The Crux of Voice (In)Security: A Brain Study of Speaker Legitimacy Detection

Ajaya Neupane*
University of California Riverside
ajaya@ucr.edu

Nitesh Saxena
University of Alabama at Birmingham
saxena@uab.edu

Leanne Hirshfield
Syracuse University
lhirshf@syr.edu

Sarah Elaine Bratt
Syracuse University
sebratt@syr.edu

***Abstract*—A new generation of scams has emerged that uses *voice impersonation* to obtain sensitive information, eavesdrop over voice calls and extort money from unsuspecting human users. Research demonstrates that users are fallible to voice impersonation attacks that exploit the current advancement in speech synthesis. In this paper, we set out to elicit a deeper understanding of such human-centered "voice hacking" based on a neuro-scientific methodology (thereby corroborating and expanding the traditional behavioral-only approach in significant ways). Specifically, we investigate the *neural underpinnings* of voice security through *functional near-infrared spectroscopy (fNIRS)*, a cutting-edge neuroimaging technique, that captures neural signals in both temporal and spatial domains. We design and conduct an fNIRS study to pursue a thorough investigation of users' mental processing related to *speaker legitimacy detection* – whether a voice sample is rendered by a target speaker, a different other human speaker or a synthesizer mimicking the speaker. We analyze the neural activity associated within this task as well as the brain areas that may control such activity.**

Our key insight is that there may be no statistically significant differences in the way the human brain processes the *legitimate speakers vs. synthesized speakers*, whereas clear differences are visible when encountering *legitimate vs. different other human speakers*. This finding may help to explain users' susceptibility to synthesized attacks, as seen from the behavioral self-reported analysis. That is, the impersonated synthesized voices may seem *indistinguishable* from the real voices in terms of both behavioral and neural perspectives. In sharp contrast, *prior studies showed subconscious neural differences in other real vs. fake artifacts* (e.g., paintings and websites), despite users failing to note these differences behaviorally.

Overall, our work dissects the fundamental neural patterns underlying voice-based insecurity and reveals users' susceptibility to voice synthesis attacks at a *biological level*. We believe that this could be a significant insight for the security community suggesting that the human detection of voice synthesis attacks may not improve over time, especially given that voice synthesis techniques will likely continue to improve, calling for the design of careful machine-assisted techniques to help humans counter these attacks.

*Work done while being a student at UAB

Network and Distributed Systems Security (NDSS) Symposium 2019
24-27 February 2019, San Diego, CA, USA
ISBN 1-891562-55-X
https://dx.doi.org/10.14722/ndss.2019.23206
www.ndss-symposium.org

I. Introduction

Voice is supposed to be a unique identifier of a person. In human-to-human conversations, people may be able to recognize the speakers based on the unique traits of their voices. However, previous studies have shown that human-based speaker verification is vulnerable to voice impersonation attacks [30]. A malicious entity can impersonate someone's voice by mimicking it using speech synthesis techniques. In particular, off-the-shelf speech morphing techniques can be used to generate the spoofed voices of people of interest (victims) [30]. The attacker can then perform social engineering trickeries using an impersonated voice to fool the users into accepting it as a legitimate one. These attacks may eventually make the users reveal sensitive and confidential information, which may hamper their security, privacy, and safety.

Such voice imitation is an emerging class of threats, especially given the advancement in speech synthesis technology, seen in a variety of contexts that can harm a victim's reputation and her security/safety [30]. For instance, the attacker could publish the morphed voice samples on social media [17], impersonate the victim in phone conversations [21], leave fake voice messages to the victim's contacts, and even launch man-in-the-middle attacks against end-to-end encryption technologies that require users to verify the voices of the callers [38], to name a few instances of such attacks.

Given the prominence and rapid emergence of these threats in the wild, it is crucial to understand users' innate psychological behavior that governs the processing of voices and their potential susceptibility to voice impersonation attacks. In this paper, we follow the neuroimaging methodology adopted in a recently introduced line of research (e.g., [8], [32], [34]) to scrutinize the user behavior in the specific context of such "voice security". Specifically, we study users' neural processes (besides their behavioral performance) to understand and leverage the *neural* mechanics when users are subjected to voice impersonation attacks using a state-of-the-art neuroimaging technique called *functional near-infrared spectroscopy (fNIRS)*.

The specific goal of this paper is to study the neural underpinnings of voice (in)security, and analyze differences (or lack thereof) in neural activities when users are processing different types of voices. We examine how the information present in the neural signals can be used to explain users' susceptibility to voice imitation attacks using synthesized voices (i.e., *speaker legitimacy detection*). Prior studies [25], [32]–[34] have shown that *subconscious neural differences*

TABLE I. SUMMARY OF REAL-FAKE ANALYSIS OBSERVED IN RELATED WORKS VS. OUR WORK

Type of Artifacts	Differences in Neural Activity	Differences in Behavioral Response
Websites under phishing ([32]–[34])	Present	Nearly absent
Paintings [25]	Present	Nearly absent
Voices (our work)	Absent	Nearly absent

exist when users are subject to real vs. fake artifacts, even though users themselves may not be able to tell the two apart behaviorally. Neupane et al. in their previous studies [32]–[34] found differences in neural activities at the right, middle, inferior, and orbitofrontal areas when users were processing real and fake websites. Similarly, Huang et al. [25] found differences in neural activation at similar areas when users were viewing real and fake Rembrandt paintings. Higher activation in the frontopolar area implicates the use of working memory and cognitive workload. Lower activation in the orbitofrontal area suggests that the users trust stimuli presented to them g [15]. of the target speakers.

Based on these prior results, we performed our study with the *hypothesis that these and other relevant brain areas might be activated differently when users are listening to the original and fake voices of a speaker*. We, therefore, set-up the fNIRS headset configuration to measure the frontal and temporo-parietal brain which overlaps with the regions reported in these previous "real-fake detection" studies. The implications of the neural activity differences, if present when processing real vs. fake voices, can be important as these differences could be automatically mined and the user under attack could be alerted to the presence/absence of the attack, even though the user may have himself failed to detect the attack (behaviorally). Neupane et al. suggested such an approach in the context of phishing attacks [33]. Our study investigates the same line in the context of voice synthesis attacks.

The neuroimaging technique used in our study, i.e., fNIRS, is a non-invasive imaging method to measure the relative concentration of oxygenated hemoglobin (oxy-Hb) and deoxygenated hemoglobin (deoxy-Hb) in brain cortex [11], [24], [26]. By examining the changes in oxy-Hb and deoxy-Hb, we can infer the activities in the neural areas of interest. We carefully selected fNIRS as our study platform as it has the unique capabilities to provide spatially accurate brain activity information better than the EEG (Electroencephalography) and similar to that of fMRI (Functional Magnetic Resonance Imaging) [26]. Therefore, we preferred fNIRS to ensure we capture the features from both temporal and spatial domains. Unlike fMRI, fNIRS also allows us to pursue the study in environments with better ecological validity since the participants do not have to be in a supine position in the fMRI scanner while making decisions.

Our Contributions: We design and conduct an fNIRS study to pursue a thorough investigation of users' processing of real and morphed voices. We provide a comprehensive analysis of the collected neuroimaging data set and the behavioral task performance data set. Contrary to our hypothesis (and contrary to the case of website/painting legitimacy detection), we do not obtain differences in the way the brains process legitimate speakers vs. synthesized speakers, when subject to voice impersonation attacks, although marked differences are seen between neural activity corresponding to a legitimate speaker vs. a different unauthorized human speaker. That is, the synthesized voices seem nearly indistinguishable from the real voices with respect to the neurocognitive perspective. This insight may serve well to explain users' susceptibility to such attacks as also reflected in our task performance results (similar to the task performance results reported in [30]). Table I captures a summary of our work versus other real-fake detection studies.

Since this potential indistinguishability of real vs. morphed lies at the core of human biology, we posit that the problem is very severe, as the human detection of synthesized attacks may not improve over time with evolution, Further, in our study, we use an off-the-shelf, academic voice morphing tool based on voice conversion, CMU Festvox [19], whereas with the advancement in the voice synthesizing technologies (e.g., newer voice modeling techniques such as those offered by Lyrebird and Google WaveNet [29], [41]), it might become even more difficult for users to identify such attacks. Also, our study participants are mostly young individuals and with no reported hearing disabilities, while older population samples and/or those having hearing disabilities may be more prone to voice synthesis attacks [16].

