Can We Use Speaker Recognition Technology to Attack Itself? Enhancing Mimicry Attacks Using Automatic Target Speaker Selection

Tomi Kinnunen, Rosa González Hautamäki, Ville Vestman, Md Sahidullah

CAN WE USE SPEAKER RECOGNITION TECHNOLOGY TO ATTACK ITSELF? ENHANCING MIMICRY ATTACKS USING AUTOMATIC TARGET SPEAKER SELECTION

Tomi Kinnunen, Rosa González Hautamäki, Ville Vestman[*]

School of Computing
University of Eastern Finland, FINLAND

Md Sahidullah

MULTISPEECH
Inria, FRANCE

ABSTRACT

We consider technology-assisted mimicry attacks in the context of automatic speaker verification (ASV). We use ASV itself to select targeted speakers to be attacked by human-based mimicry. We recorded 6 naive mimics for whom we select target celebrities from VoxCeleb1 and VoxCeleb2 corpora (7,365 potential targets) using an i-vector system. The attacker attempts to mimic the selected target, with the utterances subjected to ASV tests using an independently developed x-vector system. Our main finding is negative: even if some of the attacker scores against the target speakers were slightly increased, our mimics did not succeed in spoofing the x-vector system. Interestingly, however, the relative ordering of the selected targets (closest, furthest, median) are consistent between the systems, which suggests some level of transferability between the systems.

Index Terms— Speaker verification, mimicry, spoofing

1. INTRODUCTION

It is well known that *representation attacks* [1, 2] — also known as *spoofing attacks* — cast a shadow over the security of biometric systems. A spoofing attack involves an adversary (attacker) who aims at masquerading oneself as another targeted user to gain illegitimate access. Unprotected *automatic speaker verification* (ASV) systems can easily be spoofed using replay, voice conversion and text-to-speech attacks [3]. This has sparked research into spoofing *countermeasures* aimed at detecting the attacks from given audio. Community-driven benchmarks such as ASVspoof [4] and AVspoof [5] were launched for an organized study of countermeasures. In the context of security, continual arms race between attacks and their defenses is well known [6]: to develop effective countermeasures, it is necessary to understand the attacks. The speech synthesis community has independently launched *voice conversion challenge* [7, 8] to advance VC methods. Within the past few years, active and dynamic communities both at the 'attack' and 'defense' sides of ASV, focused on technological attacks, have emerged.

In this study we focus on a nearly-forgotten ASV attack – *mimicry* (impersonation). Unlike the technology-induced attacks, mimicry involves *human*-based modification of one's voice production. The question of recognizer vulnerability against mimicry was addressed at least around half a century ago [9, 10] and has remained a cursory topic within the ASV field [11, 12, 13, 14, 15, 16]. While ASV vulnerability caused by technical attacks is widely reported, less (reliable) information is available on effectivess of mimicry, due to adoption of small, proprietary datasets. The authors are fully aware of the difficulties in collecting mimicry data from professional artists [15], whose prevalance in the general population is arguably very low. Nonetheless, *if* mimicry attacks could be shown to be a threat to ASV, it would be conceivably challenging to devise countermeasures: natural human speech lacks processing artifacts that enable detection of technical attacks. Thus, we argue that it is important to keep mimicry also in the list of potential attacks against ASV. Of particular interest in this work are mimicry attacks against *celebrities* whose voice data is available in massive amounts in the public domain. In line with the recent EU's *General Data Protection Regulation* (GDPR), intended to protect the privacy of its citizens, it is important to assess risks associated with multimedia data in the public domain. A recent study [17] has attempted voice cloning of celebrity voices based on found data using a pre-defined target speaker. The cloned voice samples were, however, detectable as spoofed speech.

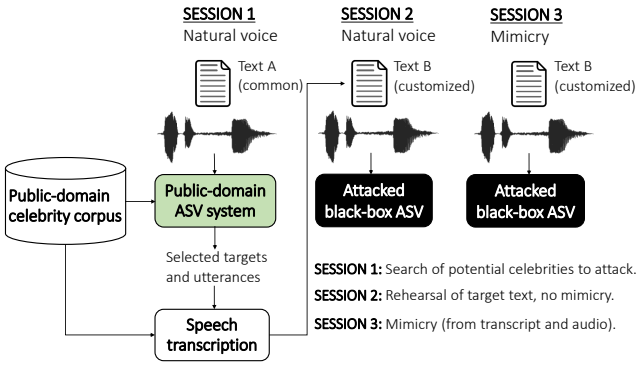

Fig. 1. Automatic speaker verification (ASV) assisted mimicry attack: attacker uses a public-domain ASV system to select target speakers matched with his/her voice from a public celebrity data. The attacker then practices target speaker mimicry, intended to attack another independently developed ASV system.

