Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 81 (2016) 30–41

www.elsevier.com/locate/specom

Synthetic speech detection using phase information

Ibon Saratxaga[a], Jon Sanchez[a,*], Zhizheng Wu[b], Inma Hernaez[a], Eva Navas[a]

[a] AHOLAB Signal Processing Lab., University of the Basque Country, UPV- EHU, Spain
[b] The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK

Received 1 July 2015; received in revised form 24 March 2016; accepted 11 April 2016
Available online 16 April 2016

Abstract

Taking advantage of the fact that most of the speech processing techniques neglect the phase information, we seek to detect phase perturbations in order to prevent synthetic impostors attacking Speaker Verification systems. Two Synthetic Speech Detection (SSD) systems that use spectral phase related information are reviewed and evaluated in this work: one based on the Modified Group Delay (MGD), and the other based on the Relative Phase Shift, (RPS). A classical module-based MFCC system is also used as baseline. Different training strategies are proposed and evaluated using both real spoofing samples and copy-synthesized signals from the natural ones, aiming to alleviate the issue of getting real data to train the systems. The recently published ASVSpoof2015 database is used for training and evaluation. Performance with completely unrelated data is also checked using synthetic speech from the Blizzard Challenge as evaluation material. The results prove that phase information can be successfully used for the SSD task even with unknown attacks.

Keywords: Synthetic speech detection; Phase; MGD; RPS.

1. Introduction

In speech processing, speech synthesis and analysis areas alike, phase information has been traditionally discarded for most conventional applications. The spectral module information is highly correlated with the perceptual features of the speech and there are well established techniques to process them. Phase information has subtler perceptual effects (Alsteris and Paliwal, 2007) (Saratxaga et al., 2012) and tricky features like wrapping make it more difficult to model and process.

This unawareness for phase information in most speech processing techniques can indeed be exploited to detect such a processing on speech, tracing the unintended perturbations of the natural phase patterns left behind by this processing. One particular case where detecting natural speech manipulations can be critical is the speaker verification field.

The first speaker verification (SV) systems tried to resolve the problem of detecting if a voice was certainly from a claimant speaker and not from other (Rosenberg, 1976). The improvement of the SV systems allowed a high success rate solving the problem of naive speaker verification, but the parallel advance in speech manipulation techniques has posed a new menace to these systems: impostors forging speech signals that imitate a particular speaker's voice. This threat was first pointed by Pellom and Hansen (1999) and Masuko et al. (2000), and has received more and more attention in literature as new voice adaptation and transformation techniques have made more feasible to mimic a speaker's voice with less and less material from the original speaker. A detailed survey is published in Wu et al. (2015a).

Nowadays two are the main speech processing techniques that allow the creation of synthetic speech spoofing signals: First, the statistical speech synthesizers (Yoshimura et al., 1999) (Tokuda et al., 2002) using voices adapted to a particular speaker (Yamagishi et al., 2009) even with minimum quality material (Yamagishi et al., 2010). Secondly, the voice conversion (VC) techniques (Jin et al., 2008) (Kinnunen et al., 2012). Both techniques can be used to generate spoofing signals that can successfully deceive state-of-the-art SV

[*] Corresponding author. Tel.: +34 946014122; fax: +34 946014259.
E-mail address: ion@aholab.ehu.eus (J. Sanchez).

http://dx.doi.org/10.1016/j.specom.2016.04.001
0167-6393/© 2016 Elsevier B.V. All rights reserved.

systems with false acceptance rates (FAR) around 80% for synthetic speech and 5% for VC.

A number of countermeasures have been proposed for these attacks. In Satoh et al. (2001), a countermeasure based on the average inter-frame difference was proposed to discriminate between natural and synthetic speech from an HMM-based speech synthesis system. Another similar countermeasure which also uses an average pair-wise distance between consecutive frames was proposed to detect voice-converted speech (Alegre et al., 2013a). Rather than capturing the inter-frame distortions, in Wu et al. (2013) and Alegre et al. (2013b), modulation-based features and local binary pattern-based features were proposed to utilize long-term spectro-temporal information for synthetic speech detection. In Sizov et al. (2015), a countermeasure which uses the same front-end as ASV was proposed to discriminate natural and voice-converted speech. All these countermeasures derive features from magnitude spectra and work well for specific previously known attack techniques.

