All Your Voices Are Belong to Us: Stealing Voices to Fool Humans and Machines

Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena *

University of Alabama at Birmingham, AL, USA
{dibya,maliheh}@uab.edu, saxena@cis.uab.edu

Abstract. In this paper, we study voice impersonation attacks to defeat humans and machines. Equipped with the current advancement in automated speech synthesis, our attacker can build a very close model of a victim's voice after learning only a *very limited* number of samples in the victim's voice (e.g., mined through the Internet, or recorded via physical proximity). Specifically, the attacker uses *voice morphing* techniques to transform its voice – speaking any arbitrary message – into the victim's voice. We examine the aftermaths of such a voice impersonation capability against two important applications and contexts: (1) impersonating the victim in a *voice-based user authentication* system, and (2) mimicking the victim in *arbitrary speech* contexts (e.g., posting fake samples on the Internet or leaving fake voice messages).

We develop our voice impersonation attacks using an off-the-shelf voice morphing tool, and evaluate their feasibility against state-of-the-art *automated* speaker verification algorithms (application 1) as well as *human* verification (application 2). Our results show that the automated systems are largely ineffective to our attacks. The average rates for rejecting fake voices were under 10-20% for most victims. Even human verification is vulnerable to our attacks. Based on two online studies with about 100 users, we found that only about an average 50% of the times people rejected the morphed voice samples of two *celebrities* as well as *briefly familiar users*.

1 Introduction

A person's voice is one of the most fundamental attributes that enables communication with others in physical proximity, or at remote locations using phones or radios, and the Internet using digital media. However, unbeknownst to them, people often leave traces of their voices in many different scenarios and contexts. To name a few, people talk out loud while socializing in cafés or restaurants, teaching, giving public presentations or interviews, making/receiving known and, sometimes unknown, phone calls, posting their voice snippets or audio(visual) clips on social networking sites like Facebook or YouTube, sending voice cards to their loved ones [11], or even donating their voices to help those with vocal impairments [14]. In other words, it is relatively easy for someone, potentially with malicious intentions, to "record" a person's voice by being in close physical proximity of the speaker (using, for example, a mobile phone), by social engineering trickeries such as making a spam call, by searching and mining for audiovisual clips online, or even by compromising servers in the cloud that store such audio information. The more popular a person is (e.g., a celebrity or a famous academician), the easier it is to obtain his/her voice samples.

* The first two authors are equally contributing

We study the implications of such a commonplace leakage of people's voice snippets. Said differently, we investigate how an attacker, in possession of a certain number of audio samples in a victim's voice, could compromise the victim's security, safety, and privacy. Given the current advancement in automated speech synthesis, an attacker can build a very close model of a victim's voice after learning only a *very limited* number of previously eavesdropped sample(s) in the victim's voice. Specifically, *voice morphing* techniques can be used to transform the attacker's voice – speaking any arbitrary message – into the victim's voice based on this model. As a result, just a *few minutes worth of audio* in a victim's voice would lead to the *cloning of the victim's voice itself*.

We show that the consequences of imitating one's voice can be grave. Since voice is regarded as a unique characteristic of a person, it forms the basis of the authentication of the person. If voice could be imitated, it would compromise the authentication functionality itself, performed implicitly by a human in human-to-human communications, or explicitly by a machine in human-to-machine interactions. As our case study in this paper, we investigate the aftermaths of stealing voices in two important applications and contexts that rely upon voices as an authentication primitive. The *first application* is a voice-based biometric or speaker verification system that uses the potentially unique features of an individual's voice to authenticate that individual. Voice biometrics is the new buzzword among banks and credit card companies. Many banks and credit card companies are striving for giving their users a hassle-free experience in using their services in terms of accessing their accounts using voice biometrics [13, 15, 18, 22, 29, 31]. The technology has now also been deployed on smartphones as a replacement to traditional PIN locks, and is being used in many government organizations for building access control. Voice biometrics is based on the assumption that each person has a unique voice that depends not only on his or her physiological features of vocal cords but also on their entire body shapes, and on the way sound is formed and articulated. Once the attacker defeats voice biometrics using fake voices, he would gain unfettered access to the system (device or service) employing the authentication functionality.

Our *second application*, naturally, is human communications. If an attacker can imitate a victim's voice, the security of (remote) arbitrary conversations could be compromised. The attacker could make the morphing system speak literally anything that the attacker wants to, in victim's tone and style of speaking, and can launch an attack that can harm victim's reputation, his security/safety and the security/safety of people around the victim. For instance, the attacker could post the morphed voice samples on the Internet, leave fake voice messages to the victim's contacts, potentially create fake audio evidence in the court, and even impersonate the victim in a real-time phone conversations with someone the victim knows. The possibilities are endless. Such arbitrary conversations are usually (implicitly) verified by humans.

Our Contributions: In this paper, we study the security threat associated with stealing someone's voice (Figure 1). We develop our voice impersonation attacks using an off-the-shelf voice morphing engine, and comprehensively evaluate their feasibility against state-of-the-art *automated* speaker verification algorithms (application 1 above) as well as *manual* verification (application 2). Our results show that the automated systems are largely ineffective to our voice impersonation attacks. The average rates for rejecting fake voices were under 10-20% for most of our victims. In addition, even human veri-

fication is vulnerable to our attacks. Based on an online study with 65 users, we found that only about an average 50% of the times people rejected the morphed voice samples of two *celebrities*, while, as a baseline, they rejected *different speakers' voices* about 98% of the times, and that 60-70% participants rated the morphed samples as being similar to original voices. We extended the same study for *briefly familiar voices* with 32 online participants, the results being slightly better than the previous study (rejection rates decrease and ambiguity of speaker verification increases).

Our work highlights a real threat of practical signficance because obtaining audio samples can be very easy both in the physical and digital worlds, and the implications of our attacks are very serious. While it may seem very difficult to prevent "voice hacking," our work may help raise people's awareness to these attacks and motivate them to be careful while sharing and posting their audio-visuals online.

2 Background and Related Work

Voice Conversion: It has always been a challenge to get a machine to talk in a human's voice. Voice synthesis (artificial creation of human voice) has a growing number of applications most dominant one is text to speech systems. There are several instances of such voice synthesizers, whose qualities are judged based on their naturalness (similarity to human voice). Some of the recent synthesizers [2, 5, 10] significantly improved quality of the speech by reducing the robotic sound that was unavoidable in earlier synthesizer. However, still the synthesized speech is distinguishable from a human voice. Besides, such systems require a huge amount of data to learn phonemes.

The other technique to create a voice is voice morphing (also referred to as voice conversion and voice transformation). Voice morphing modifies a source speaker's voice to sound like a target speaker by mapping between spectral features of their voice. Similar to the voice synthesizers the major application of voice morphing is TTS that can speak in any desired voice. Usually such techniques require smaller amounts of training data and sound more natural and fluent compared to voice synthesizers [6]. Due to these advantageous properties, voice morphing becomes an excellent tool to attack someone's voice as studied in our paper.

