“Hello, It’s Me”: Deep Learning-based Speech Synthesis Attacks in the Real World

Emily Wenger∗
ewenger@uchicago.edu
University of Chicago

Max Bronckers
mbronckers@uchicago.edu
University of Chicago

Christian Cianfarani
crc@uchicago.edu
University of Chicago

Jenna Cryan
jennacryan@uchicago.edu
University of Chicago

Angela Sha
angelasha@uchicago.edu
University of Chicago

Haitao Zheng
htzheng@uchicago.edu
University of Chicago

Ben Y. Zhao
ravenben@uchicago.edu
University of Chicago

ABSTRACT
Advances in deep learning have introduced a new wave of voice synthesis tools, capable of producing audio that sounds as if spoken by a target speaker. If successful, such tools in the wrong hands will enable a range of powerful attacks against both humans and software systems (aka machines). This paper documents efforts and findings from a comprehensive experimental study on the impact of deep-learning based speech synthesis attacks on both human listeners and machines such as speaker recognition and voice-signin systems. We find that both humans and machines can be reliably fooled by synthetic speech, and that existing defenses against synthesized speech fall short. These findings highlight the need to raise awareness and develop new protections against synthetic speech for both humans and machines.

CCS CONCEPTS
- Computing methodologies → Machine learning; • Security and privacy → Biometrics.

KEYWORDS
neural networks; speech synthesis; biometric security

ACM Reference Format:
Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y. Zhao. 2021. “Hello, It’s Me”: Deep Learning-based Speech Synthesis Attacks in the Real World. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS ’21), November 15–19, 2021, Virtual Event, Republic of Korea. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3460120.3484742

1 INTRODUCTION
Our voice conveys so much more than the words we speak. It is a fundamental part of our identity, often described as our “auditory face” [22]. Hearing our voice is often enough for a listener to make inferences about us, such as gender appearance [63], size or strength [73], approximate age [99], and even socioeconomic status [47].

But perhaps the human voice is no longer as unique as we would like to believe. Recent advances in deep learning have led to a wide range of tools that produce synthetic speech spoken in a voice of a target speaker, either as text-to-speech (TTS) tools that transform arbitrary text into spoken words [21, 36, 37, 41, 64, 76, 83, 92], or as voice conversion tools that reshape existing voice samples into the same content spoken by the target [42, 67, 69, 74, 95]. In addition to proprietary systems like Google Duplex, many others are available as open source software or commercial web services [9, 12].

Given the strong ties between our voices and our identities, a tool that successfully spoofs or mimics our voices can do severe damage in a variety of settings. First, it could bypass voice-based authentication systems (also called automatic speaker verification systems) already deployed in automated customer service phone-lines for banks and credit card companies (e.g., JP Morgan Chase and HSBC [3, 4]), as well as user login services for mobile messaging apps like WeChat [1]. It would also defeat user-based access controls in IoT devices such as digital home assistants (e.g., Amazon Alexa, Google Home) [7]. Finally, such tools could directly attack end users, by augmenting traditional phishing scams with a familiar human voice. This apparently was the case in a recent scam, where attackers used the mimicked voice of a corporate CEO to order a subordinate to issue an illegitimate money transfer [82].

These speech synthesis attacks, particularly those enabled by advances in deep learning, pose a serious threat to both computer systems and human beings. Yet, there has been – until now – no definitive effort to measure the severity of this threat in the context of deep learning systems. Prior work has established the viability of speech synthesis attacks against prior generations of synthesis tools and speaker recognition systems [28, 45, 56, 57]. Similarly, prior work assessing human vulnerability to speech synthesis attacks evaluates now-outdated systems in limited settings [57, 60].

We believe there is an urgent need to measure and understand how deep-learning based speech synthesis attacks impact two distinct entities: machines (e.g., automated software systems) and humans. Can such attacks overcome currently deployed speaker recognition systems in security-critical settings? Or can they compromise mobile systems such as voice-signin on mobile apps? Against human targets, can synthesized speech samples mimicking a particular human voice successfully convince us of their authenticity?

In this paper, we describe results of an in-depth analysis of the threat posed to both machines and humans by deep-learning speech
synthesis attacks. We begin by assessing the susceptibility of modern speaker verification systems (including commercial systems Microsoft Azure, WeChat, and Alexa) and evaluate a variety of factors affecting attack success. To assess human vulnerability to synthetic speech, we perform multiple user studies in both a survey setting and a trusted context. Finally, we assess the viability of existing defenses in defending against speech synthesis attacks.

All of our experiments use publicly available deep-learning speech synthesis systems, and our results highlight the need for new defenses against deep learning-based speech synthesis attacks, for both humans and machines.

**Key Findings.** Our study produces several key findings:

- Using a set of comprehensive experiments over 90 different speakers, we evaluate and show that DNN-based speech synthesis tools are highly effective at misleading modern speaker recognition systems (50 – 100% success).
- Our experiments find that given a handful of attempts, synthesized speech can mimic 60% of speakers in real world speaker recognition systems: Microsoft Azure, WeChat, and Amazon Alexa.
- A user survey of 200 participants shows humans can distinguish synthetic speech from the real speaker with ~50% accuracy for unfamiliar voices but near 80% for familiar voices.
- An interview-based deception study of 14 participants shows that, in a more trusted setting, inserted synthetic speech successfully deceives the large majority of participants.
- Detailed evaluation of 2 state-of-the-art defenses shows that they fall short in their goals of either preventing speech synthesis or reliably detecting it, highlighting the need for new defenses.

It is important to note that speech synthesis is intrinsically about producing audible speech that sounds like the target speaker to humans and machines alike. This is fundamentally different from adversarial attacks that perturb speech to cause misclassification in speaker recognition systems [25, 48, 53]. Such attacks do not affect human listeners, and could be addressed by developing new defenses against adversarial examples.

2 BACKGROUND

In this section, we first describe current trends in speaker recognition technology and voice synthesis systems, followed by voice-based spoof attacks. Finally, we briefly summarize defenses proposed to combat synthetic speech.

### 2.1 Voice-Based User Identification

**How Humans Identify Speakers via Voice.** The unique characteristics of each person’s vocal tract create their distinct voice. Humans use these vocal characteristics to identify people by voice [55]. Though human speaker identification is imperfect, it is highly accurate and has inspired the construction of speaker recognition systems for security purposes [75].

**Automated User Verification by Machines.** Recently, speaker recognition has become a popular alternative to other biometric authentication methods [71]. Speaker recognition systems capture characteristics of a speaker’s voice and compare them to enrolled speaker profiles. If there is a match, the recognition system grants the speaker access. Early speaker recognition systems (1970s-2010s) used parametric methods like Gaussian Mixture Models, while more recent systems (2014 onward) use deep learning models, which reduce overhead and improve accuracy [31, 70, 81, 85].

Speaker recognition is used in numerous settings, from bank customer identification to mobile app login and beyond [1, 3, 4]. Recently, virtual assistants like Alexa and Google Assistant have begun to use speaker recognition to customize system behavior [5, 7]. Speaker recognition systems are either text-dependent or text-independent [23, 35]. Text-dependent systems use the same, speaker-specific authentication phrase for both enrollment and login. Text-independent systems are content-agnostic.

### 2.2 Speech Synthesis Systems

Synthetic speech is produced by a non-human source (i.e., a computer) and imitates the sound of a human voice. Efforts to produce electronic synthetic speech go back to 1930s, where Homer Dudley developed the first vocoder [58]. Since then, systems like Festvox [20] have used Gaussian Mixture Models (GMM) to improve the quality but not the speed of speech synthesis. The recent deep learning revolution has catalyzed growth in this field.

**DNN-Based Speech Synthesis.** Numerous deep neural network (DNN) based speech synthesis systems have been proposed [21, 36, 37, 41, 42, 64, 66, 67, 69, 74, 76, 83, 92, 95]. They can be divided into two categories: text-to-speech (TTS) and voice conversion (VC).

TTS systems transform arbitrary text into words spoken in the voice of a target speaker [21, 36, 37, 41, 42, 76, 83, 92]. In contrast, VC systems take two voice samples – an attacker and target – and output a speech sample in which content from the attacker is spoken in the voice of the target [42, 66, 67, 69, 74, 95]. Both TTS and
In this work, we focus specifically on spoofing attacks against voice-based authentication and measurement studies on these attacks.

**Efficacy and Availability.** Many DNN-based speech synthesis systems report impressive speech "realism" metrics, indicating significant improvement over classical systems. Supporting evidence of DNN synthesis performance comes from real-world anecdotes. DNN-based synthetic speech has been successfully used at least one in highly profitable attack [82]. Google’s new scheduling assistant voice is so realistic that Google was instructed to announce when it was being used on phone calls [87].

Some DNN synthesis systems (and their training datasets) remain internal to companies, but many systems are available on Github [11, 13, 40, 65]. For the less tech-savvy, online services will perform voice cloning for a fee [9, 12]. This combination of speech synthesis efficacy and availability is both exciting and worrisome.

**Misuse of Speech Synthesis.** There are many positive uses for speech synthesis technology, such as giving voices to the mute, aiding spoken language translation, and increasing human trust of helper robots [19, 29, 49, 50, 84]. However, our work focuses on the “shadow side” of these uses – generating synthetic speech with malintent to deceive both humans and machines.

