Deep Features for Automatic Spoofing Detection

Yanmin Qian, Nanxin Chen, Kai Yu

PII: S0167-6393(16)30109-1
DOI: 10.1016/j.specom.2016.10.007
Reference: SPECOM 2410

To appear in: *Speech Communication*

Received date: 15 May 2016
Revised date: 15 October 2016
Accepted date: 17 October 2016

Please cite this article as: Yanmin Qian, Nanxin Chen, Kai Yu, Deep Features for Automatic Spoofing Detection, *Speech Communication* (2016), doi: 10.1016/j.specom.2016.10.007

Deep Features for Automatic Spoofing Detection

Yanmin Qian[a,*], Nanxin Chen[a], Kai Yu[a,*]

[a]*Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering*
SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Abstract

Recently biometric authentication has made progress in areas, such as speaker verification. However, some evidence shows that the technology is susceptible to malicious spoofing attacks, and thus dedicated countermeasures are needed to detect a variety of specific attack types. Inspired by the great success of deep learning in automatic speech recognition, we propose a detailed deep learning based feature engineering framework for spoofing detection in this paper. To incorporate deep learning into spoofing detection, this work proposes novel approaches for extracting and using features from deep learning models. In contrast to the traditional short-term spectral features, such as MFCC or PLP, outputs from the hidden layer of various deep models are employed as *deep features* for spoofing detection. Two frameworks are developed to extract deep features, including DNN-based frame-level feature extraction and RNN-based sequence-level feature extraction, and several structures are explored within each framework. Once the deep features are extracted, they can be used as a spoofing identity representation for each utterance, and the appropriate back-end classifier is then applied to make the final detection decision. These approaches were evaluated on the ASVspoof2015 Challenge data corpus. Experiments show that deep feature based systems achieve good performance, even without using any designed features such as phase and cochlea features common in spoofing detection, and obtain significant performance improvements compared to the traditional baselines. The EER of the best deep feature system achieves nearly 0.0% for all attack types from S1 to S9, and gets 1.1% on all averaged conditions (plus S10), which is very promising performance in ASVspoof2015 Challenge task.

Keywords: Automatic spoofing detection; Deep features; Deep neural network; Recurrent neural network

1. Introduction

Recently there has been substantial progress on biometric recognition systems, of which automatic speaker verification (ASV) system is one important type. ASV systems take claimed identity and speech samples as input and decide whether to accept or reject the claim. *Text-dependent* and *text-independent* are two categories of typical ASV systems. Text-dependent ASV systems use the fixed phrases within the enrollment and verification stages, while text-independent accepts any speech. Benefiting from a number of technical advances (Dehak et al., 2011; Chen et al., 2015b; Liu et al., 2015a), such as progress on channel and noise compensation (Solomonoff et al., 2005; Burget et al., 2007; Vair et al., 2006; Hubeika et al., 2008), current ASV systems give impressive results on many tasks (Greenberg et al., 2013).

When applying automatic speaker verification to real scenarios, security and robustness become more important. To be applied in scenarios such as mobile payment systems, automobiles or mobile phone unlocking systems, it is crucial to know the robustness of the ASV system against spoofed attacks. As with other biometric recognition systems, these attacks can be generally grouped into two categories:

- **Direct attacks**, also referred to as *spoofing attacks*, can be directly applied without any prior knowledge from the ASV system. For example, the attacker may use generated or altered speech to login as the target speaker.
- **Indirect attacks**, generally require system-level access. For instance, the attacker may attempt to change the running state of any component in the system, such feature extraction, models and decisions.

1.1. Spoofing Approaches

Since indirect attacks require system-level access, they are relatively difficult to implement. Accordingly, direct attacks are the greatest threat and are the main focus of this paper. Direct spoofing attacks can be commonly categorized into the following four types:

Impersonation: a speaker mimics another person, which requires professional expertise. These attacks have seen little interest in recent research and are generally seen as being less potent. Skilled impersonators, although being potentially serious threats, generally do not occur often. Recent work in (Wu et al., 2015a) suggested that there are no consistent results on the detection of impersonation.

Replay: an already recorded utterance of a speaker is played into the speaker verification system. Since smartphones are widely available and can easily be used to record and replay speech, this approach is an increasingly practical attack. The

*Corresponding author
Email addresses: yanminqian@sjtu.edu.cn (Yanmin Qian), bobchennan@gmail.com (Nanxin Chen), kai.yu@sjtu.edu.cn (Kai Yu)