We employed the CMU Festvox voice converter [6] (reviewed in Section 4) to attack machine-based and human-based speaker verification. We used Mel-Cepstral Distortion (MCD)[1] to measure the performance of conversion engine for different size of training dataset. The smaller the MCD, the better the quality of the conversion. MCD values between 5-8 dB are generally considered acceptable for voice conversions [9]. As a crucial component of our attacks, we found that the conversion quality is very good (within the desired range of 5-8 dB) even with very small amount of training data. Our MCD analysis is reported in Sections 5.1 and 6.3.

Machine-based Speaker Verification: Speaker verification is the biometric task of authenticating a claimed identity by means of analyzing a spoken sample of the claimant's voice. It is a 2-class problem in which the claimant must be recognized as the true

[1] MCD is a metric used to measure the similarity of the converted voice and the original voice by calculating the different between the feature vectors of the original and converted voice [26, 32, 33]

speaker or as an impostor [35]. To recognize a known target speaker, a speaker verification system goes through a prior speaker enrollment phase. In the speaker enrollment phase, the system creates a target model of a speaker from his/her speech samples so that they can be verified during the test phase in future.

A speaker verification system extracts certain spectral or prosodic features from a speech signal to enroll the model of the target speaker. After extracting the features from a speech signal, model enrollment or "voice print" generation of the target speaker is done using different modeling techniques.

With the emergence of advanced speech synthesis and voice conversion techniques, the automatic speaker verification systems may be at risk. De Leon et al. have studied the vulnerabilities of advanced speaker verification systems to synthetic speech [23–25], and proposed possible defenses for such attacks. In [16], the authors have demonstrated the vulnerabilities of speaker verification systems against artificial signals. The authors of [44] have studied the vulnerabilities of text-independent speaker verification systems against voice conversion based on telephonic speech.

In our paper, we pursue a detailed analysis of the vulnerabilities of a speaker verification system employing two advanced algorithms against voice conversion. Although some of the prior papers tested the same set of speaker verification algorithms we are testing in our paper, they did not evaluate the Festvox conversion system, which claims to require only few sentences for training [6]. Noticeably, a key difference between our work and previous studies lies in the number/length and type of samples required to build a good voice conversion model. We use very less amount of training samples (e.g., 50-100 sentences of length 5 seconds each) for voice conversion collected using unprofessional voice recording devices such as laptops and smartphones. Such short-size audio samples giving rise to a victim's voice, sets a fundamental premise of how easily a person's voice can be attacked or misused. While the prior papers do not seem to clearly specify the sizes of their voice conversion training data sets, they employ spectral conversion approaches that typically require a large amount of high-quality training data [28, 42].

Human-based Speaker Verification: Manual speech perception and recognition is a complex task, which depends on many parameters such as length/number of different samples, samples from familiar vs. famous people, and combinations thereof [38]. There exists a considerable volume of literature on how speech is recognized [20, 21, 38]. Linguistics researches show that the shorter the sentence, the more difficult it is to identify the source [27]. Based on the study conducted by Shirvanian et al. [39], it appears that the task of establishing the identity of a speaker is challenging for human users, especially in the context of short random strings (numbers or phrases). In our paper, we study the capability of human users in recognizing the speaker for an *arbitrary speech* of famous celebrities and briefly familiar speakers.

3 Our Attacks on Human Voices

3.1 Overview

In this paper, we study the attacks against human-based and machine-based speaker verification. Our attack system consists of three phases (visualized in Figure 1). The *first phase* involves the collection of voice samples $O_T = (t_1, t_2, \cdots, t_n)$, previously

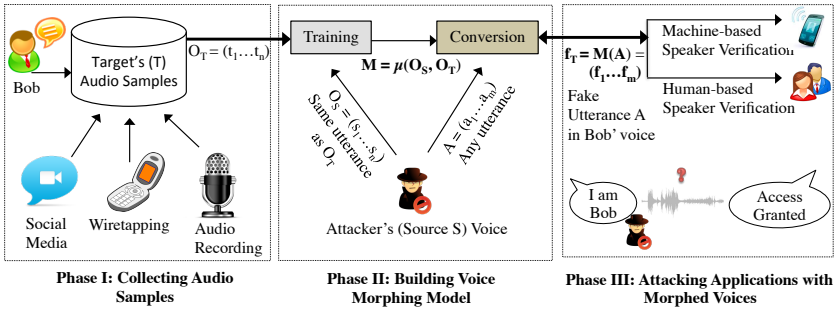


Fig. 1: An overview of our attack system

spoken by the target victim. At this point, the audio (content) privacy may have been compromised as the victim gives away (willingly or unwillingly) his/her audio samples to the attacker. The *second phase* of our attack focuses on the creation of the victim's voice based on the audio samples collected in the first phase. The attacker (source) first speaks the same sentences $O_S = (s_1, s_2, \cdots, s_n)$ the victim (target) has spoken in the recorded audio, and then feeds O_S and O_T to the morphing engine to create a model $M = \mu(O_S, O_T)$ of the victim's voice. At this point, the attacker has at its disposal essentially the voice of the victim. The *third phase* involves the use of this voice imitation capability to compromise any application or context that utilizes the victim's voice. The target applications that we study in this paper are: machine-based and human-based speaker verification systems. The attacker can speak any new arbitrary sentence $A = (a_1, a_2, \cdots, a_m)$, as required by the attacked application, which the model built in the second phase will now convert into the victim's voice as $f_T = M(A) = (f_1, f_2, \cdots, f_m)$. The morphed samples will then be fed-back to the speaker verification systems (to authenticate the morphed voice as the target victim's voice), and to people (to attack them by fooling them into believing that the morphed attacker's voice is the voice of the benign victim). The third phase of our attack system serves to demonstrate the aftermaths of the breach of voice security.

3.2 Threat Model

In our threat model, an attacker can collect a few of the victim's audio samples, for example, by recording the victim using a mobile audio recording device with or without the knowledge of the victim, or mining the previously posted samples on the web. As mentioned earlier, these samples are then used to train a morphing engine. In the training procedure, the attacker may use his own voice or could recruit other users (possibly those who can mimic the victim's voice very closely). Thus, the attacker has the ability to employ means to achieve favorable conditions for voice conversion so as to achieve the highest quality morphed samples.

Equipped with this voice morphing capability, the attacker then attempts to defeat the machine-based and human-based speaker verification systems/contexts. When attacking a machine-based speaker verification system, the attacker simply sends the morphed voices to impersonate himself as the legitimate user. In this case, we clearly assume that the attacker has permanent or temporary physical access to the terminal or device deploying voice authentication (e.g., a stolen mobile phone, a desktop left unattended during lunch-time, or a public ATM).

The attacker can defeat human-based speaker verification in many ways. Clearly in this context, faking face-to-face conversation would not be possible with voice morphing. However, the attacker can be remote and make spoofed phone calls, or leave voice messages impersonating the victim. He may even create real-time fake communication with a party with the help of a human attacker who provides meaningful conversations, which the morphing engine converts to the victim's voice on-the-fly. The attacker can also post the victim's morphed samples online on the public sites or disseminate via social networking sites, for example.