### 2.3 Voice-based Spoofing Attacks

In this work, we focus specifically on spoofing attacks against voice-based user identification, in which an attacker mimics a target’s voice to steal their identity. A parallel line of work explores adversarial attacks, in which an adversary adds inaudible perturbations to speech to fool speaker recognition systems [25, 48, 53]. While powerful, adversarial attacks differ from spoofing attacks because they do not mimic the target and so pose no threat to humans.

Figure 1 gives a high-level overview of spoofing attacks. There are several techniques the adversary could use, and these are taxonomized in Table 1. Prior work has found that all spoofing techniques – replay, impersonation, and synthesis – can reliably fool machine-based voice recognition systems, but only a few works have investigated the threat posed to humans. Here, we summarize prior work that studied these spoofing attacks.

**Spoofing Attacks Against Machines.** We first summarize prior work measuring machines’ vulnerability to spoofing attacks.

- **Human Impersonation:** Human voice actors can impersonate others’ voices to great success, and well-crafted impersonation spoofing attacks reliably fool speaker recognition systems [33, 34, 51, 78, 86]. These attacks have even defeated HSBC’s speaker recognition-based security [80]. While effective, these attacks have high overhead and limited versatility due to their dependence on human talent.

- **Machine Synthesis (Classical):** Most prior work uses GMM-based speech synthesis systems (e.g., Festvox [20]) to attack public, GMM-based speaker recognition systems [28, 45, 56, 57]. A recent work takes a “real-world” focus by testing a small set of synthetic speech generated by Festvox against five mobile apps that support voice authentication, and reports 96%+ success [79]. However, the efficacy of classical synthesis attacks against modern speaker recognition systems remains unclear.

- **Machine Synthesis (DNN-based):** To the best of our knowledge, only one work [62] has examined the performance of DNN-based synthesis attacks. It performed preliminary tests by running 10 synthesized samples for 6 speakers (generated by [41]) against three locally-trained speaker recognition prototypes. It produced vague conclusions: these speaker recognition prototypes produce more errors when running on synthesized speech compared to clean (non-synthesized) speech.

**Spoofing Attacks Against Humans.** Existing work assessing human susceptibility to spoofing only evaluates impersonation and classical synthesis attacks. The single impersonation attack paper found that humans can be fooled by actors pretending to be older or younger than they really are [34]. The first classical synthesis attack measurement paper [57] uses a traditional survey format and finds that users correctly distinguish between real and Festvox-synthesized voices (imitating the real speaker) about 50% of the time, regardless of their familiarity with the real speaker. A follow-up study to this [60] uses the same data and survey format but includes fNIRS brain scanning technology to measure participants’ neural activity. They find no statistically significant differences in neural activity when real or synthetic speakers are played.

**2.4 Defending Against Synthesized Speech**

Numerous defenses have been proposed to defend speech recognition systems against synthetic speech attacks. While most have focused on detecting synthetic speech or speakers [16, 17, 26, 32, 77, 90, 91, 98, 100], recent work has pointed towards a new defense direction: preventing unauthorized speech synthesis [38]. We discuss and evaluate representative defenses in §6.

### 3 METHODOLOGY

No comprehensive study exists today that studies the threat posed by DNN-based speech synthesis to software-based speaker recognition systems and human users. Our work addresses this critical need, and outlines future work needed to mitigate the resulting threat. Here, we describe the threat model, and the methodology, tools and datasets used by our analysis.

#### 3.1 Threat Model and Assumptions

In DNN-based speech synthesis attacks, the adversary $\mathcal{A}$’s goal is to steal a target $T$’s identity by imitating their voice. To do so, $\mathcal{A}$

| Attack Type | Description | Arbitrary Content | Measurement Studies |
|-------------|-------------|-------------------|---------------------|
| Replay      | Play pre-recorded speech from victim | No | [18, 39, 44, 96] |
| Human Impersonation | Human actor imitating victim | Limited | [33, 34, 51, 78, 86] |
| Synthesis (Classical) | Clone victim’s speech (GMM-based) | Yes | [28, 45, 56, 57, 79] |
| Synthesis (DNN) | Clone victim’s speech (DNN-based) | Yes | [62] | this work | this work |

Table I: Taxonomy of spoofing attacks against voice-based authentication and measurement studies on these attacks.
first collects a set of speech samples $S_T$ from $T$, either by secretly recording their speech in a public setting or by extracting the audio from public video/audio clips. When $A$ knows $T$ personally, these speech clips could also be obtained from private media. Next, $A$ inputs $S_T$ to a speech synthesis system, which produces synthesized or fake voice samples $S_A$. In this case, $S_A$ should sound like $T$ but contain arbitrary speech content chosen by $A$.

We make the following assumptions about the adversary $A$:

- $A$ only needs a small volume of speech samples from $T$, i.e., less than 5 minutes of speech in total;
- $A$ directly uses a publicly available, DNN-based voice synthesis system to generate synthetic speech $S_A$;
- $A$ seeks to generate fake voice samples $S_A$ that make either humans or machines believe that they are interacting with $T$.

### 3.2 Overview of Experiments

We conduct a measurement study to explore the real-world threat posed to both machines and humans by today’s publicly available, DNN-based speech synthesis systems. These include:

- Empirical experiments to examine whether synthetic speech can fool speaker recognition (SR) systems, aka machines (§4);
- User studies to explore human susceptibility to synthetic speech under multiple interactive scenarios (§5);
- Empirical experiments to assess the efficacy of existing defenses against DNN-based synthesis attacks (§6).

In the following, we describe the DNN synthesis and SR systems as well as speaker datasets used by our experiments.

### 3.3 DNN-based Synthesis Systems Studied

We consider “zero shot” systems (i.e., those requiring <5 minutes of targeted data to run synthesis) and focus on peer-reviewed, published papers with public code implementations and pre-trained models. We tested numerous synthesis systems including [21, 41, 42, 67, 74, 83], but found that many did not generalize well to unseen speakers (i.e., those not in the training dataset). Generalization is critical to low-resource attackers like the ones defined by our threat model, as it allows flexibility in target choice. In the end, we chose two systems that performed best on unseen speakers: SV2TTS [41], a TTS system based on Google’s Tacotron, and AutoVC [67], an autoencoder-based voice conversion system.

**SV2TTS.** This is a zero-shot, text-independent voice conversion system requiring only five seconds of target speech [41]. It combines three earlier works from Google: an LSTM speaker encoder [89], the DNN speech synthesis network Tacotron 2 and the WaveNet vocoder. Instead of training the model ourselves, we use a well-known public implementation [27]. In this case, the encoder is pre-trained on the VoxCeleb1/2 [59] and LibriSpeech train [61] datasets, and the synthesis network is pre-trained on LibriSpeech train, both following the original paper’s setting [41].

**AutoVC.** The second system is a zero-shot style transfer autoencoder network that performs text-independent voice conversion [67]. Its encoder bottleneck disentangles speaker characteristics from speech content to facilitate speech synthesis. Like SV2TTS, it relies on [89] and WaveNet for its speaker encoder and speech vocoder, respectively. We also use the publicly available implementation provided by [65], where the speaker encoder is pre-trained on VoxCeleb1 and LibriSpeech train, and the autoencoder is pre-trained on VCTK, again following [67]’s original setting.

### 3.4 Speaker Recognition (SR) Systems Studied

To explore the real-world threat of speech synthesis attacks on machines, we choose four state-of-the-art SR systems. These include both publicly available and proprietary systems.

**Resemblezer.** Resemblezer [6] is an open source DNN speaker encoder widely used by recent literature. It is trained on VoxCeleb 1/2 and LibriSpeech train using the generalized end-to-end loss [89]. Each speaker is enrolled into the system database using about 30 seconds of their speech, creating an embedding that represents their identity. To recognize an incoming speaker, the system compares the embedding and compares it to embeddings in the database using cosine similarity.

**Microsoft Azure.** Microsoft Azure’s cloud platform includes a text-independent speaker recognition API [10]. Speakers are enrolled using 20 seconds of speech data, and speaker verification queries are made via the API. This system is certified by several international bodies, e.g., the Payment Card Industry (PCI), HIPAA, and International Standards Organization (ISO).

**WeChat.** WeChat is a popular mobile messaging and payment platform that offers a text-dependent “voiceprint” login for authentication. Users create their voiceprint by repeating an eight digit number provided by the app. This same number is used for each subsequent voiceprint login. Users have a cap of six voice login attempts per day before the app enforces password authentication [15].

**Amazon Alexa.** Alexa is a popular virtual assistant embedded in Amazon’s smart speakers. Alexa uses “voice profiles” to customize user interactions and control access to sensitive apps like email and calendar [7]. Voice profiles control access to sensitive information in Alexa 3rd party apps, such as payment (Uber) and phone account management (Vodaphone) [2].

### 3.5 Speaker/Speech Datasets

We use four different speaker datasets to define target speakers $T$ and their speech samples $S_T$. The first three are commonly used speaker recognition datasets, and the last one is a custom dataset crafted for our experiments.