We *do not* claim that the rejection of our hypothesis necessarily means that the differences between real and morphed voices are absent conclusively – further studies might need to be conducted using other neuroimaging techniques and other wider samples of users. However, our work certainly casts a serious doubt regarding the presence of such differences (in contrast to other real-fake contexts, such as paintings or websites), which also maps well with our behavioral results, thereby explaining human-centered voice insecurity.

In light of our results, perhaps the only feasible way to protect users from such attacks would be by making them more aware of the threat, and possibly by developing technical solutions to assist the users. Even though machine-based voice biometric systems have also been shown to be vulnerable to voice synthesis attacks [30], the security community can certainly work, and has been working, towards making such techniques more secure with advanced liveness detection mechanisms, which could aid the end users against voice synthesis based social engineering scams, whenever possible.

Broader Scientific Significance: We believe that our work helps to advance the science of human-centered voice security, in many unique ways. It also serves to reveal the fundamental neural basis underlying voice-based security, and highlights users' susceptibility to advanced voice synthesis attacks. Table X gives a snapshot of all our results.

Beyond the aforementioned novel contributions, one important scientific attribute of our work lies in recreating and revalidating the findings from the prior *behavioral-only* (task performance) study of voice insecurity reported in the literature [30] through independent settings. Similar to [30], our results confirm the susceptibility of human users to voice impersonation attacks.

Security Relevance and Implications: Although our work is informed by neuroscience, it is deeply rooted in computer security and provides valuable implications for the security community. We conduct a neuroimaging-based user study and show why attackers might be successful at morphed voice attacks. Many similar security studies focusing on human neuro-physiology have been published as a new line of research in mainstream security/HCI venues, e.g., [8], [32]–[34]. How users perform at crucial security tasks from a neurological standpoint is therefore of great interest to the security community.

This line of research followed in our work provides novel security insights and lessons that are *not possible* to elicit via behavioral studies *alone*. For example, prior studies [32]–[34] showed that security (phishing) attacks can be detected based on neural cues, although users may themselves not be able to detect these attacks. Following this line, our work conducted an fNIRS study to dissect users' behavior under voice impersonation attacks, an understudied attack vector. Our results show that even brain responses cannot be used to detect such attacks, which serve to explain why users are so susceptible to these attacks.

II. Background & Prior Work

In this section, we provide an overview on fNIRS system, and discuss the related works.

A. fNIRS Overview

The non-invasive fNIRS technology has unique capabilities in that it can provide spatially accurate brain activity information in line with fMRI, but it can do so in an ecologically valid experimental environments (not inside a scanner under a supine posture). It is easy to set-up and robust to motion artifacts, and offers high spatial resolution [11], [24], [26]. The basis of fNIRS is the usage of near-infrared light (700-900 nm range), which can penetrate through scalp to reach the brain cortex. Optical fibers are placed on the surface of the head for illumination while detection fibers measure light reflected back [10], [40]. The differences in absorption spectra of the lights by oxy-Hb and deoxy-Hb allow the measurement of relative changes in hemoglobin in the blood in brain. fNIRS provides better temporal resolution compared to fMRI and better spatial resolution (approximately 5mm) compared to EEG. Based on these attractive features, we have chosen fNIRS as a platform to conduct our study reported in this paper. The hemodynamic changes measured by the fNIRS occurs at slow rate of 6-9 sec similar to fMRI, so the trial duration is relatively longer than EEG [10].

B. Related Work

Voice spoofing attacks are a serious threat to users' security and privacy. Previous studies have shown that attackers equipped with voice morphing techniques could breach machine and human speaker verification systems [28], [30], [38]. Mukhopadhyay et al. [30] attacked both machine-based and human-based speaker verification system with morphed voice samples and different speakers' (other users') voice samples. They report that both human and machine are vulnerable to such attacks. Shirvanian et al. [38] successfully launched man-in-the-middle attacks against "Crypto Phones", where users verbally exchange and compare each other's authentication codes before establishing secure communications, using morphed voice samples. Lewison et al. [28] proposed a blockchain system in which face-recognition is combined with voice verification to prevent voice morphing attacks. Bai et al. [4] proposed the use of voices to authenticate certificates. However, it can be subject to potential morphing attacks. In this light, it is important to understand why users may fail to identify the voice of a speaker under morphing attacks, and inform the designers of automated speech recognition systems as to how users may distinctively process speakers' voices.

Recently, several studies have reported the use of neuroimaging to understand the underlying neural processes controlling users' decision-making in security tasks [8], [32]–[34]. Specifically, Neupane et al. [34] analyzed brain signals when users were detecting the legitimacy of phishing websites and reading malware warnings using a state-of-the-art neuroimaging technique fMRI. Neupane et al. [32] followed up the study with EEG and eye-tracking to understand users' neural and gaze metrics during phishing detection and malware warnings tasks in a near-realistic environment.

Relevant to our study is the fNIRS study of phishing detection by Neupane et al. [33]. They used fNIRS to measure neural activities when users were identifying real and phishing websites, and built neural cues based machine learning models to predict real and phishing website. Unlike this study, we use the fNIRS system to measure the neural behavior in a different application paradigm – voice security.

There are some neuroscience studies on voice recognition (e.g., [3], [6], [7], [20]). Belin et al. [6] studied the brain areas involved in voice recognition. Formisano et al. [20] in their fMRI study showed the feasibility of decoding speech content and speaker identity from observation of auditory cortical activation patterns of the listener. Similar to our study, Bethman et al. [7] studied how human brain processes voices of a famous familiar speaker and an unfamiliar speaker. In contrast to all these studies, our research focuses on the security aspect of voice impersonation attack and explains why users fail to identify fake voices, and evaluates the privacy issues related to the BCI devices.

C. The Premise of our Hypothesis

Prior researchers [25], [32]–[34] have also conducted studies to understand why people fall for fake artifacts (e.g., fake images, fake websites) analyzing their neural signals. Huang et al. [25] had conducted a study to understand how human brain reacts when they are asked to identify if a given Rembrandt painting was a real or a fake. They reported differences in neural activation at left and right visual areas when users were viewing real and fake Rembrandt paintings (see Figure 2). Neupane et al. had also conducted studies [32]–[34] to analyze the neural activations measured using different neuroimaging devices (e.g., EEG [32], fNIRS [33] and fMRI [34]) when users were subjected to phishing attacks. The users were asked to identify if a given website was real or fake and their neural signals were concurrently measured when they were viewing these websites. They reported differences in the brain activations when users were processing real and fake websites.

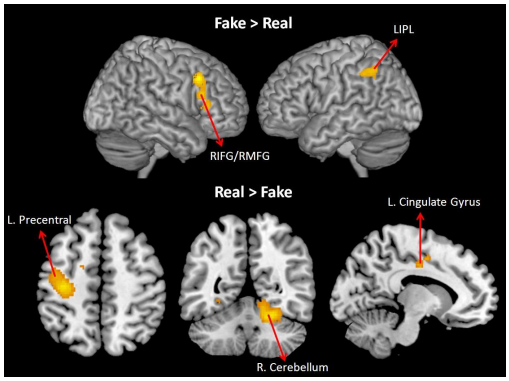

Fig. 1. Activation in right middle frontal gyri (RMFG), right inferior frotal gyri (RIFG), left inferior parietal lobule (LIPL), left precentral gyrus, right cerebellum, and left cingulate gyrus is dependent on whether or not the participant was viewing an image of a genuine website (real vs. fake) [34].

TABLE II. BRAIN AREAS ACTIVATED IN PHISHING TASK [34] AND THEIR CORRESPONDING FUNCTIONS

Brain Areas	**Functions**
Orbitofrontal area	Decision-making; judgment
Middle frontal, inferior frontal, and inferior parietal areas	Working memory
Right cerebellum and left precentral gyrus	Feedforward and feedback projections
Occipital cortex	Visual processing, search

Specifically, Neupane et al. [34] revealed statistically significant activity in several areas of the brain that are critical and specific to making "real" or "fake" judgments. For the websites those the participants identified as fake (contrasted with real), participants activated right middle, inferior, and orbital frontal gyri, and left inferior parietal lobule. The functions governed by these areas are listed in Table II. On the other hand, when real websites were identified, participants showed increased activity in several regions, such as the left precentral gyrus, right cerebellum, and the occipital cortex (see Figure 1). Neupane et al. [33] also explored a feasibility of fake website detection system based on fNIRS-measured neural cues, where they were able to obtain the best area under the curve of 76%.

Fig. 2. Activation in the visual areas, calcarine sulcus (CS), is dependent on whether or not the participant was viewing an image of a genuine Rembrandt (real vs. fake) [25].

