Voice Mimicry Attacks Assisted by Automatic Speaker Verification

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

Voice Mimicry Attacks Assisted by Automatic Speaker Verification

Ville Vestman[a,*], Tomi Kinnunen[a,**], Rosa González Hautamäki[a], Md Sahidullah[b]

[a] *School of Computing, University of Eastern Finland, FI-80101, Joensuu, Finland*
[b] *Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France*

Abstract

In this work, we simulate a scenario, where a publicly available ASV system is used to enhance mimicry attacks against another closed source ASV system. In specific, ASV technology is used to perform a similarity search between the voices of recruited attackers (6) and potential target speakers (7,365) from VoxCeleb corpora to find the closest targets for each of the attackers. In addition, we consider 'median', 'furthest', and 'common' targets to serve as a reference points.

Our goal is to gain insights how well similarity rankings transfer from the attacker's ASV system to the attacked ASV system, whether the attackers are able to improve their attacks by mimicking, and how the properties of the voices of attackers change due to mimicking. We address these questions through ASV experiments, listening tests, and prosodic and formant analyses. For the ASV experiments, we use i-vector technology in the attacker side, and x-vectors in the attacked side. For the listening tests, we recruit listeners through crowdsourcing.

The results of the ASV experiments indicate that the speaker similarity scores transfer well from one ASV system to another. Both the ASV experiments and the listening tests reveal that the mimicry attempts do not, in general, help in bringing attacker's scores closer to the target's. A detailed analysis shows that mimicking does not improve attacks, when the natural voices of attackers and targets are similar to each other. The analysis of prosody and formants suggests that the attackers were able to considerably change their speaking rates when mimicking, but the changes in F0 and formants were modest. Overall, the results suggest that untrained impersonators do not pose a high threat towards ASV systems, but the use of ASV systems to attack other ASV systems is a potential threat.

Keywords: Speaker verification, mimicry, crowdsourcing, spoofing, automatic target speaker selection, perceptual speaker similarity, prosody

1. Introduction

Security is of key importance in today's society where information processing gets increasingly digital, automated and lacks human-to-human communication. We need new ways to protect our data records from unauthorized access. Alongside with the traditional means of user authentication, biometric technology has emerged as one of the potential solutions. The use of human voice for strong user authentication is attractive especially under remote, unattended scenarios and due to the readily available infrastructure (namely, telephones) to scale it up easily.

Similar to the traditional means of user authentication, however, biometric systems are prone to malicious attacks by hackers. It is no longer news, neither to the research community nor to the general public, that biometric systems can be fooled through various *representation attacks* [1, 2], also known as *spoofing attacks*. A spoofing attack involves an adversary (attacker) who aims at masquerading oneself as another targeted user to gain illegitimate access to the targeted person's data. Unprotected *automatic speaker verification* (ASV) systems can be easily spoofed using replay, voice conversion (VC) and text-to-speech (TTS) attacks [3]. Since the attacks are typically not perfect but contain either processing artifacts or display degraded audio quality, they can be detected to a certain extent. To this end, community-driven challenges such as ASVspoof [4] and AVspoof [5] were launched for an organized study of *spoofing countermeasures*. In the context of security, the continuous arms race between attacks and their defenses is well known [6]: so as to develop effective countermeasures, it is necessary to understand the attacks. The speech synthesis community has independently launched *voice conversion challenges* [7, 8] to advance VC methods (though targeted primarily for human listeners rather than for ASV spoofing). To sum up, within the past few years, active and dynamic

[*] A part of the work of the first author was carried out during an intership at NEC.

[**] Corresponding Author

Email addresses: vvestman@cs.uef.fi (Ville Vestman), tkinnu@cs.joensuu.fi (Tomi Kinnunen), rgonza@cs.uef.fi (Rosa González Hautamäki), md.sahidullah@inria.fr (Md Sahidullah)

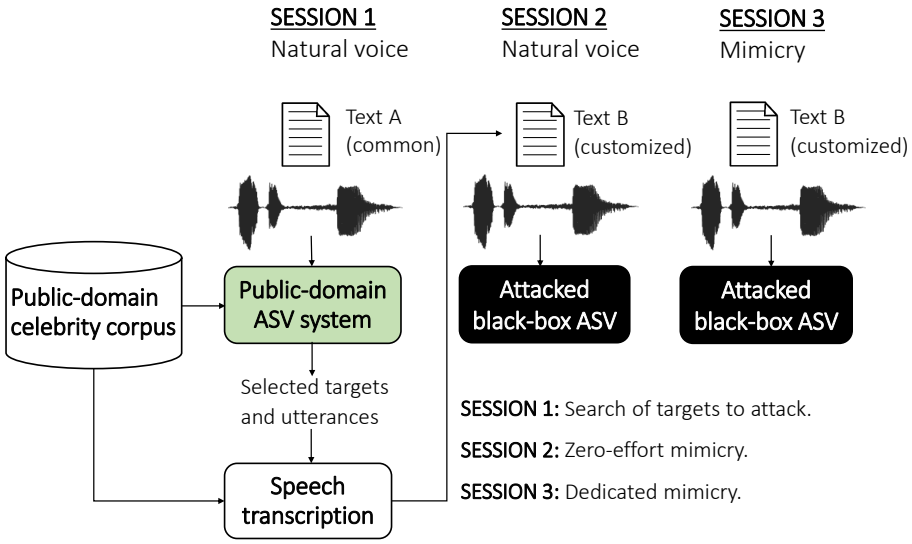

Figure 1: Automatic speaker verification (ASV) assisted mimicry attack: attacker uses a public-domain ASV system to select target speakers matched with his/her voice from a public celebrity database. The attacker then practices target speaker mimicry, intended to attack another independently developed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

communities both at the 'attack' and 'defense' sides of ASV have emerged. There is now a far better understanding of the technology-based attacks and their defenses against ASV systems than half a decade ago — see [9] for an up-to-date review.

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 [10, 11] and has remained a cursory topic within the ASV field [12, 13, 14, 15, 16, 17]. While ASV vulnerability caused by technical attacks is widely reported, less (reliable) information is available on effectivess of mimicry, primarily due to adoption of small and proprietary datasets. The only conclusions that one can possibly extrapolate from the prior studies on mimicry effect against ASV is that the results depend on a specific study. This suggests that mimicry is less consistent attack compared to replay, VC and TTS that are repeatable reported to be successful in spoofing ASV systems.

The authors are aware of the difficulties in collecting mimicry data from professional artists [16], whose prevalence 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. Besides the security aspect, mimicry could potentially help us in the design of better ASV methods for voice comparison.

Of particular interest in this work are mimicry attacks against persons whose voice data is exposed in a public domain in large quantities — such as celebrities or anyone streaming or uploading massive amounts of his/her videos to the Internet. In line with the recent EU's *General Data Protection Regulation* (GDPR) [18], intended to protect the privacy of individuals, it is important to assess potential risks associated with multimedia data in the public domain; we elaborate on this emerging problem further in Section 2. Differently from most prior studies, we focus on *technology-assisted* mimicry attacks. In specific, we use the ASV technology itself to identify potential target speakers to be subjected to mimicry attacks. The idea is to identify targets whose voice is *a priori* similar to that of the attacker's voice in terms of acoustic parameters. The assumption is that nearby target speakers might be easier to mimic due to potentially fewer articulatory or voice source modifications required. Two related prior studies are [12] and [19] which involve search of either targets [12] or attackers [19] from a pool of candidates. The authors of [12] used a Gaussian mixture model (GMM) system to find closest, intermediate and furthest target speakers from YOHO corpus for two naive impersonators, leading to substantially increased false acceptance rate for the closest targets. In [19], the authors selected impersonators (rather than targets) through a commercial crowd-sourcing platform based on self-judgment and further 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 [20] and VoxCeleb2 [21]). Besides the order of magnitude larger target speaker pool and adoption of state-of-the-art ASV systems, there is a key difference in the research methodology as well: unlike [12] that used a *single* GMM recognizer, we include two *different* ASV systems as illustrated in Fig. 1. We argue that it is unrealistic for the attacker to interact many times with the targeted ASV, as done in that past work. In our attack model, therefore, the attacker uses an offline, publicly

available *substitute* ASV system to first identify which speakers to attack; ideally, the substitute system would behave similar to the attacked ASV system. This idea bears some resemblance to *black box attacks* [22] in adversarial 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 our attacks are not strong, but we argue that they are *realistic* given the abundance of both voice data and ASV implementations in the public domain. We seek to answer the question whether the use of ASV technology itself could increase the risk of an attacker being falsely accepted by (another) ASV system.

