Crowdsourcing Attacks on Biometric Systems

Saurabh Panjwani*
Independent Consultant, India
saurabh.panjwani@gmail.com

Achintya Prakash
University of Michigan, USA
achintya@umich.edu

ABSTRACT

We introduce a new approach for attacking and analyzing biometric-based authentication systems, which involves crowdsourcing the search for potential impostors to the system. Our focus is on voice-based authentication, or speaker verification (SV), and we propose a generic method to use crowdsourcing for identifying candidate "mimics" for speakers in a given target population. We then conduct a preliminary analysis of this method with respect to a well-known text-independent SV scheme (the GMM-UBM scheme) using Mechanical Turk as the crowdsourcing platform.

Our analysis shows that the new attack method can identify mimics for target speakers with high impersonation success rates: from a pool of 176 candidates, we identified six with an overall false acceptance rate of 44%, which is higher than what has been reported for professional mimics in prior voice-mimicry experiments. This demonstrates that naïve, untrained users have the potential to carry out impersonation attacks against voice-based systems, although good imitators are rare to find. (We also implement our method with a crowd of amateur mimicry artists and obtain similar results for them.) Match scores for our best mimics were found to be lower than those for automated attacks but, given the relative difficulty of detecting mimicry attacks visá-vis automated ones, our method presents a potent threat to real systems. We discuss implications of our results for the security analysis of SV systems (and of biometric systems, in general) and highlight benefits and challenges associated with the use of crowdsourcing in such analysis.

1. INTRODUCTION

Biometric-based authentication is one of the most compelling alternatives to passwords for enabling access control in computing systems and, more generally, for identity management in systems. Even with some of the deployment difficulties associated with biometrics as compared with passwords, their usage in mainstream applications like banking and border security control is growing and new forms of biometrics are being continually experimented with for user authentication tasks [4].

Amongst many other reported advantages of biometrics, it is often claimed that they have an upper hand over passwords in their resilience to being faked or spoofed by ordinary human beings, even those who are acquainted with attack victims. This is also cited as a primary reason for preferring them over passwords or tokens in real deployments [8, 3, 25]. However, rigorous research on such claims is still lacking and even with a rich and mature literature on biometric-based authentication, there is no convincing answer to this simple question: for an authentication system $\mathcal{A}$ trained on biometric features of a set of users S, drawn from a large universe U, is it likely that users in S can be impersonated by those in U? In particular, is it likely that the biometric features of some user $u \in S$ are "similar enough" to those of another user $u' \in U$ for u' to be able to impersonate u to $\mathcal{A}$? This question, though generally relevant to biometric-based systems, is particularly interesting for behavioral biometrics, which define identification features over user actions (e.g., speaking or writing): such biometric forms can be "copied" with conscious human effort and differences in inherent characteristics could potentially be compensated for by such imitation.

In this paper, we consider the potential of imitation as a means to thwart biometric-based authentication systems with a primary focus on voice-based authentication or speaker verification. Speaker verification (SV) systems are gaining prominence in the real world because of the widespread use of mobile devices (numerous known deployments by banks and mobile operators; see Sect. 2) but security analysis of such systems has been limited to the use of automated tools and techniques (like voice conversion, record-and-replay) as attack vectors. In contrast, the ability of humans to imitate other humans' voices for the purpose of impersonation is less understood and generally assumed to be difficult in practice [11, 18]. Reflecting this contrast, defenses against automated attack techniques in SV schemes have become stronger with time but those against imitation attacks are still unknown.

We make two key contributions in this paper. First, we present a new method to execute imitation attacks on SV systems involving a large number of untrained users as imitators; and second, we analyze the effectiveness of this method with respect to a well-known and commonly-used SV scheme based on Gaussian Mixture Models (GMMs). The method

*Part of this work was done when the author was employed with Alcatel-Lucent Bell Labs.

Symposium on Usable Privacy and Security (SOUPS) 2014, July 9–11, 2014, Menlo Park, CA.

we propose is simple and generic and it essentially involves the use of crowdsourcing to search for and identify candidate mimics for users in a given target set S. It is generic in that it does not assume a specific implementation of the SV system, except that it allows black-box access to the attacker. (Black-box access is used to identify "close matches" between candidate mimics and the targets.) It is efficient in that it uses mobile phones and crowdsourcing to quickly collect speech samples from geographically-dispersed individuals and to select candidate mimics from a large set of untrained users. We do not know of any prior work which uses crowdsourcing for biometric security analysis, voice-based or otherwise, or for analyzing authentication schemes in general. The very idea of identifying candidate impersonators from a large pool of untrained users (as opposed to hand-picking them from an expert population) does not seem to have been rigorously experimented with prior to this paper.

Our analysis of the technique with respect to a GMM-based SV system yields three key outcomes. Our first learning is that mimicry is a rare skill and that the average user of Web-based crowdsourcing platforms does not have the ability to pick the right speaker to mimic from a target set *and* to mimic that speaker well, even when provided high monetary incentives. This is somewhat expected and is also aligned with prior work which argues that professional mimicry artists exhibit greater flexibility to modify their voices than amateurs (within the realm of mimicking celebrity voices) [1, 26]. What is more surprising is the second outcome, which is that the crowdsourcing technique does identify *some* users with the ability to impersonate target speakers to the system and to do so with high consistency across authentication attempts (from a pool of 176 candidates, six achieved an overall false acceptance rate of 44%). In most cases, the imitators require help in identifying the right (closely-matching) target speaker to mimic and we found only one user who was able to self-identify a target speaker successfully. We also ran parallel experiments with a crowd of amateur mimicry artists and obtained similar success rates there, although motivating these users to participate in the experiments proved harder. Our results significantly improve upon findings from prior studies [10, 11, 15, 26] and through a careful imitator selection strategy, we are able to demonstrate better impersonation success than what has been found in these studies.

Finally, we find that even the best imitators identified by our technique fare poorer than automated attack techniques in terms of attack success rates and are unable to match the mean self-scores of target speakers in impersonation attempts. While this may appear like a negative finding, it is important to view it in the light of the fact that automated attacks are becoming easier to defend against (via different forms of *liveness* detection measures) but defenses against imitation attacks are not known in the literature. The impersonation success rates we demonstrate for our crowd-based imitators are sufficient to mount online attacks on real voice-biometric systems and current defenses for automated attacks seem insufficient to prevent them. Furthermore, given the improvement our technique offers over prior work on voice mimicry, such attacks present a potent threat to SV systems and one that future systems must suitably address. We discuss implications of our results for the design of future biometric systems (voice and otherwise) and how crowdsourcing-based analysis can assist in this process.

Before we proceed with the details, we make one important high-level remark regarding the paper. Our attack implementation should be viewed as a "proof-of-concept" of mounting crowdsourced attacks on voice-biometric systems and our work is a preliminary study of the viability of such attacks. Our main goal is to investigate whether crowdsourcing platforms with naïve, untrained users *can* be used to mount imitation attacks on SV systems and how to set up the right candidate filters to enable this effectively. The scale at which such attacks might occur on a real system cannot be deduced from our results alone. We use Amazon's Mechanical Turk to implement our proof of concept (which suffices to show attack viability) but such a platform is unlikely to be the vehicle for a real attack due to the associated legal implications and sampling difficulties in attack implementation (see Sect. 5). Further studies are needed to understand how such attacks could be implemented in practice or how the attack method could be used to analyze the security of real, large-scale systems.

The rest of the paper is organized as follows. In Sect. 2, we present some background and related work on biometric security, with a specific focus on security of voice-based biometric systems. In Sect. 3, we describe our attack technique and in the following section, we describe the experimental setup we used to implement and evaluate the technique. Section 5 presents our experimental results and the paper concludes in Sect. 6.

2. BACKGROUND AND RELATED WORK

Biometrics broadly fall into two categories—physiological, which are based on physical characterisics of an individual (e.g., fingerprints, facial features) and behavioral, which are based on behavioral traits and actions (e.g., speech, typing patterns and handwritten signatures). Speech has a unique place in this categorization in that it combines elements of both physical (vocal tract structure) and behavioral (speaking style) aspects of an individual, both of which are generally regarded to have differentiating elements across humans [13].