3.3 Attacking Machine-based Speaker Verification

In this paper, we systematically test the advanced speaker verification algorithms that can be used for the purpose of user authentication, in the following scenarios:

Different Speaker Attack: This attack refers to the scenario in which, the speaker verification system trained with the voice of speaker A is attacked with another human speaker B's voice samples. If the system fails to detect this attack, then the system is not good enough to be used for the purpose of speaker verification. This is conceivably the simplest and the most naive attack that can be performed against an automatic speaker verification system. So, this attack might be used as a baseline to measure the security performance of the target speaker verification system.

Conversion Attack: This attack scenario refers to the one in which the speaker verification system is attacked by the morphed samples of an impostor replacing the legitimate user's samples. Such an attacker might have the capability to attack a speaker-verification system that gives a random challenge each time a victim user tries to login or authenticate to the system.

3.4 Attacking Human-based Speaker Verification

In this scenario, the attacker would simply create arbitrary morphed speech in the victim's voice and use it to communicate with others remotely. As mentioned earlier, some of the real life applications of this attack might include leaving fake voice-mails in the victim's voice to harm victim's family or friends, or broadcasting a morphed voice of a celebrity victim to defame him/her. While the attack itself is relatively straight-forward in many cases, the key aspect is whether the "human verifier" would fall prey to it or not. This is the primary aspect we study in our work via two user studies. Similar to our study on machine-based speaker verification, we evaluate the performance of the conversion attack contrasted with the different speaker attack as a baseline against human-based speaker verification.

4 Tools and Systems

Festvox Voice Conversion System: Voice conversion (as reviewed in Section 2) is an emerging technique to morph voices. For implementing our attacks, we have used Festvox [6], a speech conversion system developed at Carnegie Mellon University.

Festvox employs acoustic-to-articulatory inversion mapping that determines the positions of speech articulators of a speaker from the speech using some statistical models. Toda et al. proposed a method of acoustic-to-articulatory inversion mapping based on Gaussian Mixture Model in [41] that is independent of the phonetic information of the speech. The next phase in this system is spectral conversion between speakers for

transforming the source speaker's to the target speaker's voice. The authors developed a spectral conversion technique [42], in which they have used maximum likelihood based estimation taking into account the converted parameter for each utterance. The evaluation results of this unique spectral conversion technique show that this technique has fared better than the conventional spectral conversion techniques [42]. For our experiment, we fed Festvox with recordings of some prompts spoken by the source (attacker) and the target (victim) speakers. Once the system is trained, any given arbitrary recording from the source speaker can generate the corresponding speech in the target's voice.

Bob Spear Speaker Verification System: In our experiment, we have used Spear verification toolbox developed by Khoury et al. [30] The Spear system is a open source speaker verification tools that has been evaluated with standard datasets like Voxforge [12], MOBIO [7] and NIST SRE [8]. Also, it represents the state-of-the-art in speaker verification systems having implemented the current well-known speaker verification algorithms, which makes it a representative system to evaluate our attack.

The input to this system, a set of clips spoken by a number of speakers, is split into 3 sets namely: *training set*, *development set* (Dev set) and *evaluation set* (Eval set). The training set is used for background modeling. The development and evaluation sets are further divided into two subsets, namely, Enroll set (Dev.Enroll, Eval.Enroll) and Test set (Dev.Test, Eval.Test). Speaker modeling can be done using any one of the given modeling techniques, namely, Universal Background Modeling in Gaussian Mixture Model (UBM-GMM) [37] and Inter-Session Variability (ISV) [43].

UBM-GMM is a modeling technique that uses the spectral features and then computes a log-likelihood of the Gaussian Mixture Models for background modeling and speaker verification [19,34,36]. ISV is an improvement to UBM-GMM, where a speaker's variability due to age, surroundings, emotional state, etc., are compensated for, and it gives better performance for the same user in different scenarios [40, 43].

After the modeling phase, the system is then tuned and tested respectively using the Dev.Test and Eval.Test sets from Development and Evaluation sets. All the audio files in the Dev.Test and Eval.Test sets are compared with each of the speaker models for development and evaluation sets, respectively, and each file is given a similarity score with respect to each speaker in the corresponding set. The scores of the Dev.Test files are used to set a threshold value. The scores of the Eval.Test set are then normalized and compared with this threshold, depending on which each file is assigned to a speaker model. If the audio file actually belong to the speaker to whom it got assigned, then the verification is successful otherwise the verification is not successful.

5 Experiments: Attacking Machine-based Speaker Verification

We now present the experiments that we conducted to attack the well-known speaker verification algorithms using voice conversion techniques.

5.1 Setup

Datasets: We used MOBIO and Voxforge datasets, two open source speech databases that are widely used for testing different speech recognition tools. Voxforge is a much more standard dataset in terms of the quality and the length of the speech compared to MOBIO. Voxforge samples are better quality samples of about 5 secs each while

MOBIO dataset is recorded using laptop microphones and also the length of the speech samples varies from 7 to 30 secs. The reason behind choosing these two datasets was to test our attacks against both standard and sub-standard audio samples. We have chosen a set of 28 male speakers from Voxforge, and 152 (99 male and 53 female) speakers from the MOBIO. For the purpose of the experiment, this speaker set is divided into 3 subsets. These three subsets are used separately for background modeling (Train set), development (Dev set) and evaluation (Eval set) of the toolkit. The Dev and Eval sets contain both labeled and unlabeled voice samples. The labeled samples are used for target speaker modeling while the unlabeled samples are used for testing the system.

For the Voxforge dataset, the *development set* (Dev.Test) contains 30 unlabeled voice samples for each of the 10 speakers, i.e., a total of 300 voice samples. In contrast, for the MOBIO dataset, the Dev.Test subset contains 105 unlabeled samples of 24 male and 18 female speakers. The samples in the Dev.Test set are used for tuning the parameters of the system such that the system performs well on the evaluation set. The MOBIO dataset contains both male and female speakers and are modelled separately in two separate systems. Since we are using the speaker recognition tool specifically in a speaker verification scenario, our *evaluation* (Eval) set always contains a single speaker. For Voxforge, we test for 8 (male) speakers, and for MOBIO, we test for 38 male and 20 female speakers.

Metrics Used: The performance of a speaker verification system is evaluated based on the *False Rejection Rates (FRR)* and *False Acceptance Rates (FAR)*. A *benign setting* is defined as a scenario in which, the test samples are all genuine samples. That is, the samples fed to the system are spoken by the original speaker (the one whose samples were used during the training phase). If the system accepts a given test sample, then the system was successful in recognizing the speaker, while a rejection means that the system has wrongly rejected a genuine sample, and this is counted as a *false reject*.