A preliminary version of this work appears in [23]. Our preliminary findings in that work suggested a *negative* result — *i.e.* that mimicry attempts, even when the target speakers were selected with automatic speaker identification, would not have left the attacked ASV systems vulnerable. We are not entirely content with just this finding, however — we are interested to understand the reasons. To this end, the present work substantially extends [23] by contrastive automatic, perceptual, prosody, and formant analyses. In particular, we include (i) **analysis of domain mismatch in ASV score domain** (presented in Section 6), (ii) a **human benchmark** of speaker similarity (presented in Section 7), and (iii) **prosody and formant** analysis (presented in Section 8). Additionally, (iv) Section 2 provides a broad background context to our work. None of the above were provided in [23]. The score domain analysis seeks to answer whether the negative finding might have been due to condition differences across our attacker and celebrity corpora. The human benchmark, implemented via crowdsourcing, serves for a reference point to the automatic methods. Finally, the prosody and formant analyses serve to study changes in the speaking rate, fundamental frequency (F0), and formants induced by mimicry. Our hypothesis is that some of these 'broad' speech parameters might be among the prominent cues that a naive mimic attempts to primarily modify towards the target speaker. While this article is intended to be as self-contained as possible, the interested reader may consult additional online material [24] for further details about our text prompts and target speakers.

2. Attacks on speaker verification systems with found data

The amount of personal data that people upload to the Internet increases year by year. Enabled by popular social media platforms and other picture/video sharing services, people upload (or stream) their self-portraits (selfies), voice samples and video clips much more easily — perhaps more carelessly — than in the past. The general public may be unaware that their face photos, videos and voice samples contain biometric traits and form potentially their 'unique' identifiers[1]. Somewhat paradoxically, of a specific concern is the rapidly advancing biometric technology *itself*. The aim of biometric technology, similar to the traditional ways of user authentication, is to regulate access to a restricted domain. The basic premise is that a biometric database administrator (such as the police, a border control officer, or a bank) has sufficient security countermeasures to protect their biometric database and systems from being hacked or tampered. But what if the user decides to voluntarily expose his or her biometric data to the public? Very few of us would purposefully upload our credit card number or a photo-copy of our passport to a public website, but uploading our face and voice data does not seem to concern many. It is important to address the potential risk scenarios of misuse of personal data, and to make the general public aware of the potential risks of uploading their data to a public domain. Awareness on the potential risks among the professional community has increased due to initiatives such as EU's IC1206 COST action[2] that focused on de-identification and privacy protection of multimedia data (see [25] for a review). The overall picture is not yet complete, however, and human voice has received far less attention than image-based biometric traits in this context.

One potential risk is that biometric data that is not searchable or indexable using today's technology might become so tomorrow. Imagine a search engine that uses face or speaker recognition to cross-link someone's sensitive personal multimedia data — such as sexually explicit photographs shared confidently with one's partner but leaked to a porn website; or a video portraying someone under the influence of drugs — with his or her personal website or social media profile. Other risks could include fabricating a 'digital

[1] The authors argue that 'unique' is a misleading term in the context of biometrics where decisions are not based on exact pattern matching but probabilistic reasoning.

[2] https://www.cost.eu/actions/IC1206

clone' of someone using machine learning — recent warning examples are provided by the so-called *deepfakes* [26, 27, 28], realistic-appearing but fabricated or tampered videos portraying a targeted person created with the aid of deep learning (the interested reader is pointed to to [29] for a detailed review of potential societal, ethical and legal implications of deepfakes). In the context of speaker verification in specific, [30] addressed voice cloning of a well-known celebrity (the former US president Barack Obama). Even if the result was essentially negative (the cloned voice samples were detectable as artificial ones using a spoofing countermeasure), machine learning, including voice cloning techniques, do not stand still.

As current machine learning models require large training sets, one may argue that persons who have more (and of technically higher-quality) data in the Internet might become more easily exposed to novel, yet unforeseen, types of attacks and misuse in the future. Our present study is framed in the context of *celebrity* voices (due to the adoption of the VoxCeleb corpus) but we intend it as a proxy to address a specific risk associated with anyone having large quantities of biometric data in a public domain, often referred to as *found data*. In specific, we carry out empirical assessment of attacks on voice biometric system with the help of found audio data. This type of attacks have received surprisingly little attention in the literature. Unlike the use of publicly available tools for voice cloning of a specific target, we look for a speaker with the most similar voice and use him/her as an imposter. We use target speaker's publicly available voice *data* and publicly available ASV tool for the voice similarity search.

The potential threat of natural impersonated voice, also known as mimicry [16], has been studied in a limited number of target speakers and mimickers [10, 12, 16, 31]. The present work is related to the study on the impact of the voice impersonation in ASV where the impersonator and potential target speakers are selected from large set of speakers. This enables us to choose the those impersonator-target pairs who are already similar in their natural voice. Surprisingly, the studies involving the search of potential attackers and the assessment of their ability to break the biometric security system are very limited. For other behavioral biometric traits (than voice), perhaps the only related study is done with *shoulder surfing* attack in the context of touch input implicit authentication [32]. This demonstrated that when potential attackers are selected and

trained to perform targeted mimicry, this authentication method is highly prone to such attacks.

The closest prior work in spirit to our study is [13] where the authors studied the effect of mimicry in ASV with two professional imitators and four non-professional imitators. The closest speaker for each imitator was chosen from YOHO corpus of 138 speakers using Gaussian mixture model (GMM) based likelihood. The study indicated that, when mimicking the most similar speaker, the professionals did not achieve better mimicry performance than non-professional imitators. On the other hand, the professional imitators were more successful at mimicry when the target speaker is different from the most similar speaker. In another study crowdsourcing is used to select the best imitator for a set of 53 target speakers [19]. The authors used GMM-based ASV system for finding the imitators from a set of 176 participants. As a first step, the participants were asked to speak in natural and mimicked voices. Then an ASV system was used to filter the candidates by assessing the closeness of their voice samples to the target speakers. Finally, a set of good imitators were confirmed based on the performance of filtered candidates on multiple imitation tasks.

In contrast to the studies in [12, 13, 19] with limited number of target speakers (and use of a single ASV system only), the current work uses two large publicly available datasets, VoxCeleb1 and VoxCeleb2, consisting of more than 7,000 speakers to search the targets corresponding to the six recruited participants who are native Finnish speakers. In addition to the impersonator-specific closest, median, and furthest targets, we also consider a common celebrity target. This is to evaluate the impersonator's natural ability to mimic a known person. Further, the target speakers are chosen from both Finnish and non-Finnish speakers to assess impersonator's success rate for native and non-native targets.

3. ASV-assisted mimicry attacks

3.1. 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. The aim of an attacker $A \in \mathcal{A}$ is to masquerade him/herself as a specific target $T \in \mathcal{T}$ that he/she pre-selects

using automatic speaker recognition technology. We assume that $J \gg K$ — that is, an attacker is relatively infrequent, but there are many natural persons who have their voice samples available in a public domain. Celebrities and anyone actively uploading or streaming their video or voice data to social media platforms are representative examples.

Given a pair of speech utterances (or a pair of *collections* of multiple utterances), (U_i, U_j), an **automatic speaker verification** (ASV) system (speaker detector), $\mathcal{D}(U_i, U_j)$ computes a *detection score*, $s_{ij} \in \mathbb{R}$, typically a *log-likelihood ratio* (LLR),

$$s_{ij} = \log \frac{p(U_i, U_j|H_0)}{p(U_i, U_j|H_1)}, \tag{1}$$

where the null hypothesis H_0 states that U_i and U_j originate from the same speaker and its complement H_1 states they originate from two different speakers. In this work, utterances are represented as fixed-sized *embeddings* using either *identity vectors* (i-vectors) [33] or *x-vectors* [34]. If either U_i or U_j consist of multiple utterances, their embeddings are averaged. The LLR computation uses *probabilistic linear discriminant analysis* (PLDA) [35] scoring. The higher the LLR score, the stronger the support for the null hypothesis. We consider two different types of ASV systems. The first one, **attacker's ASV** ($\mathcal{D}_{\text{pub}}$), is a public-domain ASV implementation while the latter, **black-box ASV** ($\mathcal{D}_{\text{black}} \neq \mathcal{D}_{\text{pub}}$), is the system which the attacker attempts to hack into as a specific target. The attacker does not have access to the internal workings of $\mathcal{D}_{\text{black}}$ or its outputs to optimize mimicry attacks. The attack proceeds as follows:

ASV-assisted target speaker selection for mimicry attack

1. Attacker $A \in \mathcal{A}$ records his/her natural voice sample, $\mathcal{U}_{\text{nat}}$ (one or several utterances).
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 domain. A picks the **closest target**, $j^* = \arg\max_{j=1}^{J} \mathcal{D}_{\text{pub}}(U_{\text{nat}}, U_j)$, where U_j contains all the public recordings of target T_j.
3. A further uses $\mathcal{D}_{\text{pub}}$ to pick the top-scoring utterances of T_{j^*} similarly.
4. A listens to the selected utterance(s) and tries to adjust his/her voice towards the target. Once completed practicing, A submits a mimicked test utterance U_{mimic} to $\mathcal{D}_{\text{black}}(U_{\text{mimic}}, U_{j^*})$ with identity claim T_{j^*} (aiming to be accepted as T_{j^*}).