- **VCTK** [97] contains short spoken phrases from 110 English speakers with varied accents. Phrases are read from a newspaper, the Rainbow passage [30], and a dataset-specific paragraph.
- **LibriSpeech** [61] is derived from the open-source LibriVox audiobook project. We use the test-clean subset, which contains spoken phrases from 40 English speakers.
- **SpeechAccent** [94] contains the same set of spoken phrases (in English) from each of 2140 speakers. Speakers come from 177 different countries and represent 214 native languages.
- **Our Custom** dataset contains phrases from the Rainbow passage spoken in English by 14 speakers (details in §4.4). This dataset allows us to synthesize speech for real-world tests on WeChat and Alexa in §4.4.
3.6 Ethics
All of our user study protocols received approval from our local IRB board and were carefully designed to protect the privacy and well-being of our participants. We kept only audio recordings of the interview sessions, which were anonymized and stored on secure servers. Given our goal of bringing attention to the significance of this attack vector, we have also proactively contacted Microsoft Azure, WeChat, and Amazon to disclose our findings.

4 SYNTHESIZED SPEECH VS. MACHINES
We begin by asking “how vulnerable are machine-based SR systems to synthetic speech attacks?” While prior work has explored this question using classical (non-DNN) synthesis systems, the efficacy of DNN synthesis attacks against real-world SR systems remains unknown. In this section, we answer this question by evaluating the robustness of four modern SR systems (§3.4) to DNN-based synthesis attacks.

Specifically, our study consists of the following experiments:

- §4.2 attacks the widely used SR model (Resemblyzer), showing that DNN-based synthesis attacks reliably fool such systems.
- §4.3 and §4.4 attack three real-world SR deployments (Azure, WeChat, and Amazon Alexa), showing that all three are vulnerable to DNN-based synthesis attacks.

We measure attack performance using the attack success rate (AS), which denotes the average percent of synthesized samples identified as the target speaker. We design our experiments to not only evaluate the attack success rate against various SR systems, but also explore whether a target’s speech samples and personal attributes (e.g., gender/accent) impact the attack outcome.

4.1 Baseline: Classical Synthesis Attacks against Modern SR
As reference, we evaluate the efficacy of prior classical (non-DNN) synthesis attacks against today’s SR systems. A 2015 paper [57] demonstrated that synthetic speech created with Festvox [20] fooled the Bob Spear SR system [43] with > 98% success rate. We recreate this attack and find that it fails on more recent SR systems (Table 2). A detailed description of our experiments is in the Appendix.

| Experiment Methodology | Attack Success Rate (AS) |
|------------------------|--------------------------|
| Size of $S_T$ | Generate synthetic speech, varying number of target samples $N$ | Low AS when $N < 5$ |
| Quality of $S_T$ | Add Gaussian noise to target samples before speech synthesis | AS = 0 even for small noise |
| Phonetic Similarity | Vary the phonetic distance between target samples and synthesis output | No significant effect |
| Target Gender | Test male/female target speakers from VCTK and LibriSpeech separately | AS for female targets > AS for male targets |
| Target Accent | Test native/non-native English speakers from SpeechAccent dataset separately | AS for native English speakers > AS for non-native speakers |

Table 3: Overview of content- and identity-specific attribute experiments (and results) on Resemblyzer and Azure.

In total, our experiments consider 90 target speakers randomly chosen from three speaker datasets (20 randomly chosen from VCTK, 20 from LibriSpeech test−cLean, and 50 from SpeechAccent). For each target $T$, we use their speech sample set $S_T$ as input to either SV2TTS or AutoVC to produce fake voices of $T$ that contain arbitrary speech content chosen to mimic normal conversation (listed in Table 10 in the Appendix). Since AutoVC also requires a source recording, we choose the source of the same gender of the target speaker as suggested by [67]. For each test, we synthesize 10 spoken phrases per target speaker.

We note that Resemblyzer requires a threshold to detect whether two speech embeddings are from the same speaker. We configure this threshold by first enrolling the target speakers in Resemblyzer using their real speech samples, then computing their embeddings and choosing the threshold that minimizes Resemblyzer’s equal error rate (EER) on these speakers, using cosine similarity as the distance metric. When launching a synthesis attack, the attack is considered successful if the similarity between the attack and enrolled embeddings is above the threshold. For each attack, we repeat the enrollment process 10 times (using different speech samples) and report the average attack success rate and standard deviation.

Results. In total, we test 13,000 synthesized speech instances targeting 90 speakers on Resemblyzer. The results show that SV2TTS based attack is highly effective against Resemblyzer, while AutoVC is ineffective. The size and quality of $S_T$, the speaker gender and accent do impact the attack success rate, but the phonetic similarity has minimum impact. Next we report these results in more detail.

1) Attack success rate under the default setting: We start from “ideal” cases where the attacker targets native English speakers with US or British accents, and has plenty of high-quality speech samples per target. For this we consider target speakers from VCTK and LibriSpeech, and configure $S_T$ to include $N = 10$ utterances per target. As such, $S_T$ includes 30-120 seconds of clean audio, far more than the amount required to run zero-shot synthesis as claimed by [41, 67] (roughly 20 seconds). We hereby refer to $N$ as the number of target speech samples.

Next we test DNN-based speech synthesis attacks against Resemblyzer, a modern SR system widely used in recent literature. We use the official implementation of Resemblyzer provided by [6].
As shown in Figure 2a, fake speech synthesized by SV2TTS successfully fools Resemblyzer, while AutoVC fails. We think that the success of SV2TTS (particularly on LibriSpeech) is likely because Resemblyzer is trained using the same loss function used by SV2TTS’ speaker encoder [89].

2) Impact of \( T \) size: We repeat the above experiment but vary the size of target speech samples, i.e., \( N=1, 5, 10, 20, 30, 40 \) speech samples. As Figure 2b shows, the attack success rate for SV2TTS grows with \( N \) but levels off before \( N \) reaches 10. For AutoVC, the attack remains ineffective when varying \( N \).

3) Impact of \( S_T \) quality: This question has bearing on real-world attack settings, since an attacker might not always obtain high-quality audio recordings of the target. To emulate low-quality data, we add four different levels of zero-mean Gaussian noise to the original clean audio. We vary the signal-to-noise ratio from 4 dB (noise quieter than the speaker’s voice) to -15 dB (noise louder than the speaker’s voice). We find that noisy target speech samples decimate synthesis attack performance. For both SV2TTS and AutoVC, the attack success rate reduces to 0% at all four noise levels.

4) Impact of phonetic similarity between \( S_T \) and \( S_R \): This factor also has strong real-world implications – if the content similarity affects the attack success rate, the attacker may be largely constrained by which \( S_T \) they obtain. Since SV2TTS generates synthesized speech from arbitrary text, we use it to explore this question. Details of our experiments are in the Appendix.

Interestingly, we find that phoneme similarity of \( S_T \) and \( S_R \) does not have any visible effect – the attack success rate remains stable as we vary the normalized phoneme similarity from 0 to 1.

5) Impact of target gender: We now consider a target’s personal attributes which may affect the attack outcome. The first is speaker gender, which can come into play if, for example, synthesis or SR models lack sufficient gender diversity. To study this factor, we separate the results of our SV2TTS experiments by gender. We find that synthesized female speakers have higher AS on average than male speakers (Figure 2c). When we test clean (non-synthesized) speech from these target speakers on Resemblyzer, the SR accuracy is 100% for both male and female speakers.

6) Impact of target accent: Most public speech datasets consist of native English speakers with US or UK accents (i.e. VCTK, LibriSpeech, VoxCeleb 1/2). Speech synthesis systems trained on these datasets may fail to recreate the unique prosody of speakers with different accents. To test this, we choose 50 target speakers from the SpeechAccent dataset, including male/female native English speakers and male/female native speakers from the top 21 most spoken languages.

When comparing results from native/non-native English speakers, we observe a higher attack success rate among native English speakers (100%) compared to non-native English speakers (65%) for synthesized speech produced by SV2TTS. As before, attacks using synthesized speech from AutoVC are unsuccessful.

4.3 Azure (Open API, Real-World SR)

We run the same experiments of §4.2 on Azure, a real-world SR deployment. Azure’s open API allows us to enroll speakers and run numerous tests against their enrolled speaker profiles. But unlike §4.2, there is no need to configure any threshold. We enroll each of our 90 target speakers from 4.2 into Azure and use these enrolled profiles for all tests. We generate and test 10 synthesized phrases (as in 4.2) per target for each experiment. Since Azure reports the SR acceptance result per sample, we report the average success rate of all synthesized samples in each experiment.

Disclosure. We followed standard disclosure practices and reported the result of DNN-synthesized speech attacks to Microsoft.

Results. We test 13,000 synthesized speech instances targeting 90 speakers on Azure. These results show that Azure is also vulnerable to DNN-synthesized speech. Our findings on the impact of various factors mirror those from Resemblyzer.

1) Attack success rate under default settings: Figure 3a lists the overall attack success rate. We see that DNN-synthesized speech can easily fool Azure, although the attack success rate is less than those with Resemblyzer. Interestingly, for 62.5% of target speakers, at least 1 out of 10 synthesized phrases (generated by SV2TTS) was accepted by Azure as the target speaker. Thus a persistent attacker could make multiple attempts to eventually fool Azure API (assuming there is no limit on authentication attempts).

Another interesting finding is that the attack success rate displays significantly higher variance than those observed on Resemblyzer. This is particularly visible for SV2TTS. When we dig deeper to understand this high variance, we find that for the above mentioned 62.5% targets (with at least 1 success attack instance out of 10 trials), the attack success rate was 49.2 ± 23.5% for VCTK speakers and 33.1 ± 21.4% for LibriSpeech speakers. These results indicate that the attack performance against Azure is non-uniform across target speakers.