An *attack setting* is defined as a scenario in which, the test samples are fake or morphed. That is, these samples are not spoken by the original legitimate speaker, but are either spoken by some other speaker (another user) or generated using voice conversion. For simulating an attack setting, we replaced the genuine test samples with our fake samples in the Eval.Test set. So, the success of our attacks is directly proportional to the number of accepts, i.e., *false accepts*, by the system.

Different Speaker Attack Setup: For testing Voxforge dataset in this scenario, we swapped the voice samples of the original speakers with that of 4 CMU Arctic speakers [4] that have spoken the same samples as the Voxforge speakers, and tested the performance of the system. For testing with the MOBIO dataset, we replaced each speaker with all the other speakers in the Test set to see if the system could determine that the original speaker has been swapped. As discussed in Section 3, this attack is a rather naive attack, and serves as a baseline for our conversion-based attacks.

Conversion Attack Setup: In this attack scenario, we tested how robust the Spear system is to voice conversion. For implementing this attack, we changed the genuine test samples with converted ones. The voice conversion was done by training the Festvox conversion system with a set of samples spoken by both an attacker and the victim speakers. In case of Voxforge, one CMU Arctic [4] speaker posed as attackers and the 8 speakers in the Test set were the victims. For the MOBIO dataset, we chose 6 male and

3 female speakers in the Test set as attackers, and the remaining 32 male and 17 female speakers were the victims.

In case of the Voxforge dataset, we used 100 samples of 5 seconds each (so approximately 8 minutes speech data), to train the conversion system. In the MOBIO dataset, the speakers have independently recorded free speech in response to some questions asked to them. However, there were some specific common text that all the speakers have recorded. We used 12 such samples of about 30 secs each (so approximately 6 minutes of speech data) to train the conversion system. The converted voices thus generated were swapped with the genuine samples of the victim test speakers.

The MCD value after conversion in case of MOBIO speakers was about 4.58 dB (for females) and about 4.9 dB (for males) for 12 training samples (of average length 30 secs), which decreased by 0.16 dB (for females) and by 0.1 dB (for males) for about 100 training samples (of 15-30 secs length). In case of Voxforge, the MCD values were on average 5.68 dB, 5.59 dB, 5.56 dB for 50, 100, 125 training samples (of average length 5secs each) respectively. The negligible improvement in MCD of about 3% for MOBIO females, about 2% for MOBIO males, about 0.53% for Voxforge speakers led us to use 12 training samples for MOBIO and 100 training samples for Voxforge. Thus, its confirmed that voice conversion works well with only a small training dataset (a fundamental premise of our attack).

5.2 Results

Benign Setting Results: This experiment was done to set the baseline for the performance of the system being studied. The original clips of 8 Voxforge speakers, 38 male and 20 female MOBIO speakers, were used to evaluate the system. This test was done for both the algorithms: UBM-GMM and ISV. The results are summarized in the 2nd, 5th and 6th columns of Table 1. The results show that the rate of rejection of genuine (original speaker) samples (i.e., FRRs) is pretty low, less than 2% in case of Voxforge speakers, and in the range of around 7%-11% in case of MOBIO speakers. The MOBIO speakers, both male and female, have a standard deviation of more than 10% in case of UBM-GMM and that of about 8%-9% for ISV. The variation in the quality of speech across different speakers in the MOBIO dataset might be the reason behind this result.

Different Speaker Attack Results: The results for this attack is given in the 3rd, 7th and 8th columns of Table 1. From the results, we can see that the FAR (success rate of the attack) is less than 1% for the Voxforge speakers, around 10% for the male MOBIO speakers and around 16% for the female MOBIO speakers. Both UBM-GMM and ISV algorithms seem to perform similarly in case of this attack. However, the FAR of female speakers, being higher, can be attributed to the level of similarity of their voices in the MOBIO dataset. The acceptance rate is significantly low for both the datasets, which proves that Spear is robust against the naive different speaker attack, and can successfully detect, with at least 94% accuracy (for Voxforge) and with at least 84% accuracy (for MOBIO), that the speaker has been changed. This makes Spear a worthwhile system to challenge with respect to more sophisticated attacks.

Conversion Attack Results: The results of this attack are shown in the 4th, 9th and 10th columns of Table 1. The FAR in this case is above 98% for Voxforge, about 70-85% for male MOBIO speakers and about 60% for female MOBIO speakers.

Table 1: Performance of machine-based speaker verification system against our attacks. Reported numbers represent error rates *mean (standard deviation)*.

	Voxforge Dataset			*MOBIO Dataset*					
Algorithm	**Original Speaker (FRR)**	**Different Speaker Attack (FAR)**	**Conversion Attack (FAR)**	**Original Speaker (FRR)**		**Different Speaker Attack (FAR)**		**Conversion Attack (FAR)**	
				Male	**Female**	**Male**	**Female**	**Male**	**Female**
UBM-GMM	1.25% (2.48%)	0.21% (0.60%)	98.54% (2.07%)	11.32% (15.31)	11.71% (11.95%)	11.78% (7.04%)	16.70% (7.22%)	86.60% (20.16%)	64.42% (36.73%)
ISV	1.67% (1.78%)	0.73% (1.50%)	98.29% (2.61%)	9.12% (9.93%)	7.85% (7.94%)	10.79% (8.00%)	16.10% (9.52%)	73.54% (28.08%)	62.24% (37.04%)

However, the standard deviation corresponding to the speakers in the MOBIO dataset seems to be pretty high (about 28% in male and 36% in females). Hence, we analyze the distribution of the FAR values across all the Test users in the MOBIO dataset (Figure 3 of Appendix C). For MOBIO male speakers, we see that for 60% of speakers with UBM-GMM and more than 30% of speakers with ISV have 90% FAR. Overall, about 88% (for UBM-GMM) and about 85% (for ISV) of the male speakers have more than 50% FAR. For female speakers, about 52% (for UBM-GMM) and about 47% (for ISV) speakers have above 90% success rate. Overall, around 70% (for UBM-GMM) and 65% (for ISV) of the speakers have above 50% success rate. Thus, it is fair to say that the tested algorithms failed significantly against our voice conversion attacks.

Conversion Attack vs. Different Speaker Attack: We compared the mean FAR of the conversion attack with the mean FAR of the different speaker attack using the Wilcoxon Signed-Rank test, and found that the difference is statistically significant[2] with a p-value $= 0$ (for Males in case of both the algorithms), with a p-value $= 0.0015$ (for Females with UBM-GMM), with a p-value $= 0.0019$ (for Females with ISV) for MOBIO, and with a p-value $= 0.0004$ for Voxforge in case of both the algorithms. Thus, the conversion attack is significantly more successful than the different speaker attack.

UBM-GMM vs. ISV: In the two attacks, the ISV algorithm performs equally well, and in some cases, better than the GMM algorithm. We compared the two algorithms using the Wilcoxon Signed-Rank test, and noticed that the result is statistically significant for the conversion attack on male MOBIO speakers with a p-value $= 0.0147$, and in the benign setting for female MOBIO speakers with a p-value $= 0.0466$. This was an expected result since ISV has session variablitiy parameters that makes it perform better than UBM-GMM and also the MOBIO dataset was recorded in various sessions.