Note that in our model, the attacker uses the public-domain ASV system only to

select the target speakers. In some prior work, such as [31], ASV score was provided as feedback for the impersonators to improve their mimicry skills. We do not provide ASV (or other) feedback signals to our attackers. The main reason is that the ASV score is not necessarily intuitive to humans. For instance, a low attacker-to-target ASV score does not suggest *how* to modify one's voice production so as to improve the score. Providing intuitive feedback, for instance in terms of suggested articulatory or voice source modifications, would require a different system (and user interface) design. In our model, the attacker uses a readily-available public-domain ASV system to rank and select potential target speakers, but *without* any further numerical feedback or system optimization. Such 'passive' ASV system could be, for instance, a voice search service that finds most similar speakers to the user's voice from a public video archive — see [36, 37] as examples.

Both the attacker's and the attacked ASV systems are *text-independent*, *i.e.* none assumes the spoken contents of the compared enrollment and test utterances to match. Even if properly-optimized text-dependent ASV systems can provide higher recognition accuracy, text-independent ASV systems provide more flexibility and are justifiable in certain authentication applications, such as secure teleconferencing and telephone banking. The use of text-independent ASV systems in this study was, in fact, *necessary* as we have no control over the text content in the celebrity corpus (VoxCeleb).

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

The attacker's ASV system uses i-vector front-end [33] and probabilistic discriminant analysis (PLDA) [35] back-end to compute speaker similarity scores. The system's acoustic front-end[3] extracts 20 mel-frequency cepstral coefficients (MFCCs) per frame using 20 filters, leading to 60 features per frame after including deltas and double-deltas. The chosen MFCC configuration is commonly used in speaker recognition experiments [38, 33]. The features are processed with RASTA filtering [39] and cepstral mean and variance normalization (CMVN). Non-speech frames are omitted using energy-based speech activity detector (SAD) (described in Section 5.1 of [40]).

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

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 system, and hence the attacker's system differs from the attacked system.

	Attacker's ASV system ($\mathcal{D}_{\text{pub}}$)	**Attacked ASV system** ($\mathcal{D}_{\text{black}}$)
Type	Text-independent	Text-independent
Implementation	MSR Identity Toolkit (MATLAB)	Kaldi (c++)
Sampling rate	16 kHz	16 kHz
Acoustic features	60 MFCCs (20 static+20-Δ+20-$\Delta\Delta$), RASTA, SAD, CMVN	30 MFCCs (no deltas), Sliding CMN normalization, SAD
Embedding type	i-vector (400-D)	x-vector (512-D)
Back-end / scoring	LDA (250-D)+PLDA (simplified, 200-D)	LDA (200-D)+PLDA (2-cov)
Development data	Librispeech (train-clean-360 and train-clean-100 subsets), WSJ0 and WSJ1	VoxCeleb2, training part of VoxCeleb1
Data augmentation	None	Reverberation, noise, music, babble
EER*	12.84 (%)	3.11 (%)

* EER for VoxCeleb1 test protocol

The universal background model (UBM), i-vector extractor, linear discriminant analyzer (LDA), and PLDA, are trained using Wall Street Journal (WSJ) and Librispeech corpora. LDA is used to reduce 400-dimensional i-vectors to 250 dimensions before centering, whitening, and length normalization. Simplified PLDA with 200-dimensional speaker subspace is used for scoring. For further details, refer to Table 1 of the current work and Section 2.2 of [23].

3.3. Attacked ASV system

In our experiments, we regard the x-vector system [34], based on pre-trained Kaldi [41] recipe, as the ASV system 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 system in terms of feature extractor set-up, embedding type, and development corpora (Table 1). The attacked system is the Kaldi x-vector recipe for VoxCeleb, while the attacker's system uses i-vectors. Unlike the i-vector extractor, the x-vector extractor is trained discriminatively using speaker labels.

4. 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 [20] and Voxceleb2 [21] to each of the locally recruited subjects (described in Section 5). 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, most of which is active speech. Both VoxCeleb corpora were collected using automated pipeline exploiting face verification and active speaker verification technologies [21].

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 rather than non-Finnish targets. According to the VoxCeleb1 metadata, there are no Finnish speakers in VoxCeleb1. VoxCeleb2 did not include nationality metadata but we extracted the nationalities automatically using Google's *Knowledge Graph* API[4]. This way we identified a total of 44 Finnish speakers from VoxCeleb2.

5. Locally recruited attackers

5.1. Speakers and recording gear

We recruited $K = 6$ voluntary local speakers (4M + 2F) to serve as 'attackers'. The selected terminology, 'attacker', is made for convenience to reflect the focus of ASV vulnerability study; it should be understood that all speakers took part voluntarily and were not asked to 'hack' any computer systems in the sense understood in the security field. In fact, most of our speakers are considered *naive* to the study aims: two of the male subjects knew the specific goals of the study but the remaining four subjects were not informed that the text and target speakers were tailored for them, nor where the target voices were obtained from. The speakers were not informed that the study relates to ASV vulnerability, but were asked to mimic the target speakers as accurately as they could. All the subjects signed an informed consent form to use their speech data for research, and were rewarded with movie and coffee tickets.

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

All six attackers are native Finnish speakers with an age range between 24 to 44 years old. They are *naive* impersonators who lack formal training in mimicry. We adopt the same recording setup from [42] and text prompts are described in detail in [24]. 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 set-up 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 and electroglottograph) were also collected, but are not used in this study.

5.2. 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 [42]. Their corresponding English versions were added for this study. We have approximately six minutes of speech (before speech activity detection) per speaker from Session 1. Detailed description of the material used in data collection can be found in the online supplementary material [24].

5.3. 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. The selected target speakers per attacker are presented in Tables 2 and 3.

Table 2: Target speakers (closest, median and furthest) per attacker. Selection of potential targets from 44 Finnish celebrities in VoxCeleb2.

Attacker ID	**Celebrity**	**Profession**	**Spoken language**
M1	Samuli Edelmann	Actor, singer	Finnish, English
	Paavo Väyrynen	Politician	Finnish
	Antti Tuisku	Pop singer	Finnish
M2	Samuli Edelmann	Actor, singer	Finnish, English
	Paavo Väyrynen	Politician	Finnish
	Mika Kojonkoski	Ski jumper, politician	Finnish, English
M3	Joni Ortio	Ice hockey player	Finnish, English
	Elastinen	Rap musician	Finnish
	Perttu Kivilaakso	Musician	English
M4	Samuli Edelmann	Actor, singer	Finnish, English
	Tuomas Holopainen	Musician	Finnish, English
	Jyrki Katainen	Politician	Finnish, English
F1	Anna Puu	Pop singer	Finnish
	Karita Mattila	Opera singer	Finnish, English
	Tarja Halonen	Politician	Finnish, English
F2	Sofi Oksanen	Writer	Finnish, English
	Kaisa Mäkääräinen	Biathlete	Finnish, English
	Tarja Halonen	Politician	Finnish, English

Table 3: English speaking celebrities (closest, median and furthest) per attacker. Selection from 7321 potential targets in VoxCeleb1 and VoxCeleb2. * indicates speakers from VoxCeleb1.

Attacker ID	**Celebrity**	**Profession**	**Spoken language**
M1	Valentin Inzko	Politician	English (Austrian)
	Elijah Cummings	Politician	American English
	Chris Colfer *	Actor	American English
M2	Jeremy Irons *	Actor	British English
	Karan Tacker	Actor	Indian English
	Ryan Ochoa *	Actor	American English
M3	Éric Boullier	F1 manager	English (French)
	Guillaume Canet *	Actor, director	English (French)
	Bill Gilman	Singer	American English
M4	Ciarán Hinds	Actor	Irish English
	Ian Kinsler	Baseball player	American English
	Phil Mickelson	Golf player	American English
F1	Jessie J *	Singer	British English
	Candace Cameron *	Actress	American English
	Lin Shaye *	Actress	American English
F2	Fay Ripley	Actress, author	American English
	Belcim Bilgin	Actress	English (Turkish)
	Anne Hathaway *	Actress	American English

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 speaking targets are Päivi Räsänen (female, politician) and Ilkka Kanerva (male, politician), and the common English speaking targets are Hillary R. Clinton (female, politician) and Leonardo DiCaprio (male, actor). The choice of the common targets is arbitrary but based on a loose, subjective criterion *as famous as possible*. We first identified a short-list of VoxCeleb celebrities that we thought are well-known. We then ran an e-mail survey among our friends and colleagues (23 responded), asking each one to indicate the three most famous persons (in their opinion). We combined their votes to select the common targets. Even if the selected targets are well-known, from the viewpoint of ASV they are *random* target speakers with no strong presuppositions how similar their voices are to our attackers.

In summary, for each of our four male and two female subjects, we select six customized targets (three ASV-ranks $\times$ two languages) and two common gender-matched ones (one Finnish, one English). This gives a theoretical total of $3 \times 2 \times 4\,\text{male} + 2\,\text{common male} + 3 \times 2 \times 2\,\text{female} + 2\,\text{common female} = 40$ target speakers. But as the reader can see from Table 2, not all of the ASV-selected targets are unique: one Finnish male celebrity (Edelmann) was the closest target for three attackers, one Finnish male celebrity repeated as the median speaker for two male attackers (Väyrynen), and one Finnish female celebrity (Halonen) is the furthest speaker for both female attackers. These collisions might be explained by the the limited number of Finnish celebrities (30M, 14F) in VoxCeleb. The total number of unique celebrity targets is 36.