Similar to the tasks of identifying real and fake websites or images, the speaker legitimacy detection task also involves real-fake decision making and hence we hypothesized that similar areas should be activated in the speaker legitimacy detection task. We set-up our study to test the *hypothesis that the brain areas related to the real-fake decision making should be activated differently when users are listening to the original and fake voices of a speaker as well*. We, therefore, set-up the fNIRS headset configuration to measure the frontal and temporo-parietal brain regions reported in these previous "real-fake detection" studies. Next, we also try to automatically infer the type of voice (real or morphed voice) a user is listening to based on the fNIRS-measured neural signals.

D. Voice Synthesis

Voice synthesis is commonly used in text-to-speech systems. The quality of the synthesized voice is judged by its intelligibility and its similarity to a targeted human voice. Traditionally, it is known to be challenging for a voice synthesizer to produce a natural human speech (without noticeable noise) artificially and make it indistinguishable from the human voice [18]. Additionally, voice synthesizers require a huge amount of data from a target speaker to learn the phonemes and generate a synthesized speech of the target speaker.

However, newer techniques have emerged that may do a much better job of synthesizing a voice. Voice morphing (also referred to as voice conversion and voice transformation) is one such technique to generate a more naturalistic (human-type) voices with fewer samples. Voice morphing software takes some samples of a speech spoken by a source speaker and produces a sound as if it was spoken by the target speaker by mapping between the spectral features of the source speaker's and the target speaker's voice [42]. Previous research [30], [38] has reported via behavioral studies that the morphed voice attack may be successful against users with high probability. In our study, we followed a methodology described in [30] and used the CMU Festvox voice converter [19], an academic off-the-shelf voice morphing tool, to generate the morphed voices of the target speakers.

E. Threat Model

In our work, we study how access to a few voice samples of a speaker can be used to launch attacks on human-based speaker verification systems. We assume that an attacker has collected a few voice samples previously spoken by the target victim with or without her consent. The attacker can get such voice samples from any public speeches made by the victim or by stealthily following the victim and recording audio as he speaks with other people. The attacker then uses the voice samples to train a model of a morphing engine (we followed the procedures mentioned in [30] for morphing engine), such that the model creates the victim's voice for any (new) arbitrary speech spoken by the attacker. This morphed speech can now be used to attack human-based speaker verification systems. The attackers can either make a fake phone call, or leave voice messages, or produce a fake video with the victim's voice in it and post it online.

F. Terminology and Attack Description

Victim Speakers: These are the speakers whose voices were manipulated to launch voice impersonation attacks on participants in our study.

Familiar Speakers: In our study, we used the voice samples of Oprah Winfrey and Morgan Freeman to represent the set of familiar speakers. The reason for choosing these celebrities in our case study is to leverage their distinct and unique voice texture and people's pre-existing familiarity with their voices. The participants were also asked to answer if they were familiar with the voice of these celebrities as "a yes/no question" before their participation.

Briefly Familiar Speakers: In our study, these are the speakers whom the participants did not know before the study and were only familiarized during the experimental task only to establish brief familiarity. The briefly familiar speakers represent a set of people with whom the users have previously interacted only for a short-term (e.g., a brief conversation at a conference).

Different Speaker Attack: Different speakers are the arbitrary people who attempt to use their own voice to impersonate as the victim speaker. In our study, using a different speaker's voice, replacing the voice of a legitimate speaker to fool the participants, is referred to as the different speaker attack.

Morphed Voice Attack: Attackers can create spoofed voice using speech morphing techniques to impersonate the voice of a victim user, referred to as a morphed voice. In our study, the use of such a voice to deceive other users to extract their private information is therefore called the morphed voice attack.

Speaker Legitimacy Detection: In our study, this represents the act of identifying whether the given voice sample belongs to the original speaker or is generated by a morphed engine. The *different speaker attack is used as a baseline for the morphing attack*, since it is expected that people might be able to detect the different speaker attack well.

III. Study Design & Data Collection

In this section, we present the design of our experimental task, the set-up involving fNIRS, and the protocol we followed for data collection with human participants.

A. Ethical and Safety Considerations

Our study was approved by our university's institutional review board. We ensured the participation in the study was strictly voluntary and the participants were informed of the option to withdraw from the study at any point in time. We obtained an informed consent from the participants and made sure they were comfortable during the experiment. We also followed the standard best practices to protect the confidentiality and privacy of participant's data (task responses and fNIRS data).

B. Design of the Voice Recognition Task

The study design for our voice recognition task followed the one employed in recent task performance study of speaker verification [30]. However, unlike [30], our study captured neural signals in addition to the task performance data. We designed our experiment to test the following hypothesis:

Hypothesis 1 *The activation in frontopolar and temporoparietal areas, which covers most of the regions activated in previous studies (see Section II-C), will be high when participants are listening to the morphed voice of a speaker compared to the original voice of the speaker.*

To test this hypothesis, we used original, morphed, and different voices for two types of speakers – familiar speakers and briefly-familiar speakers(see Section II-F). The participants were asked regarding their familiarity with these speakers' voices as "a yes/no question" before the experiment. During the experiment, the participants were asked to identify the real (legitimate) and fake (illegitimate) voice of a speaker. We assumed the real voices may impose more trust compared to the fake voices. The failure of users in detecting such attacks would demonstrate a vulnerability of numerous real-world scenarios that rely (implicitly) on human speaker verification.

Stimuli Creation: Voice Conversion: We followed the methodology similar to the one reported in [30] to create our dataset. For familiar speakers, we collected the voice samples of two popular celebrities, Oprah Winfrey and Morgan Freeman, available from the Internet. For unfamiliar speakers, we recruited participants via Amazon Mechanical Turk (MTurk). We asked twenty American speakers in MTurk to record the speech of these two celebrities in their own voices. They were asked to read the same sentences the celebrities had spoken in the recorded audio and also to imitate the celebrities' speaking style, pace and emotion. We fed these original voice samples of the celebrities and the recorded voice samples from one male and one female speaker among these twenty speakers to the CMU Festvox voice converter [19] to generate the morphed voices of Morgan Freeman and Oprah Winfrey.

In line with the terminology used in [30], in our study, the original recording of a victim speaker's voice is referred to as a "real/original voice". For each speaker, the fake voices created using speech morphing techniques is referred as a "morphed voice", and recorded speech in other speaker's voices is referred as a "different speaker".

Experiment Design: We used an event-related (ER) design for our study. In ER design, each trial is presented as an event with longer inter-trial-interval and we can isolate fNIRS response to each item separately [37]. We familiarized the participants with the voice a of victim speaker for 1-minute. We could not let the participants replay the original audio during familiarization as it would have made the experiment longer. As the experiment gets longer, fatigue effects may set in, eventually affecting the quality of the brain signals recorded.. After familiarization, we presented 12 randomly selected speech samples (4 in original speaker's voice voice, 4 in morphed voice and 4 in different speaker's voice and asked participants to identify legitimate and fake voices of the victim speakers. The process was repeated for four victim speakers (2 familiar speakers, and 2 briefly familiar speakers) randomly. Following [30], our study participants were not informed about voice morphing to prevent *explicit priming* in the security task. In real life also, people may have to decide a speaker's legitimacy without knowing the voices may have been morphed.

The experiment started with the Firefox browser loading the instructions page (specifying the tasks the participants are supposed to perform) for 30 seconds. This was followed by a rest page of 14 sec (+ sign shown at the center of a blank page) during which participants were asked to relax. We next

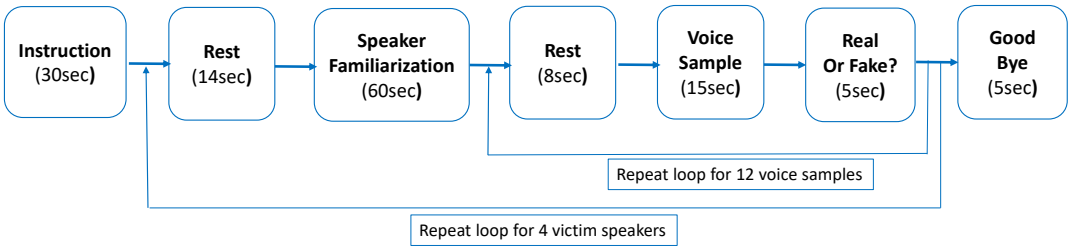

Fig. 3. The flow diagram depicts the presentation of trials in the experiment. The participants were familiarized with a speaker, and were asked to recognize the short voice samples presented next as real or fake.

played a speech sample of a victim speaker for 60 sec, and then showed a rest page for 8 sec. This was followed by 12 trials, each 20 sec long. Each trial consisted of a voice (corresponding to a fake/real voice of the speaker) played for 15 sec, followed by a 5 sec long response page. The response page had a dialog box with the question, "Do you think the voice is of the original speaker?" and the "Yes" and "No" buttons. The participants answered the question using mouse input. A rest page of 8 sec was loaded after each trial. Rest trials are considered as windows of baseline brain activity. The process was repeated for four speakers and the experiment ended with the goodbye note of 5 sec. The system recorded the participant's neural data, responses and response time. Figure 3 depicts the flow diagram.