Biometric-based authentication systems of all types have a common structure: there is a *training* component, wherein each user submits her identity u and a set of biometric samples $\gamma_1, \ldots, \gamma_k$ to the system and the system uses these samples to prepare a "model" for u; and a *testing* component, wherein each user submits a fresh sample γ', along with her identity u, and the system checks for a "match" between γ' and the model that it prepared for u. A successful match implies successful authentication to the system. Matching is a binary classification problem—a user either classifies as u or classifies as "not u". This is different from biometric-based *identification* wherein user labels are not provided during testing and the classification task is n-ary (which of the users $u_1, \ldots, u_n$ is the closest match to γ'?). Much of the work on biometric-based authentication is around defining the right approach for modelling and matching users, which differs significantly across biometric forms.

2.1 Security of Biometric-Based Authentication

The fuzzy nature of biomterics (γ' may differ across tests even for the same user) presents new security challenges for the system designer: an adversary need not compute an exact biometric sample of u in order to impersonate as u to

the system; an "approximate" sample suffices. The system could be tuned to limit the acceptable level of approximation but this is also constrained by the fact that strict limits inconvenience real users, especially if the underlying biometric suffers from high variability across time and context (what is often referred to as *session variability*). The challenge is to come up with suitable matching thresholds which enable the right users to authenticate often enough but which cause all adversarial ways to create approximate samples to fail.

Broadly, there are two approaches to security analysis that have been considered in the literature. One involves the consideration of *automated attacks*, which use computing machinary to "create" fake biometric samples that can impersonate users to the target system. The classical automated attack is record-and-replay—digitally record samples from a user u and replay them to the system to authenticate as u. Record-and-replay is the Achille's heel of biometric-based authentication, particularly so for physiological biometrics [16, 19] which have limited scope of system-imposed dynamic variations. To defend against them, system designers normally introduce an element of freshness in the biometric capture process (e.g., for voice, have the user speak a different phrase for every authentication attempt). In the recent past, newer forms of automated attacks, like generative [2] and conversion [6] attacks, have emerged which try to defeat freshness impositions in systems by learning to generate new samples for a user u based on past samples of u and auxiliary data.

As automated attacks have grown in complexity, so have the defenses against them. Most real-world biometric systems today implement some form of *liveness* detection measures [22], which are automated ways to detect whether biometric samples provided during authentication originate directly from a human (are "live") or not. For fingerprint-based authentication, a common measure is to detect pulsation or temperature gradients in the biometric-providing object. For voice, measures range from challenge-response to the use of multi-modal techniques (e.g., capture lip-movement during authentication [7]). An emerging trend in voice-based authentication is the use of *human-mediated* liveness detection: in applications where the user is required to converse with a trusted human agent and the authentication process is incidental (e.g., phone banking), delegate the task of detecting liveness to the agent, and have the machine focus on matching[1]. Since human listeners are usually better with distinguishing machine-generated speech from human speech (and since automated techniques are not known to generate "natural-sounding" human speech yet), this approach is the best defense for automated attacks in such applications.

Besides automated attacks, security analysis of biometric systems may also consider *human attacks* i.e., the faking of biometric samples for a user u by another user u'. Unlike automated attacks, these attacks (if shown to be feasible) seem harder to defend against (particularly, in remote authentication scenarios) and liveness detection is unlikely to work against them. Some researchers question the feasibility of such attacks based on the position that they require specialized skills [2] and finding skilled people is expensive. Recent work has demonstrated that this position does not hold up for some biometric forms like keystroke dynamics [17] but this work only applies to biometrics for which the notion of a "match" (and particularly, "closeness" of a match between two samples and their temporally-corresponding parts) is visually representable to human attackers. This assumption does not hold for all biometric forms, including voice biometrics, which make limited use of temporal data in creating biometric templates. Furthermore, while [17] studies the question of designing appropriate feedback mechanisms to *train* unskilled users in biometric mimicry, we consider the question of *finding* appropriate mimics in a large universe (e.g., an online crowdsourcing platform) in a manner such that they can succeed with minimal training. We expect that this approach will apply to a broader class of biometric systems and investigate it for voice in the current paper.

[1]Nuance's FreeSpeech system implements this technique: `http://www.nuance.com/landing-pages/products/voicebiometrics/freespeech.asp`

2.2 Speaker Verification Primer

Before we describe relevant literature on security of speaker verification (SV), we provide an overview of SV methods. Broadly, there are two types of speaker verification systems—*text-dependent* [12], which require users' training and test samples to have the same (or similar) text; and *text-independent*, which do not have such a requirement. Both types have multiple real-world deployments, but text-independent systems are gaining popularity because they tend to offer relatively better usability (no human memory requirements) as well as security (greater amenability to liveness detection) trade-offs. At the same time, text-independent techniques are harder to implement and less efficient: unlike their counterpart, they cannot rely on temporal relations between speech frames when modeling speakers and have to work harder to extract features from speech. We focus on text-independent systems in this paper although our method could equally well be applied to text-dependent ones.

Most text-independent SV systems work as follows. To process any input speech, they first create its frequency spectrum (using one of many variants of the Fourier transform) and based on certain properties of the spectrum, extract, what are called, *spectral features* from it. These features are generated by averaging out values across the entire length of the sample i.e. they do not contain temporal data. Spectral features extracted from the training data could either directly be mapped to a biometric template or, what is more common, a *generative model* is learnt over them. Standard machine learning approaches like expectation maximization (EM) are applied to learn such models. The most commonly-used generative models are *Gaussian Mixture Models (GMMs)* which represent speech features in the form of a collection of Gaussian distributions. The process of matching a test speech sample γ to a speaker u involves extracting spectral features from γ and testing the likelihood of these features being generated from the GMM linked with u. Some systems also try to model prosody in speech when representing users but the use of spectral features is more common. We refer the reader to [13] for a good overview of the text-independent SV literature.

In this paper, we focus entirely on one kind of SV scheme—the GMM Universal Background Model (GMM-UBM) scheme [20]—which is the most widely-studied, and possibly, the most widely-deployed, text-independent SV scheme. The key characteristic of this scheme is the use of a "background" model which is meant to model the universe of all human speech and is a GMM, say Λ_B, trained prior to creating speaker models using samples from outside the target set.

The speaker model of a user u, say Λ_u, is then built by "adapting" the background model Λ_B based on features extracted from u's training samples. Matching a sample γ to u involves comparing the likelihood that γ's features were generated from Λ_u and the likelihood that they were generated from Λ_B. A high match score is assigned to γ if the former likelihood is much greater than the latter and the sample is accepted as u's sample if and only if the match score exceeds a pre-set threshold. In UBM-based systems, the better the quality of the background model (more variety in background speech samples), the better is the performance of the system. Besides GMM-UBM, there is a variety of other GMM-based schemes in the speaker recognition literature and some of the more recently-developed ones also provide greater resilience to session variability than GMM-UBM. But these schemes are less standardized (in terms of parameter settings) and stable, well-documented implementations for academic research are not widely available.

In general, there seems to be an upward trend in the adoption and deployment of SV systems worldwide [8], although rigorous data on this is missing. Multiple banks (e.g., bank Leumi in Israel[2]) and telecom operators (e.g., Bell Canada in Canada[3] and Turkcell in Turkey [3]) have already deployed SV systems in their phone-based support services and banks elsewhere in the world are also moving in that direction [25][4]. Conceivably, a good number of these systems are text-independent [3] although accurate penetration statistics are hard to find. In India, we are aware of one company [23] which supplies voice biometric technology for on-site authentication to a large BPO with over 100K customers and has also piloted their technology with multiple financial service providers; one of our future goals is to study usability-security trade-offs in SV systems in collaboration with this company.

2.3 Security of SV Systems

As with other types of biometrics, the literature has largely focused on automated attacks when analyzing speaker verification security. Several papers analyze susceptibility of SV systems against replay and conversion attacks [6, 10, 14] but there is no evidence that these attacks work against the liveness detection measures that have been proposed for voice biometrics. In particular, human mediation and challenge-response seem sufficient to defeat them.

