From One Stolen Utterance: Assessing the Risks of Voice Cloning in the AIGC Era

Kun Wang, Meng Chen, Li Lu*, Jingwen Feng, Qianniu Chen, Zhongjie Ba, Kui Ren, and Chun Chen
State Key Laboratory of Blockchain and Data Security, Zhejiang University
Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
{*kkwang, meng.chen, li.lu, jingwen.feng, qianniuchen, zhongjieba, kuiren, chenc*}*@zju.edu.cn*

***Abstract*—The advent of voice cloning has fundamentally threatened the role of voice as a unique biometric. Many criminal incidents have already been reported to demonstrate its significant risks of identity forgery. Previous works explored the risks of voice cloning in constrained settings, which require victim speakers to either be already seen in the training data of voice cloning models, or leak dozens of minutes of their speech samples to adversaries. However, with the rapid progress of voice cloning in AIGC (Artificial Intelligence Generated Content) era, these requirements have largely been released, leaving the exact risks of state-of-the-art (SOTA) voice cloning techniques shrouded in a dense fog. To uncover it, this paper conducts a large-scale study in real-world scenarios to assess the risks of advanced voice cloning techniques. This study involves 5 SOTA voice cloning techniques (open-source and commercial), across 8 SOTA voice authentication systems (open-source and real-world) and 30 human listeners, using voice data of over 7,000 speakers (public and custom). By experimental and theoretical analysis, this study reveals that 1) state-of-the-art voice cloning techniques pose severe threats in spoofing voice authentication systems and human listeners; 2) demographic factors such as age and gender of victim speakers have a subtle impact on voice cloning attacks; 3) human listeners' subjective opinions and background about voice cloning play an important role in their susceptibility to attacks; 4) advanced detection methods still fail to identify voice cloning samples as expected.**

1. Introduction

Throughout human evolution, voice has served as a fundamental medium for individual identification. Modern times have further witnessed voice biometrics being integrated as an access control tool in various applications including messaging apps (e.g., WeChat [1]) and mobile banks (e.g., JPMorgan Chase [2], HSBC [3]), or to serve for personalized services on smart speakers (e.g., Amazon Alexa [4], Alibaba TmallGenie [5]) and personal voice assistants (e.g., Apple Siri [6], Samsung Bixby [7]). However, the dramatic rise of voice cloning techniques that can replicate a target speaker's voice with remarkable accuracy, has challenged the reliability of voice as a biometric identifier. The malicious use of these techniques has caused severe economic damage, even interfering with human cognition. For example, in 2019, cybercriminals used voice cloning to deceive a CEO into transferring $243,000 to a fraudulent account [8]; similarly, attackers authorized a $35 million bank transfer using voice cloning in 2021 [9]. On the other hand, cloned voices have been demonstrated to influence elections worldwide [10]. During the 2024 U.S. elections, robocalls featuring a cloned voice of President Biden were sent to voters widely before the New Hampshire primary, to mislead their voting [11].

Given the increasing number of incidents involving voice cloning, it is essential to evaluate it for a better understanding of public and countermeasures development. Previous work examined the security of GMM-based voice authentication (VA) systems using early voice conversion tools such as Festvox [12]. Shirvanian *et al.* [13] quantified the vulnerability of five mobile VA apps using this tool, and a more recent study [14] evaluated early deep learning-based voice cloning techniques like AutoVC [15] and SV2TTS [16] on real-world platforms such as Azure and Alexa.

However, all these studies either evaluate voice cloning attacks only on speakers already seen in the training set for voice cloning models or require $5min \sim 1h$ of speech samples from victim speakers to fine-tune the models. This situation has changed as we are ushered into the AIGC (Artificial Intelligence Generated Content) era, when voice cloning techniques has rapidly evolved. State-of-the-art (SOTA) techniques, like ElevenLabs [17], VALL-E [18], and NaturalSpeech3 [19], have allowed an adversary to clone any victim speaker's voice with just a single utterance as reference. This rapid iteration of techniques leaves the exact risks of voice cloning in the current era hidden in a dense fog.