C. Study Protocol

Our study followed a *within-subject design*, whereby all participants performed the same set of (randomized) trials corresponding to the voice recognition task.

Recruitment and Preparation Phase: We recruited twenty healthy participants from the broader university community (including students and staff) by distributing the study advertisements across our university's campus. We asked the participants about their familiarity with Morgan Freeman's and Oprah Winfrey's voices, i.e., if the participants have heard the celebrities' voices before and could recognize those voices, along with participants' age, gender, and educational background in the pre-test questionnaire. Of the 20 participants, 10 were male and 10 were female. All were English speaking participants in the age range of 19-36 years with a mean age of 24.5 years. Table III summarizes the demographic information of our participants.

Previous power analysis studies have found 20 to be an optimal number of participants for such studies. For instance, statistical power analysis of ER-design fMRI studies have demonstrated that 80% of clusters of activation proved reproducible with a sample size of 20 subjects [31]. Both fMRI and fNIRS are based on hemodynamic response of the BOLD principle, so we assumed similar power analysis for fMRI and fNIRS. Also previous fNIRS studies have shown that fNIRS measurements are reliable with 12 participants [35].

TABLE III. DEMOGRAPHICS

Participant Size N = 20	
Gender(%)	
Male	50%
Female	50%
Age(%)	
18-22	20%
22-26	55%
27-31	20%
31+	5%
Handedness(%)	
Right-Handed	90%
Left-Handed	10%
Background(%)	
High School	10%
Bachelor's	20%
Masters	55%
Doctorate	10%
Others	15%

Our participant demographics is also well-aligned with prior neuroimaging security studies [7], [8], [32], [34]. Each participant was paid $10 upon completion.

Task Execution Phase: To execute the experiment, we used two dedicated computers, one in which the experimental task was presented and the task responses and the response times were logged, namely stimuli computer, and the other in which fNIRS data measured during the experiment was recorded, namely data collection computer. We synchronized the data between these two computers by placing a marker in the fNIRS data at the beginning of the task.

We recorded hemodynamic responses, the rapid delivery of blood to active neuronal tissues [23], in the frontal cortex and the temporoparietal cortex on both hemispheres using fNIRS system developed by Hitachi Medical (ETG 4000) in our experiment. These areas cover the brain regions activated in previous studies [33], [34] (see Section II-C). We took the measurement of each participant's head to determine the best size of probe cap that would fit the participant. The fNIRS optodes were then placed on participant's head using fNIRS probe cap which ensured standardized sensor placement

according to the well established 10-20 system. The inter-optode distance was set to 30mm and the data was acquired at the frequency of 10Hz. Figure 4 depicts the experimental set-up used in our study.

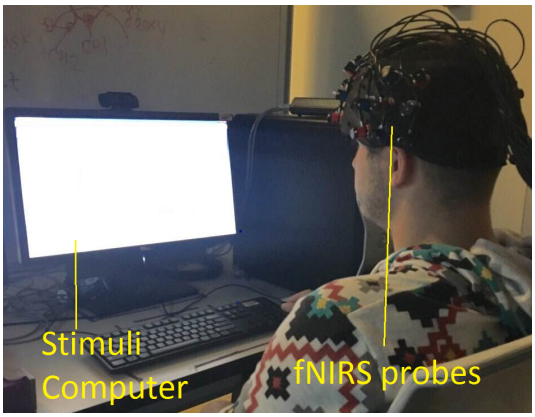

Fig. 4. Experimental setup used in our voice impersonation attack study. The instructions were shown on the stimuli computer's screen. Audio was played back through external PC speakers (not shown in the figure).

Calibration is needed to ensure the quality of the data recorded is good. The participant was then moved to the stimuli computer for performing the speaker voice recognition task. Next, the participants were instructed on the tasks they were performing in our study. They were asked to use the mouse to enter the responses and were requested to minimize the body movements during the experiment. The participants then performed the task discussed in Section III-B.

In our study, we recorded the changes in the concentration of oxygenated hemoglobin (oxy-Hb) and deoxygenated hemoglobin (deoxy-Hb) through the fNIRS device, and task performance metrics (response and response times) while the participants performed the tasks as the repeated measures of our study.

IV. Analysis Methodology

In this section, we provide the steps we followed to process the neural signals for our analysis. We also provide an overview of the statistical tests we conducted on our data.

A. Neural Data Processing

Raw light intensity fNIRS data collected in the experiment were processed to remove high-frequency noise and motion artifacts using Hitachi's data acquisition software. A bandpass filter saving the frequencies between .5 and .01 Hz was used to attenuate signatures representing respiratory and cardiac fluctuations. We used the modified Beer-Lambert law [14] to transform the resulting light intensity data into relative concentration changes of oxygenated hemoglobin (oxy-Hb) and deoxygenated hemoglobin (deoxy-Hb) [40]. We then designed an in-house software to extract the average oxy-Hb and average deoxy-Hb from each channel for each trial (15 second voice sample) presented in the experimental task.

The hemoglobin is called oxy-Hb when it transports oxygen to cerebral tissue and is called deoxy-Hb when it releases oxygen after metabolism. The difference in oxyHb and deoxyHb concentrations at the baseline (rest) and the task performance is said to determine the location in the cortex of where activities are happening [27].

The signal measured by each channel was related to the part of the brain above which the channel was placed. These fNIRS channels were virtually registered onto the stereotactic brain coordinate system using Tsuzuki's 3D-digitizer free method [39]. This method allowed us to place a virtual optode holder on the scalp by registering optodes and channels onto reference brains. We then identified the brain area represented by each channel and grouped these channels based on the majority of the Brodmann Area [9] they covered. We considered five broadman areas, namely, dorsolateral prefrontal cortex, orbitofrontal gyrus, frontopolar area, superior temporal gyrus and middle temporal gyrus, found to be activated in previous studies [25], [33], [34] for real-fake artifacts to measure differences in neural activities between real and fake voices. These areas are referred as our regions of interest (ROIs). As mentioned in III-B, the hypothesis of our study was that the areas related to decision making, trust, working-memory, and familiarity will have different activations for real and fake voices as well, similar to other real-fake artifacts, like paintings and websites. The oxy-Hb and deoxy-Hb measured by the channels in each group was then separately averaged for each trial.

B. Statistical Tests

We used IBM SPSS [36] for the purpose of statistical analysis reported in our study. Kolmogorov-Smirnov test used for normality detection revealed that oxy-Hb and deoxy-Hb data were non-normal. Thus, Friedman's test and Wilcoxon Singed-Rank Test (WSRT) were used for measuring differences in the means of different groups of trials underlying our analysis. Following the correction methodology adopted in [33], and since our analysis focused on pre-established ROIs per our hypotheses, the comparisons at each of the ROI were considered separately and were corrected using Holm-bonferroni correction [1]. We also report the effect size of WSRT which was calculated using the formula $r = Z/\sqrt{N}$, where Z is the value of the z-statistic and N is the number of observations on which Z is based. Cohen criteria [12] reports effect size $>$ .1 as small, $>$ .3 as medium and $>$ .5 as large.

V. Task Performance Results

To recall, in our experimental task, participants were asked to answer if the voice trial played was of the "real" speaker or a "fake" speaker. We had logged the participants' responses and response times during the experiment. A participant's response was marked as correct if she had identified the original speaker's voice as real, and the other speakers' (morphed and different speaker) voice as fake. We then calculated the average accuracy and response time (RTime) the participants spent on providing answers for each type of trial. Accuracy is defined as the ratio of the total number of correctly identified instances to the total number of samples presented to each participant.