There is prior work on imitation attacks, too, but most of this work is either restricted to studying mimicry of celebrity voices [26] or mimicry performed by professional or semi-professional imitators [1, 15] or else, a combination of the two [10, 11, 26]. The general picture portrayed by these works is that mimicry specialists are good at imitating prosodic elements of speech but tend to perform poorly (false acceptance rates (FARs) of 10% or less) when trying to attack GMM-based SV systems. The work of Lau *et al.* [15] is the only one we are aware of which reports FARs of greater than 30%, but they too seem to consider "amateur imitators" (two in number) with some experience in mimicry[5]. Our work significantly expands the space of amateurs through the use of Web-based crowdsourcing and we incorporate people without any experience in drama or mimicry to play the role of impostors. Prior studies [10, 11, 15, 26] use at most six potential imitators whereas we consider nearly two hundred and carefully narrow down to the most promising candidates from this set. Despite our relatively low-skilled sample space, we are able to find users who can perform successful imitation attacks on SV systems and often with performance better than what has been demonstrated for the case of experienced imitators.

3. THE ATTACK METHOD

Throughout the paper, we assume text-independent SV systems implemented over cellular networks (i.e., we assume all voice communication happens using mobile phones). While this assumption is not necessary for the implementation of our method, it arguably leads to the most convenient implementation of it. Authentication over mobiles forms one of the most compelling application scenarios for speaker verification and many real deployments operate in this scenario.

We now describe our method at an abstract level. Let $\mathcal{A}$ be the SV system being analyzed and let S be the speaker set for which the system is trained. Our attack method involves setting up a telephony server which runs an IVR system for voice data collection. The attack occurs in three steps:

1. *Imitator solicitation:* First, we use a crowdsourcing platform $\mathcal{P}$ to solicit candidate imitators for speakers in S. Workers associated with $\mathcal{P}$ are asked to perform two tasks: (a) submit natural (i.e. unmodified) speech samples to the telephony server and (b) given recorded speech samples of speakers in S, listen to these samples, select some speakers who the worker believes he can feasibly copy and submit "mimicked" speech samples for each selected speaker. We assume an IVR interface which allows workers to listen to their recordings and to re-submit a sample, if the worker perceives a previous recording to be unsuitable. Suitable incentive and disincentive schemes can be used with $\mathcal{P}$ to attract workers to these tasks.

   The mimicry task is meant to identify imitators based on their own judgement of which speakers they are capable of mimicking and their perceived similarity with such speakers. There may be few people who possess the skill to make such judgements accurately but in a large crowd of workers, finding such people is not an impossible outcome. Note that we also collect natural samples per worker, which enables us to match workers to speakers in S based on natural closeness in voice.

[2] `http://www.businesswire.com/news/home/20100415005768/en/Top-3-Israeli-Banks-Roll-Customer-Facing`

[3] IBM's 2012 Case Study titled *World's Largest Voice Authentication Deployment Makes Privacy Protection More Convenient for Bell Customers* discusses this deployment: `http://www-304.ibm.com/partnerworld/gsd/showimage.do?id=24252`

[4] `http://www.biometricupdate.com/201301/mobile-devices-to-drive-bank-adoption-of-voice-biometrics`

[5] The definition of "amateur imitators" is ambiguous in [15]. Based on communication with the authors, it seems that these imitators were less experienced than those used in prior works [10, 11] but it is unclear whether they had prior mimicry experiences or not. FARs from [15] are higher than those from other studies plausibly because the imitators were matched to targets selectively (based on voice similarity) before FAR-computation; however, the study did not use candidate filtering techniques to identify good mimics, the way we do in the current work.

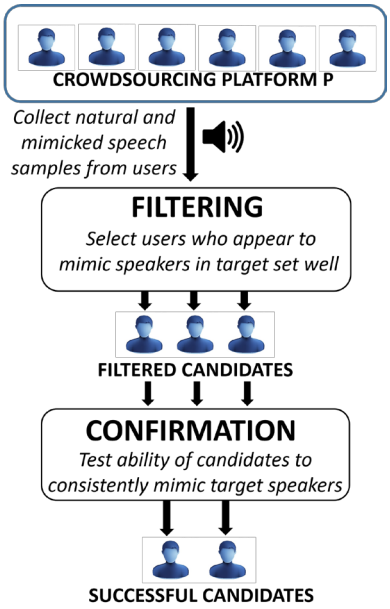

Figure 1: Pictoral depiction of our attack method.

2. *Candidate filtering:* Of all the workers who participate in the above crowdsourcing tasks, we select a few *candidate* imitators based on their performance on these tasks. For each worker w who participates, we determine whether w is a candidate imitator or not using two tests: *(a) do w's mimicked samples for any speaker $u \in S$ successfully authenticate u to $\mathcal{A}$?* and *(b) do w's natural or mimicked speech samples successfully authenticate u' to $\mathcal{A}$ for some user $u' \in S$ (not necessarily a speaker attempted to be mimicked by w)?* A worker is declared a candidate imitator if either of the tests return true for him. If he satisfies the first condition, we refer to him as a *deliberate candidate*; if he satisfies the second one, we call him a *emergent candidate*. Both conditions involve black-box invocation of the test procedure of $\mathcal{A}$. (Since the system is assumed to be text-independent, it is reasonable to test for the second condition using it.) For each condition, different implementations based on different notions of "success" can be used. For example, one implementation of type-2 candidacy testing could be: for any n natural speech samples uttered by w, do at least $n/2$ samples authenticate u' to $\mathcal{A}$ for some $u' \in S$?

   Our objective in including the second test for candidacy is to account for the potential incapability of workers to select good targets and the possibility of a "natural" match between a worker's voice—or mimicked variants of it—and a target's voice. Our technique could be generalized to capture other types of voice variations from each worker (e.g., "fake your voice by raising your pitch") and use the collective information from all variations to decide a worker's ability to mimic users in S, instead of relying only on his mimicry attempts. We restricted ourselves to the above approach for simplicity and even with this approach were able to achieve some success.

3. *Confirmation:* In this step, we try to increase our confidence in candidate imitators being good imitators. For each candidate imitator w identified above and a corresponding matching speaker u, we invite w to perform the following task: listen to the speech samples of u and submit multiple mimicked samples for that speaker. As the worker performs the task, he may also be given instantaneous feedback about his performance in order to help him create future samples better. We evaluate imitators based on their ability to successfully authenticate u to $\mathcal{A}$ in this task multiple times.

In a real implementation, there is also a fourth step in which the adversary selects the top performers in the confirmation step and has them authenticate as their corresponding speakers directly to $\mathcal{A}$. In this paper, we ignore that step since our goal is only to understand attack possibility, not in mounting an attack on a real system.

The assumption about black-box access to the attacked system $\mathcal{A}$ has some advantages. First, it makes the attack simple to implement and powerful from the perspective of proving negative results. (Insecurity against a black-box attacker implies insecurity against arbitrary attackers.) Second, it leads to a generic approach to security analysis; so, for example, the exact same technique can be applied to a different implementation of $\mathcal{A}$ with no change in the individual steps. Finally, it models the real possibility that the adversary may not have enough information about system implementation, and still be interested in breaking it. In practice, there may be limits on the number of black-box calls the adversary can make to the system (which could affect attack efficiency) but it is conceivable that the adversary can "simulate" such black-box access using other means (e.g., by computing matches on an identical copy of the system available as, say, commercial software or by working with a different system but one based on a similar algorithm). Future work is needed to determine how feasible black-box simulation is for real systems.

4. EXPERIMENTAL SETUP

This section presents the experimental setup we used to analyze our attack technique. We used an Asterisk-based IVR server[6] for all our speech data collection from users. Experiments were conducted from a laboratory in Bangalore, India and we chose to use Indian voices for both speakers and impostors in order to ease communication with users.