The Voxforge dataset being a better and standard dataset has a higher attack success rate compared to MOBIO. The quality of voice conversion plays an important role here. The attacker (source) samples for voice conversion in case of the Voxforge dataset were from the CMU Arctic database that contains samples recorded in a professional recording environment. However, in case of MOBIO, attacker samples were chosen from the Test set of MOBIO itself and that affected the quality of conversion adversely. Moreover, the style of speaking among the speakers varied widely in case of MOBIO that can also be one of the factors affecting the voice conversion training.

[2] All significance results are reported at a 95% confidence level.

6 Experiments: Attacking Human-based Speaker Verification

We now investigate the extent to which humans may be susceptible to the voice conversion attacks in arbitrary human-to-human communications.

6.1 Setup

To evaluate the performance of our voice impersonation attacks against human users, we conducted two web-based studies with 65 and 32 Amazon Mechanical Turk (M-Turk) online workers. In the first study, referred to as the *Famous Speaker Study*, we investigate a scenario, where the attacker mimics a popular celebrity, and posts, for example, the morphed fake samples of his/her speech on the Internet or broadcasts them on the radio. The primary reason for choosing celebrities in our case study was to leverage people's pre-existing familiarity with their voices. In our second study, called the *Briefly Familiar Speaker Study*, we consider a scenario, where humans are subject to a briefly familiar person's (fake) voice (e.g., someone who was introduced at a conference briefly). The failure of users in detecting such attacks would demonstrate a vulnerability of numerous real-world scenarios that rely (implicitly) on human speaker verification.

Our studies involved human subjects, who participated to provide their voice samples for building our morphing engine, and to evaluate the feasibility of our attacks (human verification of converted voices). Their participation in the study was strictly voluntary. The participants provided informed consent prior to the study and were given the option to withdraw from the study at any time. Standard best practices were followed to protect the confidentiality/privacy of participants' responses and audio samples acquired during the study as well as the morphed audio samples that were generated for in the study. Our study was approved by our University's Institutional Review Board.

The demographic information of the participants of our two studies is summarized in Appendix A Table 3. Most of the participants were young and well-educated native English speakers with no hearing impairments. For the first and second studies, each participant was paid $1 and $3, respectively, for his/her effort, which took about 30 minutes and 45 minutes, respectively, for completion.

6.2 Dataset

To build the dataset for our studies, we developed an application to collect audio samples from a group of American speakers, and posted the task on M-Turk. This job required mimicking two celebrities namely, *Oprah Winfrey* and *Morgan Freeman*.

We collected some audio samples of these celebrities available on the Internet, and, using our application, played back those samples to the speakers (who posed as attackers) and asked the male speakers to repeat and record the clips of Morgan Freeman and the female speakers to record the clips of Oprah Winfrey. While collecting these audio samples, speakers were categorically instructed to try their best to mimic the speaking style and emotion of the celebrity that they are listening to (our threat model allows the attack with such a leverage). There were around 100 samples for both the male (Morgan) and the female (Oprah) celebrities that took each user approximately an hour to record. Each participant was paid $10 for this task. Over a period of two weeks, we collected samples from 20 speakers. Among these, we picked 5 male and 5 female speakers, who could record all clips successfully in a non-noisy environment and with a similar style and pace of the original speakers. The demographic information of the final 10 participants has been given in Appendix A Table 3.

6.3 Conversion Processes

The audio data collected from the M-Turk participants acted as the source voice to synthesize the voices of Oprah and Morgan (Famous Speaker Study). The same dataset was used to generate the voices of 4 briefly familiarized target speakers (Briefly Familiar Speaker Study).

We converted attacker's voice to target voice using the CMU Festvox voice converter, and observed that for 25, 50, 100 and 125 sentences, in the training dataset, the average MCD values are 7.52 dB, 7.45 dB, 7.01 dB, and 6.98 dB, respectively. This shows an improvement of only 1%, 6% and less that 1% when increasing the training dataset size from 25 to 50, 50 to 100 and 100 to 125 sentences, respectively. This result confirms the suitability of the conversion system even with a small training dataset. Because of only a slight MCD improvement across different sample sizes, we fixed our training dataset size to only 100 sentences, each with an average duration of 4s.

We converted the voice of each of the 5 female speakers to Oprah's voice and 5 male speakers to Morgan's voice (Famous Speaker Study). We also generated 2 female and 2 male voices by converting female attackers' voices to female targets' voices, and male attackers' voices to male targets' voices (Briefly Familiar Speaker Study).

6.4 Famous Speaker Study

In this study, we asked our participants to first listen to a two-minute speech of each of our victim celebrities (Oprah and Morgan) to get to recall their voices. After familiarization, the participants had to listen to several audio clips and complete two set of tasks, namely, "Speaker Verification" and "Voice Similarity", as defined below.

Speaker Verification Test: In the first set of questions, we played 22 audio clips of around 15 seconds each, and asked the participants to decide if the speaker is Oprah. In each question, they had the choice of selecting "Yes" if they could identify Oprah's voice, "No" if they could detect that the voice does not belong to Oprah, and "Not Sure" if they could not distinguish precisely whose voice is being played. 4 of the presented samples were Oprah's voice collected from different speeches, 8 samples were from a "different speaker" in our dataset described earlier, and 5 samples were from our converted voice dataset, generated by performing voice conversions on our dataset. Similar set of questions was asked about Morgan. Morgan's challenges, consisted of 4 voice of Morgan selected from different speeches and interviews, 6 samples of different speakers, and 6 converted voices picked from our voice conversion dataset.

Voice Similarity Test: In the second set of questions, we played several samples (original speaker, different speaker, and converted voice) and asked users to rate the similarity of the samples to the two target speakers' voices. We defined five ratings to capture the similarity/dissimilarity – "exactly similar", "very similar", "somehow similar", "not very similar" and "different". For each audio sample, participants could select one of the 5 options according to the similarity of the challenge voice to the celebrities' own voices. 4 original speaker, 5 converted voice and 6 different speaker samples were presented in Oprah's set of questions. Similarly, for Morgan's questions, 4 original speaker, 6 converted, and 7 different speaker samples were played.

In both tests, we categorized audio clips into three groups, namely, Original Speaker (benign setting), Different Speaker Attack and Conversion Attack.

Table 2: Performance of human-based speaker verification against our attacks (Famous Speaker and Briefly Familiar Speaker studies). The accuracy of detecting the original as well as different speaker is around 90%, but the accuracy of detecting the conversion attack is around 50%.