We focus on *technology-assisted* mimicry attacks that uses ASV itself to identify potential target speakers to be subjected to mimicry attacks. The idea is to identify targets whose voice is similar to that of the attacker's voice, as this could potentially involve fewer articulatory or voice source modifications. Two related prior studies are [12] and [18] which involve search of either targets [12] or attackers [18] from a large set of candidates. The authors of [12] used a Gaussian mixture model (GMM) system to find closest, furthest and median targets from YOHO corpus for a few naive impersonators, leading to substantially increased false acceptance rate. In [18], the authors selected impersonators (rather than targets) through a commercial crowd-sourcing platform based on self-judgment and further

[*]The work was supported by Academy of Finland (project no. 309629).

refinement using ASV.

Our study can be seen as an attempt to reproduce the findings of [12] using up-to-date ASV technology and a far larger target candidate set (7,365 celebrities pooled from VoxCeleb1 [19] and VoxCeleb2 [20]). A key methodological difference, however, is that unlike [12] that used a *single* GMM recognizer, we include two *different* ASV systems (Fig. 1). We argue that it may be unrealistic for the attacker to interact many times with the targeted ASV, but he/she may develop an offline *substitute* ASV that, after optimization, hopefully behaves similar to the attacked one. Our work bears some resemblance to *black box attacks* [21] in adversial machine learning [6], though our adversary is not a machine learning algorithm but a human. Further, those methods use either classifier output score or decision to optimize the attacks, while we assume that the attacker receives no feedback from the attacked system in any form. Thus, we expect that such attacks are not strong, but we argue that they are *realistic* given the abundance of multimedia data of celebrity speakers and various public-domain ASV implementations. We seek to answer the question in the title of this work, using a specifically collected 'attacker' data and VoxCeleb celebrity targets.

2. ASV-ASSISTED MIMICRY ATTACKS

2.1. Attack implementation

Let $\mathcal{T} = \{T_j\}_{j=1}^{J}$ denote a set of unique, publicly known **target speaker** identities and let $\mathcal{A} = \{A_k\}_{k=1}^{K}$ denote a set of **attacker** identities. Given a pair of arbitrary utterances (or a pair of collections of many utterances), (U_i, U_j), a **automatic speaker verification** (ASV) system (speaker detector), $\mathcal{D}(U_i, U_j)$ computes a score $s \in \mathbb{R}$ with an arbitrary scale but higher relative values indicating stronger support that the source of U_i and U_j is the same speaker.

We consider two different types of ASV systems. The first one, **attacker's ASV** ($\mathcal{D}_{\text{pub}}$), is a public-domain, known ASV implementation, and the latter, **black-box ASV** ($\mathcal{D}_{\text{black}}$), is the system an attacker attempts to spoof but whose internal workings, scores and decisions are inaccessible to the attacker. The proposed attack proceeds as follows:

ASV-assisted mimicry attack

1. Attacker $A \in \mathcal{A}$ records his/her natural voice sample, $\mathcal{U}_{\text{nat}}$.
2. A uses $\mathcal{D}_{\text{pub}}$ to compute scores $\{s_j\}_{j=1}^{J}$ between $\mathcal{U}_{\text{nat}}$ and all the targets in a public database. A picks the **closest target**, $j^* = \arg\max_{j=1}^{J} \mathcal{D}_{\text{pub}}(U_{\text{nat}}, U_j)$, where U_j contains the known recordings of T_j.
3. A continues to use $\mathcal{D}_{\text{pub}}$ to pick the best-matching utterances of T_{j^*}.
4. The attacker listens to the selected utterance(s) and attempts to adjust his/her voice towards the target. Once completed practicing, A submits a mimicked test utterance U_{mimic} to $\mathcal{D}_{\text{black}}(U_{\text{mimic}}, U_{j^*})$, the aim of being authenticated as T_{j^*}.