For each of the 36 target speakers, we selected multiple short utterances so that, when combined, each target would have at minimum 30 seconds of active speech. The selected utterances were used to evaluate the ASV system attacks. We selected only short utterances for two reasons. First, the duration of most of the VoxCeleb excerpts varies between five to ten seconds. Second, we deemed shorter utterances to be easier for our attackers to imitate. Detailed description of these utterances is provided in an online supplementary material [24].

The selection of the VoxCeleb excerpts was done by utilizing attacker's ASV system. For the closest and furthest targets we selected, respectively, the highest and lowest scor-

ing utterances. For the median speakers, we selected the utterances closest to the mean. This was further accompanied by manual inspection: if the audio quality (determined subjectively by listening) in a given utterance was not deemed high enough, we discarded it and moved on to the next ones in the ranked list.

5.4. Speech transcription and the mimicry recordings

Unlike the first recording session (common to all subjects), the second and third sessions were tailored for each subject. This process involved 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 Finnish speakers. The 35 MTurk crowdworkers and the two 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. All the final transcriptions are provided in the supplementary material [24].

In Session 2, which took place five to six 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. 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 [14]. In that study, the target speakers were public personalities that each impersonator knew. Each impersonator completed three scenarios with an increasing level of detail about the target speakers. The impersonator was first asked to produce prototypical target speech without knowledge of text (other than common category, *e.g.* everyday sentences). The impersonator was then revealed the target speaker texts to be impersonated and, finally, he would be provided audio reference of target.

In the last session, which took place two to six days after Session 2, the subjects were provided with the same transcript as in Session 2. Additionally, they were now provided access to the actual target speaker audio excerpts. The transcripts were provided on a

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

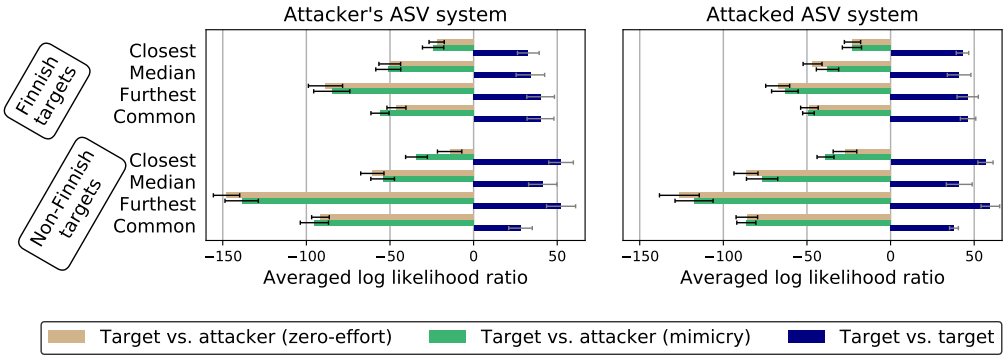

Figure 2: Comparison of attackers' ASV scores (log likelihood ratios) to the targets' scores for both of the ASV systems involved in the study. The scores are averaged over all attackers and all speech segments. The error bars represent 95 % confidence intervals for the means.

printed paper and the audio was presented through headphones connected to a tablet computer with an interactive webpage. The subject was allowed to interact with the audio samples and could listen to 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.

Following standard convention in the context of spoofing and countermeasure studies [3], we refer to the speech recordings of the second session as *zero-effort*. This is to signify that the attackers were instructed to produce target speaker texts in their own modal voice, *i.e.* without dedicated effort to sound like the target. The recordings from the last session, in turn, are simply referred to as *mimicry* utterances.

6. Results: mimicry attacks against automatic verification system

In the following, we evaluate the effectiveness of mimicry attacks against ASV systems. The target speaker models used in the experiments were enrolled using all available segments except those selected for testing as described in Section 5.3.

Figure 2 displays how the PLDA scores of genuine and attack trials compare to each other. 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

Table 4: Score differences between attacks with impersonated voices and attacks with natural voices. Differences are averaged over attackers, target nationalities, and utterances. ± indicates 95 % confidence intervals. In the case of the closest target speakers, impersonation attempts are counterproductive.

ASV system	Closest	Median	Furthest	Common
Attacker's ASV	-9.7 ± 5.2	2.2 ± 4.3	5.9 ± 7.1	-7.2 ± 4.3
Attacked ASV	-5.2 ± 3.9	9.2 ± 3.3	6.1 ± 4.3	-0.5 ± 3.8

ASV system, implying that the ASV-assisted speaker selection *can* help in ASV attacks. Second, in general, the attackers' natural and mimicry scores are significantly (by a wide margin) below the target scores. Additionally, we find no significant difference between the zero-effort and mimicry attacks (except for the closest category). Finally, as the recruited attackers are Finnish, attackers' scores against the Finnish targets are higher than for the non-Finnish targets (within each rank category).

We further display the difference of mimicked and natural speech scores in Table 4. 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 (as 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, if the targets are not close to the attackers (*i.e.*, median and furthest categories) or are less well known, impersonation is potentially helpful (though, not by a statistically significant margin).

Our attackers are native Finnish speakers recorded with a specific set-up which may differ from the target domain (VoxCeleb) conditions. This raises a question whether our mimicry attacks might have been unsuccessful due to *domain mismatch*. To address this question, we studied *target-domain*, *attacker-domain*, and *cross-domain* non-target score distributions as well as target-domain and attacker-domain target score distributions. It was not possible to construct cross-domain target trials as we do not have speakers common to both domains. The main interest in this specific study is to compare target-domain non-target scores to cross-domain non-target scores. If the cross-domain scores (the case of attacks) do not fall below the target-domain scores, it suggests that the

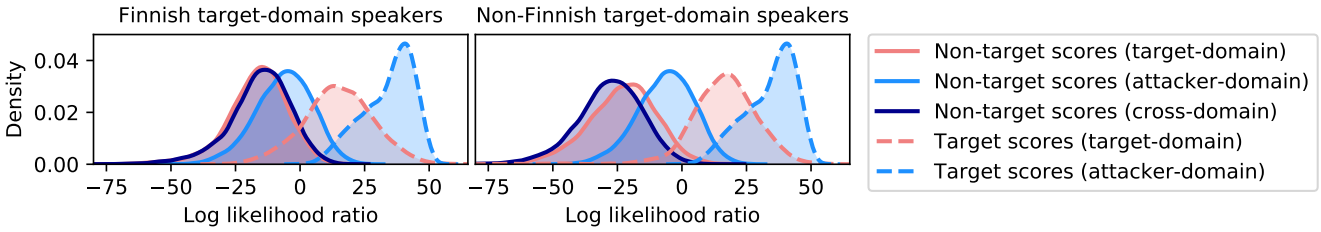

Figure 3: Distributions of target and non-target scores in different domains. *Cross-domain* non-target scores are obtained by scoring speakers from the attacker domain against the speakers from the target (VoxCeleb) domain. The simulated mimicry attacks in this work fall under the category of cross-domain trials. As the cross-domain score distributions overlap almost perfectly with the target-domain non-target distributions, the domain mismatch does not seem to make attacking more difficult, at least when the targets are Finnish.

attacker does not get penalized by the domain mismatch. The scores for the study were obtained from the attacked x-vector based ASV system.

Figure 3 indicates that when the nationality mismatch is present (non-Finnish target-domain speakers), the cross-domain non-target scores are, on average, slightly lower than the the target-domain non-target scores. If, however, the target-domain speakers are Finnish, like our recruited attackers are, the non-target speaker distributions overlap almost perfectly. This suggests that the Finnish attackers attacking the Finnish VoxCeleb targets did not seem to get penalized by the domain mismatch. The domain mismatch can be observed by comparing target and non-target scores of attacker-domain and target-domain. As the attacker-domain is has much less variability in the conditions, the scores in attacker-domain tend to be higher.

7. Perceptual evaluation of mimicry attacks

Next, we evaluated how ASV assisted mimicry attacks perform against human listeners. Further, we compared the findings of perceptual test to those obtained from the attacks against the ASV system. To avoid nationality mismatch between targets and attackers, we restricted our experiments to Finnish targets only.

7.1. Listening test setup

In total, we had 625 pairs of speech samples (trials) to be evaluated by the listeners. These trials can be divided into five groups of 125 trials (4 to 7 trials for each of the 24 attacker-target combinations). The first three groups are related to the mimicry attacks: 1) target vs. target (reference point), 2) target vs. attacker (zero-effort mimicry), and 3) target vs. attacker (mimicry). For each set of three trials, the same target enrollment utterance is used. The speech content of the test utterances is the same in all three cases, but different from that of the enrollment utterance (*i.e.* text-independent speaker comparison). The two last types of trials focus on the attacker. They are 4) attacker (zero-effort) vs. attacker (zero-effort) and 5) attacker (zero-effort) vs. attacker (mimicry). These two cases are included, respectively, to study the listeners' performance for the same-speaker trials with fixed recording conditions, and to study how much the attackers modify their voices relative to their natural voices when mimicking. In the cases 4) and 5), the enrollment utterances are selected from the English part of the data described in Section 5.2. Similarly as above, for each set of two trials, the enrollment utterance is fixed and the two test utterances have the same content. In all of the cases, the enrollment utterance was selected from the available utterances so that its duration is close to the duration of the test utterances.