	Famous Speaker						*Briefly Familiar Speaker*		
	Oprah			**Morgan**			**Averaged Over 4 Speakers**		
	Yes	**No**	**Not Sure**	**Yes**	**No**	**Not Sure**	**Yes**	**No**	**Not Sure**
Original Speaker	89.23%	9.62%	1.15%	91.54%	8.46%	0.00%	74.69%	22.81%	2.50%
Different Speaker Attack	4.04%	95.19%	0.77%	1.28%	97.95%	0.77%	14.06%	82.81%	3.13%
Conversion Attack	45.85%	45.85%	8.31%	29.33%	54.10%	16.67%	27.81%	47.81%	24.38%

Results: The results for the speaker verification test are summarized in Table 2. The success rate for users in answering original speaker challenges is shown in the first row (column 2 and 5) of Table 2, which is 89.23% for Oprah and 91.54% for Morgan (averaged across all samples across all participants). These results show that the participants were pretty much successful in detecting the original speaker's voice.

The second row (columns 3 and 6) of Table 2 depicts the accuracy of detecting the different speaker attack. The results show that a majority of participants were able to distinguish a different speaker's voice from the original speaker's voice. The rate of correctly identifying a different speaker was 95.19% for Oprah and 97.95% for Morgan (averaged across all different speaker samples across all participants). The results imply that the participants were somewhat more successful in detecting a different speaker than verifying the original speaker.

However, the participants were not as successful in detecting the conversion attack (row three of Table 2; shaded cells). The rate of successfully detecting the presence of conversion attack was around 50% (averaged across all morphed samples across all participants). Interestingly, ambiguity increased while detecting the conversion attack (which is inferred from the increase in "Not Sure" answers). This shows that participants got confused in identifying the converted voice compared to the original speaker's voice samples and different speaker's voice samples. In a real life setting, participants' confusion in recognizing the speaker might highly affect their accuracy of verifying the identity of a speaker. The reason is that, while in our experiment, participants had the choice of answering "Not Sure", in a real life application, where the users should either accept or discard a conversation (e.g., a voice message), they might rely on a random guess, possibly accept an illegitimate conversation or reject a legitimate conversation.

We compared the two attacks (different speaker and conversion attacks) using Wilcoxon Signed-Rank Test and noticed that the result is statistically significant for both of our familiar speakers (p-value $= 0$).

The results from the voice similarity test show that a majority of our participants found the original speaker samples as being "exactly similar" or "very similar" to the original speaker's voice. Only with negligible rates, participants found original samples different or not very similar to the original speaker. This is well-aligned with the speaker verification test results, and shows that people can successfully detect similarity of different samples of the same speaker. 88.08% found samples of Oprah's voice exactly similar or very similar to her voice while 95.77% found samples of Morgan's voice exactly similar or very similar to his voice.

As expected, the users could detect dissimilarity of a different speaker's voice to the original speaker. 86.81% found different speaker's voices as "different" and "not

very similar" to the Oprah's voice; this result was 94.36% for Morgan's voice. Very few users considered a different speaker's voice similar to an original speaker's voice. In line with the speaker verification test, the voice similarity test shows the success of participants in detecting a different speaker.

Our study shows that most of the users rated converted voice as "somehow similar" or "very similar" to the original speaker. 74.10% detected converted voice "very similar" and "some how similar" to Oprah's voice, while this result is 59.74% for Morgan's voice. The voice conversion makes the attacker's voice sound similar to the original target voice. The conversion depends on many parameters, including similarity between source and target before the conversion, and the level of noise present in initial source and target recordings. Since we used same samples of the target voice (Oprah, Morgan) for all conversions, difference between the conversions is mainly due to the ambient noise present in source (attacker) recordings. The source recordings, being better in quality, worked better for conversion. The attacker is assumed to have the ability to improve quality of his recordings to improve the conversion. In our study, Oprah's converted voice was more similar to her original voice than Morgan's converted voice was to his voice. However, we cannot generalize this result to all speakers and all conversions. Figure 2 (Appendix B) summarizes the results of the voice similarity test.

6.5 Briefly Familiar Speaker Study

Similar to the Famous Speaker Study, we conducted a study evaluating the performance of human users in recognizing a briefly familiar speaker. For this study, we picked two female and two male speakers from our dataset as victims, and two female and two male speakers as the attackers from the same dataset mentioned in Section 6.2. We asked the participants to first listen to a 90 seconds recording of a victim's voice to get familiar with the voice, and then answer 15 speaker verification challenges and 15 voice similarity challenges about each speaker (each audio sample was about 15 seconds long). As in the previous study, in the speaker verification test, participants were asked to verify the speaker, and, in the voice similarity test, the participants were asked to rate the level of similarity of the audio clips to the original speaker's voice. Audio clips were categorized as 5 original speaker, 5 different speaker and 5 converted voice. Moreover, we asked the participants their opinion about the tasks, and the qualitative basis for their judgment. To discard possibly inattentive participants, we included dummy questions as part of the challenges that asked the user to pick the right most option from the answers.
Results: Table 2 includes the result of the Briefly Familiar Speaker study, and Figure 2 (Appendix B) summarizes the result of the similarity test, averaged over all participants and all speakers. The participants show an average success rate of 74.68% in recognizing the original speaker correctly averaged over the four speakers (row 1, column 8). Average success rate of users in distinguishing a different speaker is 82.81% (row 2, column 9). These results show that, on an average, participants seem less successful in verifying a briefly familiar speaker compared to a famous speakers.

Importantly, the average success rate of detecting the conversion attack is 47.81% (row 3, column 9). This shows that over 50% of users could not detect the conversion attack. That is, they either misidentify the converted voice as the original speaker's voice or were not able to verify the speaker. We compared the two attacks (different speaker and conversion attacks) using the Wilcoxon Signed-Rank Test, and noticed that

the result is statistically significant (p-value $= 0.0038$), which means the conversion attack works significantly better than the different speaker attack.

The results of the similarity test shows that majority of the participants found the samples in the benign settings exactly similar to the original speaker's voice, and majority of participants found the samples in the different speaker attack setting different from the original speaker's voice. The converted voice similarity is rated as somehow similar to the original speaker's voice, which stands between the different speaker's voice rate and original speaker's voice rate.

At the end of the survey we asked the participants how easy/difficult they found the task of recognizing a speaker, on what basis they made their decisions, and what would possibly improve the participant's accuracy. In general, they found speaker verification to be a challenging task, and quality of the voice to be a prominent factor in verifying the speaker. A summary of their answers is presented in Appendix D.

6.6 Briefly Familiar Speaker vs. Famous Speaker Verification

We compared the performance of the attacks between the two settings (Famous and Briefly Familiar Speaker). Although the result of the Mann-Whitney U test does not show statistical significance between the two settings in case of the conversion attack, the result is significant for the different speaker attack, for both Oprah and Morgan combined (p-value $= 0$). This shows that people can detect the different speaker attack better in the Famous Speaker setting compared to the Briefly Familiar Speaker setting.

The fraction of participants, who could not distinguish the speakers, seems to have increased compared to the Famous Speaker study (as reflected in the last column of Table 2). This suggests that the ambiguity in recognizing a speaker increases as the familiarity with the speaker decreases. The result of the Mann-Whitney U test confirms that this increase is significant for the conversion attack (p-value $= 0.0076$), but not significant for the other two type of settings (original speaker and different speaker) for both Oprah and Morgan combined.

