Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language ■■ (2016) ■■–■■

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Spoofing voice verification systems with statistical speech synthesis using limited adaptation data

Ali Khodabakhsh *, Amir Mohammadi, Cenk Demiroglu

Electrical and Computer Engineering Department, Ozyegin University, Istanbul, Turkey

Received 9 October 2015; received in revised form 16 August 2016; accepted 24 August 2016
Available online

Abstract

State-of-the-art speaker verification systems are vulnerable to spoofing attacks using speech synthesis. To solve the issue, high-performance synthetic speech detectors (SSDs) for attack methods have been proposed recently. Here, as opposed to developing new detectors, we investigate new attack strategies. Investigating new techniques that are specifically tailored for spoofing attacks that can spoof the voice verification system and are difficult to detect is expected to increase the security of voice verification systems by enabling the development of better detectors. First, we investigated the vulnerability of an i-vector based verification system to attacks using statistical speech synthesis (SSS), with a particular focus on the case where the attacker has only a very limited amount of data from the target speaker. Even with a single adaptation utterance, the false alarm rate was found to be 23%. Still, SSS-generated speech is easy to detect (Wu et al., 2015a, 2015b), which dramatically reduces its effectiveness. For more effective attacks with limited data, we propose a hybrid statistical/concatenative synthesis approach and show that hybrid synthesis significantly increases the false alarm rate in the verification system compared to the baseline SSS method. Moreover, proposed hybrid synthesis makes detecting synthetic speech more difficult compared to SSS even when very limited amount of original speech recordings are available to the attacker. To further increase the effectiveness of the attacks, we propose a linear regression method that transforms synthetic features into more natural features. Even though the regression approach is more effective at spoofing the detectors, it is not as effective as the hybrid synthesis approach in spoofing the verification system. An interpolation approach is proposed to combine the linear regression and hybrid synthesis methods, which is shown to provide the best spoofing performance in most cases.

Keywords: Statistical speech synthesis; Hybrid speech synthesis; Spoofing verification systems; Speaker adaptation; Synthetic speech detection

1. Introduction

Text-independent voice verification (VV) systems have made tremendous progress in recent years (Martin et al., 2012). Most of the currently popular systems are based on the total variability space (TVS) approach that is based on

This work is supported by TUBITAK 1001 program under Contract Number 112E160.

* Corresponding author at: Electrical and Computer Engineering Department, Ozyegin University, Istanbul, Turkey.
E-mail address: ali.khodabakhsh@ozu.edu.tr (A. Khodabakhsh).

http://dx.doi.org/10.1016/j.csl.2016.08.004
0885-2308/© 2016 Published by Elsevier Ltd.

representing a speech signal with a low-dimensional i-vector, which is then used for verification of claimed speaker identity (Dehak et al., 2011). Performance of those systems is now acceptable for use in many real-life applications such as call centers.

Even though the speaker verification technologies have improved, they are known to be vulnerable to spoofing attacks, which is an important concern in their deployment (De Leon et al., 2012; Kinnunen et al., 2012; Wu et al., 2015a, 2015b). Moreover, improvements in the concatenative and statistical speech synthesis systems (SSS) as well as the voice conversion systems have further spurred the concerns (Wu et al., 2015a). As a result, more effective ways to attack the verification systems and protecting the system from attacks have become increasingly important areas of research (Evans et al., 2014).

One approach that is effective at spoofing attacks is voice conversion (Wu et al., 2012). In Alegre et al. (2012), Gaussian Mixture Model (GMM) based voice transformation using parallel data is found to be effective at spoofing the voice verification systems. To increase the effectiveness of the attacks, segments of speech that get high scores from the voice verification system are repeated, which can be considered as attacking with artificial data. Two countermeasures are also proposed in Alegre et al. (2012). In one approach, distributions of Gaussian components are used to detect repetitions of Gaussians in speech. In a second approach, automatic voice quality assessment tools are used to detect synthetic speech. Spoofing performance of a joint density Gaussian mixture model (JD-GMM) voice conversion system is analyzed as a function of the training data for text-dependent voice verification systems in Wu and Li (2015).

Most parametric speech codecs use minimum-phase filters since the human auditory system is assumed to be insensitive to phase (Quatieri, 2002). If such a speech codec is used during an attack, unnatural phase spectrum can be used to detect the synthetic speech as proposed in De Leon et al. (2012) and Wu et al. (2012).

Detection performance when the synthetic speech detector (SSD) is trained with different kinds of voice conversion techniques is reported in Wu et al. (2012). Besides the magnitude and phase features that rely on a single speech frame, modulation of those features over longer durations is investigated and found to be complimentary to magnitude and phase features in Wu et al. (2013).

A second approach used in spoofing is speech synthesis (Wu et al., 2015a). There are two major approaches to speech synthesis: unit selection and statistical parametric synthesis (Black et al., 2007). Unit selection synthesis requires availability of large amounts of data from the target speaker. In many real-life scenarios, the attacker attempts to spoof into the accounts of many users. For such attacks, it is impractical to collect large amounts of data for each speaker, and the attacker may only have at most few utterances from each speaker. Thus, although effective spoofing attacks can be performed with unit selection synthesis (De Leon et al., 2012), SSS is a more effective way to attack when very limited amount of data are available, since SSS can achieve rapid speaker adaptation with only a couple of utterances (Black et al., 2007; Yamagishi et al., 2009, 2010). Because spoofing attacks with only a few utterances are investigated here, SSS is used as the baseline synthesis technique.

Most of the previous spoofing literature focused on detectors that are designed to detect known spoofing techniques (Wu et al., 2015a, 2015b). However, how effective those detectors are for unknown spoofing attack types remains a serious question. Hence, speech synthesis techniques that are specifically designed for spoofing attacks should be investigated so that detectors that can generalize better and produce more secure voice verification systems can be developed.

We propose three strategies for effective spoofing attacks when limited adaptation data are available to the attacker. In the first approach, a hybrid concatenative/statistical speech synthesis method is proposed. The proposed hybrid system takes advantage of the rapid adaptation capability of the statistical systems while using the available natural speech segments from the speaker as much as possible. Even though spoofing with a unit selection system is hard to detect (Wu et al., 2015), it cannot be deployed in a limited data case as discussed above. Still, here, we show that effectiveness of the attacks can be significantly improved with the proposed hybrid approach that exploits the available units in the database while using SSS when units are not available.

In the second approach, linear regression (LR) is done to transform synthetic speech parameters closer to natural ones. Transformation matrices are learned from a speaker-independent speech database. Even though the resulting features are more natural and more effective than the hybrid approach at spoofing the SSD, they are not as effective in spoofing the verification system. To further boost its effectiveness, in a third approach, we propose an algorithm to combine the hybrid features and transformed features, which is found to be the most effective system for spoofing attacks in most cases. The proposed algorithms were tested using three state-of-the-art synthetic speech detectors (SSD).

Please cite this article in press as: Ali Khodabakhsh, Amir Mohammadi, Cenk Demiroglu, Spoofing voice verification systems with statistical speech synthesis using limited adaptation data, Computer Speech and Language (2016), doi: 10.1016/j.csl.2016.08.004