The listening trials were accompanied with a question *"How similar the two speakers in the two voice samples sound to you?"*, to which the listeners answered using a 4-point scale with options *Very dissimilar*, *Dissimilar*, *Similar*, and *Very similar*. The 4-point scale was selected to enforce the listeners to make up their mind regarding speaker similarity. When presenting the trials, the order of the two voice samples in a trial was randomized so that the enrollment utterance was not always played the first. Each trial was presented individually and their order was randomized as well. For each of the 625 trials, we asked opinions from five different listeners, so in total we collected 3125 responses from the listeners.

We recruited the listeners using the Amazon's MTurk service. All the listeners were either native English speakers or had advanced English skills. In total, 225 crowdworkers participated the listening trials. Five workers rated more than 100 trials, whereas 130 completed less than five. On average, a crowdworker completed $3125/225 \approx 14$ trials.

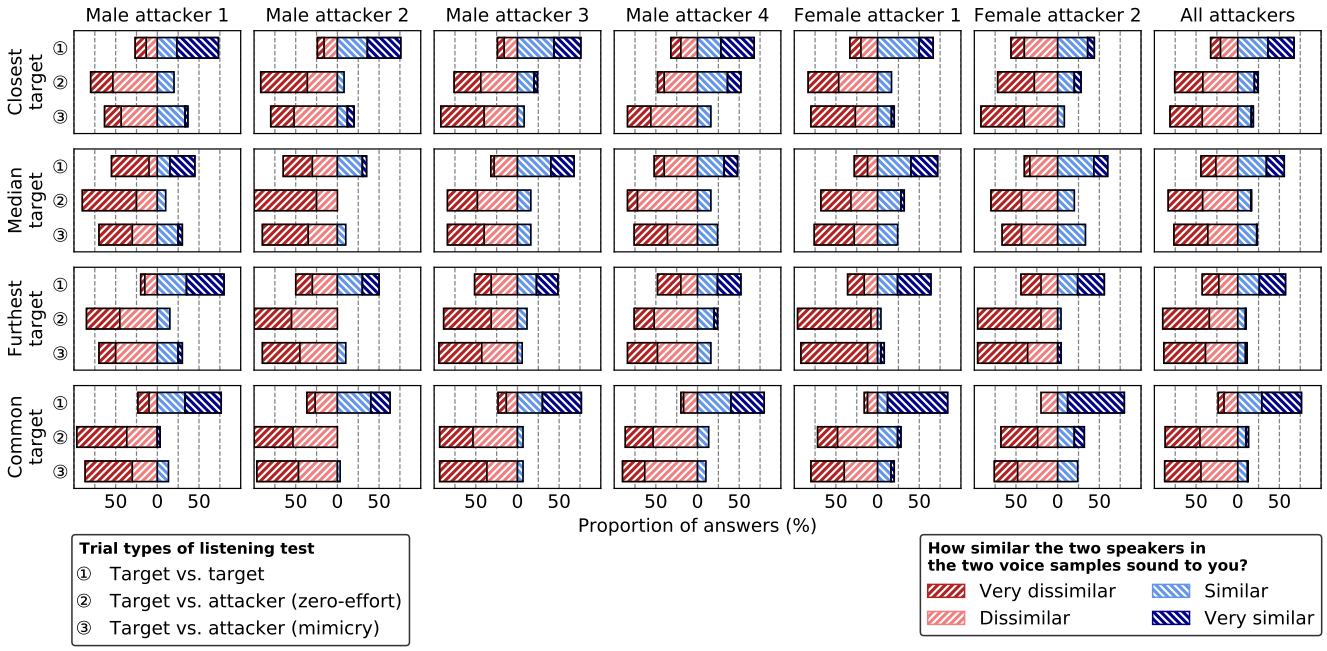

Figure 4: Results from the listening test (target speaker enrollment vs. test segment). Each attacker (in columns) has 4 targets speakers (in rows: closest, median, furthest, common). For each attacker-target combination, there are three different trial types (denoted by circled digits) as described in the left-hand side legend. The last column shows the results when trials from all the attackers are combined.

Out of the 225 listeners, 40 provided information about their mother tongue: 26 English, 4 Italian, 4 Portuguese, 2 German, 2 Spanish, 1 Estonian, 1 Tamil.

7.2. Listening test results

We present the main results of the listening test in Figure 4, which presents the listener judgements of speaker similarity for all the studied attacker-target combinations. First, the listeners regard the two samples from the same target speaker (target vs. target cases) similar or very similar to each other, as expected. However, there are individual cases that turned out to be difficult for the listeners. For example, the median target of the male attacker 1 was considered dissimilar or very dissimilar sounding to himself in most of the answers. Informal listening of the utterances of this target revealed that the target's voice sounded different each time mostly due to differences in speaking style, recording conditions, and audio processing. For example, in one sample, the target speaker (Finnish

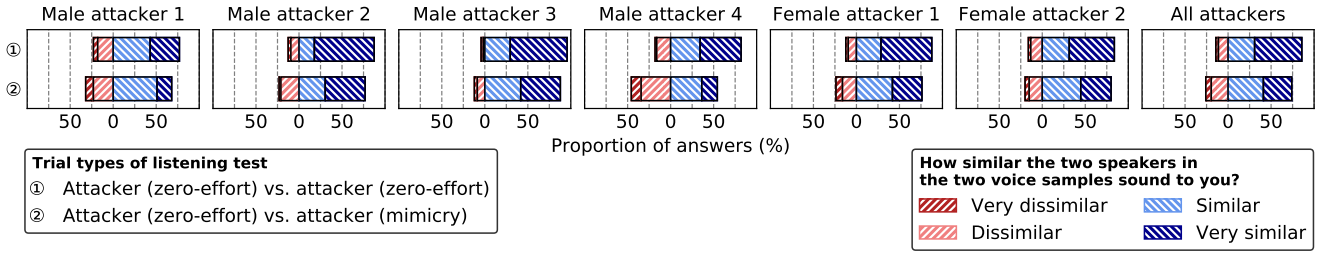

Figure 5: Results from the listening test (attacker enrollment vs. attacker test segment). Listeners evaluate each attacker's enrollment samples against attacker's zero-effort and mimicry-effort attack samples. The voice modification induced by mimicry attempt makes the attackers sound less like themselves.

politician) is being interviewed in a talk show, whereas in another sample he is giving a public speech in very different conditions.

How are the listeners opinions affected by mimicry? On average (see the last column of Figure 4), mimicry does not seem to help to make the attackers sound more like the targets. At the individual level, we find, however, that male attackers 1 and 2 got higher ratings for their mimicked speech. Further, we find that ASV assisted target speaker selection can help in choosing attacker-target pairs that sound similar to each other. That is, the furthest targets get lower similarity ratings than the closest targets. Even if automatic systems and humans based their speaker similarity judgments differently, the broad rank categories seem consistent.

Figure 5 displays listening test results for those trial types where attacker's enrollment utterances are compared to attacker's test segments with and without mimicry effort. The same-speaker trials have higher similarity ratings in comparison to those in Figure 4). This is expected since our attacker corpus is practically free from channel variation and background noise unlike the VoxCeleb collections. In addition, we find that when the attackers are trying to mimic the voices of the target speakers, they sound a little bit less like themselves.

7.3. Comparison of human listeners and automatic speaker verification system

To compare human opinions to ASV system scores, we scored the same trials using both the attacker's ASV system and the attacked ASV system. All the individual scores for three different trial types are displayed in Figure 6. The scores for the content