4.1 Speech Materials Used

While there are many standard speech datasets available for conducting speaker recognition experiments (e.g., the NIST SRE datasets which are updated on a regular basis), there are none with Indian voices that we have found available for free, which is why we decided to create our own dataset[7]. Our target set S consisted of 53 male users employed in a Bangalore-based IT company. Each speaker

[6] Asterisk is an open-source platform for building voice-based applications: `http://www.asterisk.org`

[7] Using standard datasets would have introduced effects of accent mismatch in the mimic selection process which we wished to avoid. Besides, such datasets are available under restricted usage licenses (e.g., use only for evaluating certain new SV techniques), which didn't fit our experiment goals.

provided two training samples (20-30 secs each) and multiple test samples (4-10 sec each) containing a combination of spoken digits and English sentences. Training and test samples were not phonetically matched, although test samples had some repetitions. Across speakers, training (resp. test) samples contained identical text, modulo some minor differences based on speaker identity (e.g., samples contained the name and occupation of the speaker). All speakers provided informed consent for using their speech data for our experiments. Our target set is admittedly small, but this only helps us strengthen our claims regarding the possibility of crowdsourced attacks on SV systems.

Speech was recorded via calls made to our IVR system from one out of two experimental handsets. Speakers spoke in a laboratory environment with limited ambient noise (modulo the sound of air conditioners and PCs). We spent about 5 minutes collecting speech samples per speaker. We focused on male speakers because we expected the task of finding male imitators to be easier than that for female ones (most Indian crowd-workers are male [21]). As our work is a proof of concept for crowdsourced attacks, focussing on males is sufficient to establish the viability of such attacks. Future work is needed to extend our results to female speakers. Our dataset is freely available upon request to the authors.

4.2 SV Settings

For our experiments, we used an implementation of GMM-UBM in the open-source package Alize [5], which is the only open-source package for speaker recognition with an active developer community today. The system was set up to operate on spectral features, as is standard in the GMM-UBM method. Waveforms were sampled at a frequency of 8 KHz and processed in 20ms frames with intervals of 10ms. The feature set consisted of 16th order mel frequency cepstral coefficients (MFCCs) and a log-energy term, augmented with corresponding first order derivatives, to result in a $(16+1)\times 2 = 34$ dimensional feature vector per frame. Standard normalization and energy filtering techniques were deployed to fine-tune the features.

For training the background model, we used a set of 424 speech samples (all male Indian voices) obtained from a set of crowdsourced data collection tasks posted on Amazon's Mechanical Turk (MTurk). These samples were conceivably submitted using a variety of mobile handsets, which would imply a good degree of variance and hence representativeness of the background model. Speaker models for all speakers in S were trained using a maximum a posteriori (MAP) based trainer. As is common in SV implementations, GMMs with 512 distributions were used and for computing test scores, we used average log-likelihood ratios (computed for the "top ten" distributions in the speaker models), refined with a standard normalization technique (T-Norm). We refer to the resulting SV system as $\mathcal{A}$ in what follows.

We evaluated the system's performance using a set of $10s$ test samples, one sample per speaker. (The same samples were used for score normalization across speakers.) Even with relatively short input speech, we recorded a small Equal Error Rate (EER)[8] of 2.31% for our 53-speaker dataset, not a surprising finding given that our data collection took place in a controlled environment. The Detection Error Trade-off (DET) plot for our baseline setup is shown in figure 2. The EER threshold score was determined to be $t_e = 2.04$. In our experiments, we used this threshold to determine success of matching test speech samples to speaker models: when we say that a sample γ *authenticates* u *to* $\mathcal{A}$, we mean that the result of matching γ to the speaker model of u using $\mathcal{A}$ produces a score greater than t_e. We assume that $\mathcal{A}$ also provides an interface to query for the *closest match* to a given test sample γ i.e. the speaker label u for which $\mathcal{A}$'s matching procedure produces the highest matching score between γ and u's speaker model, when compared across models. This interface is required in our attack implementation below. The interface may not exist for a real SV deployment, but it is generally possible for an attacker to simulate it—via standard log-likelihood computations—once he has acquired speech samples from the target speakers. (This is not an onerous task for a determined attacker given that speech is a frequently revealed biometric across users.) Finally, in our descriptions below, when we say that a sample γ *strongly* authenticates u to $\mathcal{A}$, we mean that it authenticates u to $\mathcal{A}$ and u is the closest match to γ in S.

In our settings, we did not make an attempt to optimize the parameters to get the best EER value, and chose standard parameters that are recommended by the literature on Alize. Since our interest is mainly in understanding the *relative* performance of imitators (compared with the baseline), this optimization does not seem necessary.

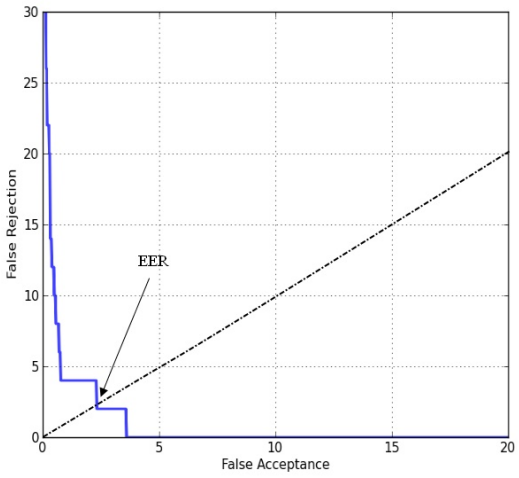

Figure 2: DET plot for our baseline SV system. The EER is 2.31% and the corresponding threshold is 2.04.

4.3 Crowdsourcing Apparatus

We implemented our attack technique using tasks posted on Amazon's MTurk platform. We randomly selected 20 speakers from our dataset and published their test and training samples via a Web interface[9]. Workers were asked to do

[8] Based on the threshold t set for match scores by the system's decision-making procedure, two types of errors can arise: *false rejection rate (FRR)*, the fraction of legitimate test-sample/speaker-model pairs which fail to score greater than t; & *false acceptance rate (FAR)*, the fraction of non-matching test-sample/speaker-model pairs which obtain a score greater than t. The EER is the value of FAR determined for the threshold t_e for which FAR equals FRR.

[9] We chose to publish speech samples of a subset of speakers instead of the entire set in order to ease target selection for our crowd-workers and to ease call management at our end. For studying attack "possibility", this approach is sufficient

the following task: listen to the 20 speakers' samples, select the speakers you want to mimic, practise mimicking them and then call our IVR server from any mobile phone[10]. On each call, the worker was required to provide three speech samples. The first was a natural speech utterance of two sentences which would involve stating their worker ID (a 13-15 character alphanumeric string) and a "task ID" associated with the speaker they are trying to mimic in that call. This was to be followed by two mimicry attempts corresponding to text spoken by the speaker in the published test samples (and lasting 4-10s each in the recordings). Workers could re-record mimicked samples in the same call but only one call per task ID from the same worker was considered for the evaluation. (We still maintained data from the other calls.) We explicitly stated that the task was restricted to male workers in India.

We expected workers to make multiple calls to the system, which would enable us to capture multiple natural speech samples per worker attempting our task. In effect, we recorded, for most workers, at least two mimicked utterances per selected target and two or more natural speech utterances. We emphasized the possibility of a big bonus (5 USD per target) for the best mimicry attempts but provided limited guarantees of payment to bad attempts. We used a subjective, tiered definition for "goodness" of a mimicry attempt. Workers who completed calls as required and who we perceived as making a well-intentioned voice modification (even though unsuccessful) were rewarded with a payout of at least 5 cents for each such call. If the worker made a remarkably good attempt (again, subjectively judged) or if the test scores for at least one of the mimicked sample was at least 50% of the self-scores of the target speaker, the payout ranged between 10 cents to 1 USD, based on our subjective evaluations. The ten best attempts received the bonus in the experiment[11].

Our candidate filtering process was as follows. First, we test whether both of w's mimicked speech samples for a target speaker u authenticate u to $\mathcal{A}$, in which case w is declared a deliberate candidate imitator. The second test, for checking emergent candidacy, involved two sub-tests: we test whether, for some $u' \in S$, either

- a majority of the natural speech samples of w strongly authenticate u' to $\mathcal{A}$; or
- both of w's mimicked speech samples for one target speaker (other than u') strongly authenticate u' to $\mathcal{A}$

If so, we accept him as a candidate. A worker could potentially be both a deliberate and an emergent candidate imitator or satisfy more than one conditions in the emergence test. The motivation for using strong authentication (instead of plain authentication) for the latter test is to increase the likelihood that the observed match between w and a speaker model was not a serendipitous event. All candidate imitators identified in this manner received at least 10 cents each and 5 USD in case they proved to be a deliberate candidate.

and only helps strengthen our results.