From Table V, we observe that the overall accuracy of correctly identifying the voice of the speaker is around 64% which is only slightly better than the random guessing (50%). It seems highest for the original speaker and the lowest for

TABLE IV. REGIONS OF INTEREST (ROI): THE BRAIN AREAS COVERED BY OUR FNIRS PROBE-CAP

#	ROI Name	Acronym	Brodmann Area #	Functionality
2	Dorsolateral Prefrontal Cortex	DLPFC	9	Working memory, attention
3	FrontoPolar Area	FPA	10	Memory recall, executive functions
7	Superior Temporal Gyrus	STG	22	Primary auditory cortex, auditory processing
8	Middle Temporal Gyrus	MTG	21	Recognition of known faces
9	Orbitofrontal Area	OFA	11	Cognitive processing, decision making, trust

TABLE V. ALL SPEAKERS: ACCURACY (%), PRECISION (%), RECALL (%) AND F-MEASURE (%) AND RESPONSE TIME (SECONDS)

Trial	Acc	Prec	Rec	FM	RTime
	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)
Original	82.1 (16.6)	50.63 (12.5)	83.2 (16.1)	61.31 (9.0)	2.57 (0.5)
Morph	42.8 (24.1)	46.71 (19.7)	42.8 (24.1)	43.40 (19.8)	2.58 (0.5)
Different	67.2 (21.5)	47.82 (16.0)	68.1 (21.2)	55.82 (16.0)	2.51 (0.5)
Average	64.2 (11.5)	48.39 (15.3)	64.7 (20.7)	53.51 (15.0)	2.54 (0.5)

the morphed speakers. Also, we notice that the participants reported 58% of the morphed speakers and 33% of different speakers as real speakers. This shows that the morphed speakers were more successful than the different speakers in voice impersonation attacks. Our results are in line with the task performance results of voice impersonation attacks reported by Mukhopadhyay et al. [30].

The Friedman's test showed a statistically significant difference in mean accuracies across original, morphed and different speaker voices ($\chi^2(20)=17.71$, $p<.0005$). On further contrasting the accuracy rates across different types of trials with WSRT, we found that the participants identified original voices with a statistically significantly higher accuracy than morphed voices ($p<.001$) with large effect size ($r=.76$), and identified different speaker's voices with statistically significantly higher accuracy than morphed voices ($p<.0005$) with a large effect size ($r=.78$). We did not see other statistically significant results.

The users failing to identify morphed voices shows the quality of the converted voice generated by the morphing engine. The morphed voices might have sounded so similar to the original speaker's voice that the participants failed to identify them most of the times. This suggests that the attacker with a sophisticated tech speech morphing engine may successfully launch voice impersonation attacks on users, which is a concern to the security and privacy community.

VI. NEURAL RESULTS

In this section, we analyze the neural activations when users are listening to the original, morphed, and different speakers voice with the baseline (rest condition), and with respect to each other.

A. Voice Trial vs. Rest Trial

To recall, in our experimental task, participants were instructed to identify the voice of the speaker when played back and to relax when the rest sign was displayed. To evaluate the brain areas activated when participants were listening to original, morphed and different speakers, we contrasted the brain activation during these trials with the rest trial as a ground truth. This analysis provided the neural signatures of detecting the legitimacy of speakers.

We ran Wilcoxon-Signed Rank Tests (WSRT) to evaluate the differences in mean oxy-Hb and mean deoxy-Hb at each regions of Interest (ROIs) between the original trial vs. the rest trialVI. We found statistically significant differences in oxy-Hb at the dorsolateral prefrontal cortex, frontopolar area, superior temporal gyrus and middle temporal gyrus for the original speaker trial than the rest trial (such differences are listed in Table VI, rows 1-4 and Figure 5(a)). Statistically significant differences in deoxy-Hb at the dorsolateral prefrontal cortex, frontopolar area, middle temporal gyrus and orbitofrontal area for the original vs. rest trial (Table VI, rows 5-8) was also observed.

TABLE VI. NEURAL ACTIVATIONS: ORIGINAL SPEAKER VS. REST

#	ROI-Type	Hb-Type	p-value	Effect Size
1	DLPFC	oxy	.009	.60
2	FPA	oxy	.008	.59
3	STG	oxy	.002	.70
4	MTG	oxy	.045	.44
5	DLPFC	deoxy	.035	.47
6	FPA	deoxy	.000	.87
7	MTG	deoxy	.000	.82
8	OFA	deoxy	.022	.51

TABLE VII. NEURAL ACTIVATIONS: MORPHED SPEAKER VS. REST

#	ROI-Type	Hb-Type	p-value	Effect Size
1	DLPFC	oxy	.035	.47
2	FPA	oxy	.002	.73
3	STG	oxy	.024	.50
4	FPA	deoxy	.007	.60
5	STG	deoxy	.041	.45
6	MTG	deoxy	.000	.88

Similarly, on contrasting changes in mean-oxy and mean-deoxy at different ROIs between the morphed speaker and the rest trial using WSRT, we noticed statistically significant differences in oxy-Hb at the dorsolateral prefrontal cortex, frontopolar area, and superior temporal gyrus (Table VII, rows 1-3 and Figure 5(b)). Also, at frontopolar area, superior temporal gyrus, and middle temporal gyrus, statistically significant differences in deoxy-Hb were observed for the morphed speaker than in the rest trial (Table VII, rows 4-6).

Applying WSRT to measure differences in mean oxy-Hb

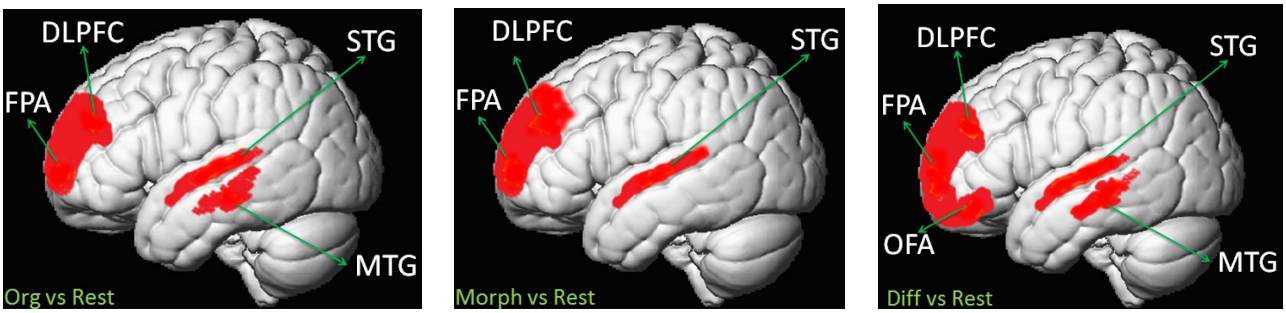

Fig. 5. Activation regions with statistically significant oxy-Hb changes: (a) Original vs Rest; (b) Morphed vs Rest; (c) Different vs Rest.

TABLE VIII. NEURAL ACTIVATIONS: DIFFERENT SPEAKER VS. REST

#	ROI-Type	Hb-Type	p-value	Effect Size
1	DLPFC	oxy	.007	.60
2	FPA	oxy	.001	.71
3	STG	oxy	.000	.83
4	MTG	oxy	.004	.63
5	OFA	oxy	.042	.45
6	DLPFC	deoxy	.027	.49
7	FPA	deoxy	.001	.73
8	MTG	deoxy	.000	.81
9	OFA	deoxy	.046	.44

and mean deoxy-Hb between the different speaker trial and the rest trial at various regions of interest revealed statistically significant differences in oxy-Hb at dorsolateral prefrontal cortex, frontopolar area, superior temporal gyrus, middle temporal gyrus and orbitofrontal gyrus for different speaker trial than in the rest trial (Table VIII, rows 1-5 and Figure 5(c)). It also rendered statistically significant differences in deoxy-Hb at dorsolateral prefrontal cortex, frontopolar area, middle temporal gyrus, and orbitofrontal area for the different speaker trial compared to the rest trial (Table VIII, rows 4-6).

Interpretation: Listening, attention and decision-making are critical components of higher cognitive function. The detection of voice impersonation attacks in our study involves all these elements in helping participants figure out the original and the fake voices of the speakers. The statistically significant activity in the dorsolateral prefrontal cortex and the frontopolar area at the neural level (Tables VI, VIII and VII) is indicative of the involvement of working memory and executive cognitive functions in this task. The previous studies [5], [13] have found interaction among the dorsolateral prefrontal cortex, frontopolar area and orbitofrontal area to play a critical role in conflict-dependent decision-making. The activation in orbitofrontal area portrays that the participants were being suspicious while identifying the different voices presented to them. They were sometimes trusting the voice of the real speaker as real and sometimes distrusting the voice as fake. A study by Dimoka et al. [15] found that trust was associated with lower activation in orbitofrontal area. The activation in orbitofrontal cortex is observed only when users are listening to voice of different speaker indicating that they were being suspicious when they were asked to identify different speaker's voice sample. This level of suspicion is also reflected in the behavioral results where participants were able to detect the different speakers much better than the morphed speakers. The activation in superior temporal gyrus and middle temporal gyrus, related to auditory processing [43], shows that the participants were carefully processing the voice of the speakers to decide their legitimacy. This shows that the participants were actively trying to decide if the given voice was real or fake.