2) Impact of \( S_T \) size, quality, and phonetic similarity of \( S_T \) and \( S_R \): Our results from these experiments mirror those of Resemblyzer: Figure 3b shows that the attack performance levels off when \( N \) reaches 10; none of the speech synthesized from noisy versions of \( S_T \) was accepted by Azure; and the phonetic similarity of \( S_T \) and \( S_R \) does not affect the attack outcome.
4.4 WeChat and Amazon Alexa (Closed API, Real-World SR)

Finally, we experiment with two additional real-world SR systems: WeChat and Amazon Alexa. In contrast to Azure, both employ closed-API SR, largely limiting our experimental bandwidth. Since WeChat and Alexa’s SR systems link to individual accounts, we must test synthesis attacks with real users. Note that our goal is not to test the (in)security of WeChat or Alexa platforms, but to use them as case studies of deployed SR systems to illustrate the potential impact of DNN-based speech synthesis attacks.

Experimental Setup. We conduct an IRB-approved user study to evaluate synthetic speech attacks (IRB information omitted for anonymity). Specifically, we collect speech samples from study participants, synthesize speech imitating each participant, and give these speech samples to each participant to test their WeChat and Amazon Alexa apps. Given the poor performance of AutoVC on Resemblizer and Azure, we only use SV2TTS for these experiments.

We recruited 14 participants with different linguistic background (1 native Marathi speaker; 1 native Dutch speaker; 3 native Mandarin speakers; 9 native English speakers) and gender (10 female/4 male). All participants signed written consent forms to participate in our user study and were compensated $10 for their time. We asked our participants to submit a small set of their voice recordings. Each participant used a voice memo recording app to record themselves speaking 20 sentences in English from the Rainbow Passage. The Rainbow Passage is commonly used in linguistic studies because it contains most of the phoneme combinations in the English language [30] (the full passage is in the Appendix). In this study, 7 participants recorded on a Macbook Pro, 4 recorded on an iPhone 11+ phone, and 1 recorded on a Google Pixel phone.

For each participant, we use their submitted speech recordings as the target speech sample set \( \mathcal{S}_T \), and input them to SV2TTS to generate synthetic speech imitating \( \mathcal{T} \). The content of the synthetic speech \( \mathcal{S}_R \) is designed to match the context of the SR system, which we describe below (also see Table 4).

- **WeChat** uses a **text-dependent** speaker verification system that asks for stating the same eight-digit login number for each SR attempt. Each participant consented to share their unique login number with our user study administrators, and these were used to generate synthesized login speech samples. To ensure participant privacy and security, login numbers were password-protected, anonymized, and deleted when the study ended. For each participant, we generate six synthesized login samples.

- **Alexa** employs a **text-independent** speaker verification system, but its specific uses of voice profiles constrain the samples we can test. We create a short list of Alexa commands that Amazon explicitly states should be linked to a user’s voice profile, restricting our attention to native Alexa skills [7] (see Table 4).

After setting up their WeChat voice login and Alexa voice profile (using the Alexa smartphone app), all 14 participants verified that they could use their real voices to log into WeChat and access specified applications with Alexa. They were then given their synthesized speech samples (6 samples of login for WeChat, 7 samples of commands for Alexa) and instructed to play each sample over a computer loudspeaker located six inches from their phone microphone. Samples were played as the targeted apps were set up to perform normal voice authentication. Each participant tested the WeChat samples twice and the Alexa samples once. Participants recorded how the apps responded to the synthesized samples and reported their results via a standardized form.

**Attack Evaluation.** In total, our user study tested WeChat and Alexa with 168 and 98 synthesized speech instances, respectively. Again, we use attack success rate (AS) to evaluate how effectively synthesized speech can fool both SR systems. For WeChat, each attack instance is successful if the login is approved. For Alexa, we use a different approach because Alexa does not provide clear-cut success/failure results: a synthesis attack instance \( \mathcal{S}_R \) succeeds if Alexa responds to \( \mathcal{S}_R \) the same way it responds to a clean (non-synthesized) version of the command.

**Results.** On average, our speech synthesis attacks had a 63% AS across all tests on WeChat and Alexa.

**WeChat.** 9 of the 14 participants (64%) successfully logged into their WeChat account using a synthesized speech sample. In general, this indicates that speech synthesis attacks are a viable authentication attack against WeChat. However, the number of successful fake login samples varied significantly across participants (despite using the same setup for each test). On average, 1.33 ± 1.67
fake login samples succeeded per participant. For one participant, all six login samples worked. For 8 other participants who successfully logged in, only one or two samples consistently worked.

**Alexa**: Our attacks on Alexa were similarly successful (62.2% AS on average). All 14 participants had at least 2 synthesized commands that fooled Alexa. These synthesized commands were able to access private emails, check calendar appointments, and request financial transactions. Table 4 reports these results. Furthermore, in limited tests, we found that sometimes a wrong voice (i.e. from a different person) was able to access user data purportedly protected by voice profiles.

**Effect of Loudspeaker**: Our study participants reported the device used to play their attack samples. Devices used include LG desktop monitors, Macbook Pros, Bose Soundlink Speakers, and iPhone 11s. We examined the impact of speaker hardware on attack success and found no correlation between the two. While the tested devices already cover a broad set of speaker hardwares found in today’s households and offices, more experiments are needed to quantify if attack success depends on speaker quality.

**Disclosure**: We followed standard disclosure practices and reported our attacks to both WeChat and Amazon.

### 4.5 Key Takeaways

All four modern SR systems tested are vulnerable to DNN-based speech synthesis attacks, especially those generated by SV2TTS. It is alarming that for three popular real-world SR systems (Azure, WeChat, Alexa), more than 60% of enrolled speakers have at least 1 synthesized (attack) sample accepted by these systems. This clearly demonstrates the real-world threat of speech synthesis attacks.

Another key observation is that the attack performance is speaker-dependent, e.g., the number of synthesized samples that successfully fooled the SR systems varies across speakers. For Resemblyzer and Azure, the attack success rate is consistently higher for female and native English speakers.

**Limitations and Next Steps.** Our experiments, especially those on WeChat and Alexa, involved a moderately-sized set of target speakers to demonstrate the real-world threat of speech synthesis attacks. To further evaluate the attack dependence on target human speakers, we believe viable next steps include expanding the speaker pool and testing more operational scenarios. With these two changes, we could more closely examine how an individual’s vocal characteristics (e.g., pitch, accent, tone) affect the attack success rate, and whether their impact can be reduced by improving the underlying speech synthesis systems.

Similarly, due to our focus on low-resource attackers, our experiments used two publicly available speech synthesis systems (SV2TTS and AutoVC) that were trained only on publicly available datasets. These two systems will likely underperform advanced synthesis systems trained on larger, proprietary datasets, and consequently our reported results only offer a “conservative” measure of the threat. As speech synthesis systems continue to advance, the threat (and damage) of speech synthesis attacks will grow and warrant our continuous attention.

## 5 SYNTHESIZED SPEECH VS. HUMANS

Having demonstrated that DNN-synthesized speech can easily fool machines (e.g., real-world SR systems), we now move to evaluate their impact on humans. Different from prior work that uses surveys to measure human perception of speech synthesized by classical (non-DNN) tools [57, 60], we assess the susceptibility of humans to DNN-synthesized speech in different interactive settings. For this we conduct two user studies, covering both static survey and “trusted” interaction settings. Next, we describe the methodology behind our user studies and give a preview of our key findings, before presenting both studies in detail.

### 5.1 Methodology and Key Findings

An attacker $A$ can perform a variety of attacks against human listeners using synthesized speech. Such attacks can be particularly effective if the listener has limited familiarity with the owner of the spoofed voice. For example, $A$ could use a synthesized voice to perform the classic spear-phishing attack, where an elderly victim gets a phone call from their “grandchild” they haven’t seen in months who is stuck in a foreign country and needs emergency cash to get home, or an employee gets a call from their “boss” confirming an earlier (phishing) email authorizing a money transfer [82].

With these in mind, our study on the impact of synthesized voices on human listeners has two goals: understanding human listeners’ susceptibility to synthesized speech in isolation and in trusted contexts. We designed experimental protocols for both parts of our study, giving detailed consideration to issues of ethics and impact on our participants. All protocols were carefully evaluated and approved by our institutional IRB review board. We discuss ethical considerations in §7.

**User Study A (Online Survey).** We first evaluate whether humans can discern the difference between real and DNN-synthesized (fake) speech. We conduct an online user survey and compare how well participants could identify synthesized speech for voices with which they have varying levels of familiarity (e.g., strangers vs. celebrities). We also measure the effect of priming by comparing results from 2 scenarios: one in which participants are told the speech samples will contain a mix of real and fake speech, and one without the disclosure.

**Finding:** In this “survey” setting, DNN-synthesized speech fails to consistently fool humans. Participants could more easily distinguish between real and fake speech when they were more familiar with the speaker, and when they were aware that some speech may be fake (thus tended to listen carefully with added skepticism).