[10]While we did not try to rigorously determine the nature of the calling devices, a cursory examination of caller IDs suggests that most callers used mobile phones for the calls.

[11]Some amount of subjectivity in incentivizing workers seems necessary given that mimicry, in general, is judged perceptually and the association between perceptual and quantitative judgements is unclear and opaque to the workers.

Candidate imitators were invited to take part in the confirmation step. In this step, the expectation from the invited workers was to submit more natural speech samples, two per call and at least 60 mimicked samples for the target speaker u that the worker was able to (strongly) authenticate as. Candidates were largely required to utter the same text used in the target test samples but we also collected a few recordings of speech that differed from the test samples in textual content but were similar in the number of constituent syllables. The mimicked samples were collected across multiple recording sessions of at least 20 recordings each. Candidates were promised a bonus of 5 USD per session if they "performed well" in that session, which we defined as authenticating (though not necessarily, strongly so) as u in at least 30% of their attempts. We manually interacted with the imitators during the confirmation task, giving them instantaneous feedback on their performance (over a parallel Google Hangout session) as they submitted fresh recordings to the system and injecting frequent remarks of encouragement. Candidates were instigated to listen to prior "good" recordings of theirs and to try to imitate such recordings, as a strategy to improve scores[12]. In some cases, a target speaker *other than* u emerged as the closest match for the imitator in a majority of the new mimicked recordings; when this happened, we invited the imitator to attempt to mimic the new target afresh. At the end of this interaction, the candidate responded to a questionnaire asking about his background and experience in imitation and on MTurk tasks, in general. Cumulatively, we spent at least 3 hours per candidate during confirmation.

4.4 Mimicry Artists

For the sake of comparison, we also solicited participation in our task from mimicry artists in India. We found contact details of 25 artists (through a human agent who managed their portfolios) and reached out to them by phone. The people we reached were mostly amateurs and enthusiasts of mimicry and their expertise in the art was not well-established. None reported to practise it as their primary occupation although some claimed to have performed imitation acts in competitions, as part of plays and in gatherings.

Our interactions with these artists were similar in nature to those with the MTurk workers except that we interacted over phone more than email, which enabled us to converse with them more openly. The artists used our IVR server to provide sets of natural and mimicked speech samples just like the MTurk workers. We offered incentives similar to those offered to MTurk workers (equivalent of 5 USD bonus for the best attempts) and applied the same filtering techniques. In effect, this part of our experiment was a more targeted form of crowdsourcing aimed at a specific audience which seemed to possess the skills our task demanded.

[12]More elaborate forms of feedback could also be conceived for human attacks on biometric systems. For example, Meng *et al.* [17] provide visual feedback to their imitators of keystroke biometrics on "how close" they are to their target user. Such visual feedback is currently difficult to design for text-independent SV systems because of the lack of temporality in the creation of feature vectors in these systems. (The feedback giver cannot easily depict in a picture which part of the speech is being mimicked well and which part is not.) As such, we restricted ourselves to giving overall score-based feedback to our participants and included oral forms of encouragement along the way.

5. RESULTS

Over the 8-week period that our experiment ran in, we received a total of 733 calls to our IVR server, which included calls from both MTurk workers as well as some of the artists. Out of these, about a hundred calls generated audio files which encountered errors against Alize (either due to IVR bugs or because of missing voice data) and some others had issues of missing information (e.g., missing or incoherent MTurk ID) or were from a female caller. Discarding all such cases, we were left with 493 calls from 180 unique callers—176 MTurk workers and 4 artists. In our analysis, we used data only from these calls.

Persuading the artists to sign up for our task proved more difficult than we expected. There were multiple reasons for this difficulty. Some artists indicated that they had stopped practising the art. Others were not excited about imitating non-celebrity voices. A few felt that our incentives were not sufficient (we adjusted our offering in such cases, though this did not affect the eventual callers). Finally, many indicated an intent to call but never did, conceivably for a lack of real interest or distrust. While this was disappointing in a way, a useful side-effect was that the few artists who did participate were highly motivated to perform our task, as was exhibited in their repeated calls (more than 175 recordings in the case of one artist) during the experiment.

5.1 Candidate filtering outcomes

Among the 176 valid MTurk callers, 39 were determined as candidate imitators for our system—2 deliberate candidates and 38 emergent candidates, with one overlapping the two criteria. When probed further, we learnt that one of the two deliberate mimics had used a record-and-replay technique to impersonate as his target to the system instead of self-imitating the voice[13]. The other performed better during the confirmation phase, as discussed below. Effectively, only one out of 176 MTurk workers emerged as a true deliberate candidate in our experiment.

While the finding of a single deliberate candidate may seem like a dismal outcome, it is remarkable in the light of the fact that workers selected their targets from a sample of size 20 and did not have any visibility into the workings of the SV system being tested. As a comparison, none of the artists qualified as deliberate candidates, an indication that experience in mimicry may not be a criterion for imitation attacks against speaker verification systems. We gave the artists the additional capability of accessing recordings of all speakers in S and mimicking as many as they desired; still, deliberate candidacy proved difficult for them.

Even the finding of emergent candidates is interesting. Given that we had only 53 speakers in our dataset, we find it surprising that over a fifth of the 176 workers being evaluated could match *some* speaker in this set so as to be able to strongly authenticate as that user multiple times. Not all of these workers matched to unique users: the 38 candidates were mapped to 21 target speakers, with one target speaker emerging as the closest match for *six* candidates. Our confirmation task tested the resilience of some of these matches by collecting more samples from the workers.

The artists did slightly better on emergent candidacy, 2 out of 4 (50%) of them satisfying the condition. We enroled a third artist for the confirmation task even though he did not strictly qualify as a candidate. This artist was the most enthusiastic participant amongst all mimics: he attempted mimicry on more than 20 targets, was ostensibly making significant modifications to his voice in his mimicry attempts and even though he failed the emergent criteria on all targets in the first attempt, he returned to make further attempts wherein he was successful in meeting it for one speaker.

Interestingly, most of our emergent candidates were declared emergent not because of a close match between their *natural* voice and a speaker in our dataset, but because of closeness between their *faked* (mimicked) voice and a speaker they didn't intend to mimic. Out of the 38 MTurk emergent candidates, this was true for 28 of them, which included one worker who satisfied both criteria. The same was true for all the three artists who were emergent candidates. This suggests that when evaluating the ability of a user as an imitator for speakers in a system, simply matching his natural voice to the existing speaker models, as done in past research [15], is not sufficient; requiring him to vary his voice may give better hints on who his closest matches could be. The success of some of our emergent candidates, as discussed below, in continuing to impersonate their targets further supports this hypothesis.

Overall, our key learnings from this part of the experiment were: (a) most MTurk users do not have the ability to *self-identify* which speakers from a given dataset they can mimic to an SV system, but people experienced in mimicry do not seem to possess that ability either (within the scope of "ordinary" non-celebrity voices); and (b) even though such an ability may be scarce, several users (workers as well as artists) may be able to create voice modifications which bring them unexpectedly close to such speakers.

5.2 Confirmation outcomes

Out of the 38 MTurk workers we invited to participate in our confirmation task, 13 (i.e. 34%) responded with an affirmative response. We sent multiple follow-ups to the non-respondents but this number did not change. Even amongst the respondents, 4 out of the 13 responded only after multiple invites. (See the appendix for the email template we used.) While there could have been several reasons for this poor response rate, we believe that the peculiar nature of our tasks (expectation of mimicking others' voices, doing it over phone and doing it multiple times) influenced participant behavior and likely raised a sense of suspicion or distrust amongst the workers who did not respond. Some of the respondents even expressed concern in their initial email responses, one going as far as saying:

> *I got a little concern[ed] about my privacy when going through [your email]. Can I know [what] you need the recordings for?*

Nevertheless, the thirteen workers we recruited provided us with sufficient data to demonstrate the possibility of our attack technique being successful and we did not attempt further solicitation of workers, nor did we try too hard to allay potential feelings of distrust. In real attacks, it is unlikely that the attacker would use a platform like MTurk, relying instead on more targeted platforms (with better support for subversive activities) to conduct the attack.

