Available online at www.sciencedirect.com
ScienceDirect
Computer Speech & Language xxx (2017) xxx-xxx
www.elsevier.com/locate/csl

Synthetic speech detection using fundamental frequency variation and spectral features☆

Q2 Monisankha Pal*,[a], Dipjyoti Paul[a], Goutam Saha[a]

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721 302, India

Received 18 June 2017; received in revised form 31 August 2017; accepted 6 October 2017
Available online xxx

Abstract

Recent works on the vulnerability of automatic speaker verification (ASV) systems confirm that malicious spoofing attacks using synthetic speech can provoke significant increase in false acceptance rate. A reliable detection of synthetic speech is key to develop countermeasure for synthetic speech based spoofing attacks. In this paper, we targeted that by focusing on three major types of artifacts related to magnitude, phase and pitch variation, which are introduced during the generation of synthetic speech. We proposed a new approach to detect synthetic speech using score-level fusion of front-end features namely, constant Q cepstral coefficients (CQCCs), all-pole group delay function (APGDF) and fundamental frequency variation (FFV). CQCC and APGDF were individually used earlier for spoofing detection task and yielded the best performance among magnitude and phase spectrum related features, respectively. The novel FFV feature introduced in this paper to extract pitch variation at frame-level, provides complementary information to CQCC and APGDF. Experimental results show that the proposed approach produces the best stand-alone spoofing detection performance using Gaussian mixture model (GMM) based classifier on ASVspoof 2015 evaluation dataset. An overall equal error rate of 0.05% with a relative performance improvement of 76.19% over the next best-reported results is obtained using the proposed method. In addition to outperforming all existing baseline features for both known and unknown attacks, the proposed feature combination yields superior performance for ASV system (GMM with universal background model/i-vector) integrated with countermeasure framework. Further, the proposed method is found to have relatively better generalization ability when either one or both of copy-synthesized data and limited spoofing data are available a priori in the training pool.

Keywords: All-pole group delay function (APGDF); Anti-spoofing; Constant Q cepstral coefficient (CQCC); Fundamental frequency variation (FFV); Score-level fusion; Spoofing attack

1. Introduction

Automatic speaker verification (ASV) systems that accept or reject an identity claim have a wide range of applications in banking, forensics, voice mail, etc. (Kinnunen and Li, 2010; Pal and Saha, 2015). However, the main concern with the deployment of ASV systems is their vulnerability towards *spoofing attacks* in which a fraudster tries to

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

* Corresponding Author.
E-mail address: monisankhapal@iitkgp.ac.in (M. Pal), dipjyotipaul@ece.iitkgp.ernet.in (D. Paul), gsaha@ece.iitkgp.ernet.in (G. Saha).

http://dx.doi.org/10.1016/j.csl.2017.10.001

masquerade an enrolled person's voice to get illegitimate acceptance. The major forms of spoofing attacks are *impersonation* (Hautamäki et al., 2015), *replay* (Villalba and Lleida, 2011), *speaker adapted speech synthesis* (Moulines and Charpentier, 1990) and *voice conversion* (Stylianou et al., 1998). Recent advancements in computer-assisted speech synthesis (SS) and voice conversion (VC) technology, with availability of related open source software, make SS and VC the most potent means of spoofing attacks to swindle ASV systems. To mitigate this, appropriate *countermeasure* (CM) to discriminate natural speech from the synthetic speech is essential.