**User Study B (Deceptive Zoom Interview).** We seek to better understand the impact of context on listeners’ susceptibility to fake speech. To do so, we conduct interviews over Zoom calls. Participants believed they were speaking to two (human) researchers, but in reality one of the voices was synthesized speech.

| SR System | Phrases Used for SR | AS |
|-----------|---------------------|----|
| Alexa     | Hey Alexa add an event to my calendar for tomorrow at 5 PM | 64.3% |
|           | Hey Alexa check my email | 64.3% |
|           | Alexa say who is talking with you now | 55.7% |
|           | Alexa tell me what is on my calendar | 39.1% |
|           | Tell me what is on my calendar for this week | 88.5% |
|           | Alexa make an appointment with my doctor | 72.1% |
|           | Hey Alexa make a donation to the American Cancer Institute | 31.4% |

Table 4: The phrases used in our Amazon Alexa and WeChat experiments and the corresponding attack success rate (AS) on synthesized versions of each phrase.
Finding: In this “trusted setting,” all 14 participants showed no hesitancy or suspicion during the interview, and readily responded to and complied with all requests from the “fake interviewer.” In other words, a synthesized voice consistently fooled humans in this trusted interview setting.

5.2 User Study A: Can Users Distinguish Synthesized and Real Speech?

We begin our human perception experiments with the critical question: “Can human listeners distinguish synthetic speech of a speaker from the real thing?” We deploy a survey to assess users’ ability to distinguish between real and fake speakers.

Participants. We recruited 200 participants via the online crowd source platform Prolific (https://www.prolific.co/). All self-identified as native English speakers residing in the United States. Of our participants, 57% identified as female (43% male). The participants are all 18+ years old and cover multiple age groups: 18-29 (43%), 30-39 (32%), 40-49 (14%), 50-59 (8%), 60+ (3%). The survey was designed to take 10 minutes on average, and participants received $2 as compensation. The study was approved by our local IRB.

Procedure. The participants completed an online survey consisting of several speech samples presented in pairs for side-by-side comparison. Each pair of samples contains one of the three following combinations: two real speech samples of the same speaker (referred to as “Real A/Real A” in this section); one real speech sample from a speaker and one real speech sample from a different speaker (“Real A/Real B”); or one real speech sample from a speaker and one fake speech sample imitating the speaker (“Real A/Fake A”). We generate fake speech using SV2TTS, using 30 seconds of clean speech samples $S_f$ from the speaker.

Types of Speakers: We included speakers whose (real) voices have varying levels of familiarity with our participants:

- **Unfamiliar speakers:** Speakers from the VCTK [97] dataset whose voices have (most likely) never been heard by the participants.
- **Briefly familiar speakers:** Inspired by [57], we included a set of speakers whose voices the participants hear only briefly. For each speaker, we provided participants with a short audio clip to familiarize them with the speaker’s voice. There are four briefly familiar speakers, and for each one we provided a different length audio clip – 30 seconds for the first speaker, 60 seconds for the second, 90 seconds for the third, and 120 seconds for the fourth.
- **Famous speakers:** We used the voices of two American public figures: Donald Trump and Michelle Obama. We asked participants whether they have heard these voices outside the context of this survey, and over 90% responded “yes.”

Task. Participants listened to pairs of speech samples and reported if both samples were spoken by the same person.

Conditions. We deploy two versions of the survey. Both versions ask participants to assess the identity of the speaker and quality of speech samples. The first version does not mention fake speech at all. The second version of the survey mentions fake speech, both in its title and in its description of the task.

Results. We seek to answer the following questions:

1) Do participants think generated fake speech was spoken by the original speaker? As shown in Table 5 (bottom row), about half of participants were fooled, i.e. they responded “yes” or “not sure,” when asked this question about unfamiliar or briefly familiar speakers. For famous speakers whose voice participants are generally familiar with, this number drops to 20%.

2) Does hearing more samples from a speaker (i.e., knowing a speaker better) make fake speech more detectable? Results in Table 5 suggest that greater familiarity with a speaker will lead to increased skepticism of a fake voice. Compared to a similar user study performed 6 years ago [57], proportion of participants who correctly identified the fake voice for unfamiliar or briefly familiar speakers is consistent (50% for our work vs. 48% in [57]). However, participants in our survey were more accurate at identifying fake speech from famous speakers (80% vs. 50% in [57]), perhaps reflecting a higher general awareness of speech synthesis attacks.

3) Does mentioning fake speech in the survey description change participants’ perceptions of the fake speech samples? Mentioning fake speech in the survey description showed a statistically significant effect on survey responses. Figure 4 shows how the responses to the survey version mentioning fake speech reflect an apparent increased skepticism of fake speakers.

Using a chi-squared test for independence, we compared responses from each speaker familiarity category between the two survey versions to see if this change is statistically significant.

- For unfamiliar speakers, all but one speaker has a significant ($p < 0.05$) difference in responses.
- For somewhat familiar speakers, again all but one speaker has a significant ($p < 0.05$) difference in responses.
- For famous speakers, only Trump has a statistically significant difference in responses.

4) Do participant demographics (age, gender) affect responses? Women and younger people were more likely to correctly identify real and fake speakers. Using a chi-squared test for independence, we compared the responses of men vs. women and younger people (age $< 25$) to older people (age $> 45$). For unfamiliar speakers, there were statistically significant ($p < 0.05$) differences in response between genders and age groups. For somewhat familiar and famous speakers, the differences in response were not statistically significant.
speakers, statistically significant differences are observed for some, but not all, speakers.

5.3 User Study B: How Do Users Interact with Synthesized Speech in Trusted Settings?
Our user study A confirms that DNN-synthesized speech fails to consistently fool humans in a survey setting. Looking beyond this, we wonder how, if at all, the context in which users were exposed to fake speech influences their susceptibility to these attacks. Specifically, how would participants behave in a setting where they were predisposed not to think critically about the voices they hear? Examples of such “trusted settings” include phone or Zoom meetings with colleagues or calls with one or more people they know (or think they know). Human behavior in these so-called “trusted settings” may differ from behavior in a survey-based setting. When humans are primed by their setting to think they are speaking to a real person, they may be more likely to accept fake speech as real.

Study Design. To understand the impact of trusted settings on human interactions with fake speech, we conduct a user study involving deceptive interviews.

Ethics: This study was approved by our institutional IRB. Participants submitted a signed consent form prior to the interview and received a full debriefing afterwards to inform them of the deception and true purpose of the study. No personal information about participants was retained after the interview, and interview recordings were anonymized to protect participant privacy.

Participants. Interviewees were recruited from among the students in our institution’s computer science department. We conducted a total of 14 interviews. Twelve interviewees were male (2 female). All were between the ages of 20 and 35 years old, with varied ethnic/racial backgrounds (American, Chinese, Indian, Indonesian, Turkish). The interviews were approximately 10 minutes, and participants were compensated with a $10 Amazon gift card.

Procedure. The recruitment call asked for participation in an interview study about use of speech recognition systems (e.g., Siri) and their perceptions of privacy with respect to these systems. Each interview took place over a Zoom call, with two paper authors functioning as “interviewers.” One of the interviewers (hereafter referred to as the real interviewer) used their real voice throughout the staged interview, while the other (referred to as the fake interviewer) used only fake speech samples based on their real voice. All fake speech samples were generated using the SV2TTS synthesis system and less than 5 minutes of real voice samples from the fake interviewer. Throughout the call, the fake speech samples are played from an iPad Pro, held close to the fake interviewer’s computer microphone.

After the conclusion of the deceptive portion, we revealed the use of fake voice samples and disclosed our research objectives before asking a few additional questions. Participant responses were categorized and coded separately by each interviewer, who later met to combine the codes and resolve discrepancies. All themes in responses described below were expressed by >= 3 participants, unless otherwise noted. More details about the post-deception interview and analysis procedure can be found in the Appendix.

Because all of the interviewees were members of the authors’ academic division, they had different levels of familiarity with the interviewers, ranging from general knowledge to frequent social interaction. At the end of each interview, participants were asked to rank their familiarity with the interviewers’ voices prior to the interview on a scale of 1 (“not at all familiar”) to 5 (“extremely familiar”). Table 6 lists the distribution of familiarity rankings.

| Familiarity | Real Interviewer | Fake Interviewer |
|-------------|------------------|------------------|
| Not at all  | 1                | 7                |
| Slightly    | 2                | 2                |
| Moderately  | 4                | 2                |
| Very        | 5                | 1                |
| Extremely   | 1                | 1                |

Table 6: # of participants and their declared familiarity with the two interviewer’s voices before the Zoom interview.

Task. The staged interview itself consists of 8 questions about use of automatic speech recognition systems and perceptions of privacy (see Table 7). Five are asked by the real interviewer, and three are asked by the fake interviewer. The three fake interviewer questions are designed to solicit three different types of behavior from participants: conversational response (Q2), website access (Q5), and personal information (Q7).

| #       | Interviewer | Question                                                                 |
|---------|-------------|--------------------------------------------------------------------------|
| 1       | Real        | Do you use automatic speech recognition systems in everyday life?         |
| 2       | Fake        | How often do you use these systems in your daily life?                   |
| 3       | Real        | What do you do in your interactions with these systems?                  |
| 4       | Real        | Do you ever think about your privacy during your interactions with these systems? |
| 5       | Fake        | Can you visit this website? I’ll put the link in the chat.               |
| 6       | Real        | Have you ever used the “voice profiles” feature of these systems?       |
| 7       | Real        | Are you ever concerned about privacy if/when you use voice profiles?    |
| 8       | Fake        | We need your student id to track your participation in this study. Can you leave it in the chat? |

Table 7: Questions asked by real and fake interviewers.