Toward this end, this work aims to assess the threat of SOTA voice cloning techniques in the AIGC era quantitatively and provide a deeper understanding of voice cloning attacks. To achieve the goals, we first need to answer (RQ1) *Attack performance:* how effective are voice cloning attacks in spoofing voice authentication systems and human listeners, even when the attacker has limited capabilities? To further understand the dynamics of these attacks and gain deeper insights, another question arises, (RQ2) *Factor impact:* how do various factors affect the success of voice cloning attacks? Finally, in terms of the rapid progress

* *Corresponding author*

2025 IEEE Symposium on Security and Privacy (SP) | 979-8-3315-2236-0/25/$31.00 ©2025 IEEE | DOI: 10.1109/SP61157.2025.00238

DOI 10.1109/SP61157.2025.00238

of voice cloning detection methods, we further investigate (RQ3) *Detection effectiveness:* how well do existing cutting-edge detection methods perform?

To answer the aforementioned questions, this paper presents a large-scale study in real-world scenarios involving 5 SOTA voice cloning techniques (open-source and commercial) across 8 SOTA voice authentication systems (open-source and real-world) and 30 human listeners, using voice data of over 7,000 speakers (public and custom). We first assess the effectiveness of voice cloning techniques in spoofing both VA systems and human listeners. The results reveal that even using a single utterance of the victim speaker as reference only, adversaries can successfully spoof VA systems and human listeners with above 96% and 68% rates, respectively. Real-world products like Amazon Echo and Alibaba TmallGenie, are highly vulnerable to voice cloning attacks, with success rates of 100%. Even more concerning, WeChat and a national-wide bank also exhibit significant vulnerability, with victim compromise rates of 80.56% and 45.37%, respectively. Next, we conduct an in-depth analysis of the factors affecting the success rate of voice cloning attacks from the perspectives of the victim speaker, the attacker, and human listeners. We find that females are more vulnerable to voice cloning attacks than males with a higher attack success rate by 3.2%. And we discover that a listener's subjective opinion and background related to voice cloning plays a significant role, with a success rate difference of 11.3%. Finally, we assess state-of-the-art detection methods designed to counter voice cloning attacks. We find that existing passive detection methods perform poorly, with Equal Error Rates (EERs) exceeding 16%. Similarly, proactive detection methods show unsatisfactory results, with EERs above 84%, and only when cloned voices are marked using a specified watermark tool do they achieve near-perfect accuracy.

Our key contributions are highlighted as follows:

- We conduct a large-scale study to assess the risks of state-of-the-art voice cloning techniques in real-world scenarios. To the best of our knowledge, we are the first to assess the risks of voice cloning regarding a real-world national-wide bank.
- Our results demonstrate that even using a single utterance as reference only, an attacker can spoof open-source VA systems, closed-source VA systems and commercial smart speakers, with over 96%, 81%, and 93% rates, respectively. For the national-wide bank, more than 69% victim users are compromised within three attack attempts.
- The in-depth analysis uncovers that the age and gender of the victim speaker have subtle impact on a successful voice cloning attack, which contrasts with previous studies. Moreover, we find that human listeners' subjective opinions towards voice cloning significantly affect their susceptibility to attacks.
- Further evaluation of existing state-of-the-art detection methods reveals that both passive and proactive methods struggle to identify cloning samples in real-world scenarios, largely due to limited generalization capabilities and rigid enforcement protocols.

We have reported our findings to Alibaba, WeChat, Amazon, and the national-wide bank following standard disclosure practices. Amazon, Alibaba, and the bank have all acknowledged the vulnerability. The bank requested a six-month delay in publication, while Alibaba has actively collaborated with our group to develop countermeasures. Note that all the experiments on human participants are validated by the Institutional Review Board (IRB) in our university, and all participant-related data were stored locally and deleted upon completion of the study.

2. Background and Related Work

In this section, we first introduce the current state of voice authentication systems, voice cloning techniques, and their detection methods, followed by a discussion on social awareness and perception of voice cloning. Finally, we review related work on assessing the risks of voice cloning and discuss their limitations.