The artists were more responsive than the workers (all

[13] Only 2 workers tried this technique to fool our system, out of which 1 passed our filtering criteria, a plausible indication that malicious intent is scarce amongst MTurk users.

three candidates completed the confirmation task) plausibly because our engagement with them was less anonymous and more conversational in nature, which may have increased trust in the activity. Overall, 16 candidate mimics attempted our confirmation task—13 workers, 3 artists—which is more than the number of mimics used in any prior research on imitation attacks [11].

Measure	MTurk	Artists
#(participants solicited)	> 200	25
#(participants who made valid calls)	176	4
#(candidate imitators filtered)	38	3
#(candidates who replied to emails)	13	3
#(candidates who were successful)	6	3

Table 1: Summary of our filtering and confirmation outcomes. We don't have data on the exact number of MTurk workers solicited but based on the calls received, it is clear that there were more than 200 of them. Through our iterative filtering process, six of these were confirmed to be successful imitators at the end. Of the 25 solicited artists, 3 were confirmed as being successful.

In post-hoc interviews, none of the MTurk workers reported to have had any training in mimicry or drama in the past although four of them claimed to have practised casual mimicry in the company of friends and family. The artists were more experienced, but not significantly so, with one artist reporting not to have done stage performances ever and another reporting to have had extensive voice-over experience but none so in celebrity mimicry. Geographically, these individuals were dispersed across India with exactly half of them from the South and the remaining from northern India. Their ages ranged from 20 to 63 (median age of 26) and their personal incomes varied from 33 USD to 1100 USD per month (median income of 42 USD per month)[14].

Our confirmation procedure ran for a cumulative period of six weeks and we collected a total of 1060 speech samples from our candidate imitators during this period. Each candidate mimic called from a mobile phone but the phone model differed across candidates. We did not attempt to control for recording environment except for a general advice to call from a quiet room. In our analysis of the confirmation data below, we use the actual scores of candidate imitators against speaker models used by our SV system $\mathcal{A}$ and not just the binary outcome of its matching procedure. This is done purely for the sake of analysis and does not affect attack implementation; a real-world adversary may not have access to scores from the SV system if it is assumed to be black-box accessible only.

5.2.1 EER-based Evaluation

Overall, the performance of our candidate imitators declined in the confirmation stage but a majority of them continued to authenticate their associated targets to the system across sessions. For each candidate w, we computed his individual false acceptance rate (FAR)—the number of w's speech samples that could authenticate his target speaker to the system (at the EER threshold) divided by the total number of speech samples evaluated for him. The FARs for the nine leading candidates were observed to be consistently over 20% across sessions. For the remaining seven, we observed FARs of less than 20% in the first two sessions and for the most part, we did not engage them beyond the second session. Our analysis below focuses on the 9 leading candidates. We refer to the workers amongst these as $\mathsf{w}_1, \ldots, \mathsf{w}_6$ and the artists as $\mathsf{a}_1, \ldots, \mathsf{a}_3$.

Each of these candidates participated in 3 to 5 well-separated recording sessions (inter-session separation of at least a day) of at least 20 recordings each. Each participated in at least two contiguous sessions with individual FARs exceeding 0.3 and we continued recording until this was accomplished by each of them (going beyond the third session where required). The mean FARs for the candidates in their *last 3 sessions* are depicted in Table 2 (in the column labeled FAR). The total number of confirmation sessions conducted per candidate is denoted n. Out of the 9 candidates, only one, namely w_1, is a deliberate candidate and of the remaining, only two (marked with a superscript nat) were identified based on the closeness of their *natural* voice to that of the target. The remaining emerged candidates due to an observed closeness between their faked (mimicked) voice and their target's voice.

Source	Label	Type	n	FAR	modifier?
MTurk	w_1	deliberate	5	0.424	N
	w_2	emergent	5	0.683	Y^+
	w_3	emergent	4	0.417	Y^+
	w_4	emergent	3	0.417	N^+
	w_5	$\text{emergent}^{\mathsf{nat}}$	3	0.367	N
	w_6	$\text{emergent}^{\mathsf{nat}}$	3	0.333	N
Artists	a_1	emergent	3	0.567	Y^+
	a_2	emergent	3	0.383	Y^+
	a_3	emergent	3	0.533	Y^+

Table 2: Overall False Acceptance Rates of the nine leading candidate imitators.

All of these nine candidates were able to authenticate as their target speakers in at least 33% of their (last 60) recordings. This is significantly greater than the 2.3% FAR rate that the system was initially calibrated for and it suffices to launch online attacks on a real system. Averaged across all candidate imitators and all speech data used in the table, we computed an FAR of 45.8% and for the six MTurk workers, this figure stood at 44%. The finding is made all the more significant by the fact that we did not control for environmental and channel effects in the voice recordings; the imitators' handsets and speaking environment could have been very different from that of the speakers, which could have made it hard for them to match the speaker voices.

The target speakers associated with these candidates are not all unique: w_4 and a_3 share the same target and so do w_6 and a_1. The target for w_1 is shared by 5 other candidate imitators, although only w_1 reached the confirmation stage of our experiment (the rest did not respond to our invitations).

The last column in the table (labeled modifier?) depicts whether a majority of a imitator's mimicked speech samples were seen to match the target speaker's voice more closely than his natural voice—a likely indication that the imitator was making a significant effort to modify his voice to match the target. This metric was computed by first creating speaker models for all the nine candidates in the system and then invoking the matching procedure on the imitator

[14] We use a USD to INR conversion rate of 1:60 for these computations.

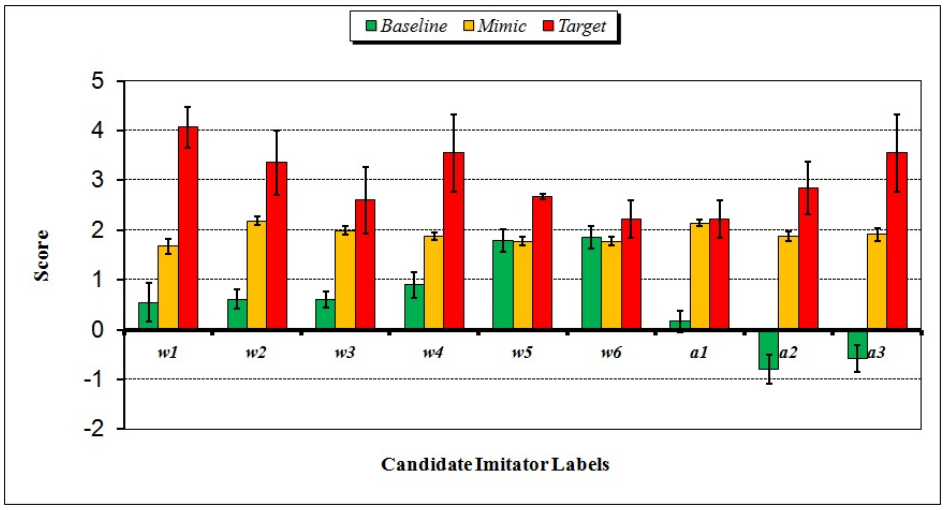

Figure 3: Comparison of candidate imitator scores with target self-scores: *Baseline* refers to the score of an imitator's natural speech samples against the target speaker model and *Mimic* is the score of his mimicked speech samples. *Target* is the target speaker's self-score. Error bars depict standard error of mean.

speech samples for both the target speaker model and the imitator speaker model. Two of the MTurk candidates and all of the artists were found to be modifiers by this definition. A superscript of + indicates candidates who reported to have had mimicry experience in the past, which was true for workers w_2, w_3, w_4 and all the artists. (This coincides almost perfectly with our modifier set.) Notice that the modifiers are also better performers on average: the mean FAR for the modifiers is 0.52, whereas that for the rest is 0.39. While mimicry skill or experience does not seem necessary for launching successful imitation attacks on SV systems (as demonstrated by the performance of non-modifiers in our set), it does seem to aid attack success.