Conditions. Participants are not told that the study is actually about perceptions of fake speech, and they do not know that one of the interviewers is using a fake voice. When the participant joins the Zoom call, the real interviewer informs them that everyone in the call is keeping their video off to preserve interviewee privacy. In reality, keeping videos off prevents the participant from observing that the fake interviewer is using a fake voice. We also asked the interviewees if we could record the interview to maintain a record of their responses.

Because of the relatively low quality of the fake interviewer voice (see §5.2), interviewees are primed to expect a low quality voice from the fake interviewer. For 10 of the 14 participants, before beginning the interview questions, the real interviewer notes that the fake interviewer is feeling unwell and will only chime in intermittently during the interview. We examine the effect of excluding this priming statement from the interview in later sections.

Results. None of the participants exhibited any suspicion or hesitancy during interactions with the fake interviewer’s voice. All 14 responded without hesitation to the three questions asked by the fake interviewer, visited the requested website, and even gave their school ID number to the interviewers. After the interview concluded and the deception was revealed, only four of the 14 participants stated that they thought something was “off” about the fake
interviewer’s voice. Importantly, these four participants had (intentionally) not been given the “priming” statement that the fake interviewer “had a cold.” Below, we explore the most interesting results from this study and highlight several key limitations.

1) Reaction to fake voice: Several themes arose during the post-deception interviews, as summarized below.

- **Complete surprise**: Four participants were visibly and audibly astonished when the deception was revealed. P5 noted that, “I really thought it was you – like 100%,” while P10, after a shocked moment of silence, said “computers just won the Turing test.”
- **Satisfied with “sick” excuse**: Seven participants noted explicitly that the “sick” excuse squashed any concerns about the fake interviewer’s voice. P4 said that “I think it totally worked – I thought you were terribly sick,” while P2 noted that “it was really kind of worrying [how sick you sounded].”
- **Silently suspicious**: Four participants (P9, P12, P13, P14) expressed suspicions after the deception was revealed. P12 and P13 said it “sounded like speaker had a cold,” and P14 supposed “it was a poor quality microphone.”

2) Why participants didn’t voice their concerns: After the deception was revealed, participants were asked to identify elements of the interview structure that increased their trust in the fake interviewer. Some, of course, were completely unsuspecting and did not think to question the fake interviewer. However, others noted that the presence of a second (obviously human) interviewer, social convention, and the origin of the interview request (from within our department) bolstered their trust.

- **Presence of real interviewer**: Several participants credited the “tag-team” nature of the interview, with real and fake interviewers colluding, as making the deception more believable. “I feel like [the real interviewer’s] obviously human presence played a big factor in [my not saying anything].” (P9).
- **Polite social convention**: Multiple participants noted that they felt it would be uncomfortable or wrong for them to say something about the fake interviewer’s voice during the interview. When asked why they didn’t say anything about the quality of the fake interviewer’s voice, P12 exclaimed, “well that would be quite the insult!”
- **Provenance of interview request**: Since we recruited from within our department, the recruitment was sent out through trusted channels only accessible by members of the department (i.e., email list-serv, Slack). P9 expressed suspensions during the debriefing, but credited the “provenance of the study... seemed like a legit source” as a reason to fully participate with our questions.

3) What would have made participants suspicious: When asked to articulate what would have made them more suspicious, participant responses varied.

- **Nothing**: Participants most surprised by the deception claimed that nothing would have caused them to question the credibility of the fake interviewer: “I’m glad you guys didn’t ask me for a bank account, because [...] I would have given it to you” (P5).
- **Requesting more personal information**: One participant noted “I don’t think the information you wanted was very sensitive [so] I don’t see why I need to be concerned about this” (P6).

IRB constraints prevented us from soliciting anything more personal than a student ID, used to access services at our university. While not public, this information is not inherently sensitive.

4) Effect of familiarity with interviewers: Seven of the participants rated their familiarity with both interviewers’ voices as a 1 out of 5 (e.g., not at all familiar with either). Their responses, though, were consistent with the other participants who had some previous familiarity with one or both of the interviewers’ voices. Only one participant (P8) mentioned that “the voice did seem pretty weird, but since I trust you both, I just went [with it].” These results suggest that the trusted setting and presence of a human likely play a larger factor than prior familiarity with the speaker’s voice.

5) Effect of priming statement: To examine the effect of the “sick” excuse on the believability of the fake voice, we conduct four interviews in which participants are not told that the fake interviewer is sick. In these interviews, participants exhibit an increased level of skepticism about the fake interviewer during the debriefing. One claimed “it was very obviously a fake voice,” (P11) but said that based on their experience in other deception studies they decided not to say anything. Others did not see through the deception but did note that “I was feeling weird” (P13) and “I just feel your voice is very strange” (P14).

5.4 Key Takeaways

Our two user studies (A & B) show that context and demographics impact the credibility of synthesized speech for human users. In study A, we found that mentioning fake speech increased participants’ skepticism of the fake speakers they heard. Additionally, women and younger participants in study A were more likely to correctly identify fake speakers.

Our key takeaway from study B is that a fake voice fooled humans in a trusted interview setting. Of particular interest is that all our study B participants were graduate students in computer science, some of whom actively research security or machine learning. Our starting hypothesis was that computer science graduate students would be among the hardest targets to fool with a fake voice. Yet, none of them expressed suspicion about the fake voice during the interview.

Limitations & Next Steps. Our participant pool for study B was largely homogenous in gender, age, and educational background. To conduct a “trusted” interview, our participants were drawn from our academic department (computer science). The gender breakdown of our participants matches that of the department, which skews heavily male. It is possible that the observed effect of gender and age on responses in study A could also extend to study B. Therefore, a viable follow-up work is to conduct larger, more diverse user studies to provide a more nuanced understanding of synthesized voice attacks in trusted settings.

On a related note, our trusted interview in study B followed a voice-only format, where voice is the only medium for interaction. Yet in real-world scenarios, interviewees could use two-factor authentication mechanisms to verify the trusted setting, e.g., requesting the interviewees to turn on their video feed, or challenging the interviewees with some verbal tests. These combined verification methods could make the attacks much more difficult, allowing human users to effectively defend against speech synthesis attacks. We believe this is an important direction for follow-up work.
6 EVALUATING EXISTING DEFENSES

Given the potency of these attacks, we now ask: "what can be done to stop them?" Numerous defenses have been proposed to mitigate voice-based spoofing attacks, many with significant assumptions that limit their practical applicability. Here, we consider a range of defenses in light of our threat model and note the limitations associated with different approaches. Finally, we experimentally evaluate two representative defenses: one that detects synthetic speech using physical artifacts from replay [16], and one that prevents voice synthesis by embedding audio perturbations [38].

6.1 Existing Defenses: Detection & Prevention

We categorize existing defenses into multiple categories in Table 8, noting approach and key limitations for each.

| Category            | Defense | Method                                                                 | Limitations                                                                 |
|---------------------|---------|------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Liveness            | [100]   | Measures human vocal tract movement using Doppler radar.               | Requires precise static calibration during enrollment/testing.             |
| Detection           | [90]    | Detects presence of human breath on mic.                              | Requires careful motion of microphone during recording.                    |
| Loudspeaker Detection | [26]   | Detects presence of magnetic fields produced by loudspeakers.         | Requires precise static calibration during enrollment/testing.             |
| Artifacts           | [16, 32, 54, 93] | Trains models to recognize spectral characteristics of synthetic speech. | Only effective when audio directly played from speaker.                    |
| Preventing Synthesis| [38]    | Corrupts speech samples to prevent unauthorized speech synthesis.      | Degrades quality of defended speech.                                       |
|                     |         |                                                                        |                                                                           |

Table 8: Taxonomy of defenses proposed to prevent speech synthesis attacks.

6.2 Detecting Synthetic Speech using Void

Void [16] protects systems against synthesized speech attacks. It identifies 97 distinct low-frequency spectral features that distinguish human speech and replayed speech. These features can be used to train a variety of detection models. Since the WeChat/Alexa attacks rely on replayed synthesized speech, this defense applies to our setting. In the original paper, Void is tested extensively on replay attacks but only cursorily on synthesis attacks.

Methodology. We recreate the feature extraction pipeline of [16] and train three models using the 2017 ASVspoof dataset [44]. Like Void, we report the equal error rate (EER) for each of the trained models, and the detection success rate (i.e. model’s ability to distinguish real/replayed speech). EER measures when the false positive and false negative rates of a system are equal and is commonly used to report performance of biometric systems. The EER of our trained models is on par with that reported in the original paper.

- **Support Vector Machine (SVM) with RBF Kernel**: This model had the best reported performance in [16].
- **LightCNN** [52]: propose a 27-layer DNN for synthetic speech detection which is also evaluated in [16]. We use the same architecture and parameters as in [52], but modify the input size to accommodate [16]'s 97 features.
- **Custom CNN**: Our last model is a custom 5-layer CNN (see Table 14 in the Appendix). We train this model for 25 epochs using the Adam optimizer with \( \epsilon = 0.001 \).

We test the models on a custom dataset of replayed synthesized samples (targeting VCTK speakers). The synthesized samples are generated on SV2TTS with \( N = 20 \) source files per speaker. They are replayed over two different devices (a UE Boom loudspeaker and LG UltraFine 4K monitor) and recorded using an iPhone 11 situated 6 inches from the audio source. Each replayed set contains 200 samples from 20 different speakers. For comparison, we add 200 clean samples from the same speakers to each replayed set.