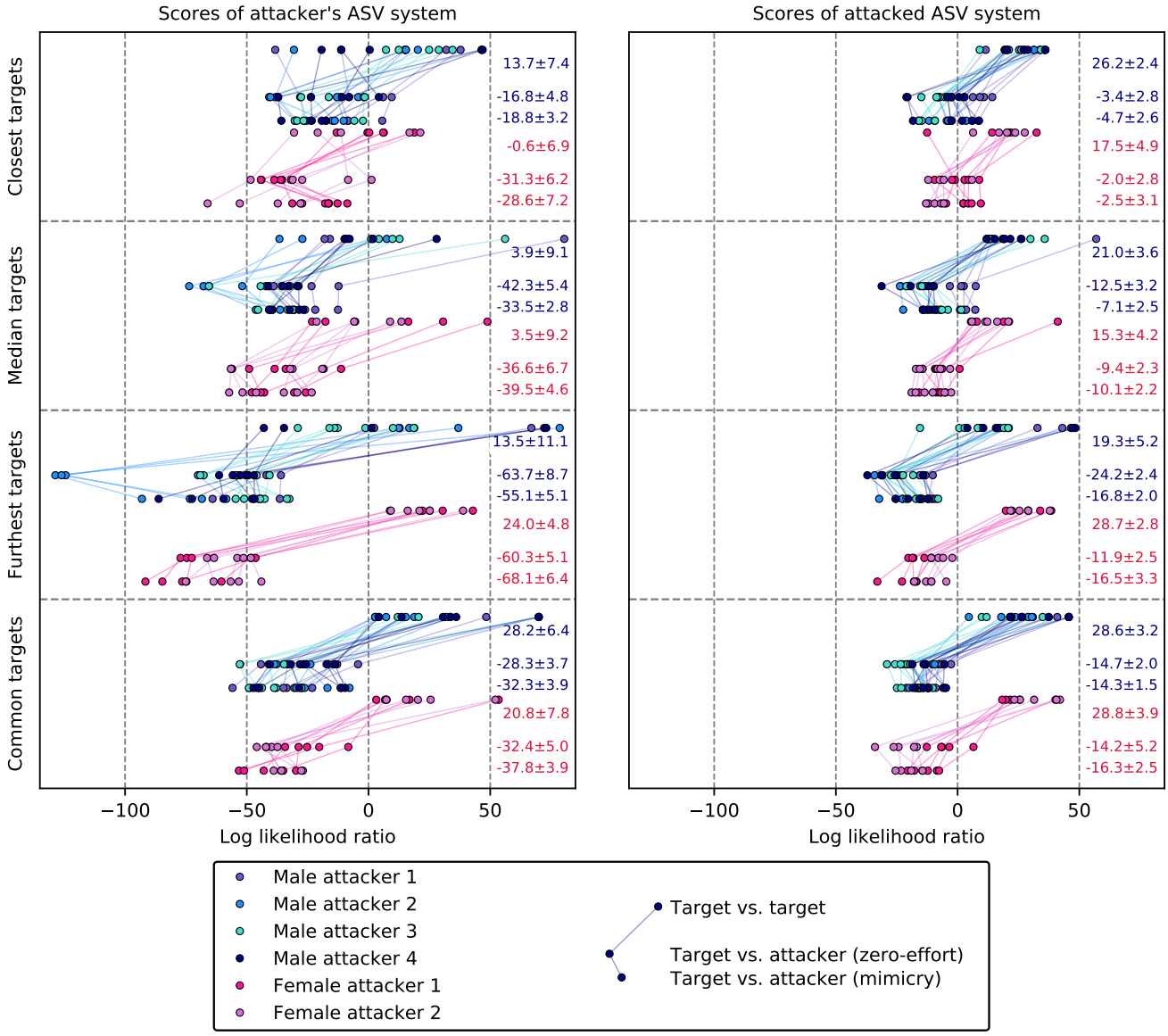

Figure 6: The scores of the ASV systems for the trials used in the listening test. The scores in each score triplet (described in the legend) are from the trials that have the same target speaker enrollment utterance and the speech content is the same in all the three test segments. Scores for male and female attackers are shown in separate groups. The right side of each graph displays the mean values of the score groups together with standard error of the mean multiplied by 1.96.

matching test utterances are connected with lines and thus form score-triplets. This allows us to see how close the attacker's scores are to the target's scores and how successful were the mimicry attempts in individual cases. The results agree with the results of Figure 2, as expected — the only difference with the earlier ASV protocol is the number of target speaker enrollment utterances, which is now only one[6].

In general, the findings from the listening test are similar to what the ASV system scores imply. The ASV-assisted target speaker selection helps to bring attacker's scores closer to the target's scores, while the mimicry attempts do not seem to help much to bring the scores closer to the target's scores.

8. Prosody and formant analysis of mimicry attacks

To gain further insight how attackers' change their voices to mimic their targets, we carried out a study of the changes in fundamental frequency (F0), speaking rate, and formants. Our main motivation to study these qualities is to see whether attackers changed more their prosody than spectral cues. If this is the case, the changes might not be reflected by ASV scores as our systems are based on spectral features.

8.1. Estimation of fundamental frequency and speech rate

Speaking rate, in terms of syllable rate (the number of syllables per second), was measured using a Praat [43] implementation [44] that automatically calculates the number of syllables per sample duration by detecting syllable nuclei [45] and pause duration. As for F0 extraction, we adopt an autocorrelation-based method [46] implemented in Praat. We use gender-specific frequency ranges set to [75, 200] Hz for males and [100, 300] Hz for females. We initially tested F0 extraction with wider F0 ranges but it was observed that the selected ranges were appropriate to exclude possible tracking errors and outliers

[6] In general, data processing capacity of ASV systems and listeners differ: ASV systems can process multiple enrollment utterances and large number of trials, but humans have limited attention span and memory and cannot process many trials (or excessively long utterances). For the maximum benefit of the ASV system, the earlier ASV protocol used in Fig. 2 used multiple enrollment utterances, while the scaled-down ASV protocol (single enrollment utterance) used in Fig. 6 was designed to facilitate perceptual speaker comparisons.

in the F0 contour. The parameters to select the F0 candidates at 10ms intervals were set at their default values in Praat: silence threshold 0.03, voicing threshold 0.45, octave cost per octave 0.01, octave-jump cost 0.35, and voiced-unvoiced transition cost 0.14.

We summarize F0 values of each utterance using two summary statistics, namely, median and standard deviation. They reflect, respectively, the average pitch range and pitch dynamics within a given utterance. We study changes in these summary statistics between the zero-effort and mimicry attempts, with the aim of studying whether or not our attackers attempt to match their broad prosody characteristics with those of their targets upon their mimicry attempts.

8.2. Estimation and alignment of formant frequencies

We performed formant analysis by comparing formant information of aligned utterances. First, we extracted formant center frequencies of the first three formants (F1, F2, and F3) using VoiceSauce [47] with Praat backend. Next, we aligned attacker's utterances (natural & mimicry) with target's utterance using *dynamic time warping* (DTW) [48]. The aligning process was done similarly as in [49]. This process involves using automatic selection of active speech frames that are well aligned and have reliable formant information. The alignment of utterances turned out to be challenging due to differences in speaking styles, acoustic conditions, and small deviations in spoken texts caused by mumbling. Thus, in addition to the automatic frame selection, we listened the aligned utterances in order to discard the the badly misaligned ones. Finally, after getting the aligned formant data, we measured the formant difference d between utterances a and b as

$$d(a,b) = \frac{1}{3T} \sum_{t=1}^{T} \sum_{n=1}^{3} \left| f_a(t,n) - f_b(t,n) \right|, \tag{2}$$

where T is the number of aligned frames and $f_a(t,n)$ is the center frequency of formant n of utterance a at frame t.

8.3. Results of prosody and formant analysis

In Figure 7a, we show the results for the analysis of speech rate differences. For each attacker-target combination, the displayed speech rates are obtained by averaging the speech rates of the available utterances (4 to 7 utterances per combination). The results

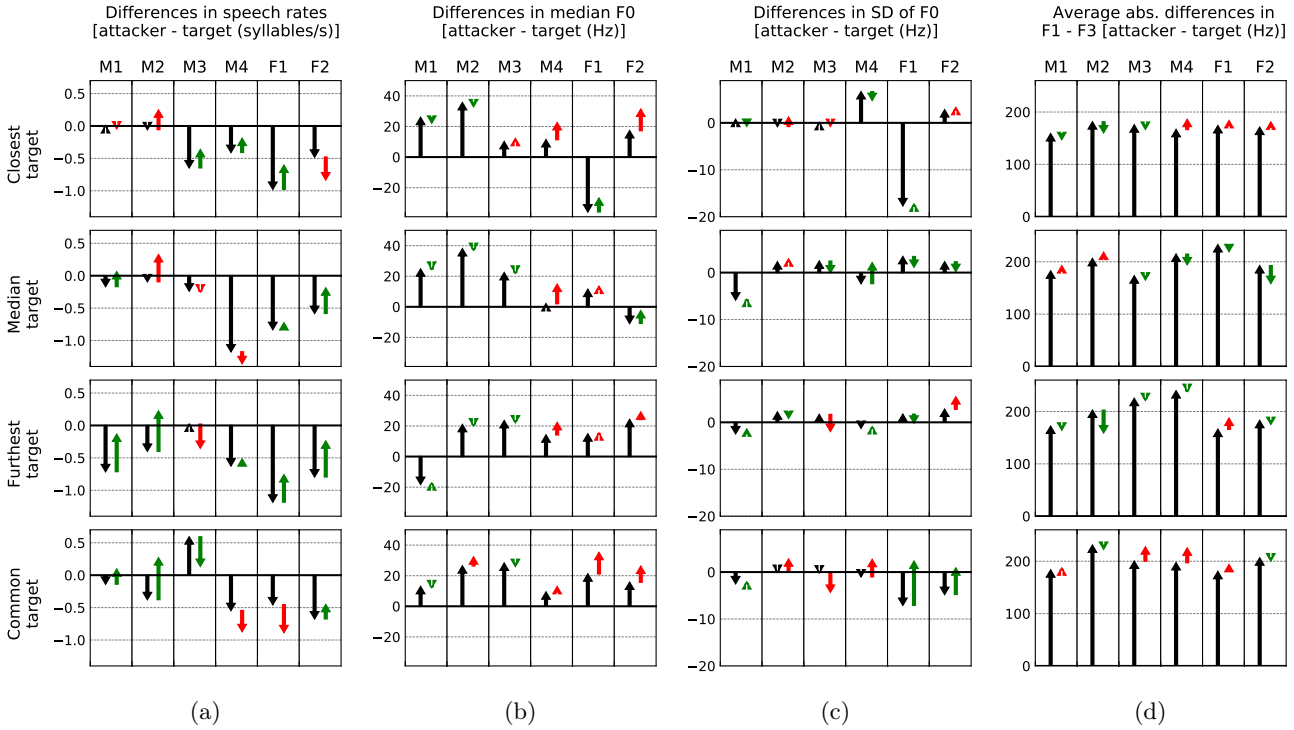

Figure 7: Differences of attacker's (M1, M2, M3, M4, F1, F2) prosodic and formant parameters to target's parameters for all attacker-target combinations. Differences are shown for non-effort speech (**black arrow**) and for mimicked speech. The effect of mimicry is displayed with a **green arrow** if it made attacker's and target's parameters closer to each other and with a **red arrow** otherwise.

