I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry

Rosa González Hautamäki[1], *Tomi Kinnunen*[1], *Ville Hautamäki*[1], *Timo Leino*[2], *Anne-Maria Laukkanen*[2]

[1]School of Computing, University of Eastern Finland, Joensuu, Finland
[2]Speech and Voice Research Laboratory, School of Education, University of Tampere, Finland
{rgonza, tkinnu, villeh}@cs.uef.fi,
{timo.leino, Anne-Maria.Laukkanen}@uta.fi

Abstract

Voice imitation is mimicry of another speaker's voice characteristics and speech behavior. Professional voice mimicry can create entertaining, yet realistic sounding target speaker renditions. As mimicry tends to exaggerate prosodic, idiosyncratic and lexical behavior, it is unclear how modern spectral-feature automatic speaker verification systems respond to mimicry "attacks". We study the vulnerability of two well-known speaker recognition systems, traditional Gaussian mixture model – universal background model (GMM-UBM) and a state-of-the-art i-vector classifier with cosine scoring. The material consists of one professional Finnish imitator impersonating five well-known Finnish public figures. In a carefully controlled setting, mimicry attack does slightly increase the false acceptance rate for the i-vector system, but generally this is not alarmingly large in comparison to voice conversion or playback attacks.

Index Terms: Voice imitation, speaker recognition, mimicry attack

1. Introduction

Speaker verification is the task of verifying the speaker's identity based on his or her speech sample [1]. Due to abundance of smartphones, speaker verification has a huge market potential as a low-cost authentication method to remote services such as e-banking, or in verifying the identity of the device user. As an example, in 2012 a new phone model that has a built-in speaker verification technology was introduced [2].

Recognition accuracy of speaker verification methods has come a long way during the last decade, largely thanks to advanced statistical methods to compensate for channel, intersession and environmental noise effects. But even the most advanced recognizers can easily be spoofed by replay attacks [3], voice conversion [4] and adaptive speech synthesis [5].

In addition to such *technical* spoofing techniques, *voice imitation* [6, 7, 8] – human-based mimicry of another speaker's voice quality and vocal behavior – represents an interesting attack scenario for several reasons. First, unlike the technical means of spoofing that can partially be counter-attacked for by searching traces of signal processing artifacts introduced by synthesis techniques (e.g. [9, 5, 10]), imitators are actual human beings that produce technically valid speech inputs. Mimic attacks cannot be detected by such methods. Second, it is unclear how modern recognizers respond to mimicry attacks. While human imitators, often in an entertainment setting, are likely to copy lexical, prosodic and idiosyncratic behavior of their target speakers, state-of-the-art recognizers typically use only low-level spectral features to recognize speakers. Imitation is also relevant in forensic settings involving speech of identical twins, for instance. To sum up, besides raising phonetically interesting basic research questions, the study of human mimicry attacks is relevant for studies of recognizer vulnerability.

Table 1: *Impersonator and target speakers. YLE=Yleisradio, Finnish national public broadcasting company.*

TARGET SPEAKERS			
Name	**Position**	**Source material**	**Duration (mins)**
Martti Ahtisaari	Former president, UN mediator	YLE Radio	6:31
Hjallis Harkimo	Politician, businessman	YLE Radio	6:20
Sauli Niinistö	Current president of Finland	YLE Radio	6:31
Jouko Turkka	Theatrical director	YLE Arch.	6:28
Matti Vanhanen	Former prime minister	YLE Radio	6:15
IMPERSONATOR			
Reijo Salminen	Impersonator, Singer	Studio recording	10:60(own) 5:52 (imp.)

This study focuses on voice mimicry attacks to analyze the vulnerability of speaker recognition systems with audio material from the speakers described in Table 1. Before proceeding to detailed description of our data, we should highlight the general challenges in studying the problem. The main constraints are:

- **There is no standard evaluation corpus:** Unlike for NIST speaker recognition evaluation campaigns, standard or public corpora to study imitation attacks are nonexisting.
- **Data is scarce and expensive to collect:** Imitators are often professional mimics, singers or voice actors whose time is expensive. There are few professional impersonators.
- **Technical mismatches are inevitable:** As the target speakers are usually politicians or other public figures, it is challenging to have their, and the imitator's voice, to be recorded in technically matched conditions.