Results. Void reliably differentiates real and synthesized samples in our two custom datasets but has high EERs (i.e. false positive/negative rate) across all models, as shown in Table 9. All models have \( \geq 88\% \) detection success rate, but EER for all models is \( \geq 5\% \). High-performing biometric systems typically have EER \( < 1\% \).

Discussion. Void’s high EER renders it less effective in practice in our setting, although the original paper reports a much lower EER when using a custom training dataset. If the custom training dataset were more widely available, Void could provide effective protection for scenarios like the WeChat/Alexa attacks (§4.4).
6.3 Preventing Speech Synthesis via Attack-VC

Attack-VC [38] is designed to protect users from having their voice copied via speech synthesis. Attack-VC adds carefully designed perturbations to speech samples that disrupt unauthorized future synthesis. The “embedding” perturbation generation method in [38] assumes full knowledge of the downstream voice synthesis model $M$ (i.e., a white-box threat model). A defender uses the speaker embedding component of $M$ to create a size-bounded perturbation $\delta$ that shifts the speaker embedding of their sample $x$ towards the embedding of a different speaker’s sample $d$. Then, an adversary $A$ who steals the victim $T$’s defended sample $x + \delta$, cannot use $M$ to successfully synthesize a fake voice sample. The synthesized samples $S_A$ should not sound like $T$.

Methodology. We perform a small-scale study using the VCTK dataset and two models – AutoVC and SV2TTS (as in §4.2). We use the same subset of 20 VCTK speakers as in §4.2. Using author-provided code [8], we generate 19 defended samples per speaker (using the other speakers as optimization targets). We test three perturbation levels, $\epsilon = 0.01, 0.05, 0.1$, following [38].

We notice that the original perturbation loss function $\mathcal{L}$ from [38] does not sufficiently constrain perturbation size. This results in large perturbations that make defended audio samples sound inhuman. To fix this, we add a term to $\mathcal{L}$ (new term is bold):

$$\mathcal{L} = \alpha \cdot \text{MSE}(x + \delta, d) - \beta \cdot \text{MSE}(x, d) + \gamma \cdot \|\delta\|.$$  \hspace{1cm} (1)

where $\text{MSE}$ represents the mean-squared error. This additional term makes the perturbation less audible but does not affect attack success. Empirically, setting $\alpha, \beta = 1$ and $\gamma = 0.1$ works best. We multiply $\gamma$ by 0.99 every 100 iterations.

Then, we use the methodology of §4.2 to synthesize speech from the defended samples. Both the defended samples and the “synthesized-from-defended” samples are evaluated against Azure and Resembyzer (see §4 for details).

Results. As Figure 5 shows, Attack-VC does thwart voice synthesis, but it also corrupts defended samples beyond reliable recognition. For both models and all speaker recognition systems, the speaker recognition accuracy for “synthesized-from-defended” samples is less than 35%, meaning that synthesis attacks after Attack-VC are less successful. However, speaker recognition accuracy for “defended” samples is at most 55%, meaning that they cannot be properly matched to the true speaker. Additionally, “defended” samples still have significant audible distortion, even with our additional constraints on perturbation size.

6.4 Combining Void and Attack-VC

Finally, we evaluate a “stronger” defense that combines Void and Attack-VC, but find that it only provides marginal benefits. In this experiment, we test Void’s detection efficacy on speech synthesized from Attack-VC protected samples, with varying perturbation levels $\epsilon = 0.01, 0.05, 0.1$. We find that speech generated from protected samples with $\epsilon < 0.1$ can only be detected 2 – 4% better (with lower EERs) than normal synthetic speech. Detailed results are in the Appendix.

6.5 Key Takeaways

Our results demonstrate a significant need for new and improved defenses against synthesized speech attacks, particularly defenses generalizable enough for real-world applications. While Void reliably detects fake speech played through speakers, its applicability is limited to replay attacks. Meanwhile, existing prevention defenses such as Attack-VC distort voices beyond recognition, and might benefit from using acoustic hiding techniques [68]. These defenses also assume perfect (white-box) knowledge of the attacker’s speech synthesis model, which is unrealistic in real-world settings.

Limitations & Next Steps. We only evaluate two representative and top-performing defenses (one for each category) and their combined effect. A more comprehensive investigation is required, especially as new defenses emerge.

We also note that current defenses focus on protecting SR systems. However, our results in §5.3 indicate an equal need for human-centric defenses against synthetic speech. One possible direction is to make synthetic speech more “obvious” to human audiences, either by corrupting its generation process to make the speech sound inhuman (i.e., Attack-VC’s yet-unreached goal) or designing parallel authentication methods (i.e. video feed or vocal challenges) that help expose fake speakers.

7 CONCLUSION

Our work represents a first step towards understanding the real-world threat of deep learning-based speech synthesis attacks. Our results demonstrate that synthetic speech generated using publicly available systems can already fool both humans and today’s popular software systems, and that existing defenses fall short. As such, our work highlights the need for new defenses, for both humans and machines, against speech synthesis attacks, promote further research efforts for exploring subsequent challenges and opportunities, while providing a solid benchmark for future research.

ACKNOWLEDGEMENTS

We thank our anonymous reviewers for their insightful feedback. This work is supported in part by NSF grants CNS-1949650, CNS-1923778, CNS1705042, and by the DARPA GARD program. Emily Wenger is also supported by a GFSD fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.
REFERENCES

[1] 2015. Announcing WeChat Voice Print. https://blog.wechat.com/2015/05/21/voiceprint-the-new-voice-password/.
[2] 2019. Personalizing Your Alexa Experience with Voice Profiles. https://developer.amazon.com/blogs/alexa/post/1ad1f4b-4f52-4e68-9187-cc2e9fa55/recognize-voices-and-personalize-your-skills
[3] 2020. Chase VoiceID. https://www.chase.com/personal/voice-biometrics
[4] 2020. HSBC VoiceID. https://www.us.hsbc.com/customer-service/voices/
[5] 2020. Link Your Voice to your Google Assistant device. https://support.google.com/assistant/answer/9071681
[6] 2020. Resemblezy. https://github.com/resemble-ai/Resemblezy
[7] 2020. What Are Alexa Voice Profiles? https://www.amazon.com/gp/help/customer-display.html?nodeId=GVCXY2AB2QWXT2X
[8] 2021. Attack-VC. Github Implementation. https://github.com/cyhuang-tw/attack-vc
[9] 2021. Lyrebird AI. https://www.descript.com/lyrebird
[10] 2021. Microsoft Azure Speaker Recognition. https://azure.microsoft.com/en-us/services/cognitive-services/speaker-recognition/
[11] 2021. Mozilla TTS. https://github.com/mozilla/TTS
[12] 2021. Resemble AI. https://www.resemble.ai/
[13] 2021. Tensorflow TTS. https://github.com/TensorSpeech/TensorFlowTTS
[14] 2021. Voxforge. http://www.voxforge.org/
[15] 2021. WeChat VoicePrint Documentation. https://help.wechat.com/cgi-bin/micromsg-bin/answercenter?opcode=2#plat=ios&lang=en&cid=1568594719Y1EYVb.
[16] 2010. Third International Symposium on Electronic Commerce and Security. IEEE, 265–269
[17] 2016. Amazon Alexa Voice. https://developer.amazon.com/alexa/voiceProfiles.
[18] 2019. Personalize Your Alexa Experience with Voice Profiles. https://developer.amazon.com/blogs/alexa/post/1ad1f4b-4f52-4e68-9187-cc2e9fa55/recognize-voices-and-personalize-your-skills
[19] 2017. You can hear but you cannot steal: Defending against adversarial attacks on speaker verification. In Proc. of ICASSP.
[20] 2003. CMU ARCTIC database s for speech Understanding Bi spectral Analysis. In CVPR Workshops.
[21] 2002. Revisit the security of speaker verification systems against replay spoofing attacks. In Proc. of ICASSP.
[22] 2013. Revisiting the security of speaker verification systems against replay spoofing attacks. In Proc. of ICASSP.
[23] 2017. Assessing the limits of replay spoofing attack detection. (2017).
[24] 2020. Speech synthesis using EEG. In Proc. of ICASSP.
[25] 2020. Speech synthesis using EEG. In Proc. of ICASSP.
[26] 2019. Who is real bob? adversarial attacks on speaker recognition. In Proc. of ICASSP.
[27] 2019. Real Time Voice Cloning. https://github.com/Corientin/Real-Time-Voice-Cloning
[28] 2013. Lanceba: A tool for creation and analyses of large speech corpora. In Proc. of ISIMP.
[29] 2020. Audio Replay Attack Detection with Deep Learning Frameworks. In Proc. of INTERSPEECH.
[30] 2020. Speech synthesis using EEG. In Proc. of ICASSP.
[31] 2007. Evidence for the reproduction of social class in brief speech. Proc. of National Academy of Sciences 114, 46 (Nov. 2019), 22998–23003.
[32] 2019. Fooling end-to-end speaker verification with adversarial examples. In Proc. of ICASSP.
[33] 2003. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. (2017).
[34] 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. In Proc. of ICASSP.
[35] 2008. Improving the security of speaker verification systems against replay spoofing attacks. In Proc. of ICASSP.
[36] 2013. Evidence for the reproduction of social class in brief speech. Proc. of National Academy of Sciences 114, 46 (Nov. 2019), 22998–23003.
[37] 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. (2017).
[38] 2008. Vulnerability of speaker verification to voice mimicking. In Proc. of ISIMP.
[39] 2013. Vocal caricatures reveal signatures of speaker identity. Scientific Reports (2013).
[40] 2009. Hello? Who Am I Talking to? A Shallow CNN Approach for Human vs. Bot Speech Classification. In Proc. of ICASSP.
[41] 2013. Vocal caricatures reveal signatures of speaker identity. Scientific Reports (2013).
[42] 2009. Hello? Who Am I Talking to? A Shallow CNN Approach for Human vs. Bot Speech Classification. In Proc. of ICASSP.
[43] 2017. Inter- and intra-speaker voice variability for speaker recognition. In Proc. of ICASSP.
[44] 2007. Vocal caricatures reveal signatures of speaker identity. Scientific Reports (2013).
[45] 2010. Computer Synthesized Speech Technologies: Tools for Aiding Impairment. IGI Global.
[46] 2019. Hierarchical generative modeling for controllable speech synthesis. Proc. of ICLR (2019).
[47] 2019. Real voice mimic attacks on speaker recognition systems. In Proc. of ICASSP.
[48] 2015. Revisiting the security of speaker verification systems against replay spoofing attacks. In Proc. of ICASSP.
[49] 2019. Revisiting the security of speaker verification systems against replay spoofing attacks. In Proc. of ICASSP.
[50] 2017. Assessing the limits of replay spoofing attack detection. (2017).
[51] 2007. Evidence for the reproduction of social class in brief speech. Proc. of National Academy of Sciences 114, 46 (Nov. 2019), 22998–23003.
[52] 2019. Fooling end-to-end speaker verification with adversarial examples. In Proc. of ICASSP.
[53] 2006. Vulnerability of speaker verification to voice mimicking. In Proc. of ISIMP.
[54] 2013. Vocal caricatures reveal signatures of speaker identity. Scientific Reports (2013).
[55] 2010. Third International Symposium on Electronic Commerce and Security. IEEE, 265–269
[56] 2019. Hierarchical generative modeling for controllable speech synthesis. Proc. of ICLR (2019).
[57] 2017. AV Speech spoofing. In Proc. of ICASSP.
[58] 2019. Vocal caricatures reveal signatures of speaker identity. Scientific Reports (2013).
[59] 2020. Large-scale speaker verification in the wild. Computer-Speech & Language (2020).
[60] 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. (2017).
[61] 2019. Hierarchical generative modeling for controllable speech synthesis. Proc. of ICLR (2019).
[62] 2020. Voice-celeb Large-scale speaker verification in the wild. Computer-Speech & Language (2020).
[63] 2018. HSBC VoiceID. https://www.hsbc.com/hsbczone.
[64] 2019. Hierarchical generative modeling for controllable speech synthesis. Proc. of ICLR (2019).
[65] 2019. Hierarchical generative modeling for controllable speech synthesis. Proc. of ICLR (2019).
[66] 2019. Speech synthesis using EEG. In Proc. of ICASSP.
[67] 2019. Speech synthesis using EEG. In Proc. of ICASSP.
[68] 2019. Speech synthesis using EEG. In Proc. of ICASSP.
8 APPENDIX