7 Summary

We explored how human voice authenticity can be easily breached using voice conversion, and how such a breach can undermine the security of machine-based and human-based speaker verification. Our voice conversion attack against the state-of-the-art speaker verification algorithms has a very high success rate, about 80-90%. This suggests that current algorithms would not be able to prevent a malicious impostor with morphing capability from accessing the authentication terminal or remote services employing voice biometrics. In our attacks against human verification, the target victims were known users (celebrities) as well as briefly familiar users. The results corresponding to both types of victims highlight that even humans can be fooled into believing, in almost 50% of the cases, that the morphed samples are from a genuine speaker. Natually, people seem to detect attacks against celebrity voices better than briefly familiar voices. In light of this result, it seems that an attacker can compromise the authenticity of remote arbitrary human-to-human conversations with a relatively high chance.

Voice conversion sits right at the heart of all our attacks. Therefore, in order to achieve the best possible outcome, an attacker should strive to improve the voice conversion quality. This could be achieved by choosing high quality audio samples of the

target (victim) when possible and by creating high quality audio samples for the source (attacker), ideally mimicking the victim's voice and verbal style as much as possible. Moreover, if required, the attacker may process the victim samples before and after performing the voice conversion to improve the voice quality (e.g., by filtering-out noise).

8 Conclusions, Limitations and Future Work

In this paper, we studied how human voices can be easily stolen and used against applications and contexts that rely upon these voices, specifically focusing on machine-based and human-based speaker verification. We showed that voice conversion poses a serious threat and our attacks can be successful for a majority of cases. Worryingly, the attacks against human-based speaker verification may become more effective in the future because voice conversion/synthesis quality will continue to improve, while it can be safely said that human ability will likely not.

Our current study has certain limitations that might affect the results when our attacks are implemented in real-life. First, we only used the known state-of-the-art biometric speaker verification system and an off-the-shelf voice conversion tool for conducting our attacks. There may be other systems, especially used in industry, that might give different (better or worse) results under our attacks. Second, our arbitrary speech attack was designed to imitate the scenario, in which an attacker posts fake audio samples of a victim over the Internet or even leaves fake voice messages to someone's phone. The current study does not tell us how the attacks might work in other scenarios such as faking real-time communication, or faking court evidences. Third, we asked the participants in our human verification study to pay close attention to the samples before responding. In real-life, however, if someone posts an audio snippet or leaves a voicemail, people may not pay as much attention. Thus, in this scenario, the possibility of accepting a morphed sample in real-life may actually increase (compared to our study). All these issues should be subject to further research, which we plan to explore.

Among these limitations, our study has certain strengths as well. The users who have participated in the study, in case of the arbitrary speech experiment, were all fairly young with no hearing problems. Older people, or those with hearing disabilities, might perform worse against our attacks. Moreover, our results may be much better if a trained mimicry artist serves the role of an attacker resulting in a better voice conversion model.

Although protecting against our attacks seems challenging, there can be ways to ensure that one's voice does not get stolen by an adversary in the first place. Such measures may include people's awareness to these attacks, and people being wary about posting their audio-visuals online. Another line of defense lies in defeating audio monitoring in public places. For example, putting in place stricter policies for audio recording in public or actively preventing audio monitoring by using high frequency audio transmitters that cloak the audio recordings (without affecting human perception). There exist commercial equipment to jam audio and jeopardize audio surveillance systems [1, 3].

Another natural defense strategy would be the development of speaker verification systems that can resist voice conversion attacks by using liveness tests for a speaker. A development in the field of speaker liveness detection is proposed by Baughman et al. [17]. In our future work, we plan to study these different defense strategies.

References

1. Atlassound – speech privacy/sound masking. http://www.atlassound.com/SpeechPrivacy-SoundMasking-376.
2. AT&T Natural Voices Text-to-Speech. http://www2.research.att.com/~ttsweb/tts.
3. Audio Jammers. http://www.brickhousesecurity.com/category/counter+surveillance/audio+jammers.do.
4. CMU Arctic Databases. http://festvox.org/cmu_arctic/index.html.
5. Festival. http://www.cstr.ed.ac.uk/projects/festival/.
6. Festvox. http://festvox.org/.
7. Mobio. https://www.idiap.ch/dataset/mobio.
8. (NIST SREs. http://www.itl.nist.gov/iad/mig//tests/spk/.
9. Statistical Parametric Synthesis And Voice Conversion Techniques. http://festvox.org/11752/slides/lecture11a.pdf.
10. The ModelTalker TTS system. https://www.modeltalker.org.
11. Voice Cards. http://www.voicecards.com/index.html.
12. Voxforge. http://www.voxforge.org/.
13. Banking on the power of speech, 2013. https://wealth.barclays.com/en_gb/internationalwealth/manage-your-money/banking-on-the-power-of-speech.html.
14. VocaliD: Donating your voice to people with speech impairment, 2014. http://www.assistivetechnologyblog.com/2014/03/vocalid-donating-your-voice-to-people.html.
15. Wells Fargo tests mobile banking voice recognition, 2014. http://www.mobilepaymentstoday.com/news/wells-fargo-tests-mobile-banking-voice-recognition.
16. F. Alegre, R. Vipperla, N. Evans, and B. Fauve. On the vulnerability of automatic speaker recognition to spoofing attacks with artificial signals. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012.
17. Aaron K Baughman and Jason W Pelecanos. Speaker liveness detection, November 19 2013. US Patent 8,589,167.
18. Jennifer Bjhorus. Big banks edge into biometrics, 2014.
19. Lukas Burget, Pavel Matejka, Petr Schwarz, Ondrej Glembek, and Jan Cernocky. Analysis of feature extraction and channel compensation in a gmm speaker recognition system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2007.
20. Joseph P Campbell Jr. Speaker Recognition: A Tutorial. *Proceedings of the IEEE*, 1997.
21. Mark Chevillet, Maximilian Riesenhuber, and Josef P Rauschecker. Functional Correlates of the Anterolateral Processing Hierarchy in Human Auditory Cortex. *The Journal of Neuroscience*, 2011.
22. Penny Crosman. U.s. bank pushes voice biometrics to replace clunky passwords, 2014.
23. Phillip L De Leon, Vijendra Raj Apsingekar, Michael Pucher, and Junichi Yamagishi. Revisiting the security of speaker verification systems against imposture using synthetic speech. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010.
24. Phillip L De Leon, Michael Pucher, and Junichi Yamagishi. Evaluation of the vulnerability of speaker verification to synthetic speech. 2010.
25. Phillip L De Leon, Michael Pucher, Junichi Yamagishi, Inma Hernaez, and Ibon Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2012.