A popular approach for spoofing detection is the use of front-end features that targets capturing the artifacts introduced during speech manipulation. Modulation features extracted from magnitude or phase spectrum, which carry long-term temporal information were incorporated to distinguish HMM-based synthetic speech from natural speech (Wu et al., 2013). Based on human speech perception and the fact that phase information is usually lost during synthesis in VC, phase-based features like cosine-normalized phase (CosPhase) and modified group delay function (MGDF) were proposed to detect VC spoofing as stand-alone system (Wu et al., 2012b) and with ASV framework (Wu et al., 2012a). The use of relative phase shift (RPS) features for reliable detection of hidden Markov model (HMM) based text-to-speech (TTS) synthesized speech was introduced in De Leon et al. (2012a). Another phase-based feature called mel regularized RPS was also explored to detect TTS synthesized speech (Sanchez et al., 2015b). A synthetic speech detector using prosodic features like pitch pattern statistics from image analysis of pitch patterns was proposed in De Leon et al. (2012b). A new countermeasure based on spectro-temporal information from local binary patterns was derived in Alegre et al. (2013), which is less reliant on prior knowledge and provides robust protection from VC as well as SS attacks. Countermeasure using high-level dynamic features and voice quality assessment was employed in Alegre et al. (2012). Most of these hand-crafted discriminative features used prior spoofing data to distinguish natural and synthetic utterances drawn from a closed set database. However, development of generalized countermeasure to act on unseen spoofing data type is gaining practical importance. Apart from focusing on front-end features, *anti-spoofing* using i-vector based representation of speech utterances as back-end was investigated in Sizov et al. (2015). This back-end generative model is promising to provide generalized countermeasure.

To address the issues arising from using closed set databases, the research community developed ASVspoof 2015 (Wu et al., 2015), which contains spoofed data from a diverse range of spoofing attacks. The winning system (Patel and Patil, 2017a) of this challenge used a combination of standard mel frequency cepstral coefficients (MFCCs) and cochlear filter cepstral coefficients (CFCCs) with the change in instantaneous frequency (IF). A significant number of countermeasures have been proposed till date on this challenge data. Among them, MGD-phase features (Alam et al., 2015; Liu et al., 2015), RPS features (Sanchez et al., 2015a; Wang et al., 2015), discriminatory sub-band features (Sriskandaraja et al., 2016), wavelet-based features (Novoselov et al., 2016), linear prediction features (Alam et al., 2015; Janicki, 2015) were explored for spoofing detection task. A comparative study showing the efficacy of various short-term power spectrum and phase features, dynamic and complementary features for anti-spoofing was presented in Sahidullah et al. (2015). In this study, linear frequency cepstral coefficient (LFCC) and inverted mel frequency cepstral coefficient features emerged as excellent discriminative features for spoofing detection. Short-term spectral features like *constant Q cepstral coefficients* (CQCCs) exhibited best results on ASVspoof 2015 data as found in Todisco et al. (2017). Recently, scattering cepstral coefficients (SCCs) which are similar to CQCCs were proposed for stand-alone spoofing detection (Sriskandaraja et al., 2017). The effectiveness of cepstral features for spoofing detection as stand-alone and integrated with ASV using inverted frequency warping scale and overlapped block transformation was demonstrated in Paul et al. (2017a). At classifier level, the usage of deep neural network (DNN) (Chen et al., 2015; Soni et al., 2016), DNN and support vector machine (Villalba et al., 2015), i-vectors (Weng et al., 2015) were used on open set ASVspoof data.

The standard approach for spoofing detection captures artifact traces introduced while doing speech manipulation by SS or VC. This is done by using efficient feature extraction techniques as front-end and standard classifiers at the back-end. The work in Hanilçi et al. (2015) suggests that more reliable spoofing detection can be achieved by discriminative features than complex classifiers. However, most of the countermeasures in literature focused on a particular type of artifacts (phase, pitch contour, high frequency spectral, temporal). ASVspoof 2015 challenge results also showed that most of the submitted systems provided poorer performance for unknown spoofed data types (Wu et al., 2015). It suggests that instead of a single feature, a bank of features may prove to be useful for spoofing detection. Therefore, the aim of this work is to propose an efficient feature combination that can capture possible sources of artifacts and can easily distinguish natural speech from synthetic speech for both known and unknown spoofing

Please cite this article as: M. Pal et al., Synthetic speech detection using fundamental frequency variation and spectral features, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.10.001