2.1. Voice Authentication

The fact that humans can identify people by voice has inspired the emergence of Voice Authentication (VA). This automatic technique, also known as Speaker Recognition (SR), enables machines to recognize a speaker's identity from voice characteristics [20]. Due to advances in deep learning and artificial intelligence technologies, voice authentication systems have made remarkable progress in the past few years and are becoming increasingly mature and robust [21], [22]. In real-world scenarios, voice authentication systems have been applied across various domains, including personalized services, digital forensics, and financial transactions. Numerous commercial products have integrated voice authentication systems, such as Apple Siri, Amazon Alexa, HSBC's VoiceId, and Barclays' Voice Security.

2.2. Voice Cloning

Voice Cloning refers to the process of creating a synthetic speech imitating the voice of a particular person. This is mainly achieved by speech synthesis systems such as Text-To-Speech (TTS) [23] and Voice Conversion (VC) [24]. TTS aims to synthesize natural and intelligible speech given text, while VC focuses on how to convert one's voice to sound like that of another without changing the linguistic content. Drawing on advancements in generative models (e.g., VAE [25], GAN [26], Transformer [27], and Diffusion Model [28]), state-of-the-art voice cloning techniques can now synthesize the voice of an unseen speaker using only a few seconds of speech. Based on that, numerous accessible online tools facilitate voice cloning for users including Elevenlabs, Coqui, Resemble AI, Speechify, and Microsoft's VALL-E 1/2 [18], [29] as well as NaturalSpeech3 [19].

2.3. Cloned Voice Detection

The mainstream detection methods against cloned voices could be classified into two categories: passive and proactive detections. Passive detection methods aim to determine whether the suspect speech is machine-generated or human voice. Classical approaches typically build features and exploit statistical differences between synthetic and human speech such as in pitch [30], phoneme transitions [31], and spectral correlations [32]. More recent approaches have incorporated explicit prosody [33], vocal [34] and perceptual models [35]. The state-of-the-art passive detection methods are all deep learning-based end-to-end models such as RawNet2 [36] and AASIST [37].

Another line of work is proactive detection, which aims to actively mark the audio content to identify it once it is released. This method has attracted considerable interest in the context of generative models including those for text [38], image [39], and audio/speech [40], as it provides a means to not only detect synthetic content but also track them. Liu *et al.* [41] proposed an end-to-end voice cloning detection framework based on audio watermarking. Recently, the research group from Meta has published a state-of-the-art proactive detection method of voice cloning, called AudioSeal [42].

2.4. Awareness and Perception of Voice Cloning

Although detection methods for voice cloning exist, most people still express their concerns and hold a negative sentiment. According to a comprehensive report on voice cloning based on a survey of 2,027 U.S. adults [43], nearly two-thirds (63.6%) of U.S. adults are aware of the term "voice clones". The report also indicates that nearly half (49.0%) of the consumers have a negative sentiment toward voice clones, while only one-third (34.3%) of the consumers have a positive sentiment. Although some consumers express strong interest in voice clones for purposes such as comedy and gaming, over 90% of U.S. adults express some concern about the potential negative impact of voice clones on them.

While consumers express significant concern about voice clones, they are most confident that institutions in banking, insurance, and healthcare have already taken steps to protect them against risks, particularly through voice authentication. The report also notes that despite concerns that voice clones may be harmful, consumers appear to trust that voice authentication can combat the risk. However, other reports have emerged indicating that financial institutions using voice authentication can now be spoofed with voice cloning [44]. Therefore, it is crucial to thoroughly assess the effectiveness of voice cloning techniques against voice authentication in real-world scenarios.

2.5. Related Work on Assessing Voice Cloning

Early studies on voice cloning assessment mainly focus on spoofing traditional voice authentication systems. Leon *et al.* [45] revisited the security of GMM-UBM and SVM-based voice authentication systems against an HMM-based speech synthesizer. Kinnunen *et al.* [46] explored the vulnerability of the GMM-JFA voice authentication system using joint density GMM-based voice conversion. Following this work, Wu *et al.* [47] further evaluated both text-dependent and text-independent voice authentication systems using the same voice conversion method. After that, Mukhopadhyay *et al.* [48], using a well-recognized voice conversion tool Festvox [12], were the first to evaluate voice cloning for spoofing both human perception and the then state-of-the-art VA system, Bob Spear [49].