B. Speaker Legitimacy Analysis

Now we present the results of contrasting neural activations between the original speaker and fake speakers (i.e., the morphed speaker and the different speaker). These comparisons of the brain activities delineate the brain areas involved in processing the voices of original and morphed speakers (synthesized voices), and original and different speakers. To recall, the different speaker scenario is used as a baseline to study the morphed speaker scenario.

Contrast 1: Original Speaker vs. Morphed Voice: This analysis provides an understanding of how the original speaker's voice and morphed speaker's voices are perceived by the human brain. To recall, our experimental task had four victim speakers. All these speakers were familiarized to participants during the experiment. In this analysis, we examined the neural activities when participants were listening to all original speakers and all morphed speakers. For the same, we ran WSRT at different ROIs to evaluate the differences in mean oxy-Hb, and deoxy-Hb between original and morphed voice. However, we did not observe any statistically significant differences.

Contrast 2: Original Speaker vs. Different Speaker: In this analysis, we compared the neural metrics when participants were listening to the voice of original speaker vs. the voice of a different speaker. We hypothesized that the original speakers – since they were familiarized to participants – will produce different neural activations than the different speakers.

For this analysis, we applied WSRT to contrast the mean oxy-Hb and mean deoxy-Hb at different ROIs between all the voices of original speaker and the voices of different speaker. It revealed statistically significant differences in oxy-Hb for original speaker than different speaker at superior temporal gyrus ($p=.029$) with medium effect size ($r=.48$). These differences are visualized in Figure 6.

We also applied WSRT to contrast the neural activity for original and different speakers' voices only corresponding to the samples correctly identified by the participants per their behavioral response. We observed statistically significant differences at superior temporal gyrus ($p<.05$).

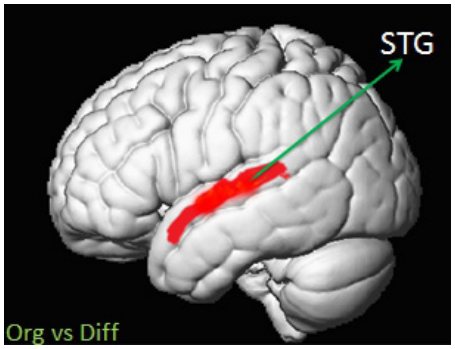

Fig. 6. Statistically significantly higher oxy-Hb observed in superior temporal gyrus (STG) when participants were listening to the voice of original speakers compared to the voice of the different speaker.

Interpretation of Contrasts 1 and 2: Based on the results of prior real vs. fake website/painting detection studies [25], [32]–[34], we expected to see differences in neural activity corresponding to real vs. morphed voices. However, contrary to our hypothesis, our results do not reveal such differences, which suggests that the original voices may have sounded identical to the morphed voices. This insight is also captured in our behavioral task performance analysis (Section V), and serves well to explain the users' susceptibility to voice morphing attacks.

Unlike original vs morphed analysis, we observed statistically significant differences in neural activities at superior temporal gyrus when users were listening to voices of original speakers vs. different speakers. The superior temporal gyrus is found to be activated for familiar voices in previous neuroscience studies [7] (the original speaker voices are the familiar voices in our case). Given this difference in neural activation, it is justified that the users were able to identify the different speakers in our voice impersonation attacks a large majority of times (Section V).

Overall, we see that users' brain activation explains why people were not able to detect the morphed speakers while they could detect the different speaker reasonably well, as shown in the task performance results.

C. Gender-based Speaker Legitimacy Analysis

Since gender may play some role in the detection of original, different and morphed voices, we pursued gender-centric analysis from our dataset. In this analysis, we divided our datasets into two categories based on the gender of the participants – male and female. We had 10 male and 10 female participants in our study. We first performed analysis on each category of participants to understand how they react to real, morphed and different voices compared to rest. Similar to our analysis in Section VI-A, on using WSRT, we observed differences in oxy-Hb in dorsolateral prefrontal cortex, frontopolar area, middle temporal gyrus and superior temporal gyrus between original voice trails and rest trials (all p-values less than $p<.05$) for both male and female participants. We also observed differences in oxy-Hb at dorsolateral prefrontal cortex, frontopolar area, and superior temporal gyrus between morphed and rest for both male and female participants (all p-values less than $p<.05$). Similarly, for different speakers' voice trials compared to rest trials on WSRT, we observed differences in oxy-Hb in dorsolateral prefrontal cortex, frontopolar area, orbitofrontal area, middle temporal gyrus, and superior temporal gyrus (all p-values less than $p<.05$). These areas of brain are activated in decision making, familiarity analysis, and real-fake judgment (detailed interpretation presented in Section VI-A).

We also compared the neural activation when female participants were listening to real and morphed voices of a speaker and did not observe any statistically significantly different result. Similarly, we did not observe any statistically significant difference in neural activation when male participants were listening to real and morphed voices. This suggests that users maybe biologically susceptible to fake voices irrespective of their gender.

D. Familiarity-based Speaker Legitimacy Analysis

To recall, we had polled the participants' about their familiarity with the two famous celebrities in our pre-test questionnaire. We compared the brain activities when participants were listening to the famous speakers' original samples to the briefly familiar speakers' original samples. This helped us understand the differences in neural activations when participants heard long-term familiarized voices vs. briefly-familiarized voices. We performed WSRT in mean oxy-Hb and deoxy-Hb between the original trials of the famous speakers and the briefly familiar speakers at different ROIs. We observed statistically significantly higher deoxy-Hb at frontopolar area ($p=.031$) with medium effect size ($r=.48$) and middle temporal gyrus ($p=.005$) with large effect size ($r=.63$). These differences have been depicted in Figure 7.

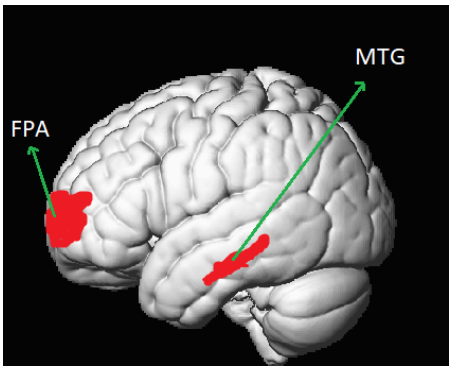

Fig. 7. Statistically significantly higher deoxy-Hb observed in middle temporal gyrus (MTG) and frontopolar area (FPA) when participants were listening to the voice of familiar speaker compared to the unfamiliar speaker.

Interpretation: The famous speakers were previously known to the participants, so their voice prints might have resided in the long-term memory. The higher activation in frontopolar area for famous speakers compared to briefly familiar speakers showed that the users were using their memory to identify the original voice samples. Similarly, we also saw higher activation in middle temporal gyrus for famous speakers voice in comparison to briefly familiar speakers. The middle temporal gyrus has been found to be more activated for familiar voices in previous studies [7]. Overall, these results illustrate that the human brain processes familiar voices differently from the unfamiliar voices.

VII. Neural Analytics: Original vs. Morphed Classification

In the previous section, we observed that the neural activity in original vs. morphed speakers was not statistically significantly different. In this section, to further validate this result, we show that the machine-learning on neural data can also not help classify these differences.

A. Feature and Performance Metrics

For extracting features, we normalized the oxy-Hb and deoxy-Hb data in each region of interest using z-score normalization. Next, at each ROI, we computed maximum, minimum, average, standard deviation, slope, variation, skew, and kurtoisis for the normalized oxy-Hb and deoxy-Hb data for each trial as features. We computed these features separately for the first and second half of each 15 second long task. We also separated the 15 seconds of data into 3 segments, and we took the average value of each of these 3 segments for the oxy-Hb and deoxy-Hb datastreams. We had one feature vector for each trial.

To build classification models, we utilized off-the-shelf machine learning algorithms provided by Weka. To this end, we used 10-fold cross-validation for estimation and validation of the models built on different algorithms including: *Trees* – J48, Logistic Model Trees (LMT), Random Forest (RF) and Random Tree (RT); *Functions* – Neural Networks, Multilayer Perceptron (MP), Support Vector Machines (SMO), Logistics (L) and Simple Logistic (SL), and *Bayesian Networks* – Naive Bayes (NB).