26. Srinivas Desai, E Veera Raghavendra, B Yegnanarayana, Alan W Black, and Kishore Prahallad. Voice conversion using artificial neural networks. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009.
27. Harry Hollien, Wojciech Majewski, and E Thomas Doherty. Perceptual Identification of Voices Under Normal, Stress and Disguise Speaking Conditions. *Journal of Phonetics*, 1982.
28. Alexander Kain and Michael W Macon. Spectral voice conversion for text-to-speech synthesis. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 1998.
29. Ashlee Keiler. Is Voice-Recognition The Future of Banking? Wells Fargo Thinks So, 2014. `http://consumerist.com/2014/01/21/is-voice-recognition-the-future-of-banking-wells-fargo-thinks-so/`.
30. E. Khoury, L. El Shafey, and S. Marcel. Spear: An open source toolbox for speaker recognition based on bob. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
31. Eran Kinsbruner. Key considerations for testing voice recognition in mobile banking applicaions, 2013. `http://www.banktech.com/channels/key-considerations-for-testing-voice-recognition-in-mobile-banking-applications/a/d-id/1296456?.`
32. John Kominek, Tanja Schultz, and Alan W Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *SLTU*, 2008.
33. Robert F Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Communications, Computers and Signal Processing, 1993., IEEE Pacific Rim Conference on*, volume 1. IEEE, 1993.
34. G Suvarna Kumar, KA Prasad Raju, Mohan Rao CPVNJ, and P Satheesh. Speaker recognition using gmm. *International Journal of Engineering Science and Technology*, 2010.
35. Hakan Melin, 2006. Automatic Speaker verification on site and by telephone: methods, applications and assessment.
36. Douglas Reynolds. An overview of automatic speaker recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.
37. Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, 2000.
38. Phil Rose. *Forensic Speaker Identification*. CRC Press, 2003.
39. Maliheh Shirvanian and Nitesh Saxena. Wiretapping via mimicry: Short voice imitation man-in-the-middle attacks on crypto phones. In *ACM CCS 2014*, 2014.
40. Andreas Stolcke, Sachin S Kajarekar, Luciana Ferrer, and Elizabeth Shrinberg. Speaker recognition with session variability normalization based on mllr adaptation transforms. *Audio, Speech, and Language Processing, IEEE Transactions on*, 2007.
41. Tomoki Toda, Alan W Black, and Keiichi Tokuda. Acoustic-to-articulatory inversion mapping with gaussian mixture model. In *INTERSPEECH*, 2004.
42. Tomoki Toda, Alan W Black, and Keiichi Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. In *ICASSP (1)*, 2005.
43. Robbie Vogt and Sridha Sridharan. Explicit modelling of session variability for speaker verification. *Computer Speech & Language*, 2008.
44. Zhizheng Wu and Haizhou Li. Voice conversion and spoofing attack on speaker verification systems. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, 2013.

A Demographics Information

Table 3: Demographics information of (a) speakers of the arbitrary speech dataset (b) participants in the human-based famous speaker verification study (c) participants in the human-based briefly familiar speaker verification study

	(a)	(b)	(c)
	N = 10	**N = 65**	**N = 32**
Gender			
Male	50%	42%	49%
Female	50%	58%	51%
Age			
18-24 years	40%	8%	19%
25-34 years	60%	61%	46%
35-44 years	0%	22%	21%
45-54 years	0%	6%	8%
55-64 years	0%	3%	6%
Education			
High school degree or equivalent (e.g. GED)	10%	17%	24%
Some college but no degree	10%	16%	19%
Associate degree (e.g. AA, AS)	0%	8%	19%
Bachelor's degree (e.g. BA, AB, BS)	80%	40%	35%
Master's degree (e.g. MA, MS, MBA)	0%	13%	3%
Professional degree (e.g. MD, DDS, JD)	0%	3%	0%
Doctorate degree (e.g. PhD)	0%	2%	0%
English as First Language			
Yes	100%	91%	89%
No	0%	9%	11%
Hearing Impairment			
Yes	10%	0%	3%
No	90%	100%	97%

B Voice Similarity Test Results

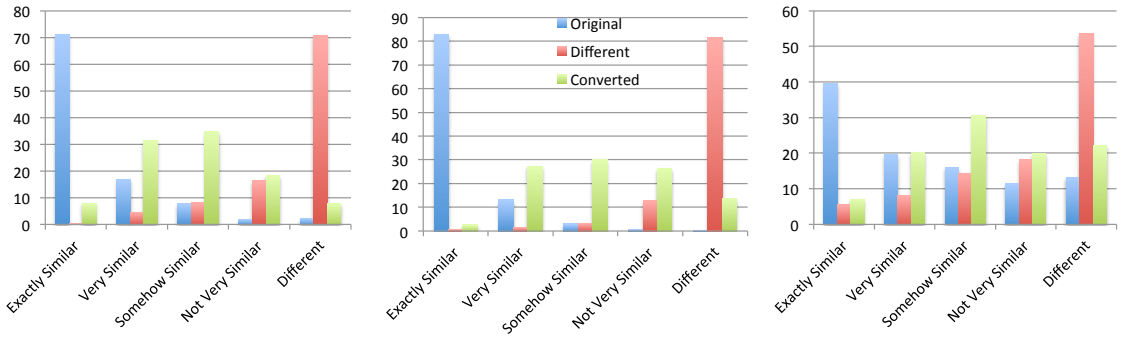
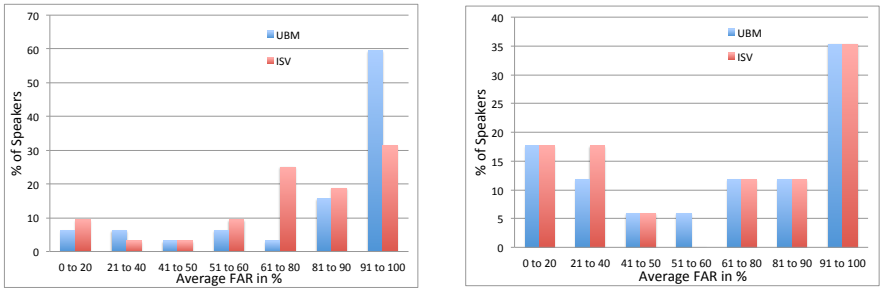


Fig. 2: The voice similarity test results for Oprah (left), Morgan (middle), and unfamiliar speakers (right)

C Voice Conversion Attack FAR distribution (MOBIO Dataset)



(a) **(b)**

Fig. 3: Distribution of FAR for Voice Conversion Attack across the (a) Male, (b) Female users in the MOBIO dataset

D Open-Ended Feedback

At the end of the second study, we asked the participants as to how easy/difficult they find the task of recognizing a speaker. Majority of participants found the task to be "fairly difficult", some believed that it was easier for some recordings and more difficult for others, and a couple of participants found the female speakers more easy to distinguish. We also asked the participants what possibly can improve the accuracy of their answers. Most of the users reported that the quality of the recordings plays an important role, others believed that associating the voice to an image (i.e., a face) helps to recognize the speaker better. Some answered that listening to multiple topics spoken by the speaker or hear the speaker sing a song can help to understand his/her particular style of speaking. The last question polled the participants about the basis behind their decisions. Each user had distinct opinion, including quality, naturalness, genuineness, pitch, tone, style, pace, accent, volume, background noise, age, and race of the speaker.