CS230

Speaker Recognition Adversarial Attack via Style Transfer

Zamir Ben-Hur
SUNet ID: zamirbh
Department of Computer Science
Stanford University
zamirbh@stanford.edu

Vladimir Tourbabin
SUNet ID: tourbabv
Department of Computer Science
Stanford University
tourbabv@stanford.edu

Abstract

With the increase popularity of deep-neural network speaker recognition systems, which becoming widely used for various applications such as smart lockers and financial services, the importance of their security has gained attention. As shown in the current work, such systems can be relatively easily deceived by carefully designed adversarial example with small perturbations of the original speech. In this work we are proposing an adversarial attack for such state-of-the-art system based on neural style transfer. The proposed method is shown to be effective and efficient by evaluating its performance and comparing to a human-level performance using a subjective listening test.

1 Introduction

Speaker recognition (SR) technology is aiming at identification and/or verification of people's identity based on their vocal characteristics. Early SR techniques relied on a human expert to make a decision on a persons' identity by comparing their speech characteristics [1]. More recently, fully automated SR systems have been developed, which are based on deep-neural networks, and are commonly used for variety of applications, such as speaker verification (authentication), speaker and event classification, speaker segmentation, diarization, detection and tracking [2]. Such applications are widely deployed in existing products such as smart vehicle-mounted systems, smart locks, and financial services. However, deep neural network-based SR systems are vulnerable to adversarial attacks, which may deceive the system to make wrong decisions by relative small perturbations usually imperceivable to a human listener.

This work is focused on designing adversarial attack for a SR systems, with the aim of challenging current state-of-the-art (SOTA) SR systems to make wrong speaker identity prediction. The approach proposed here is based on the "neural style transfer" idea [3]. "Neural style transfer" has been proposed for creating artistic imagery by separating and recombining image content and style. Similar approach was recently suggested for voice conversion[4]. The method utilizes a pre-trained network to extract embeddings from a given input voice utterance. Then, back-propagation is used to compute gradients w.r.t the input itself, thereby iteratively "optimizing" the input until it converges to match the desired embeddings.

The inputs to our algorithm are: a speech utterance of the desired speaker, and an utterance of another non-desired speaker. We then use a modified neural style transfer approach to modify the non-desired speaker speech such that the attacked network will detect it as the desired speaker voice. We had also designed and conducted a listening test with human subjects to assess the effectiveness of the method proposed here and compare to human performance.

2 Related work

One particularly popular SOTA SR system is the ECAPA-TDNN model proposed in [5], which, as of today, has more than 400 citations, and is also a part of the open-source conversational AI toolkit, SpeechBrain [6]. The toolkit is