2.2. Public-domain (attacker's) ASV system

The attacker's ASV uses i-vector front-end and probabilistic discriminant analysis (PLDA) back-end to compute speaker similarity scores. We extract 20 mel-frequency cepstral coefficients (MFCCs) using 20 filters[1], leading to 60 features per frame after including deltas and double-deltas. The features are processed with RASTA filtering and cepstral mean and variance normalization (CMVN). Non-speech frames are omitted using energy-based speech activity detector (described in Section 5.1 of [22]).

To train the i-vector extractor, we compute sufficient statistics using a universal background model (UBM) of 512 Gaussians. To train the UBM, we choose randomly 10,000 speech utterances from SI-284 subset of the Wall Street Journal corpus (WSJ0 and WSJ1) and 10,000 speech utterances from `train-clean-360` and `train-clean-100` subsets of the Librispeech [23] corpus. To train the T-matrix with 400 total factors, we randomly choose 20,000 speech utterances from the same subset of WSJ and 20,000 speech utterances from the same subset of Librispeech. Finally, the PLDA is trained on the entire SI-284 subset (from WSJ0 and WSJ1) and entire train-clean subset (from Librispeech) consisting 169,969 speech utterances from a total 1,455 speakers. The 400-dimensional i-vectors are further reduced to 250 dimensions with linear discriminant analysis (LDA) using the same data as in PLDA training, followed by centering and whitening. We use a simplified PLDA with 200-dimensional speaker subspace. We adopt MATLAB-based MSR Identity Toolkit[2] to train the attacker's ASV.

The above development data and parameter selections are based on preliminary ASV experiments on the AVOID corpus (collected in [24]) and VoxCeleb (1 and 2). For the AVOID corpus, we obtained EERs of 3.30% and 4.52% for text-dependent scenario with two English sentences separately. Further, we obtained EERs of 10.50% and 16.98% for text-independent ASV on two subsets of VoxCeleb1 and VoxCeleb2. These custom protocols were created by randomly choosing 60 speakers from VoxCeleb1 (6,163 target and 363,617 non-target trials) and VoxCeleb2 (9,118 target and 537,962 non-target trials).

2.3. Attacked ASV systems

In our experiments, we regard x-vector systems [25] based on pretrained Kaldi [26] recipes as ASV systems to be attacked. To emulate the scenario of attacker's limited knowledge of this system, the attacker's ASV is made intentionally different from the attacked ASV systems in terms of feature extractor set-up, embedding type, and development corpora (Table 1). Both attacked systems are based on Kaldi recipes for VoxCeleb and NIST Speaker Recognition Evaluation 2016, while the attacker's system uses i-vectors.

3. CORPUS OF TARGET SPEAKERS: VOXCELEB

The attacker's ASV is used as a voice search tool to find the closest speakers from the combination of VoxCeleb1 [19] and Voxceleb2 [20] to each of the locally recruited subjects (described in Section 4). The combined VoxCeleb corpus contains about 1.3 million speech excerpts extracted from more than 170,000 YouTube videos from $J = 7,365$ unique speakers. This totals to about 2,800 hours of audio material that is, for the most part, active speech. Both VoxCeleb corpora were collected using automated pipeline exploiting face verification and active speaker verification technologies [20].

VoxCeleb1 contains mostly English speech, while VoxCeleb2 is more diverse in nationalities and languages. The nationality information of the target speakers was of our interest, as the recruited local speakers are Finnish and we wanted to see if Finnish people do better job at imitating Finnish targets than non-Finnish. According to the VoxCeleb1 metadata, there were no Finnish targets in VoxCeleb1. The VoxCeleb2 did not include nationality metadata, so we

[1] http://cs.joensuu.fi/~sahid/codes/AntiSpoofing_Features.zip

[2] https://www.microsoft.com/en-us/download/details.aspx?id=52279

Table 1. Details of the speaker verification systems used to simulate targeted impersonation attack against automatic speaker verification. The attacker is assumed to not have information about the attacked systems, and hence the attacker's system differs from the attacked systems.