indicate that the speech rates of the attackers were, in general, slower than the targets' speech rates, when the attackers were not mimicking. This was anticipated, since the attackers were reading prompted text from a paper yielding slower speaking rates as opposed to those of the targets samples obtained from conversational situations. After listening to target's speech, the attackers were in most cases able to change their speech rates towards the targets' speech rates. At the individual level, we find that the male attacker 1 (M1) was good at adjusting his speech rate, while the male attacker 3 (M3) had naturally fast reading pace so that in some cases (common target) his speech rate was already too fast.

A similar comparison regarding F0 statistics is shown in Figures 7b and 7c. We find that the attackers M1, M2, and M3 did not change their F0 considerably while

mimicking, whereas attackers M4, F1, and F2 had some mimicry attempts with clearly different F0 than what their natural F0 is. We do not observe clear differences between closest, median, and furthest target categories in terms of distances in F0 parameters between attackers and targets.

Finally, in Figure 7d, we depict the formant differences between targets and attackers as defined in (2). Again, we find that the mimicking did not have major impact to the similarity of the formant frequencies. In 14 out of 24 cases, mimickers managed to get slightly closer to their targets in terms of the given metric. We further find that the formant differences are larger in the furthest category than in the closest category, which is expected as the location of formants affect the spectral features used in the target speaker selection.

9. Conclusion

Biometric data uploaded to the Internet in large quantities, including human voice samples, opens up potential for misuse whenever the same biometric identifiers are adopted for strong user authentication to regulate access to personal data records, bank accounts and other services. Our study addressed a potential risk related to combination of public-domain automatic speaker verification (ASV) technology and public-domain voice data. The former is used as a search tool to identify potential target speakers to be mimicked.

Our results suggest that human mimicry is a rather special skill and less effective in spoofing modern ASV systems compared to voice conversion, text-to-speech, and replay. In specific, none of our six attackers received high detection scores for their attacks from our simulated[7] public-domain or attacked ASV systems. Similar negative findings have been reported in earlier studies and are often speculated to be due to difficulty of humans to mimic accurately low-level spectral cues employed by ASV systems. One of our motivations was to *re-assess* whether speech mimicry — one of the weakest known attacks against ASV — might be made substantially stronger (or more practical) when the target

[7]The ASV implementations combine scripts/tools (*e.g.* MSR Identity Toolkit, Kaldi) that are all public-domain code. They should be considered as proxies of modern ASV technology, rather than end-user software.

speakers are selected using ASV. We approached this question from two perspectives. On the one hand, we wanted to find out how the score ranges associated with broad target speaker rank (closest, median, further) transfer from the attacker's ASV to the attacked ASV. This is the *technology* dimension of our attack model. On the other hand, we wanted to isolate the effect of the mimicry effort by collecting attackers' voice samples both 'before' (zero-effort attack) and 'after' (mimicry attack) listening to the target speaker's voice. This allows us to analyze the changes in attacker-to-target log-likelihood ratio (LLR) scores due to mimicry effect alone. This is the *human* dimension of our attack model. Concerning the broad target speaker rank, the score relations generalize well from the attacker's ASV system to the attacked ASV system: LLR(closest target) > LLR(median target) > LLR(furthest target) relationship was retained both for Finnish and non-Finnish targets. This suggests that one could, indeed, use one ASV system (here, i-vector PLDA) to emulate the broad speaker ranking of another, targeted ASV system (here, x-vector PLDA). We find this result interesting and worthwhile of future work. Even if the VoxCeleb corpora are among the largest (public) speaker corpora at this time, they are still tiny compared to the number of voice samples in the Internet. It would be interesting to repeat a similar study design to ours in a few years, perhaps with an order of magnitude larger target speaker corpus and, at this stage, unforeseen ASV technology. It would be important to uncover the conditions under which such emulation succeeds (or fails). With an increasing number of video and voice samples posted online, it is not only the security, but user privacy, that deserves attention.

Concerning the impact of mimicry effort, the attacker-to-target LLRs remained low, and substantially below the target-to-target LLRs in both zero-effort and mimicry scenarios. Curiously, while the LLR scores for the furthest target speakers indicated some increase between zero-effort and mimicry scenarios, for the closest targets the LLR scores *decreased* (but significantly only for the non-Finnish target speakers). To sum up, the broad target speaker rank generalized across the ASV systems, while the mimicry effect itself lead to negative (or no difference) effect. These findings reinforce the conjecture that voice mimicry by itself may not pose a strong attack against ASV; but ASV-based target speaker selection may.

We hypothesized that while our attackers' mimicry efforts did not have major impact

on the ASV scores, they might have impact on human perception. Human listeners might, to some degree, focus on different cues of speaker identity than the ASV systems, which mostly focus on spectral characteristics of speech. However, the results of our listening test did not support the above hypothesis, as the results showed similar patterns to those we saw from the ASV scores.

So as to understand better the mimicry strategies implemented by the attackers, we also analyzed changes in formant frequencies and prosody statistics (F0, speaking rate). Even if some attackers were able to adjust their average formant frequencies towards those of their target speakers, the relative change in attacker-to-target formant distance (from zero-effort to mimicry) was minor. Adjustments in F0 statistics were minor as well. The most prominent adjustments towards the targets were seen in the speaking rate.

Our study has a number of limitations that one should take into account in future studies. First, the number of attackers (six) is admittedly small. This limitation, familiar to some of the authors [16], is common to most speech mimicry studies and relates to difficulties in data collection. The number of attackers varies from 1 to half dozen (or so) [3]. Here, additional complications were caused by tailored target speaker selection, involving tedious speech transcription and several stages of data quality auditing. In future work, it might be practical to drop the transcription step and ask the attackers to impersonate their targets based on audio only. Another way to scale up the study would be attacker recruitment through crowdsourcing [19]. This will, however, introduce new uncontrolled variations (such as attacker microphone differences). All our attacks were recorded using the same gear in the same room.

The second limitation relates to the cross-domain data conditions: our attackers are native Finnish speakers, while VoxCeleb consists of many different nationalities and accents. Further, VoxCeleb consists of conversational speech while our attackers read text passages in an office environment. These differences induce style differences and might make the impersonation task harder for the attackers. This limitation is primarily due to lack of large Finnish celebrity corpus at the authors' exposure, as well as our preference to interact with the attackers conveniently. It would be interesting to repeat selected experiments using a larger target speaker corpus with matched mother tongue.

In VoxCeleb, we are limited to 44 Finnish target speakers. Future work could therefore either adopt a larger Finnish celebrity corpus, or to recruit native American English attackers. Given the nature of *found data*, controlling all the variations will be difficult.

Our attacks could also be made stronger in a number of ways. First, the attacker might use the public-domain ASV system in a more proactive way, such as optimizing its detection accuracy further in off-line experiments. Second, the attacker could potentially utilize more detailed feedback from a dedicated ASV system — in this work, attackers used ASV for speaker *ranking* while some prior work has used ASV score as a feedback signal [31]. Third, assuming there would be an actual monetary (or other strong) motivator to seriously mimic someone — similar to practicing to forge someone's signature — the attacker might use substantially more effort to get familiar with the speaking style of his or her targets. He or she might perhaps use feedback from prosody measurements in addition to ASV score. In our study, given the extensive work required to prepare the tailored targets and collect the data, all the above had to be relaxed to complete recordings in a reasonable time. The mimicry attacks (with audio reference of the target) took place in a single session and our attackers completed their mimicry tasks relatively fast. Nonetheless, in future work it would be interesting to evaluate whether mimicry attacks could be improved with further, and more proactive, training. Another interesting target would be studying combination of automatic target speaker selection with voice conversion (or other technical) spoofing attacks.

It would be also interesting to address whether, and how, one may benefit from current (or suitably modified) ASV methods to provide intuitive feedback to improve one's mimicry skills. This would be potentially helpful in suggesting specific articulatory or voice source modifications required to increase the ASV score. The present study was framed to the context of ASV attacks but such methods could be potentially useful for mimicry artists, voice actors, and language learners as well.

