SPECOM 2251
5 November 2014
ARTICLE IN PRESS
No. of Pages 24, Model 5+

ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication xxx (2014) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Spoofing and countermeasures for speaker verification: A survey

Zhizheng Wu [a,*], Nicholas Evans [b], Tomi Kinnunen [c], Junichi Yamagishi [d,e], Federico Alegre [b], Haizhou Li [a,f]

[a] *Nanyang Technological University, Singapore*
[b] *EURECOM, France*
[c] *University of Eastern Finland, Finland*
[d] *National Institute of Informatics, Japan*
[e] *University of Edinburgh, United Kingdom*
[f] *Institute for Infocomm Research, Singapore*

Received 7 April 2014; received in revised form 30 September 2014; accepted 24 October 2014

Abstract

While biometric authentication has advanced significantly in recent years, evidence shows the technology can be susceptible to malicious spoofing attacks. The research community has responded with dedicated countermeasures which aim to detect and deflect such attacks. Even if the literature shows that they can be effective, the problem is far from being solved; biometric systems remain vulnerable to spoofing. Despite a growing momentum to develop spoofing countermeasures for automatic speaker verification, now that the technology has matured sufficiently to support mass deployment in an array of diverse applications, greater effort will be needed in the future to ensure adequate protection against spoofing. This article provides a survey of past work and identifies priority research directions for the future. We summarise previous studies involving impersonation, replay, speech synthesis and voice conversion spoofing attacks and more recent efforts to develop dedicated countermeasures. The survey shows that future research should address the lack of standard datasets and the over-fitting of existing countermeasures to specific, known spoofing attacks.

Keywords: Automatic speaker verification; Spoofing attack; Countermeasure; Security

Contents

1. Introduction . . . 00
2. Automatic speaker verification . . . 00
   2.1. Feature extraction . . . 00
   2.2. Speaker modelling and classification . . . 00
   2.3. System fusion . . . 00

* Corresponding author at: University of Edinburgh, United Kingdom.
E-mail addresses: zhizheng.wu@ed.ac.uk (Z. Wu), evans@eurecom.fr (N. Evans), tkinnu@cs.uef.fi (T. Kinnunen), jyamagis@inf.ed.ac.uk (J. Yamagishi), alegre@eurecom.fr (F. Alegre), hli@i2r.a-star.edu.sg (H. Li).

http://dx.doi.org/10.1016/j.specom.2014.10.005
0167-6393/

3. Vulnerability of speaker verification to spoofing . . . 00
3.1. Possible attack points . . . 00
3.2. Potential vulnerabilities . . . 00
3.2.1. Feature extraction . . . 00
3.2.2. Speaker modelling . . . 00
4. Evaluation protocol . . . 00
4.1. Dataset design . . . 00
4.2. Evaluation metrics . . . 00
5. Spoofing and countermeasures . . . 00
5.1. Impersonation . . . 00
5.1.1. Spoofing . . . 00
5.1.2. Countermeasures . . . 00
5.2. Replay . . . 00
5.2.1. Spoofing . . . 00
5.2.2. Countermeasures . . . 00
5.3. Speech synthesis . . . 00
5.3.1. Spoofing . . . 00
5.3.2. Countermeasures . . . 00
5.4. Voice conversion . . . 00
5.4.1. Spoofing . . . 00
5.4.2. Countermeasures . . . 00
6. Discussion . . . 00
6.1. Spoofing . . . 00
6.2. Countermeasures . . . 00
6.3. Generalised countermeasures . . . 00
7. Issues for future research . . . 00
7.1. Large-scale standard datasets . . . 00
7.2. Evaluation metrics . . . 00
7.3. Open-source software packages . . . 00
7.4. Future directions . . . 00
8. Conclusions . . . 00
References . . . 00

1. Introduction

Various distinctive and measurable physiological and behavioural traits have been investigated for biometric recognition (Jain et al., 2006). As our primary method of communication, speech is a particularly appealing modality. Individual differences in both physiological and behavioural characteristics, e.g. the vocal tract shape and intonation, can be captured and utilised for automatic speaker verification (ASV) (Kinnunen and Li, 2010).

Recent advances in channel and noise compensation techniques have significantly improved ASV performance to levels required for mass-market adoption. Reliable and efficient authentication is now possible in smartphone logical access scenarios (Lee et al., 2013) and in e-commerce (Nuance, 2013) for example. Even though ASV provides a low-cost and convenient approach to authentication, however, reliability in the face of spoofing remains a concern (Evans et al., 2013; Evans et al., 2014b).

A generic biometric system may be manipulated or attacked at various stages between sample acquisition and the delivery of an authentication result (Ratha et al., 2001; Faundez-Zanuy, 2004; Galbally et al., 2010). In the specific case of ASV as illustrated in Fig. 1, attacks at both the microphone and transmission levels are generally considered to pose the greatest threat (Faundez-Zanuy et al., 2006). Here, an adversary, typically referred to as an impostor, might seek to deceive the system by impersonating another enrolled user at the microphone in order to manipulate the ASV result. Alternatively, captured speech signals can be intercepted and replaced at the transmission level by another specially crafted voice signal. Since speaker recognition is commonly used in telephony, or other unattended, distributed scenarios without human supervision or face-to-face contact, speech is arguably more prone to malicious interference or manipulation than other biometric signals; the potential for ASV systems to be spoofed is now well-recognised (Evans et al., 2013; Evans et al., 2014b; Wu and Li, 2013).

Prior to the consideration of spoofing, ASV systems were designed to distinguish between target speakers and zero-effort impostors. This research focuses on improving fundamental recognition performance, as opposed to security or robustness to spoofing and drove the community to investigate different approaches to speaker characterisation at the feature level including: (i) short-term spectral and voice source features, such as Mel-frequency cepstral coefficients (MFCCs) and glottal pulse features; (ii) prosodic and spectro-temporal features such as rhythm, pitch and other segmental information; (iii) high-level features such