In terms of inter-session variability in mimicry performance, our data does not reveal a consistent trend across imitators—some (in particular, w_1, w_2 and a_3) improved with time while others performed non-monotonically although per-session FARs remained consistently above 20%. Intra-session trends are also not monotonic and in particular, success likelihood did not necessarily increase across attempts within a session. These findings could be explained by two characteristics of our experiment: first, effects of learning could have been negated by factors like boredom and fatigue within sessions; and second, as we find below, match scores for most candidates are generally in the neighborhood of the EER threshold and could thus be more sensitive to inter- and intra-session changes than the target voices. Future work is needed to address these issues and in particular, to devise feedback mechanisms that have sustained effects on mimic performance. As indicated earlier, this is challenging for text-independent voice biometric systems because of the lack of temporality in feature creation in such systems.

Only one imitator (w_4) was asked to change his target speaker in the process of confirmation (because of greater closeness observed with that target during initial authentication attempts); the rest mimicked their initially-assigned targets. Two of the artists (namely, a_1 and a_2) attempted mimicking multiple speakers based on target suggestions provided by us but their FARs were smaller (less than 10%) for all such targets, an indication that *consistently* and *successfully* mimicking *multiple* targets is a difficult undertaking, even for people experienced in mimicry.

The gap between the nine successful candidates and the remaining seven who took part in the confirmation stage was striking. The latter provided a total of 279 speech samples across 1-2 sessions each but achieved an average FAR of only 4.9% for their respective target speakers. Individual FARs for these seven candidates ranged between 0 and 0.125 with a median of 0.044. These findings highlight the importance of including a confirmation step in the attack as a way to weed out serendipitous matches during candidate filtering.

5.2.2 Comparison with self-scores

Although EER-based analysis gives us some indication of candidate performance, by itself it does not present a complete picture of attacker capability: even by crossing the EER threshold repeatedly, an imitator could be far from the expected target score, something that an SV system could be programmed to detect. As such, it is also useful to compare the scores of these imitators against the *self-scores* of the target speakers i.e., the scores computed on their test samples against their speaker models.

For the comparison, we used test samples we initially collected from each speaker in S, half of which had deliberate background noise (sounds from a busy Indian street) included in the recording. We did not attempt to exaggerate the noise addition and even with its presence, all nine targets' scores were above the EER threshold, on average. Incorporating noisy test samples in the analysis reduces the challenge for the adversary, but also models a more realistic scenario: test samples for honest users are unlikely to be all clean in reality, but an attacker can control his test environmental conditions (e.g., avoid calling from the street). We compared three statistics: candidates' natural speech sample scores against target speaker model (*baseline*), candidates' mimicked speech sample scores against target speaker model (*mimic*) and the targets' self-scores (*target*). Results are shown in figure 3. We use unpaired, 2-tailed t-tests for

measuring significance in our statistical analyses below, with a threshold p value of 0.01.[15]

It is clear that most candidates are modifying their voice in order to mimic their target speaker, although the artists seem to be making more significant modifications. The results for artists a_2 and a_3 are particularly striking—even with natural voices whose match against the target have opposite polarity as the target's self-match, these artists were able to obtain scores that are in the proximity of the target self-scores. The MTurk workers, being relatively less skilled, are making limited modifications and seem to be relying on their inherent closeness with the target's voice. (This is consistent with prior literature on the ability of skilled imitators vs. ordinary humans in being able to modify their voices both in terms of spectral features and prosody [26].) In particular, for workers $\mathsf{w}_1, \mathsf{w}_5, \mathsf{w}_6$, the candidates claimed to be least experienced in mimicry, the difference between the natural speech scores and mimicked speech scores is statistically insignificant.

Second, even though most candidates are making noticeable jumps in moving from the baseline to the mimic conditions, their mean mimic scores never exceed their targets' self-scores, although the difference is statistically insignificant for 6 out of our 9 leading candidates, namely, workers $\mathsf{w}_2, \mathsf{w}_3, \mathsf{w}_6$ and all the artists.[16] Artist a_1, in particular, exceeds his target's mean self-score in more than 40% of his attempts. The performance of artists observed in our experiments surpasses that of mimicry specialists used in prior works like [10, 11, 15, 26], which did not deploy candidate filtering techniques like ours in selecting their mimics (instead relying purely on perceptual mimicry ability)[17]. Finally, we note that even though the workers' overall performance is poorer than that of the artists in our experiment, they compare favorably with that of the latter both in terms of EER-based measures (table 2) and in terms of the means of the raw scores (figure 3); workers $\mathsf{w}_2, \mathsf{w}_3$ and w_4, in fact, surpass a_2 on both these measures and also surpass artists used in prior works [10, 11, 26].

Overall, we learnt three key things from this part of our analysis: (a) most candidate imitators identified by our filtering techniques exhibit good mimicking capability in confirmation tests although the artists are more consistent than the MTurk workers; (b) in terms of absolute scores (EER-based evaluation), these candidates present a potent threat to the system but when viewed relative to target self-matches, they seem less competent; and (c) it is possible that *some* imitators picked from MTurk (like $\mathsf{w}_2, \mathsf{w}_3, \mathsf{w}_4$) can surpass more experienced ones (like a_2) as SV-impostors in absolute measures (which is a new finding relative to the literature) but their *overall* ability to imitate others seems less than that of the latter (which is consistent with the literature [26]).

[15] Although the *baseline* and *mimic* distributions are not strictly independent (same speaker for both), it is conceivable that the speaker applies independent techniques in generating them, which justifies the use of the unpaired test.

[16] In parallel experiments, we have also tested the performance of different forms of record-and-replay attacks on the same system and found them to be able to match, and often exceed, target self-scores for a large number of users.

[17] Most prior work except [11] does not use self-score comparisons, the way we do, which makes comparing against such works difficult. However, even based on EER-based evaluations, our results seem stronger.

6. DISCUSSION AND CONCLUSION

Prior work has contended that human attacks on biometric systems are possible only by people with skill or expertise to imitate others [2, 26]. Our work shows that this is not necessarily the case and it is the first that does so for voice biometrics. Even with a relatively small target set of fifty-three speakers, we were able to find ordinary, untrained people on an online crowdsourcing platform with the capability to impersonate (in absolute EER-based measures) some of these speakers to a well-studied SV scheme. We believe that the geographic and cultural spread of MTurk was critical in enabling us to reach this result: hand-picking candidate imitators from our vicinity may not have been as successful, at least in discovering a deliberate candidate like w_1 as we did in our experiment. Furthermore, the strategy of matching imitators to the right targets based on the formers' faked voices rather than their natural ones seems to have helped—most of the emergent candidates in our experiment were matched to their targets via this approach.

Even for people with experience (artists or professionals), our work provides an improvement over what has been shown in the literature till now. Our artist imitators were recruited after deliberately contacting 25 individuals over phone and carefully mapping them to "close" targets; in the process, we may have ended up selecting some of the more intrinsically-motivated individuals than others did in their mimicry experiments [10, 11, 26]. Our mimicry artists not only exhibited good FARs with respect to their target speakers, but they did excellently in terms of matching target self-scores as well, which is better than what prior work reports (e.g., the professional imitator used in [11] managed an FAR of only about 10% and did not match any of his target's self-scores).

The observation that MTurk workers could not surpass their targets' mean self-scores on average (even after diluting the latter with ambient noise) is our main negative finding from the perspective of attackers. A potential defense against MTurk-based imitators could thus simply be to require every user's test sample to match his or her prior test sample scores in expectation (or do this for the more vulnerable users identified from the analysis). However, not all systems may be in a position to implement this defense for their users, especially in order to be able to handle unexpected session variabilities. To the extent that there remain SV systems with EER-based decision procedures in deployment, the threat from crowdsourcing-based imitation attacks to these systems will also remain.

The overall performance of MTurk workers was worse than of the artists in our experiments and the fraction of successful workers was also considerably smaller (only 6 out of the 176 valid callers i.e. about 3%, succeeded in the EER-based evaluation). But these findings should be viewed in the light of the fact that MTurk workers are easier to find and cheaper to recruit than typical artists and professional imitators. The lower success rate of the workers could potentially be compensated for by expanding the sample space of crowd workers (e.g., assuming a rate of 3% for finding successful mimics, adjust the sample space size based on the number of successful mimics required[18]). Adjusting the

[18] We caution that the sample space and the target set used in our experiments are small; more experiments are needed to confirm the rate of finding successful mimics on MTurk.

sample space for professional mimics is harder because they are difficult to find in the first place.