As performance measures, we report the precision ($Prec$), recall (Rec), F-measure (FM) or F1 Score for machine learning classification models. $Prec$ refers to the accuracy of the system in rejecting negative classes and measures the security of the proposed system. Rec is the accuracy of the system in accepting positive classes and measures the usability of the proposed system. Low recall leads to high rejection of positive instances, hence unusable, and low precision leads to high acceptance of negative instances, hence insecure. FM represents the balance between precision and recall.

True positive (TP) represents the total number of correctly identified instances belonging to the positive class while *true negative*(TN) is the number correctly rejected instances of negative class. Similarly, *false positive* (FP) is the number of negative instances incorrectly accepted as positive class and *false negative* (FN) represents the number of times the positive class is rejected.

B. Speaker Legitimacy Detection Accuracies

Our statistical analysis of neural data related to original speaker, morphed speaker voices showed some, albeit minimal, significant differences in oxy-Hb and deoxy-Hb. Taking this into account, we extracted features from these underlying neural differences and built a 2-class classifiers as discussed in Section VII-A to identify original and morphed speakers. In this classification task, the positive class corresponds to original speaker and negative class corresponds to the morphed. For the speaker legitimacy detection, we observed that none of the classifiers performed well. The best F-measure of identifying the voice of morphed speaker vs. original speaker obtained was 53% (see Table IX). To improve the results, we also evaluated the classification model with the best subset of features selected using correlation-based feature selection algorithm [22], and the best F-measure we achieved was 56.2%. We did not see statistically significant difference in these F-measures when compared to those of human detection and random guessing. These results show that, similar to human behavioral performance, even neural patterns and machine learning may not be successful to identify voice impersonation attacks.

TABLE IX. Speaker Legitimacy Detection: Precision, recall and F-measure (highlighted best classifier)

	Prec	**Rec**	**FM**
RandomTree	49.5 (9.5)	48.8 (10.9)	48.9 (9.7)
Logistic	49.4 (11.2)	50.0 (11.9)	49.4 (10.9)
J48	48.8 (10.7)	48.8 (12.1)	48.6 (11.2)
NaiveBayes	46.7 (10.1)	43.8 (11.2)	44.8 (9.5)
MultilayerPerceptron	50.0 (11.3)	48.8 (11.1)	49.0 (10.3)
LMT	48.3 (11.1)	54.4 (12.4)	50.8 (10.9)
SimpleLogistic	49.4 (10.3)	59.1 (13.5)	53.2 (10.1)
SMO	49.0 (12.6)	47.2 (12.9)	47.8 (12.0)
RandomForest	47.1 (9.5)	52.8 (11.4)	49.6 (9.8)

VIII. Discussion and Future Work

In this section, we summarize and discuss the main findings from our work, outline the strengths/limitations of our study and point to future work. Table X provides a summary of our overall results.

Summary and Insights: The participants in our study showed increased activation in dorsolateral prefrontal cortex, frontopolar cortex and orbitofrontal gyrus,the areas associated with decision-making, working memory, memory recall and trust while deciding on the legitimacy of the voices of speakers compared to the rest trials. They also showed activation in superior temporal gyrus, which is the region that processes the auditory signals. Overall, these results show that the users were certainly putting a considerable effort in making real vs. fake decisions as reflected by their brain activity in regions correlated with higher order cognitive processing. However, our behavioral results suggests that users were not doing well in identifying original, morphed and different speakers' voices. Perhaps the poor behavioral result was because the participants were unaware of the fact that the voices could be morphed (a real-world attack scenario). Another reason could be that the quality of voice morphing technology is very good in capturing features of the original speaker/victim.

We also analyzed the differences in neural activities when participants were listening to original voice and morphed voice of a speaker. However, we did not see any statistically significant differences in the activations in brain areas reported in previous real–fake studies [34]. Although the lack of statistical difference does not necessarily mean that the differences do not exist, our study confirms that they may not be present always, if at all present. Our task performance results also showed that people were nearly as fallible to the morphed voice as they were to the real voices. The results show that the human users may be inherently incapable of distinguishing between real and

TABLE X. RESULTS SUMMARY: NEURAL FINGERPRINTS OF SPEAKER LEGITIMACY DETECTION, AND CORRESPONDING TASK PERFORMANCE AND MACHINE LEARNING (ML) RESULTS.

Task	Condition	Activation Regions	Task Result	ML Result	Implications
Voice Trial vs Rest Trial	Original vs. Rest	DPLFC; FPD; STG; MTG; OFA	N/A	N/A	Regions associated with working memory, decision-making, trust, familiarity, and voice processing are all activated during the experimental task.
	Morphed vs. Rest	DPLFC/ FPA; STG; MTG	N/A	N/A	
	Different vs. Rest	DPLFC; FPA; STG; MTG; OFA	N/A	N/A	
Speaker Legitimacy Detection	Original vs. Morphed	No difference	43%	53%	Both users and their brains seem to fail at detecting voice morphing attacks
	Original vs. Different	STG	55%	N/A	Brain differentiates original speaker and different speakers voice

morphed voices. Consequently, they may need to rely on other external mechanisms to help them perform this differentiation (e.g., automated detection tools). Nevertheless, even current voice biometric solutions have been shown vulnerable to voice impersonation attacks [30].

In line to this, we built and tested an automated mechanism to identify original and morphed voice extracting features from neural data corresponding to each of the two types of voices. We found that it could only predict the voice correctly with the accuracy slightly better than random guessing. This shows that both the explicit responses of the users and the implicit activity of their brains are indicative that morphed voices are nearly indistinguishable from the original voices. This neuro-behavioral phenomenon serves to show that users would be susceptible to voice imitation attacks based on off-the-shelf voice morphing techniques. We believe this to be an important finding given the already emergence of voice imitation attacks. More advanced state-of-the-art voice modeling techniques than the one we used in our study, such as those offered by Lyrebird and Google WaveNet [29], [41], could make people even more susceptible to voice imitation attacks.

This same finding also bears positive news in other domains. The fact that "listeners" can not differentiate between the original voice of a speaker from a morphed voice of the same speaker, at both neural and behavioral level, seems to suggest that current morphing technology may be ready to serve those who have lost their voices.

In our study, the participants were not trained to look for specific forms of fabrications in the fake voice samples. Such an explicit training of users to detect fabrications could make an impact on users' ability to detect fabricated voices. Future studies should be conducted to analyze the effect of such training against users' performance in identifying morphed voices. This will be an interesting avenue for further research.

Study Strengths and Limitations: In line with any other study involving human subjects, our study had certain limitations. The study was conducted in a lab setting. So, the performance of the users might have been affected since they might not have been concerned about the security risks. Even though we tried to emulate a real-world scenario of computer usage in a lab setting and used a light-weight fNIRS probe caps designed for the comfort level, performing the task with the headset on might have affected the performance of some of the participants. In real-world voice impersonation attacks, the participants are not explicitly told to identify the real and fake voices of the speaker unlike our study. However, this could actually be seen as a strength of our work. Our results show that the participants were unable to detect these attacks despite being asked explicitly, and hence the result in the real-world attack may be even worse, where the users have to make the decision implicitly. Also, the participants in our study were mostly young, so the results of the study may not represent the success rate of voice spoofing attacks against older people or people with declination in hearing ability (the attack success rates against such populations may actually be higher in practice [16]). However, our sample size and diversity is well-aligned with those of prior studies (e.g., [33], [34]). Also, our participants belonged to the broader university community, including students, employees, and others.

One limitation of our study pertains to the number of trials. We asked the users to identify forty eight voice samples, each presented for 16 seconds, in about thirty minutes of the study. Although multiple long trials are a norm in neuroimaging studies for desired statistical power [2], the users may not have to face these many challenges in a short span of time in real-life. Another limitation pertains to the fNIRS device we used for the study. fNIRS captures the brain activities mostly close to the cortex of the brain. So, it might have missed the neural activities in the inner core of the brain. The users' decision making process towards the end of 15 seconds might also not be captured by fNIRS. It is an inherent limitation in measuring the BOLD signal. Finally, the fact that the voices of the speakers, Oprah and Freeman used in our study were distinctive and familiar to the speakers, might have confounded the neural activity in the frontal cortex. Future studies might be needed to explore a more realistic task set-up. Nevertheless, we believe that our work provides a sound first study of vocal security from a neuro-physiological perspective with important take-aways and insights that future studies can build upon.