In the context of the present study, we cannot do much about the first two. Our imitated speech material is taken off-the-shelf from an earlier study carried out at the University of Tampere

Table 2: Previous studies on mimicry attack to speaker verification systems.

Study	Target language	Target speakers	Impersonators	Speaker verification	FAR or IER
Lau *et* al.(2004) [11]	English	Closest, average and furthest targets from YOHO corpus selected with automatic system	2 naïve	GMM	n/a
Lau *et* al. (2005) [12]	English	Similar to [11]	2 professional linguist, 4 naïve	GMM	30 - 40 %
Mariethoz & Bengio (2006) [13]	Swiss	3	1 professional, 1 intermediate and 1 naïve	GMM	n/a
Zetterholm (2007) [8]	Swedish	9	3	Auditory analysis by a panel	n/a
Farrús *et* al.(2010) [6]	Spanish	5	2 Professional	Prosodic parameters	5 - 22% (IER)
This study	Finnish	5	1 Professional	ivector-cosine, GMM	9 - 12%

[14], while the target speaker material is mostly collected from public radio and TV resources in Finland. The amount of data is comparable to other studies on the topic, see Table 2. Regarding the third challenge, mismatches in channel, recording environment and session problems are common within the context of NIST speaker recognition evaluations. State-of-the-art recognizers involving channel and intersession normalization can tackle these relatively well [15], and therefore, provide a suitable test bench to study imitation spoofing. But the state-of-the-art techniques also require massive quantities (thousands of hours) of development speech to train their hyperparameters such as the universal background model [16]. The hyperparameters are typically trained from English utterances and it is not obvious whether such recognizer can be meaningfully configured for a different language, which is the case here.

The primary contribution of our study is to compare two high-performance speaker recognition systems, traditional *Gaussian mixture model - universal background model* (GMM-UBM) recognizer [16], and state-of-the-art *i-vector* recognizer [17]. Our work extends the prior art on the topic [6, 7, 8] involving experiments on a previously unstudied, phonetically rich language, Finnish. Importantly, for the first time it compares the accuracies of GMM-UBM and i-vector recognition systems. Additionally, Section 2 provides a survey of related literature that helps us in interpreting the results.

2. Imitation and speaker verification

A few earlier studies have analyzed the effects on voice imitation and its effects to recognition of a speaker either by prosody or acoustical methods [6, 7, 8]. Professional impersonators, specially in entertainment, mimic certain characteristics related to prosody, pitch, voice quality, dialect and speech style of a target speaker. In the works by [6, 8], different techniques used by professional imitators are studied. The studies defined that the impersonators are able to adapt the fundamental frequency and the formant frequencies of the target voices. This presents a potential vulnerability to automatic speaker recognition systems that mainly utilize spectral features.

In [6], the authors try to quantify how much a speaker is able to approximate others' voices, by focusing in the selection of prosodic and acoustic features from two professional impersonators that imitated well-known politicians. In contrast to our study, the authors use an automatic speaker recognition system based on both prosodic and acoustic features. Prosodic parameters used in the experiment included words' duration, word segments, means and ranges of the fundamental frequency, as well as jitter and shimmer measurements. In their imitation experiment, the identification error rate increased when the score level fusion of the 12 prosodic features was performed.

