Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 67 (2015) 143–153

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Playback attack detection for text-dependent speaker verification over telephone channels ☆

Jakub Gałka *, Marcin Grzywacz, Rafał Samborski

AGH University of Science and Technology, Faculty of Computer Science, Electronics and Telecommunications, Poland

Received 21 May 2014; received in revised form 1 December 2014; accepted 3 December 2014
Available online 11 December 2014

Abstract

Playback attacks constitute one of the biggest threats in biometric speaker verification systems, in which a previously recorded passphrase is played back by an unprivileged person in order to gain access. This paper features a description of the playback attack detection (PAD) algorithm, designed to protect text-dependent speaker verification systems from the aforementioned spoofing attacks. The paper also describes the usage of spectral landmarks and score normalization methods in the playback detection algorithm. Different factors are discussed in terms of the performance of the algorithm. The authors investigate two issues: (1) extracting the PAD features which are robust against channel variations and (2) the robustness of the algorithm in adverse acoustical environments (e.g. office, street, cocktail party noise). The experiments are performed on a prepared speech corpus containing 4187 occurrences of a passphrase spoken by 175 speakers. The results of the experiment show the equal error rate (EER) to be as low as 1.0%. These findings demonstrate that such spoofing-oriented playback attacks can be effectively detected and should not be considered a significant argument against applications of text-dependent speaker verification.

Keywords: Speaker verification; Playback attack detection; Telephone channel; Spectral landmarks

1. Introduction

The task of biometric speaker verification is to accept or reject the identity claim of the speaker based on a sample of the speaker's voice. Telephone-based automatic speaker verification (by use of telephone channel) has already been a subject of research (Murthy et al., 1999; Kinnunen et al., 2012). Despite the fact that such systems perform very well, reaching relatively low EERs in demanding testing scenarios (NIST, 2012), consumers and organizations still have their doubts in the context of high-security applications (e.g. e-banking). One of the prevailing arguments against voice biometry concerns common passphrase text-dependent systems, in which the passphrase uttered by the speaker does not change from one login attempt to another. This enables the possibility of breaking into such systems by playing back a recording obtained earlier, using a microphone or any other eavesdropping method (e.g. malicious mobile software). This type of attack is called a playback attack and is available to anyone with minimal signal processing knowledge.

One of the solutions to this problem is to use a text-prompted system in which the user is asked to speak a randomly selected phrase for each access attempt. It is worth noting that such systems are more sensitive to other types of attacks (such as the concatenation of previously recorded digits) (Lindberg and Blomberg, 1999), and, due

☆ This work was supported by the Polish National Centre for Research and Development – Applied Research Program under Grant PBS1/B3/1/2012 titled "Biometric voice verification and identification".
* Corresponding author.
E-mail addresses: jgalka@agh.edu.pl (J. Gałka), mar.grzyw@gmail.com (M. Grzywacz), rafal.samborski@gmail.com (R. Samborski).

http://dx.doi.org/10.1016/j.specom.2014.12.003
0167-6393/

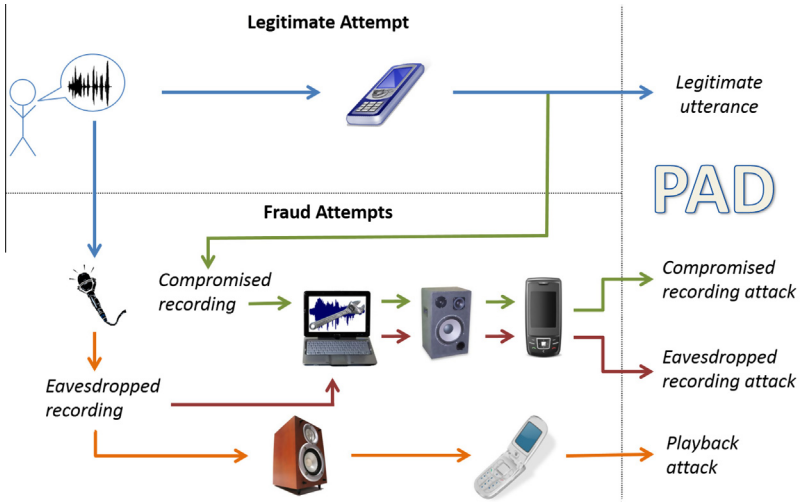

Fig. 1. Illustration of both legitimate and fraud verification attempts described in this document.

to the fact that the system cannot use lexical knowledge in its assessment, they achieve higher error rates, as compared to text-dependent solutions (Boves and den Os, 1998). The presented work focused on safeguarding text-dependent systems against playback attacks.

Several methods of playback attack detection are described, the bibliography on this subject not being very extensive. Employing direct spectral features such as the low frequency ratio was investigated in Villalba and Lleida (2011). The comparison of maps containing the highest peaks of the magnitude spectrum was described in Stevenson (2008). Another method, making use of a specific channel pattern, was presented in Wang et al. (2011). Despite the possibility of the normalization of similarity scores dramatically increasing the effectiveness of any of the aforementioned methods, there is little existing research on the subject. Shang and Stevenson (2010) successfully used the relative similarity score, which resulted in a reduction of the EER from 11.94% to 6.81%. None of the authors presented any kind of analysis of the impact of noise present in the attacking recording on detection performance. This is one of the reasons for the existence of the aforementioned algorithm. One of the objectives of this work was to achieve high noise robustness in a wide range of signal-to-noise ratio (SNR) of root mean square values. Another goal was taking advantage of the features available in devices which require high-speed data processing and have limited memory resources, such as embedded systems, physical biometric locks or other small-scale consumer electronics.

The method described in this paper uses both spectral features and score normalization to obtain a robust algorithm that addresses the issues of operating in an adverse acoustic environment, such as the one mentioned above. The paper is divided as follows: In Section 2, the core PAD algorithm is described, the corpus recorded for use in the experiments is described in Section 3, Section 4 provides the results of the conducted evaluations and covers the method of score normalization, in Section 5 conclusions are made and future work is discussed.

To improve the clarity of the paper, verification scenarios are presented in Fig. 1 and the following terms are defined:

Target: a privileged user, owner of data protected by biometric security.
Impostor: an unauthorized person claiming to be the owner of protected data, who attacks the system by modifying a previously acquired recording of the privileged (*Target*) user.
Legitimate: a non-playback-based verification attempt of the target.
Fraud: a playback attack by an impostor.
Authentic recording: a recording of a successful target verification attempt acquired server side.
Eavesdropped recording: a recording intercepted by an impostor on the client-side of the telecommunication channel during a legitimate verification attempt.
Compromised recording: a recording intercepted by an impostor on the server side of the telecommunication channel during a legitimate target's verification attempt, or a recording stolen from the server's user database.
Playback recording: the eavesdropped recording played back by the impostor and received at the server side of the telecommunication channel.

2. PAD algorithm

In this section, the PAD algorithm is presented. The concept of this solution is based on the music recognition system presented in Wang (2003) and Ellis (2009). Wang's idea of the algorithm is based on comparing recordings on the basis of the similarity of the local configuration of maxima pairs extracted from spectrograms of verified and reference recordings. According to the author of Wang (2003), the