Phase based parameters are good candidates to detect synthetic speech due to the usual phase information neglect of many speech processing techniques. Phase information can be analyzed in many ways (instantaneous phase, short-term group delay (Banno et al., 1998), anticausal cepstrum (Drugman et al., 2011), and others), but not all the parameters are suitable for statistical modeling as required by a classifier. Phase-based countermeasures proposed by the authors of this work have been used for both synthetic and voice-converted speech detection. In Wu et al. (2012) synthetic speech detectors (SSD) based on cosine normalized phase and modified-group delay (MGD) (Yegnanarayana and Murthy, 1992) are evaluated with converted spoofing signals. In Wu et al. (2013), modulation spectrum derived from the modified group delay spectrum was used for synthetic speech detection. These works have confirmed the effectiveness of phase information in detecting synthetic speech with matched vocoder.

Relative Phase Shift (RPS) representation (Saratxaga et al., 2009) for the harmonic phase has also been used to build SSD systems aimed to detect spoofing signals created with adapted synthetic voices (De Leon et al., 2011) (De Leon et al., 2012) with good results. The initial works were focused on evaluating the actual capability of the RPSs to detect the phase modifications due to the synthetic generation of the spoofing signals. Consequently synthesized impostors were used to model the spoofing attacks. This approach has the double downside of requiring the adaptation of synthetic voices to generate the spoofing samples, and, more important, using particular attacks to train the synthetic models yields that their performance will be attack-dependent, and they will not be able to detect spoofing signals created with another attacking technique.

Once the validity of the RPS based SSD was demonstrated, the problem of avoiding attack dependence of the SSD was addressed in Sanchez et al. (2014) and Sanchez et al. (2015b). In these works, the authors analyze the use of copy-synthesized signals to create the imposter models. This way, the models are not dependent on the particular features of a specific synthesizer, but they can detect any signal created with a vocoder. Multi-vocoder models trained and tested with completely unrelated signals were evaluated with good results.

Recently, the use of phase for synthetic speech detection has been widely adopted, either alone or combined with other parameters, and using different classifiers. Many systems include group delay derived parameters like MGD or all-pole group delay function (APGD) (Sahidullah et al., 2015)(Alam et al., 2015). Other reported phase parameters are cosine phase (Liu et al., 2015), relative phase (Wang et al., 2015), instantaneous frequency (i.e. time derivative of the phase) (Patel and Patil, 2015), baseband phase difference (BDP) and phase at the CGI (pitch synchronous phase) (Xiao et al., 2015) or the RPS (Villalba et al., 2015) (Sahidullah et al., 2015)(Sanchez et al., 2015a).

In this paper we review and evaluate two phase based SSD systems known for their good performance in statistical modeling and classification: a MGD based and a RPS based SSD system, benchmarking them against a spectral module based (MFCC) baseline system. In this work we especially analyze the optimal use of training material comparing the strategy of using “real” spoofing signals versus using copy-synthesis signals from the natural ones.

Recently the work in this area has been promoted by the ASVSpoof2015, the Automatic Speaker Verification Spoofing and Countermeasures Challenge (Wu et al., 2014). The participants were invited to submit the results of independent SSD modules for evaluation. Spoofing detection systems were tested with a database (the so-called ASVSpoof database), containing different spoofing techniques such as speech synthesis and voice conversion. The performance of the different systems was assessed by the organization using standard metrics. This database has been made available to the public, and we are using it in this work to evaluate our SSD systems.

The performance of the systems with unknown signals is also evaluated using a completely unrelated set of signals from the Blizzard Challenge (Black and Tokuda, 2005). This is the most popular international event for TTS system evaluations, where independent participants build synthetic voices using a common speech corpus and send some samples to be evaluated. They are, undoubtedly, a representative sample of the current technology in speech synthesis, and, consequently, of the kind of likely spoofing technique.

Furthermore, the tests with a completely unrelated database, as the Blizzard Challenge one, introduce the channel-mismatch issue for spoofing detection. While in the ASVSpoof Challenge the same recording channel is assumed for every signal, the channel information of Blizzard Challenge data is different from ASVSpoof data. The robustness to the channel of the different SSDs has been little studied in literature and will be analyzed in this work for the proposed systems.

The paper is organized as follows. First, the phase representation and parameterization methods – RPS and MGD – are described. Then, in Section 3, the Synthetic Speech De-