However, the voice cloning methods assessed in the aforementioned works all required a re-training process, making them inaccessible to most real-world adversaries without a professional background. On the other hand, all evaluated spoof targets were open-source voice authentication systems without any real-world applications. To assess the risks of voice cloning in the real world, Shirvanian *et al.* [13] quantified the breakability of five mobile voice authentication apps using Festvox [12]. Further, Wenger *et al.* [14] evaluated deep learning-based voice cloning attacks, including AutoVC [15] and SV2TTS [16], which require only a few minutes of the victim's data for adaptation, on real-world applications such as Azure, WeChat and Alexa. However, these studies still have the following limitations: 1) The results and insights are now inapplicable due to the outdated voice cloning methods employed ; 2) Experiments were conducted with only tens of speakers and in limited settings, making the assessment results less convincing; 3) The victim speakers were already seen during the training procedure of voice cloning, which does not accurately reflect real-world scenarios.

Different from the aforementioned works, we revisit and evaluate the state-of-the-art, out-of-the-box voice cloning methods, based on one stolen utterance of an unseen speaker, in real-world scenarios with a scale of over 7,000 speakers. And, we further study and analyze the influencing factors in the attack process and evaluate the state-of-the-art detection methods against voice cloning.

3. Methodology

In this section, we first introduce our threat model and assumptions on the adversary in this work. We then detail the employed voice clone techniques and the evaluated spoof targets. Finally, we describe the construction procedure of the dataset used in our experiment.

3.1. Threat Model and Assumptions

Figure 1 shows the general three-phase pipeline of a voice cloning attack considered in this work. In **Phase I**, the adversary (*Mallory*) first collects one speech utterance from the victim (*Bob*) through passive exposure or active sharing. For instance, *Mallory* can record *Bob*'s utterances during public speaking engagements like presentations or online meetings. Alternatively, *Mallory* could scrape the

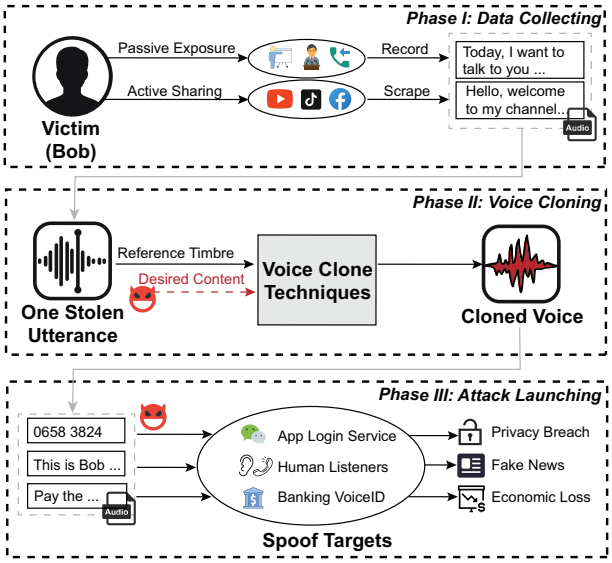


Figure 1: General pipeline of a voice cloning attack.

audio actively shared by *Bob* on social media. In **Phase II**, *Mallory* employs out-of-the-box voice cloning systems to replicate *Bob*'s voice. Specifically, *Mallory* inputs the stolen utterance along with her desired text content into the voice clone system to produce cloned voices. Based on merely one stolen utterance, the advanced voice cloning techniques allow *Mallory* to generate voices that resemble *Bob*'s voice with any desired text content. In **Phase III**, *Mallory* replays the cloned voices to spoof voice authentication systems or deceive human listeners, which could probably compromise Bob's privacy and property, such as accessing *Bob*'s social media accounts, conducting transactions through Banking Apps, or committing financial fraud.

This would be a low-cost, high-return attack, especially with mature voice cloning tools readily accessible online. In real-world scenarios, the adversary may have different levels of knowledge and capability to launch such an attack. To assess the threats posed by an attacker that anyone (even without a professional background) could easily become, we make stringent yet realistic assumptions about the adversary as follows:

- **Unseen victim speaker.** The adversary can only use voice cloning techniques in a zero-shot setting where the victim speaker is unseen by the underlying models, i.e., out of the training set of these models.
- **Limited voice samples.** The adversary can obtain only a single 15*s* utterance from the victim, either through digital scraping or physical recording.
- **Restricted cloning operations.** The adversary can only use out-of-the-box voice cloning tools, whether open-source models or commercial platforms, and lacks the expertise to manually fine-tune the cloned voices, such as adjusting pitch, tone, or accent.

