Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 72 (2015) 13–31

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Automatic versus human speaker verification: The case of voice mimicry

Rosa González Hautamäki[a,*], Tomi Kinnunen[a], Ville Hautamäki[a], Anne-Maria Laukkanen[b]

[a] Speech and Image Processing Unit, School of Computing, University of Eastern Finland, P.O. Box 111, FI-80101 Joensuu, Finland
[b] Speech and Voice Research Laboratory, School of Education, University of Tampere, FI-33014 Tampere, Finland

Received 22 July 2014; received in revised form 1 May 2015; accepted 3 May 2015
Available online 8 May 2015

Abstract

In this work, we compare the performance of three modern speaker verification systems and non-expert human listeners in the presence of voice mimicry. Our goal is to gain insights on how vulnerable speaker verification systems are to mimicry attack and compare it to the performance of human listeners. We study both traditional Gaussian mixture model-universal background model (GMM-UBM) and an i-vector based classifier with cosine scoring and probabilistic linear discriminant analysis (PLDA) scoring. For the studied material in Finnish language, the mimicry attack decreased lightly the equal error rate (EER) for GMM-UBM from 10.83 to 10.31, while for i-vector systems the EER increased from 6.80 to 13.76 and from 4.36 to 7.38. The performance of the human listening panel shows that imitated speech increases the difficulty of the speaker verification task. It is even more difficult to recognize a person who is intentionally concealing his or her identity. For Impersonator A, the average listener made 8 errors from 34 trials while the automatic systems had 6 errors in the same set. The average listener for Impersonator B made 7 errors from the 28 trials, while the automatic systems made 7 to 9 errors. A statistical analysis of the listener performance was also conducted. We found out a statistically significant association, with $p = 0.00019$ and $R^2 = 0.59$, between listener accuracy and self reported factors only when familiar voices were present in the test.

Keywords: Voice imitation; Speaker recognition; Mimicry attack; Listening test

1. Introduction

Speaker verification (Campbell, 1997; Reynolds, 2002) is the task of recognizing persons from their voices. The accuracy of speaker verification systems has steadily improved in the recent years due to advances in channel, noise and inter-session compensation techniques, making the technology available for tailored applications. Automatic speaker verification (ASV) technology is generally used under three scenarios. Firstly, *authentication* applications involve verifying the identity of a cooperative user who demands physical or logical access. Secondly, a *forensic* scenario involves comparing two speech samples to determine whether they originate from the same or different subject. Finally, *screening* and *indexing* applications involve searching a particular target speaker from large amounts of unlabeled data.

One of the increasing concerns in practical uses of ASV technology is vulnerability of the recognizers to intentional circumvention (Wu et al., 2015). In the first case, authentication, this refers to dedicated effort to manipulate one's speech so that an ASV system would misclassify the attacker's sample to originate from the target (client). There are

[*] Corresponding author.
E-mail addresses: rgonza@cs.uef.fi (R. González Hautamäki), tkinnu@cs.uef.fi (T. Kinnunen), villeh@cs.uef.fi (V. Hautamäki), anne-maria.laukkanen@uta.fi (A.-M. Laukkanen).

http://dx.doi.org/10.1016/j.specom.2015.05.002
0167-6393/© 2015 Elsevier B.V. All rights reserved.

four main types of such *spoofing attacks* (Evans et al., 2013; Wu et al., 2015): mimicry, replay (Villalba and Lleida, 2011), speaker-adapted speech synthesis (De Leon et al., 2012) and voice conversion (Kinnunen et al., 2012). A common feature of all spoofing attacks is that the attacker uses non-zero effort to circumvent an ASV system, for instance, with financial motivation. This is different from the latter two use cases, forensics and screening, where the person in question may desire *not* to be detected as him/herself, and is therefore being considered to be non-cooperative. This type of circumvention, with an intention to provoke false rejections (misses), is known as *evasion* or *obfuscation* (Alegre et al., 2014). Similar to spoofing, evasion could be achieved by both technical means (for instance, by adding reverberation) and by disguising one's speech by, for instance by raising F0 or imitating a foreign accent (Zhang and Tan, 2008; Kajarekar et al., 2006). We should also point out that some speakers, without any voluntary effort to spoof or evade recognizers, tend to be confused with other users (Doddington et al., 1998; Yager and Dunstone, 2010). In this study, we focus on scenarios with intentional speech modification, namely, mimicry.