In [11], the authors used YOHO corpus like [18]. The authors used two naïve impersonators (native Chinese, living in Australia more than 7 years) with no experience in mimicry. Having the natural voices of these targets, the authors used a spectral GMM system to pick 3 different speakers, the *closest, intermediate and furthest* speaker from YOHO. Then the imitators read all 40 training utterances from the three speakers, listened to the samples and tried to imitate them. There were four recording sessions for both impersonators because the authors wanted to see whether the imitators become better with more training and concluded that the verification errors indeed increased as a function of training times. An interesting observation is that both of these "naïve" impersonators could be accepted by the system as that speaker they were imitating. However, this was true *only* for the closest speaker. Neither imitator was able to be accepted as the intermediate or further speaker. This seems to suggest that speakers whose vowel space is similar to that of the imitator tend to be easily imitated, likely due to similar articulatory constraints. If the articulators are very different, it will be difficult or impossible to modify the voice sufficiently towards the target.

In a similar study [12], the same authors tested two groups of imitators: professional and non-professional, where the professional group consisted of one female linguist and one male linguist. Four other naïve imitators (two Chinese male/female, two Australian male/female) were in the non-professional group. Like in their other study [11], some speakers to be imitated were selected from YOHO corpus, this time only the most similar speakers in the sense of GMM likelihood. Three recording sessions for each imitator were taken. For the first professional (female linguist) the *false acceptance rate* (FAR) increased from practically 0 % to 60 %; for the male linguist, to (only) 10 % FAR using the same threshold setting. For the amateur female imitators (1 Chinese, 1 Australian), the numbers were around 20 % to 30 % FAR. The Australian male achieved similar result to the male linguist. However, the Chinese male could achieve as high as 60 % FAR. The study suggests that, independently of whether professional or not, the error rates were increased, and that linguists impersonators are not necessarily better. At least in the case when the voices of the target speakers are similar to the source speakers.

In a different study, the authors of [19], describe a technique to evaluate the quality of mimicked speech using prosodic features. The material used consists of text dependent and text indepedent utterances from 15 professional mimicry artists impersonating 7 celebrities. The effectivenes of their evaluation technique was measured with a perceptual study including 15 listeners.

Besides these studies with mimicry attack, voice conversion by means of speech synthesis and by playing a recording with the voice of the target speaker, or replay attack, have been studied to investigate the performance of speaker recognition systems. In [4], the authors applied voice conversion techniques to simulate a spoofing attack and in that way measure the vulnerability of speaker recognition systems. The study concludes that with a simple voice conversion the systems showed degradation in their performance. Nevertheless it is obvious that a human listener is able to judge that the samples sound unnatural. In [3], the authors defined a replay attack detection system and how to incorporate it to a speaker recognition system to reduce the recognition error.

3. Material

The speech material, described in Table 1, consists of one male Finnish impersonator imitating 5 well-known Finnish public figures. The language is Finnish. The target speakers data was collected from public radio interviews and TV programs. The impersonator's audio samples were recorded in a studio environment. All the speech samples were down-sampled to 8kHz from 44.1 kHz and converted to mono.

3.1. Corpus design

For the enrollment phase, the target speakers training material includes 5 minutes of active speech after performing human-based voice activity detection. For the experiments, the test segments were chunked from long recordings to 20 seconds segments, considering the amount of imitated speech available in our corpus. The impersonator's own voice was recorded reading text fragments from interviews of the target speakers. The imitation samples include 2 recordings per target speaker. The test segments and trial lists are prepared in a way to directly analyze the performance of the automatic systems in a NIST-like speaker recognition evaluation setting.

4. Experiments

The baseline test includes the 5 target speakers with 27 test segments, making equal number of genuine trials. The *baseline* case does not include imitation samples, instead the impersonator's *natural* voice recordings are used as impostor trials. All the test segments have a duration of 20 seconds. For the *mimicry attack* test, the genuine trials were maintained and the impostor trials consisted of imitations by the impersonator of the target speakers with 31 test segments, adding to 155 trials in total. Table 3 shows the trial statistics. In both settings, the impersonator is the same speaker impostor. In this way, the effect of imitation, when the impostor uses his natural voice (baseline case) and when he tries to sound like a target speaker (mimicry attack case), can be studied.

