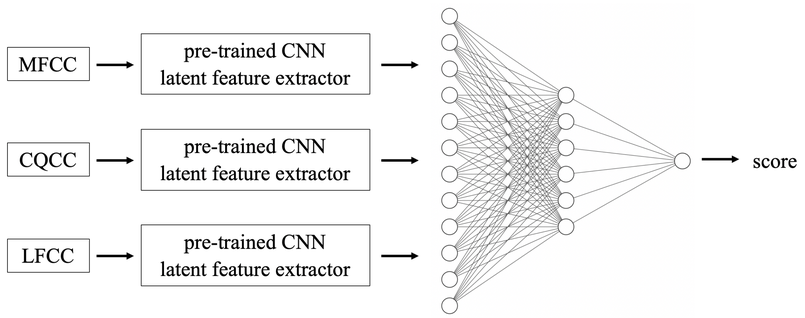


Fig. 4: One-class Neural Network.

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.

Evaluation

1.Equal Error Rate (EER)

Equal Error Rate is a threshold for the decision score where false alarm rate is equal to the miss rate.

$$P_{fa}(\theta) = \frac{\#\{\text{ spoof trials with score } > \theta\}}{\#\{\text{ total spoof trials }\}}$$
$$P_{\text{miss}}(\theta) = \frac{\#\{\text{ human trials with score } \leq \theta\}}{\#\{\text{ total human trials }\}} \qquad (2)$$

2. Tandem detection cost function (t-DCF)

The tandem detection cost function assess the influence of countermeasure system on the reliability of ASV system.

Results

Table 1: Comparison of baseline methods:

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

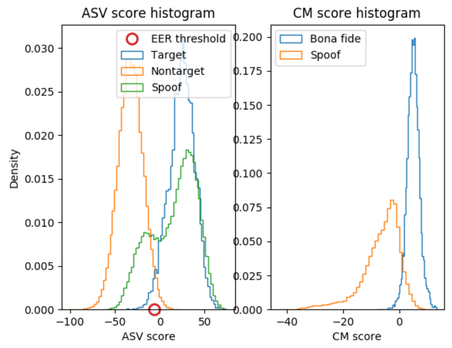


Fig. 5: ASV score and CM score for LFCC+CNN.

References

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

[2] Tomi Kinnunen et al. "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection". In: (2017).

[3] Massimiliano Todisco et al. "ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection". In: *arXiv preprint arXiv:1904.05441* (2019).

[4] Jesus Villalba et al. "Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

[5] Zhizheng Wu et al. "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

[6] Zhizheng Wu et al. "Spoofing and countermeasures for speaker verification: A survey". In: *speech communication* 66 (2015), pp. 130–153. ISSN: 0167-6393.