	Attacker's ASV	**Attacked ASV 1**	**Attacked ASV 2**
Sampling rate	16 kHz	16 kHz	8 kHz
Acoustic features	60-dimensional MFCCs (20 static+20-Δ+20-$\Delta\Delta$), RASTA, SAD, CMVN.	30 MFCC coeffs (no deltas), Sliding CMN normalization	23 MFCC coeffs (no deltas), Sliding CMN normalization
Embedding type	i-vector	x-vector	x-vector
Back-end / scoring	PLDA	PLDA	PLDA
Development data	WSJ0 and WSJ1, Librispeech	VoxCeleb2, training part of VoxCeleb1	Switchboard 2 (P. 1, 2, 3), Switchboard Cellular, NIST SREs 04 – 10, Mixer 6
Data augmentation	None	Reverberation, noise, music, babble	Reverberation, noise, music, babble
EER (VoxCeleb1)	12.84 (%)	3.11 (%)	9.91 (%)

made a script to automatically obtain nationalities using Google's *Knowledge Graph* API[3]. With this strategy we found 44 Finnish speakers from VoxCeleb2.

4. LOCALLY RECRUITED ATTACKERS

4.1. Speakers and recording gear

We recruited $K = 6$ voluntary local speakers (4M + 2F) to serve as 'attackers'. All are native Finnish speakers and one of them is a co-author of this study. All the six speakers, selected based on their availability, took part in 2015 to the recordings of [24], currently under release with the name AVOID corpus[4]. We adopt the same recording setup and part of text prompts from [24] but otherwise the two studies are unrelated. All the subjects signed an informed consent form to use their speech data for research, and were rewarded with movie and coffee tickets. Two of the male subjects knew the specific goals of our study while the remaining four subjects were not informed that the text and target speakers were tailored for them, nor where do the voices originate from. They were not informed that the study relates to ASV vulnerability, but were merely asked to mimic the targets as accurately as they could.

As illustrated in Fig. 1, the subjects took part to three recording sessions. The first session, produced in the subject's natural voice, is used for VoxCeleb target speaker selection, while the remaining two sessions serve for vulnerability analysis of the attacked systems. The tasks in the recording sessions differed, while the recording setup was the same: recordings took place in a silent laboratory room with a portable Zoom H6 Handy Recorder using an omnidirectional headset mic (Glottal Enterprises M80) with 44.1 kHz sampling and 16-bit quantization. Three other channels (two smartphones, electroglottograph) were also collected, but are not used in this study.

4.2. The first recording session (data for target search)

The first session, used for the targeted VoxCeleb speaker search, consists of four tasks in the speaker's natural voice. The tasks consisted of spontaneous speech and read text (13 sentences) in both Finnish and English. The read texts in Finnish are the same used in [24] and their corresponding English versions were added for this study. We have approximately 6 minutes of speech (before speech activity detection) per speaker from Session 1.

4.3. Attacked target speaker search and utterance selection

For the purpose of targeted speaker search, we compute a single averaged i-vector for each of the six speakers resulting from 28 individual utterances from Session 1. Similar to [12], we use the ASV system to pick for each attacker the **closest**, **median**, and **furthest** speakers among the VoxCeleb speakers. The closest one is most relevant for vulnerability analysis while the other two serve for reference purposes. We do this ASV-assisted search separately for *all* the VoxCeleb speakers (unconstrained search from 7,365 speakers) and for the subset of 44 Finnish speakers. We pool all the speech data of the VoxCeleb speakers to compute average i-vector per target.

In addition to the three ASV-selected targets, we include **common target** matched with the speaker's gender, in both Finnish and English. The common Finnish targets are Päivi Räsänen (F, politician) and Ilkka Kanerva (M, politician), and the common English targets are Hillary R. Clinton (F, politician) and Leonardo DiCaprio (M, actor). Even if well-known persons, from the viewpoint of ASV they are *random* targets with no strong presuppositions how similar their voices are to our attackers.

In summary, for each of our 4 male and 2 female subjects, we select 6 customized targets (3 ASV-ranks $\times$ 2 languages) and 2 common gender-matched ones (one Finnish, one English). This gives a theoretical total of $3 \times 2 \times 4$ male $+ 2$ common male + $3 \times 2 \times 2$ female $+ 2$ common female $= 40$ target speakers. Not all of the ASV-selected targets are unique, however: one male Finnish celebrity was the closest target for three attackers, one male Finnish celebrity repeated as the median speaker for two male attackers, and one female Finnish celebrity is the furthest speaker for the two female attackers. The final number of unique celebrity targets is 36.