8.1 Methodology for §4.1

Here, we describe the methodology used to recreate [57] in §4.1.

Systems Used. We use the following synthesis and SR systems.

- **Festvox** [20] is a text-dependent voice conversion system that uses GMMs [70] to model the vocal characteristics. Training the system requires 4-8 minutes of speech from both the source and target speakers and takes approximately 5 - 10 minutes.

- **Bob Spear** [43] is an open-source speaker recognition system. It uses classical statistical techniques (Universal Background Model Gaussian Mixture Models (UBM-GMM) and Inter-Session Variability (ISV)) to perform speaker recognition [88].

- **Resemblyzer, Azure** See §3.

Datasets Used. The prior attack uses the CMU ARCTIC and Voxforge datasets. CMU ARCTIC [46] contains 1150 spoken phrases from male and female English speakers. Voxforge [14] is an open-source dataset of transcribed speech. The subset we use contains 6561 spoken phrases, each approximately 5 seconds long, from 30 English speakers.

Attack Implementation. We recreate the exact setup of the voice conversion attack from [57]. We use Festvox to convert speech from two CMU ARCTIC speakers (one male, one female) to imitate 10 speakers of the Voxforge dataset. The synthesized attack samples are tested against both of Bob Spear’s speaker recognition algorithms (UBM-GMM and ISV), along with Resemblyzer and Azure. We measure attack performance using the attack success rate (AS), which denotes the percent of synthesized samples identified as the target.

Attack Results. Our attack results on Bob Spear align with prior work, with an average AS of 93.1% for Spear’s UBM-GMM method and 97.1% for Spear’s ISV method (see Table 2). Our synthesized samples have an average Mel-Cepstral Distortion (MCD) of 4.79 dB, compared to the average MCD of 5.59 dB reported in [57] (lower is better).

However, the Festvox attack fails on current SR systems, including Resemblyzer and Azure. Out of the 38 speakers tested, only 8 have one or more attack samples accepted (21%). Moreover, Festvox requires a significant amount (around 6-8 minutes) of content-specific speech from both the victim and attacker to synthesize fake speech, making the attack impractical. Modern voice conversion systems like SV2TTS are text-independent and need much fewer victim speech samples to successfully synthesize speech.

8.2 The Rainbow Passage

Participants in our user study of §4.2 were asked to record the Rainbow Passage. The Rainbow Passage is commonly used in linguistic studies, since it contains nearly all the phonemes combinations of the English language. The full text of the Rainbow Passage is below:

> When sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries men have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from the earth to their home in the sky. Other men have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun’s rays by the rain. Since then physicists have found that it is not the reflection, but refraction by the raindrops which causes the rainbow. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the water drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of superposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

8.3 Phrases for Synthesis

Table 10 lists the phrases used for synthetic speech generation with SV2TTS and AutoVC.

| Phrases used for SV2TTS Speech Synthesis |
|----------------------------------------|
| We control complexity by establishing new languages for describing a design, each of which emphasizes particular aspects of the design and deemphasizes others. |
| An interpreter raises the machine to the level of the user program. |
| Everything should be made as simple as possible, and no simpler. |
| The great dividing line between success and failure can be expressed in five words: ‘I did not have time.’ |
| When your enemy is making a very serious mistake, don’t be impolite and disturb him. |
| A charlatan makes obscure what is clear; a thinker makes clear what is obscure. |
| There are two ways of constructing a software design, one way is to make it so simple that there are obviously no deficiencies, and the other way is to make it so complicated that there are no obvious deficiencies. |
| The three chief virtues of a programmer are: Laziness, Impatience and Hubris. |
| All non-trivial abstractions, to some degree, are leaky. |
| XML wasn’t designed to be edited by humans on a regular basis. |

Table 10: Phrases used for SV2TTS speech synthesis attacks in §4.2 and 4.3.
3.9% EER 94.6% 91.3%

Table 11: Changes to a reference text (top line) as phoneme error rate (PER) increases. Changes are generated by randomly selecting phrases from CMUDict [93] which match the specified PER.

| Metric | Detection Success | EER |
|--------|-------------------|-----|
| Source | SVM               | LightCNN | Custom DNN | SVM | LightCNN | Custom DNN |
| e     | 0.01              | 91.3% | 94.1% | 90.2% | 4.8% | 3.1% | 9.8% |
|       | 0.05              | 93.0% | 94.6% | 91.4% | 3.9% | 4.4% | 8.3% |
|       | 0.10              | 90.8% | 80.1% | 84.4% | 6.2% | 19.1% | 17.9% |

Table 13: Void detection success rates and EER for combined defenses. Speech is first synthesized from Attack-VC defended samples (different ε levels indicated by row), replayed using the UE Boom loudspeaker, then run through the Void models.

- What, if anything, would have made you suspicious of <fake interviewer>?
- Do you have any additional comments about your experience?

Qualitative Analysis Procedure. Following best practices for analysis of open-ended questions, as described in [72], the interviewers made notes during and/or immediately following the post-deception portion of the interview. After the conclusion of all interviews, each interviewer independently reviewed their notes and/or rewatched the interviews as needed. For some questions, i.e., level of familiarity with interviewers, the responses already had discrete categories. For open-ended questions, using a bottom-up approach, the notes were used to categorize responses into general themes regarding the interviewees level of suspicion, and reasons why. Upon completing independent coding of all interviews, the researchers met to consolidate the code books into a single code book with consistent categories, and resolve any discrepancies. In the end, each coding category represents sentiments expressed by 3 or more participants, unless otherwise noted in §5.3.

8.6 Results on Combined Defenses from §6
We also tested if Void can detect speech synthesized from AttackVC-protected samples. We first generated protected samples using ε = 0.01, 0.05, 0.1 for AttackVC (as in §6.1), then recorded them using the UE Boom loudspeaker for playback as before. When tested, all three Void models (SVM/LightCNN/CustomDNN) show high detection rates (91-94%) on protected samples at ε = 0.01, 0.05 levels but slightly lower detection rates at ε = 0.1. Detection rates for protected samples with ε < 0.1 are higher than those for unprotected samples (see Table 9).

8.7 DNN Architecture from §6.2
Table 14 lists the architecture used for our custom DNN trained for Void in §6.2.