Acknowledgement

The work has been supported by Academy of Finland (proj. no. 309629 entitled "NOTCH: NOn-cooperaTive speaker CHaracterization") and by the Doctoral Programme in Science, Technology and Computing (SCITECO) of the UEF. A part of the work of

the first author was supported by NEC internship program. The work of Md Sahidullah was made with the support of Region Grand Est. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

References

[1] N. Ratha, J. Connell, R. Bolle, Enhancing security and privacy in biometrics-based authentication systems, IBM Systems Journal 40 (3) (2001) 614–634. doi:10.1147/sj.403.0614.

[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, Online; accessed 22-February-2018 (2016).

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

[4] Z. Wu, T. Kinnunen, N.E., J. Yamagishi, C. Hanilçi, M. Sahidullah, A. Sizov, ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge, in: Proc. INTERSPEECH, 2015, pp. 2037–2041.
URL http://www.isca-speech.org/archive/interspeech_2015/i15_2037.html

[5] S. Ergunay, E. Khoury, A. Lazaridis, S. Marcel, On the vulnerability of speaker verification to realistic voice spoofing, in: Proc. Seventh International Conference on Biometrics Theory, Applications and Systems (BTAS), IEEE, 2015, pp. 1–6.

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

[7] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, J. Yamagishi, The voice conversion challenge 2016, in: Proc. INTERSPEECH, 2016, pp. 1632–1636. doi:10.21437/Interspeech.2016-1066.
URL https://doi.org/10.21437/Interspeech.2016-1066

[8] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, 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] M. Sahidullah, H. Delgado, M. Todisco, T. Kinnunen, N. Evans, J. Yamagishi, K. Lee, Handbook of Biometric Anti-Spoofing: Presentation Attack Detection, 2nd Edition, Advances in Computer Vision and Pattern Recognition, Springer, 2018, Ch. Introduction to voice presentation attack detection and recent advances.

[10] J. Luck, Automatic speaker verification using cepstral measurements, The Journal of the Acoustic Society of America 46 (4) (1969) 1026–1032.

[11] W. Endres, W. Bambach, G. Flösser, Voice spectrograms as a function of age, voice disguise, and voice imitation, The Journal of the Acoustical Society of America 49 (6) (1971) 1842–1848.

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

[13] Y. Lau, D.T., M. Wagner, Testing voice mimicry with the YOHO speaker verification corpus, in: Proc. 9th Int. Conf. Knowledge-Based Intelligent Information and Engineering Systems KES, Part IV, Melbourne, Australia, 2005, pp. 15–21.

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

[15] A. Eriksson, The disguised voice: imitating accents or speech styles and impersonating individuals, in: C. Llamas, D. Watt (Eds.), Language and Identities, Vol. 8, Edinburgh University Press, 2010, pp. 86–96.

[16] R. González Hautamäki, T. Kinnunen, V. Hautamäki, A.-M. Laukkanen, Automatic versus human speaker verification: The case of voice mimicry, Speech Communication 72 (2015) 13–31. doi:https://doi.org/10.1016/j.specom.2015.05.002.

[17] M. Farrús, Voice disguise in automatic speaker recognition, ACM Comput. Surv. 51 (4) (2018) 68:1–68:22. doi:10.1145/3195832.

[18] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (General Data Protection Regulation), OJ L 119 (1) (2016) 1–88.
URL http://data.europa.eu/eli/reg/2016/679/oj

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

[20] A. Nagrani, J. Chung, A. Zisserman, VoxCeleb: A large-scale speaker identification dataset, in: Proc. INTERSPEECH, 2017, pp. 2616–2620. doi:10.21437/Interspeech.2017-950.
URL http://dx.doi.org/10.21437/Interspeech.2017-950

[21] J. Chung, A. Nagrani, A. Zisserman, VoxCeleb2: Deep speaker recognition, in: Proc. INTERSPEECH, 2018, pp. 1086–1090. doi:10.21437/Interspeech.2018-1929.
URL http://dx.doi.org/10.21437/Interspeech.2018-1929

[22] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, 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, 2017, pp. 506–519.

[23] T. Kinnunen, R. González Hautamäki, V. Vestman, M. Sahidullah, Can we use speaker recognition technology to attack itself? enhancing mimicry attacks using automatic target speaker selection, in: Proc. ICASSP 2019 (to appear), IEEE, Brighton, UK, 2019.

[24] V. Vestman, T. Kinnunen, R. González Hautamäki, M. Sahidullah, Voice mimicry attacks assisted by automatic speaker verification. Additional material, Published online (2019).
URL http://cs.uef.fi/~rgonza/papers/Additional_material_attacker-impersonator.pdf

[25] S. Ribaric, A. M. Ariyaeeinia, N. Pavesic, De-identification for privacy protection in multimedia

content: A survey, Sig. Proc.: Image Comm. 47 (2016) 131–151.

[26] J. S. Chung, A. Jamaludin, A. Zisserman, You said that?, in: Proc. British Machine Vision Conference, 2017.

[27] M.-Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, in: Proc. Advances in Neural Information Processing Systems, 2017, pp. 700–708.

[28] S. Suwajanakorn, S. M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing Obama: Learning lip sync from audio, ACM Transactions on Graphics (TOG) 36 (4) (2017) 95.

[29] R. Chesney, D. K. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security, 07 California Law Review (2019, Forthcoming); U. of Texas Law, Public Law Research Paper No. 692; U. of Maryland Legal Studies Research Paper No. 2018–21. (July 2018).
URL http://dx.doi.org/10.2139/ssrn.3213954

[30] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, 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.

[31] E. Zetterholm, M. Blomberg, D. Elenius, A comparison between human perception and a speaker verification system score of a voice imitation, in: Proc. Tenth Australian International Conference on Speech Science & Technology, Macquarie University, Sydney, Australia, 2004, pp. 393–397.

[32] H. Khan, U. Hengartner, D. Vogel, Targeted mimicry attacks on touch input based implicit authentication schemes, in: Proc. of the 14th Annual International Conference on Mobile Systems, Applications, and Services, ACM, 2016, pp. 387–398.

[33] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Trans. on Audio, Speech, and Language Processing 19 (4) (2011) 788–798. doi:10.1109/TASL.2010.2064307.
URL https://doi.org/10.1109/TASL.2010.2064307

[34] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust DNN embeddings for speaker recognition, in: Proc. ICASSP, IEEE, 2018, pp. 5329–5333.
URL http://www.danielpovey.com/files/2018_icassp_xvectors.pdf

[35] S. J. D. Prince, J. H. Elder, Probabilistic linear discriminant analysis for inferences about identity, in: Proc. Eleventh IEEE International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007, 2007, pp. 1–8. doi:10.1109/ICCV.2007.4409052.
URL https://doi.org/10.1109/ICCV.2007.4409052

[36] V. Vestman, B. Soomro, A. Kanervisto, V. Hautamäki, T. Kinnunen, Who do I sound like? showcasing speaker recognition technology by YouTube voice search, in: Proc. ICASSP 2019 (to appear), IEEE, Brighton, UK, 2019.

[37] Intelligent Voice, Who do you sound like? (2019).
URL https://celebsoundalike.com/

[38] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, D. O'Shaughnessy, Multitaper MFCC and PLP features for speaker verification using i-vectors, Speech Communication 55 (2) (2013) 237 – 251.

[39] H. Hermansky, N. Morgan, RASTA processing of speech, IEEE Trans. Speech and Audio Processing 2 (4) (1994) 578–589.

[40] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: From features to supervectors, Speech Communication 52 (1) (2010) 12–40.

[41] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, The kaldi speech recognition toolkit, in: Proc. IEEE ASRU, 2011, pp. 1–4.

[42] R. González Hautamäki, M. Sahidullah, V. Hautamäki, T. Kinnunen, Acoustical and perceptual study of voice disguise by age modification in speaker verification, Speech Communication 95 (2017) 1–15. doi:10.1016/j.specom.2017.10.002.
URL https://doi.org/10.1016/j.specom.2017.10.002

[43] P. Boersma, D. Weenink, Praat: doing phonetics by computer [Computer program], version 5.4.09, retrieved 15 June 2015 from http://www.praat.org/ (2015).

[44] N. H. De Jong, T. Wempe, Praat script to detect syllable nuclei and measure speech rate automatically, Behavior Research Methods 41 (2) (2009) 385–390.

[45] D. Wang, S. S. Narayanan, Robust speech rate estimation for spontaneous speech, IEEE Trans. on Audio, Speech, and Language Processing 15 (8) (2007) 2190–2201.

[46] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Proc. of the Institute of Phonetic Sciences, Vol. 17, 1993, pp. 97–110.

[47] Y.-L. Shue, P. Keating, C. Vicenik, K. Yu, VoiceSauce: A program for voice analysis, in: Proc. Seventeenth International Congress of Phonetic Sciences, 2011, pp. 1846–1849.

[48] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. on Acoustics, Speech, and Signal Processing 26 (1) (1978) 43–49.

[49] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, T. Kinnunen, Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction, Speech Communication 99 (2018) 62–79.