For each of the 36 target speakers, we selected at minimum 30 seconds of active speech to evaluate the ASV system attacks. This duration of speech was collected from multiple shorter utterances for two reasons: First, the segments in VoxCeleb corpora are typically about 5 to 10 seconds long. Secondly, as we were going to ask the recruited attackers to imitate the target speakers, shorter utterances would be easier to impersonate.

We utilized Attacker's ASV also for the utterance selection. For the closest targets, we selected the highest scoring utterances, while for the furthest targets, the utterances with the lowest scores where selected. For the median speakers, we selected the utterances close to mean. This was further accompanied by manual inspection: if the audio quality (determined subjectively by listening) was not deemed high enough, we discarded it and moved on to the next ones in the ranked list.

4.4. Speech transcription and the mimicry recordings

Unlike the first recording session (common to all subjects), the second and third sessions were customized for each subject. This process involves the use of speech transcripts of the selected target utterances. To this end, we used Amazon's Mechanical Turk[5] (MTurk), a commercial crowdsourcing service, to transcribe the English language audio. The Finnish transcripts were produced by two native

[3] https://developers.google.com/knowledge-graph/

[4] http://urn.fi/urn:nbn:fi:lb-2018060621

[5] https://www.mturk.com/

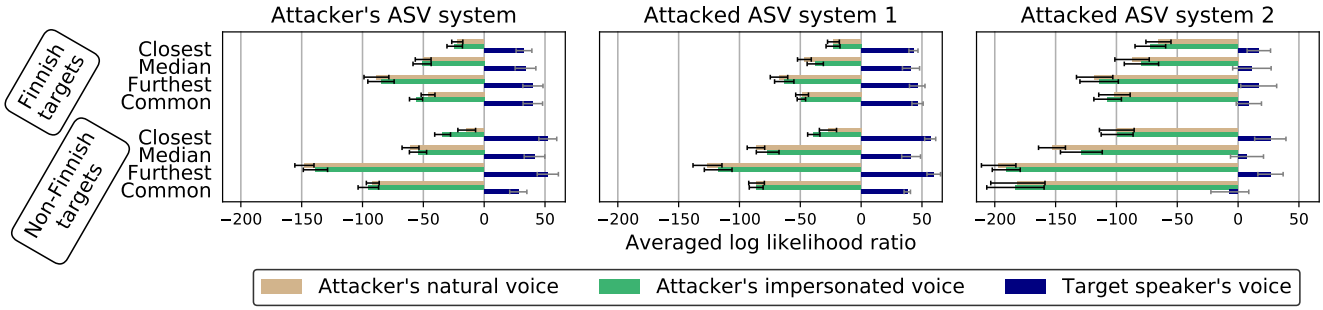

Fig. 2. Comparison of attackers' ASV scores (log likelihood ratios) to the targets' scores for all the three ASV systems involved in the study. The scores are averaged over all attackers and all speech segments. The error bars represent 95 % confidence intervals for the means.

Finnish speakers. The 35 MTurk crowdworkers and the 2 Finnish transcribers were asked to transcribe all the nuances of conversational speech, including repetitions, hesitations, filler words *etc*. Finally, two reviewers audited the quality of all the transcripts.

In Session 2, which took place 5 to 6 weeks after Session 1, the subject was provided with the transcripts of the selected target utterance(s) and was asked to read the sentences twice in his or her natural voice. The speaker was not informed whose speech the transcripts corresponded to; even the coauthor taking part to the study was unaware whose transcripts he was reading. The rationale of including this session was to familiarize each attacker with the target speaker sentences. We adopted the general idea to include a session with reference text only and another one with audio from the design used in [13].

In the last session, which took place 2 to 6 days after Session 2, the subjects were provided with the same transcript as in Session 2, but in addition they had now access to the actual target speaker audio excerpts played through headphones. The transcripts were provided on a printed paper and the audio reference from a tablet computer that the subject was able to interact with; he/she could play the target utterance(s) as many times as needed, and he/she then tried to mimic the voice according to their best skills. Again, the subject was asked to mimic each sentence twice. In the experiments, we use only the second recording of each sentence.