Speech mimicry is an interesting research phenomenon for several reasons. Firstly, most readers are likely to be familiar with talented impersonators (often stand-up comedians) in their mother tongue who are able to create funny, yet convincing-sounding impersonations of politicians or other public figures. We, as ASV researchers, are frequently asked whether such impersonators would be able to spoof ASV systems; a general belief is that human listeners can be fooled but ASV system accuracy is not affected by mimicry attacks. Table 1 summarizes some of the previous speaker recognition studies for mimicry data. Secondly, studying mimicry as a potential spoofing technique is also relevant. Detection of technical spoofing attacks, such as speech synthesis and voice conversion, can already to a certain extent be achieved by designing discriminative features known to differentiate synthetic and natural utterances (De Leon et al., 2012; Alegre et al., 2013; Wu et al., 2012). Clearly, such countermeasures are inapplicable for detection of impersonation produced by a real human being, making mimicry a challenging test case for spoofing countermeasure development, and particularly interesting for forensic and speech security applications. Thirdly, looking from the perspective of the impersonator, ASV technology could be used as an objective feedback tool to evaluate the similarity of one's impersonations against the intended target speaker. Such technology might help, for instance, actors to help practicing idiosyncratic speech of their characters.

The general challenges related to studies that involve mimicry include lack of a standard corpus for evaluation and technical mismatches. While there are standard and public corpora to benchmark speaker verification systems under zero-effort imposture, this is not the case regarding mimicry attacks; professional impersonators are not easily available to provide speech samples, and target speakers are often public figures whose samples are collected from public sources. Naturally, mismatches of audio recordings arise when professional impersonators' speech is collected in a studio environment and the target speakers' recordings from TV and radio interviews. An alternative way to analyze the mimicry attack is to include a perceptual test as a benchmark parallel to automatic system analysis. A human benchmark, compared to automatic systems in a zero-effort imposture setting, has been used in previous studies (Schmidt-Nielsen and Crystal, 2000; Hautamäki et al., 2010). In terms of human assisted speaker verification system (Greenberg et al., 2011; Hautamäki et al., 2010; González Hautamäki et al., 2013a), such as a forensic system, it is important to know how a non-cooperative subject could either mimic some other speaker or disguise his or her voice.

In the present study, we analyze voice mimicry attacks with audio material from the speakers described in Section 3, extending our preliminary analyses reported in González Hautamäki et al. (2013b, 2014). The current study extends these preliminary studies both regarding data and analyses. Firstly, we have collected fresh data from a new impersonator who mimics four additional target speakers presented in neither González Hautamäki et al. (2013b) nor in González Hautamäki et al. (2014). Secondly, a new human benchmark involving a large listening panel was also added.

Overall, our major contribution is an up-to-date analysis of mimicry attacks against state-of-the-art automatic speaker verification systems accompanied by a relatively large-scale human benchmark. Earlier studies on mimicry attacks (Table 1) have included classical spectral

Table 1
Some of the previous studies on mimicry data and the present study. Previous studies concentrate on acoustical analysis and Gaussian Mixture Model (GMM) with and without universal background model (UBM).

Study	Target language	Target speakers	Impersonators	Speaker verification
Lau et al. (2004)	English	6	2 naïve	GMM
Lau et al. (2005)	English	6	2 professional linguists, 4 naïve	GMM
Mariéthoz and Bengio (2005)	French	3	1 professional, 1 intermediate and 1 naïve	GMM-UBM
Zetterholm (2007)	Swedish	9	2 professional, 1 amateur	Auditory analysis by a panel
Farrús et al. (2010)	Spanish-Catalan	5	2 professional	Prosodic system
Panjwani and Prakash (2014)	English	53	3 professional and 13 naïve	GMM-UBM
This study	Finnish	8	2 professional	GMM-UBM, i-vector cosine and i-vector-PLDA, perceptual test