The fact that six of our workers' performance came so close to that of the artists, and exceeded the latter in a few cases, encourages the continued usage of online crowdsourcing for large-scale impostor search. Future work is needed to understand how best to set incentives for online crowd workers so as to attract more of them to complete such tasks without compromising on work quality. Future work is also needed to understand how varying authentication criteria (like requiring phonetically-richer test samples from users) or the biometric modeling process can affect the performance of candidate imitators discovered by our method. We believe that increasing the phonetic complexity of either the test samples or training samples (or both) is likely to increase resistance to mimicry attacks, but note that this also affects usability for honest users. Achieving imitation-resistance while maintaining system usability is where the challenge lies.

6.1 Implications for Real Systems

Although we have demonstrated the possibility of crowdsourcing-based attacks on SV schemes, the *feasibility* of these attacks and the scale at which they can be mounted on real systems is still unresolved. Will the most capable imitators discovered using the method be able to successfully impersonate their target as they converse with a real system that implements liveness detection? Is it likely that "many" imitators can be found to do this? And, most importantly, are there crowdsourcing platforms where it is possible to find sufficiently many people with the motivation to help break a real system? Recent work [24] shows that for some types of malicious objectives like online vandalism and fake account creations, systematic use of crowdsourcing platforms has already evolved but to the best of our knowledge, such practice is not yet prevalent for attacking biometric (or other forms of) authentication yet. The attack analysis presented in this paper suggests that when this practice shapes up, the resulting attacks are also likely to be quite successful.

While it is important to extend our study to understand attack feasibility, we believe that the most immediate implication of our work to real systems is that it gives system developers a new tool to *analyze* the security of their systems with. Using our method, they can simulate imitation attacks on their systems more easily and, arguably, more cheaply than they could by hiring mimicry artists, which we experienced to be an excruciating and slow process in our study. It also helps them get a better perspective on which speakers in their dataset are more vulnerable to imitation attacks (the so-called "lambs" in the system [9]) than they would by trying such attacks with a handful of professional mimics or considering within-dataset impostors only. For example, in our own attack implementation we observed one target speaker emerge as the closest match for six different candidate imitators during the candidate filtering process (Sec. 5.1). Even though we could confirm this closeness with only one candidate (our deliberate candidate, w_1), it is plausible that more of the others would also have proven as consistent imitators with respect to that speaker, had we recruited them for confirmation. Such vulnerability assessment of individual speakers is impossible if one restricts the analysis to speakers within the target speaker-set; in our case, this "lamb" speaker was found *not* to be a consistent closest, or even second-closest, match for any of the speakers in S. Assessing speakers in this manner could inform the customization of system parameters for preventing imitation attacks on individuals in the dataset.

Other biometric forms could potentially also benefit from crowdsourcing the search for impostors, if not now, then at least in the near future. As more people in the developing world go online and join the crowdsourcing workforce, and as sensing devices like fingerprint scanners and cameras become more ubiquitous, new possibilities for large-scale, cheap and efficient biometric data collection will open up. Such data collection and subsequent analysis can lead to new insights on system vulnerabilities as we discovered in the case of voice in our experiment. Of course, the question of feasibly translating such data collection into real attacks will still remain, but independent of it, existing systems can benefit from the search for impostors in crowdsourced data and prepare better for attacks that may occur in the future.

7. REFERENCES

[1] G. Ashour and I. Gath. Characterization of speech during imitation. In *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, pages 1187–1190, 1999.

[2] L. Ballard, F. Monrose, and D. Lopresti. Biometric authentication revisited: Understanding the impact of wolves in sheep's clothing. In *Proc. of Usenix Security*, pages 29–41, 2006.

[3] BBC. Hello, is that really you? `http://www.bbc.com/news/business-24898367`, 2013.

[4] BiometricUpdate.Com. Coursera looks to verify online student identity with photo, keystroke dynamics. `http://www.biometricupdate.com/201301/coursera-looks-to-verify-online-student-identity-with-photo-keystroke-dynamics`, 2013.

[5] J. Bonastre, F. Wils, and S. Meignier. Alize: A Free Toolkit for Speaker Recognition. In *Proc. of ICASSP*, 2005.

[6] J.-F. Bonastre, D. Matrouf, and C. Fredouille. Artificial impostor voice transformation effects on false acceptance rates. In *Proc. of Interspeech*, pages 2053–2056, 2007.

[7] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp. Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading. *Transaction on Image Processing*, 15(10), 2006.

[8] CIO Journal. Banks Eye Voice Biometrics to Verify Customers. `http://blogs.wsj.com/cio/2013/05/09/banks-eye-voice-biometrics-to-verify-customers/`, 2013.

[9] G. R. Doddington, W. Liggett, A. F. Martin, M. Przybocki, and D. A. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *Proceedings of the Fifth International Conference on Spoken Language Processing*, 1998.

[10] M. Farrus, M. Wagner, D. Erro, and J. Hernando. Automatic speaker recognition as a measurement of voice imitation and conversion. *The International Journal of Speech, Language and the Law*, 17(1):119.

[11] R. G. Hautamaki, T. Kinnunen, V. Hautamaki,

T. Leino, and A.-M. Laukkanen. I-vectors meet imitators: on vulnerability of speaker verification systems against voice mimicry. In *Proc. of Interspeech*, 2013.

[12] M. Hebert. Text-dependent speaker recognition. In *Springer handbook of speech processing (Heidelberg, 2008)*, pages 743–762. Springer Verlag, 2008.

[13] T. Kinnunen and H. Li. An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. *Speech Communication*, 52(1):52.

[14] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li. Vulnerability of speaker verification systems against voice conversion spoofing attacks: the case of telephone speech. In *Proc. of ICASSP*, 2012.

[15] Y. W. Lau, M. Wagner, and D. Tran. Vulnerability of Speaker Verification to Voice Mimicking. In *Proc. Int. Symp on Intelligent Multimedia, Video and Speech Processing (ISIMP '04)*, pages 145–148, 2004.

[16] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. Impact of Artificial Gummy Fingers on Fingerprint Systems. In *Proc. of SPIE, Optical Security and Counterfeit Deterrence Techniques IV*, volume 4677, pages 275–289, 2002.

[17] T. C. Meng, P. Gupta, and D. Gao. I can be You: Questioning the use of Keystroke Dynamics as Biometrics. In *Proc. of NDSS*, 2013.

[18] Nuance Communications. Private communication, 2013.

[19] Reuters. German group claims to have hacked Apple iPhone fingerprint scanners. `http://www.reuters.com/article/2013/09/23/us-iphone-hackers-idUSBRE98M01X20130923`, 2013.

[20] D. Reynolds, T. Quatieri, and R. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, Jan. 2000.

[21] J. Ross, I. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the Crowdworkers? Worker Demographics in Amazon Mechanical Turk. In *CHI Extended Abstracts 2010*, pages 2863–2872, 2010.

[22] B. Toth. Biometric Liveness Detection. *Information Security Bulletin*, 10, 2005.

[23] Uniphore. `http://uniphore.com/`, 2008.

[24] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and Turf: Crowdturfing for Fun and Profit. In *Proc. of WWW*, 2012.

[25] ZDNet. Indian banks explore voice biometrics for added security. `http://www.zdnet.com/in/india-banks-explore-voice-biometrics-for-added-security-7000018883/`, 2013.

[26] E. Zetterholm. Detection of speaker characteristics using voice imitation. In *Speaker Classification II*, LNAI 4441, pages 192–205, 2007.

APPENDIX

A. INVITATION EMAIL SENT TO MTURK WORKERS

Given below is the email template we used to invite MTurk workers identified as candidate imitators to participate in the confirmation task:

Dear MTurk worker ——,

Based on your performance on our HIT, you have been selected to do a bonus task for us in which the minimum pay is 5 USD. In this bonus task you will be required to make at least 20 voice recordings on our system.

Kindly email —— if you are interested in doing this bonus task. Please specify your name and MTurk ID in the email. Based on your response, we will send you more details about the bonus task.

Looking forward to hearing from you!