5. RESULTS

In the following, we evaluate effectiveness of the mimicry attacks. The target speaker models used in the experiments were enrolled using all available segments except those selected for testing as described in Section 4.3.

Figure 2 displays how the attacker's PLDA scores compare to the target scores. The general findings are as expected. First, the order of the closest, the median, and the furthest speakers transfers from the attacker's ASV system to the attacked ASV systems implying that the ASV-assisted speaker selection *can* help in ASV attacks. Second, in general, the mimicry attempts were not successful as the attacker's natural and mimicry scores are significantly (and substantially) below the target scores. Additionally, we find no significant difference between the natural and impersonated versions. Finally, as the recruited attackers were Finnish, attackers' scores against the Finnish targets were higher than for non-Finnish targets.

To look at the effect of mimicry closer, we analyze the difference of mimicked and natural speech scores (Table 2). Interestingly, and contradictory to what we assumed, if the target speaker's voice is already close to the attacker's voice, the impersonation attempts *degrade* the score. The same finding was noted in situations where the target is a well known public figure like the targets in the common category are. We suspect that the effect might be due to people having higher tendency to overact someone they already know well. However, in the case of the targets that are not close to the attackers and, on average, are not so well known (median and furthest categories), impersonation is helpful. Figure 3 shows a sample of the best and worst attackers for the common targets, with similar findings as above.

Table 2. Score differences between attacks with impersonated voices and attacks with natural voices. Differences are averaged over attackers, target nationalities, and utterances. $\pm$ indicates 95 % confidence intervals. In the case of the closest target speakers, impersonation attempts are counterproductive.

ASV system	Closest	Median	Furthest	Common
Attacker	-9.7 $\pm$ 5.2	2.2 $\pm$ 4.3	5.9 $\pm$ 7.1	-7.2 $\pm$ 4.3
Attacked1	-5.2 $\pm$ 3.9	9.2 $\pm$ 3.3	6.1 $\pm$ 4.3	-0.5 $\pm$ 3.8
Attacked2	-3.7 $\pm$ 5.5	15.0 $\pm$ 7.0	4.7 $\pm$ 7.4	-4.0 $\pm$ 7.7

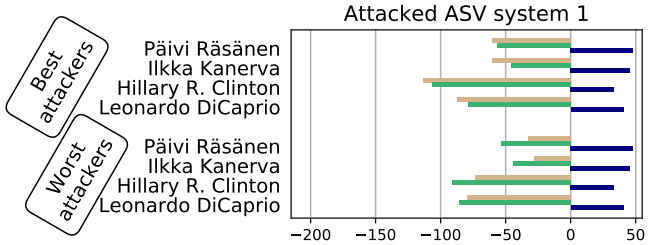

Fig. 3. Scores against the common celebrity targets for the best and the worst impersonators (Color codes are equal to Fig. 2).

Our attackers are native Finnish speakers recorded with a specific set-up which could be different from the VoxCeleb conditions. This raises a question whether our mimicry attacks were unsuccessful simply due to domain mismatch. To address this concern, we ran an additional experiment, in which we scored attacker's test segments against attacker's own speaker models. If our ASV systems are able to cope with domain mismatch, we expect high detection scores similar to the VoxCeleb target scores). The results shown in Figure 4 confirm this as, similarly to Figure 2, the averaged log likelihood ratios are close to 50 for the attacked ASV system 1. From Figure 4 we also find that impersonation lowers the scores, showing that the ASV system is not robust against *disguise* (the act of attempting to be not recognized as oneself). This finding was not a surprise to us [24].

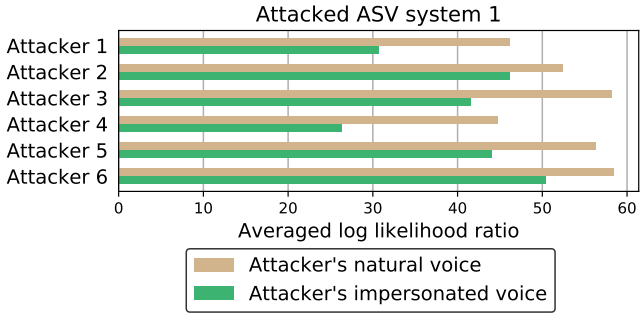

Fig. 4. Attacker's scores in the case where attacker's test sentences are tested against attackers' own speaker models instead targets' speaker models as in Figure 2.

