Abstract

Human voices can be used to authenticate the identity of the speaker, but the automatic speaker verification (ASV) systems are vulnerable to voice spoofing attacks, such as impersonation, replay, text-to-speech, and voice conversion. Recently, researchers have developed anti-spoofing techniques to improve the reliability of ASV systems against spoofing attacks. However, most methods encounter difficulties in detecting unknown attacks in practice, which often have different statistical distributions from known attacks. Especially, the fast development of echo voice spoofing algorithms is generating increasingly powerful attacks, putting the ASV systems at risk of unseen attacks. In this work, we propose an anti-spoofing system to detect synthetic voice spoofing attacks (i.e., text-to-speech or voice conversion) using one-class learning. The key idea is to compact the bona fide speech representation and inject an angular margin to separate the spoofing attacks in the embedding space. Without resorting to any data augmentation methods, our proposed system achieves an equal error rate (EER) of 2.19% on the evaluation set of ASVspoof 2019 Challenge logical access scenario, outperforming all existing single systems (i.e., those without model ensemble).

Introduction

Speaker verification plays an essential role in biometric authentication; it uses acoustic features to verify whether a given utterance is from the target person. However, ASV systems can be fooled by spoofing attacks, such as impersonation (mimics or twins), replay (pre-recorded audio), text-to-speech (converting text to spoken words), and voice conversion (converting speech from source speaker to target speaker).

While much progress has been made, existing methods generally suffer from generalization to unseen spoofing attacks in the test stage. Due to the development of speech synthesis techniques, the synthetic spoofing attacks in the training set may never be able to catch up with the expansion of the distribution of spoofing attacks in practice. This distribution mismatch between training and test for the fake class actually makes the problem a good fit for one-class classification. 

Method

In this section, we first briefly introduce and analyze the widely used binary classification loss functions, then propose our one-class learning loss function for voice spoofing detection.

The original Softmax loss for binary classification can be formulated as:

$$L_{SOFT} = \frac{1}{N} \sum_{i=1}^{N} \log \left( \frac{e^{x_i}}{1 + e^{x_i}} \right).$$

(1)

AM-Softmax improves upon this by introducing an angular margin to make the embedding distributions of both classes more compact, around the weight difference vector's two directions:

$$L_{AMS} = \frac{1}{N} \sum_{i=1}^{N} \log \left( 1 + e^{a(x_i - m)} \right) - \frac{1}{N} \sum_{i=1}^{N} \log \left( 1 + e^{a(m - x_i - m_0)} \right).$$

(2)

In voice spoofing detection, it is reasonable to train a compact embedding space for bona fide speech. However, if we also train a compact embedding space for the spoofing attacks, it may overfit known attacks. To address this issue, we propose to introduce two different margins for better compacting the bona fide speech and isolating the spoofing attacks. The proposed loss function One-class Softmax (OC-Softmax) is denoted as:

$$L_{OCS} = \frac{1}{N} \sum_{i=1}^{N} \log \left( 1 + e^{a(x_i - m_0)} \right),$$

(3)

Comparing losses

To demonstrate the superiority of our proposed method, we compared our system with other existing single systems (no model fusion) without data augmentation. It can be seen that our proposed system significantly outperforms all of the other single systems.

The three losses perform similarly on the development set, showing that they have good discrimination ability to detect known attacks. For the evaluation set, where all the attacks are not included in the development set, our one-class learning system with the proposed OC-Softmax surpasses the binary classification losses (Softmax and AM-Softmax). The relative improvement on EER is up to 33%.

Results

The dimension-reduced embedding visualization is shown in Figure ???. The same t-distributed Stochastic Neighbor Embedding (t-SNE) and Principle Component Analysis (PCA) projections are applied to development and evaluation datasets of ASVspoof 2019 LA scenario. In other words, the visualizations of the two sets use the same coordinating systems. The figure verifies our problem formulation and shows the effectiveness of our proposed OC-Softmax loss.

Acknowledgements

This work was supported by National Science Foundation grant No. 1741472 and funding from Voice Biometrics Group.

References