Table 3: *Imitation attack test trials.*

	Total	Duration
Genuine trials	27	7 mins.
Impostor trials	155	30 - 40 mins

4.1. Speaker recognition systems

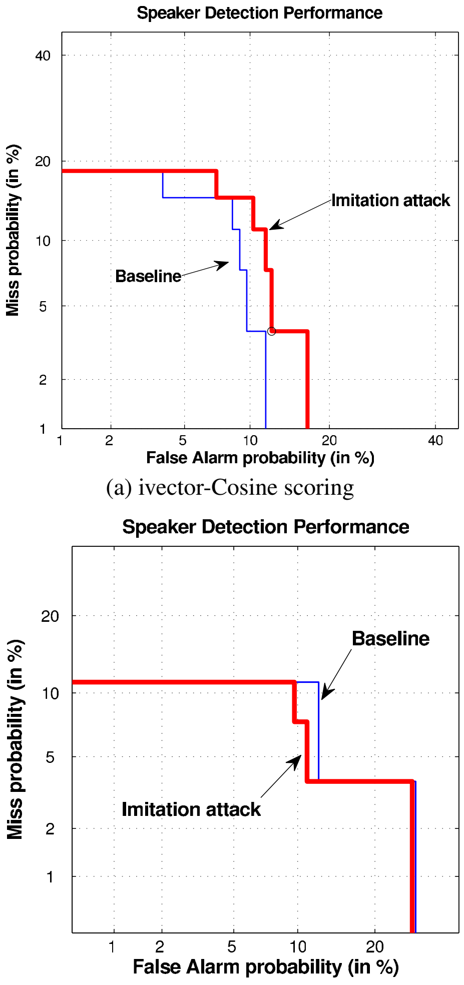
This study considers two speaker recognition systems: a traditional **Gaussian mixture model with universal background model** (GMM-UBM) [16] and a state-of-the-art **i-vector with cosine scoring** [17].

The *GMM-UBM* is the standard system with UBM trained with NIST 04, 05, 06 and 08 data and 512 Gaussians components.

For the *i-vector with cosine scoring*, given two utterances represented by two vectors in the *i-vector* space [20], the angle between the two vectors, or *cosine similarity*, is considered as a measure to compare them and use it to make decisions in the recognition process. The cosine similarity for i-vector based speaker recognition was introduced in [21]. 54-dimensional MFCCs are extracted and a gender-dependent universal background model (UBM) with 512 Gaussian components is trained from NIST 04, 05, 06 and 08 data. The i-vector extractor is 400-dimensional, T-matrix is trained using the same data set plus Fisher and Switchboard.

5. Results

Figure 1 shows the standard *detection error trade-off* (DET) curves for the baseline and the mimicry attack test. It was observed that at the *equal error rate* (EER) point for i-vector-Cosine system, only 2 target trials were classified as non target and 18 out of 155 non-target trials were classified as target. For the GMM-UBM, no significant differences are observed from the DET curves.



(a) ivector-Cosine scoring

(b) GMM - UBM

Figure 1: DET curves for both systems.

To better analyze the effect of imitation spoofing, we set the decision threshold to the EER threshold on the baseline data, similar to [4]. The *false acceptance rate* (FAR) is then measured on the imitation data. Table 4 shows that FAR slightly increases in the imitation case for the i-vector system, but decreases for the GMM-UBM system. When imitation data is included in the test, recognition accuracy decreases in a range acceptable for its performance. This means that the imitation attack scenario studied here does not affect the performance of the selected systems.