6. CONCLUSION

Can one use ASV technology to attack itself? To answer this, we collected 6 native Finnish mimics and used ASV to locate customized targets from VoxCeleb. Our preliminary analysis reveals that, unlike in [12], our mimicry attempts were unsuccessful. In fact, the ASV scores even degraded, specifically when the impersonator's natural voice was already close to the target speaker's voice. Further work is required to analyze the reasons, specifically in terms of acoustic modifications implemented by our naive impersonators. The relative ordering of the closest, median and furthest speaker was, however, preserved across the attacker's and attacked ASV systems, with higher relative scores obtained for Finnish targets. Though our assisted attacks did not succeeed to spoof state-of-the-art x-vector technology, selection of imposters from a larger set of speakers (e.g. using crowd-sourcing [18]) may help in spoofing ASV systems.

7. REFERENCES

[1] N.K. Ratha, J. Connell, and R.M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM Systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

[2] ISO/IEC 30107-1:2016, "Information technology – biometric presentation attack detection – part 1: Framework," https://www.iso.org/obp/ui/#iso:std:iso-iec:30107:-1:ed-1:v1:en, 2016, Online; accessed 22-February-2018.

[3] Z. Wu, N.W.D. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[4] Z. Wu, T. Kinnunen, N.E., J. Yamagishi, C. Hanilçi, M. Sahidullah, and A. Sizov, "ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge," in *Interspeech 2015*, 2015, pp. 2037–2041.

[5] S.K. Ergunay, E. Khoury, A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *BTAS*. 2015, pp. 1–6, IEEE.

[6] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[7] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Proc. Interspeech*, 2016, pp. 1632–1636.

[8] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey: the Speaker and Language Recognition Workshop*, Les Sables d'Olonne, France, 2018, pp. 195–202.

[9] J.E. Luck, "Automatic speaker verification using cepstral measurements," *Journal of the Acoustic Society of America*, vol. 46, no. 4, pp. 1026–1032, October 1969.

[10] W. Endres, W. Bambach, and G. Flösser, "Voice spectrograms as a function of age, voice disguise, and voice imitation," *The Journal of the Acoustical Society of America*, vol. 49, no. 6, pp. 1842–1848, 1971.

[11] Y.W. 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.W. Lau, D.T., and M. Wagner, "Testing voice mimicry with the YOHO speaker verification corpus," in *Knowledge-Based Intelligent Information and Engineering Systems, 9th International Conference, KES 2005, Melbourne, Australia, September 14-16, 2005, Proceedings, Part IV*, 2005, pp. 15–21.

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

[14] Anders Eriksson, "The disguised voice: imitating accents or speech styles and impersonating individuals," in *Language and identities*, Carmen Llamas and Dominic Watt, Eds., vol. 8, pp. 86–96. Edinburgh University Press, 2010.

[15] R. González Hautamäki, T. Kinnunen, V. Hautamäki, and An.-M. Laukkanen, "Automatic versus human speaker verification: The case of voice mimicry," *Speech Communication*, vol. 72, pp. 13–31, 2015.

[16] Mireia Farrús, "Voice disguise in automatic speaker recognition," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 68:1–68:22, July 2018.

[17] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the internet? initial investigation of cloning obama's voice using GAN, WaveNet and low-quality found data," in *Proc. Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 240–247.

[18] S. Panjwani and A. Prakash, "Crowdsourcing attacks on biometric systems," in *Tenth Symposium on Usable Privacy and Security, SOUPS 2014*, 2014, pp. 257–269.

[19] A. Nagrani, J.S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[20] J.S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[21] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM on Asia Conf. on Computer and Comm. Security, AsiaCCS 2017*, Abu Dhabi, UAE, April 2017, pp. 506–519.

[22] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, Brisbane, Australia, April 2015, pp. 5206–5210.

[24] R. González Hautamäki, M. Sahidullah, V. Hautamäki, and T. Kinnunen, "Acoustical and perceptual study of voice disguise by age modification in speaker verification," *Speech Communication*, vol. 95, pp. 1–15, 2017.

[25] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. IEEE ASRU*, Dec. 2011.