IX. CONCLUDING REMARKS

In this paper, we explored voice security through the lens of neuroscience and neuroimaging, running an fNIRS study of human-centered speaker legitimacy detection. We dissected the brain processes that control people's responses to a real speaker's voice, a different speaker's voice, and a morphed

voice. We showed that there are differences in neural activities when users are listening to real vs. different speakers' voices. However, we did not notice significant differences when users were subject to real vs. morphed voices, irrespective of their behavioral response. We believe that this key insight from our work helps justify the users' susceptibility to morphing attacks as also demonstrated by our task performance results as well as prior studies.

ACKNOWLEDGMENT

We would like to thank Micah Sherr (our shepherd) and NDSS'19 anonymous reviewers for their constructive comments and guidance. We are also grateful to Rajesh Kana, Abhishek Anand, Maliheh Shirvanian and Kiavash Satvat for critiquing previous drafts of the paper. This work has been funded in part by a Department of Justice (DOJ) GRF-STEM fellowship.

REFERENCES

[1] H. Abdi, "Holms sequential bonferroni procedure."

[2] E. Amaro and G. J. Barker, "Study design in fmri: basic principles," *Brain and cognition*, vol. 60, no. 3, pp. 220–232, 2006.

[3] S. R. Arnott, C. A. Heywood, R. W. Kentridge, and M. A. Goodale, "Voice recognition and the posterior cingulate: an fmri study of prosopagnosia," *Journal of neuropsychology*, vol. 2, no. 1, pp. 269–286, 2008.

[4] X. Bai, L. Xing, N. Zhang, X. Wang, X. Liao, T. Li, and S. M. Hu, "Staying secure and unprepared: Understanding and mitigating the security risks of apple zeroconf," in *2016 IEEE Symposium on Security and Privacy (SP)*, May 2016, pp. 655–674.

[5] J. V. Baldo and N. F. Dronkers, "The role of inferior parietal and inferior frontal cortex in working memory." *Neuropsychology*, vol. 20, no. 5, p. 529, 2006.

[6] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, "Voice-selective areas in human auditory cortex," *Nature*, vol. 403, no. 6767, pp. 309–312, 2000.

[7] A. Bethmann, H. Scheich, and A. Brechmann, "The temporal lobes differentiate between the voices of famous and unknown people: an event-related fmri study on speaker recognition," *PloS one*, vol. 7, no. 10, p. e47626, 2012.

[8] Bonnie Brinton Anderson and C. Brock Kirwan and Jeffrey L. Jenkins and David Eargle and Seth Howard and Anthony Vance, "How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study," in *ACM Conference on Human Factors in Computing Systems, CHI*, 2015.

[9] K. Brodmann, *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth, 1909.

[10] S. C. Bunce, M. Izzetoglu, K. Izzetoglu, B. Onaral, and K. Pourrezaei, "Functional near-infrared spectroscopy," *IEEE engineering in medicine and biology magazine*, vol. 25, no. 4, pp. 54–62, 2006.

[11] B. Chance, Z. Zhuang, C. UnAh, C. Alter, and L. Lipton, "Cognition-activated low-frequency modulation of light absorption in human brain." *Proceedings of the National Academy of Sciences*, vol. 90, no. 8, pp. 3770–3774, 1993.

[12] J. Cohen, "Statistical power analysis for the behavioral sciences (revised ed.)," 1977.

[13] C. E. Curtis and M. D'Esposito, "Persistent activity in the prefrontal cortex during working memory," *Trends in cognitive sciences*, vol. 7, no. 9, pp. 415–423, 2003.

[14] D. T. Delpy, M. Cope, P. van der Zee, S. Arridge, S. Wray, and J. Wyatt, "Estimation of optical pathlength through tissue from direct time of flight measurement," *Physics in medicine and biology*, vol. 33, no. 12, p. 1433, 1988.

[15] A. Dimoka, "What does the brain tell us about trust and distrust? evidence from a functional neuroimaging study," *Mis Quarterly*, pp. 373–396, 2010.

[16] J. R. Dubno, D. D. Dirks, and D. E. Morgan, "Effects of age and mild hearing loss on speech recognition in noise," *The Journal of the Acoustical Society of America*, vol. 76, no. 1, pp. 87–96, 1984.

[17] "Bellesouth: Facebook scammers use voice-imitation to prey on users relatives." http://bellesouthblogs.com/facebookscam/. , 2012, accessed: 5-12-2018.

[18] "Festival," http://www.cstr.ed.ac.uk/projects/festival/, 2017, accessed: 5-12-2018.

[19] "Festvox," http://festvox.org/, 2014, accessed:05-11-2018.

[20] E. Formisano, F. De Martino, M. Bonte, and R. Goebel, "" who" is saying" what"? brain-based decoding of human voice and speech," *Science*, vol. 322, no. 5903, pp. 970–973, 2008.

[21] "Grandparents scam," http://www.michigan.gov/ag/0,4534,7-164-18156-205169–,00.html, 2017, accessed: 5-12-2018.

[22] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[23] "Hemodynamic response," https://en.wikipedia.org, 2017, accessed: 5-12-2018.

[24] L. M. Hirshfield, R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, R. Ward, T. Williams, and R. Jacob, "This is your brain on interfaces: enhancing usability testing with functional near-infrared spectroscopy," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 373–382.

[25] M. Huang, H. Bridge, M. J. Kemp, and A. J. Parker, "Human cortical activity evoked by the assignment of authenticity when viewing works of art," *Frontiers in human neuroscience*, vol. 5, 2011.

[26] K. Izzetoglu, S. Bunce, B. Onaral, K. Pourrezaei, and B. Chance, "Functional optical brain imaging using near-infrared during cognitive tasks," *International Journal of human-computer interaction*, vol. 17, no. 2, pp. 211–227, 2004.

[27] J. León-Carrión and U. León-Domínguez, "Functional near-infrared spectroscopy (fnirs): principles and neuroscientific applications," *Neuroimaging methods. Rijeka, Croatia: InTech (2012): 47-74*, 2012.

[28] K. Lewison and F. Corella, "Backing rich credentials with a blockchain pki," 2016.

[29] "Lyrebird," https://lyrebird.ai/, 2017, accessed: 5-12-2018.

[30] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines," in *European Symposium on Research in Computer Security*. Springer, 2015, pp. 599–621.

[31] K. Murphy and H. Garavan, "An empirical investigation into the number of subjects required for an event-related fmri study," *Neuroimage*, vol. 22, no. 2, pp. 879–885, 2004.

[32] A. Neupane, M. L. Rahman, N. Saxena, and L. Hirshfield, "A multimodal neuro-physiological study of phishing detection and malware warnings," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 479–491.

[33] A. Neupane, N. Saxena, and L. Hirshfield, "Neural underpinnings of website legitimacy and familiarity detection: An fnirs study," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 1571–1580.

[34] A. Neupane, N. Saxena, K. Kuruvilla, M. Georgescu, and R. Kana, "Neural signatures of user-centered security: An fMRI study of phishing, and malware warnings," in *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2014, pp. 1–16.

[35] M. Plichta, M. Herrmann, C. Baehne, A.-C. Ehlis, M. Richter, P. Pauli, and A. Fallgatter, "Event-related functional near-infrared spectroscopy (fnirs): are the measurements reliable?" *Neuroimage*, vol. 31, no. 1, pp. 116–124, 2006.

[36] I. Released, "Ibm spss statistics for windows. 20." *Armonk, NY: IBM Corp*, 2013.

[37] B. R. Rosen, R. L. Buckner, and A. M. Dale, "Event-related functional mri: past, present, and future," *Proceedings of the National Academy of Sciences*, vol. 95, no. 3, pp. 773–780, 1998.

[38] M. Shirvanian and N. Saxena, "Wiretapping via mimicry: short voice imitation man-in-the-middle attacks on crypto phones," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 868–879.

[39] D. Tsuzuki and I. Dan, "Spatial registration for functional near-infrared spectroscopy: from channel position on the scalp to cortical location in individual and group analyses," *Neuroimage*, vol. 85, pp. 92–103, 2014.

[40] A. Villringer and B. Chance, "Non-invasive optical spectroscopy and imaging of human brain function," *Trends in neurosciences*, vol. 20, no. 10, pp. 435–442, 1997.

[41] "Google wavenet," https://cloud.google.com/text-to-speech/docs/wavenet, 2017, accessed: 8-3-2018.

[42] H. Ye and S. Young, "High quality voice morphing," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–9.

[43] R. J. Zatorre, A. C. Evans *et al.*, "Lateralization of phonetic and pitch discrimination in speech processing," *Science*, vol. 256, no. 5058, p. 846, 1992.