Table 4: *Effect of mimicry attack to false acceptance rates (FAR %). Decision threshold is set to EER point on the baseline data.*

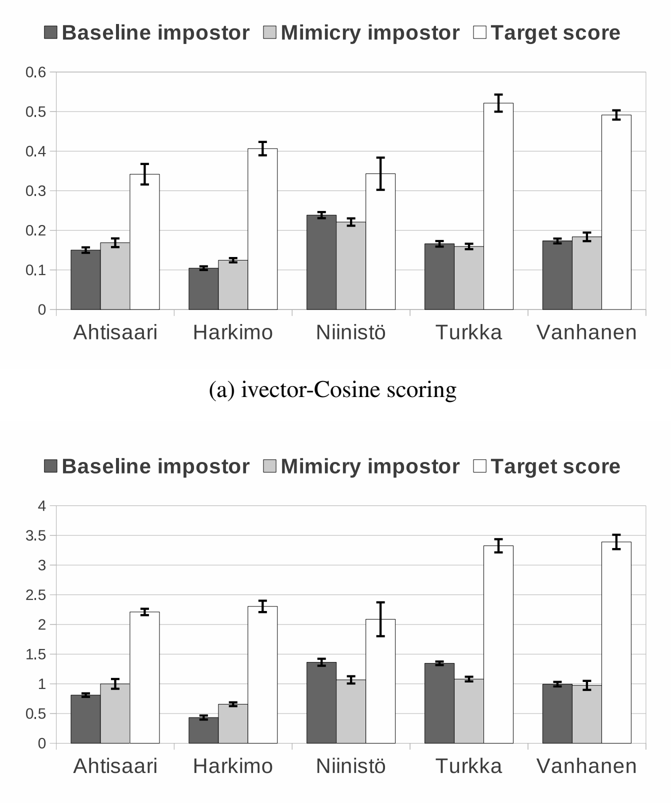
Test	i-vector Cosine	GMM-UBM
Baseline	9.03	11.11
Mimicry attack	11.61	9.68

The DET plots in Fig. 1 are presented given that they are a standard tool for assessing speaker verification accuracy. Admittedly, the usefulness of DET plot for the present analysis is limited due to sparse data, which is typical in imitation studies. Therefore, a possibly more insightful analysis could be obtained by studying the response of the recognition systems to individual target speakers; this is shown in Fig. 2. This graph displays the average recognizer score per target speakers before (baseline) and after the attack (mimicry). The *standard errors of the mean* (SEM), with 95% confidence range are also given.

Several interesting observations can be made. Firstly, comparing the heights of the baseline graphs – a measure of the similarity of our imitator's *natural* voice against a particular target – Niinistö appears to be the most similar to the imitator's voice, while Turkka and Vanhanen have lower recognizer scores. The same pattern holds for both recognizers. Previous literature [8, 12] has suggested that imitation attacks against "similar" target speakers might be easier than against speakers with very different voice quality. According to Fig. 2, this is *not* the case here; the imitation scores against the most similar target, Niinistö, in fact *lower* the scores, while the relative increase is largest with the most dissimilar target, Harkimo. The difference of scores in GMM-UBM system is more visible than in i-vector system. In the case the recognition threshold is carefully set between the impostor's natural voice scores and his imitations of Ahtisaari and Harkimo, he maybe accepted as the target speaker. It must be kept in mind that the recognizers used in [11] had high-quality clean input signal and controlled text passages to select similar and dissimilar speakers, while our study deals with a scenario with free text-inputs and involve nonperfect and unknown target recordings. While our speaker verification systems are able to cope with this variability (the baseline error rates are on typical range of GMM-based recognizers), it is likely that the additional effect due to impersonation gets masked under the problems introduced in both lexical and channel differences. In order to exclude these effects and focus solely on the success of imitator, ideally all the speech samples, target speakers and impersonator, should be collected under the same, clean conditions. However, as discussed above, this is neither practical nor presents a realistic case what it comes to recognizer vulnerabilities "out in the wild".

6. Conclusions

In this paper, a study of the vulnerability of speaker verification systems against voice mimicry attacks is presented. Two high-performance speaker recognition systems were compared with mimicry attack data in Finnish language. The comparison of the accuracy of the recognition systems, GMM-UBM and i-vector system respectively, was possible even with the limited data that characterizes imitation attacks material in general. Our results suggest that the professional impersonator was not able to degrade the performance of our state-of-the-art automatic speaker recognition systems. While there are voices that are possible to imitate by a skillful person, the system is able to detect correctly a good portion of the imitation trials. This implies that even when auditory perception dictates that the impersonator is good at mimicking other voices, the system is able to recognize the impersonator's own voice characteristics and reject it as an impostor. For future work, incorporating more than one impersonator to the test as a impostor subject, and including placing a lexical constraint on the test segments, from the impersonator and the target speakers, could give us an additional knowledge on when impersonator is able to successfully attack the system. Especially, as a future work we are looking into how well our current results generalize to a wider range of impersonators.



(a) ivector-Cosine scoring

(b) GMM - UBM

Figure 2: Score distribution comparison per target. The bars also show the standard error of the mean with 95% confidence.

7. Acknowledgments

This work was supported by the Academy of Finland projects number 253120 and 253000.

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52 (1), pp. 12–40, January 2010.

[2] "Speak to unlock," 2012. [Online]. Available: http://www.i2r.a-star.edu.sg/files/documents/416/BIRC_Brings_First_Speaker_Verification_Technology_Into_Smartphones_With_Built-In_Voiceprint_Feature_Final.pdf

[3] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," *Biometrics and ID Management*, pp. 274–285, 2011.

[4] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech," in *Proc. ICASSP*, Kyoto, Japan, March 2012, pp. 4401 – 4404.

[5] P. DeLeon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," in *Odyssey 2010: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, pp. 151–158 (paper 28).

[6] M. Farrús, M. Wagner, D. Erro, and F. J. Hernando, "Automatic speaker recognition as a measurement of voice imitation and conversion," *The Int. Journal of Speech, Language and the Law*, vol. 1, no. 17, pp. 119–142, 2010.

[7] P. Perrot, G. Aversano, and G. Chollet, "Voice disguise and automatic detection: Review and perspectives," in *Progress in Nonlinear Speech Processing*, ser. Lecture Notes in Computer Science, 2007, pp. 101–117.

[8] E. Zetterholm, "Detection of speaker characteristics using voice imitation," in *Speaker Classification II*, ser. Lecture Notes in Computer Science, 2007, pp. 192–205.

[9] A. Ogihara, H. Unno, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE Trnas. Fundamentals*, vol. E88-A, no. 1, pp. 280–286, Jan. 2005.

[10] Z. Wu, T. Kinnunen, E. Chng, H. Li, and E. Ambikairajah, "A study on spoofing attack in state-of-the-art speaker verification: the telephone speech case," in *Proc. 2012 Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC 2012)*, Hollywood, USA, December 2012, pp. 1–5.

[11] Y. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Proc. Int. Symp on Intelligent Multimedia, Video & Speech Processing (ISIMP'2004)*, Hong Kong, October 2004, pp. 145–148.

[12] Y. Lau, D. Tran, and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems (KES 2005)*, Melbourne, Australia, September 2005, pp. 15–21.

[13] J. Mariéthoz and S. Bengio, "Can a professional imitator fool a gmm-based speaker verification system?" IDIAP, Idiap-RR, 2005.

[14] J. Leskelä, "Changes in $f0$, formant frequencies and spectral slope in imitation," Master's thesis, University of Tampere, 2011.

[15] P. Kenny, "Joint factor analysis of speaker and session variability: theory and algorithms," technical report CRIM-06/08-14, Montreal, CRIM, 2006.

[16] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *DSP*, vol. 10, no. 1, pp. 19–41, Jan 2000.

[17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *ITASLP*, vol. 19, no. 4, pp. 788–798, May 2011.

[18] B. Pellom and J. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *ICASSP*, Phoenix, Arizona, USA, March 1999, pp. 837–840.

[19] L. Mary, A. B. K. K, A. Joseph, and G. M. George, "Evaluation of mimicked speech using prosodic features," in *ICASSP 2013*, Vancouver, May 2013, pp. 7189 – 7193.

[20] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech 2011*, Florence, Italy, August 2011, pp. 249–252.

[21] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2010.