Under these constrained conditions, the risks of voice cloning attacks remain unclear. To address this gap, we conduct a large-scale study to assess the performance of state-of-the-art voice cloning techniques, examine the impact of various factors in voice cloning attacks, and evaluate the effectiveness of existing detection methods. In the following sections, we detail the voice cloning techniques, spoof targets, and speaker datasets used in this study.

3.2. Voice Clone Techniques

Based on the threat model, we selected five state-of-the-art, out-of-the-box voice cloning techniques, including three text-to-speech techniques and two voice conversion techniques. Each of these systems is capable of cloning voices using a single utterance of less than 15*s*. Additionally, their user-friendly interfaces make them accessible to the public.

XTTS is a text-to-speech model built on Tortoise [50] and supports voice cloning in different languages using as little as a 3-*s* audio clip. It is one of the most popular open-source online voice cloning tools, and in our experiments, we directly use the API provided on GitHub[1]. **ElevenLabs**[2] is a pioneering software company in the voice cloning industry, recognized as one of the major companies behind the ongoing AI boom [51]. Their voice cloning technology is highly mature and people can easily access their services for approximately $5 per 30*min* of audio. We registered for their voice cloning service and generated cloned voices via their API. **VALL-E X** [52] is a multilingual text-to-speech model proposed by Microsoft as an extension of VALL-E [53]. Note that Microsoft did not release any official open-source models, we used an unofficial open-source implementation on GitHub[3].

In addition to the aforementioned text-to-speech systems, we also considered the following voice conversion systems. **FreeVC** is built on the end-to-end framework [25]. By imposing an information bottleneck to WavLM [54] features and a spectrogram-resize-based data augmentation method, FreeVC can improve the purity of extracted content information and generate high-quality cloned voices. The official implementation of FreeVC is open-source on the Coqui platform [55]. Although Coqui has shut down, its source code and pretrained models remain available on GitHub[1], and we directly used their pretrained model for our experiments. **DDDM-VC** [56] introduces decoupled denoising diffusion models (DDDMs) into voice conversion tasks and outperforms other publicly available voice conversion models. Specifically, they use a self-supervised representation to disentangle the speech representation such as linguistic information, intonation, and timbre, then apply DDDMs to resynthesize the speech from disentangled representations. The official implementation of this work is also open-source, so we directly used their pretrained models on GitHub[4].

1. https://github.com/coqui-ai/TTS
2. https://elevenlabs.io/voice-cloning
3. https://github.com/Plachtaa/VALL-E-X
4. https://github.com/hayeong0/DDDM-VC

3.3. Spoof Targets

As described in Section 3.1, the adversary can attempt to spoof voice authentication systems and deceive human listeners. To comprehensively assess the risks of voice cloning in real-world scenarios, we consider both machine-based systems and human listeners as spoof targets, as summarized in Table 1.

Open-Source VA Systems. We selected three open-source voice authentication systems including *Ecapa SR* [22], *ResNet SR* [57], and *Resemblyzer* [58]. Ecapa SR and ResNet SR are composed of Ecapa TDNN models and ResNet TDNN models, respectively, achieving SOTA performance in speaker recognition tasks. Both systems are trained on Voxceleb 1/2 [59], [21] datasets using AM-Softmax Loss [60]. Resemblyzer, another widely used open-source speaker recognition system, is trained on both Voxceleb 1/2 and LibriSpeech using generalized end-to-end loss [61]. In these systems, each speaker is enrolled using approximately 30*s* of audio data, from which a speaker embedding is extracted and stored. To authenticate a claimed speaker, these systems derive the embedding of the input audio and compare it to the enrolled embedding of the claimed identity using cosine similarity. In our assessments, we use the pre-trained Ecapa and Resnet SR systems implemented by SpeechBrain [62] as well as the official implementation of Resemblyzer on GitHub.

Closed Service Provider. As a corporation focusing on intelligent speech technology, *iFLYTEK* creates voice authentication software as well as over 10 voice-based internet and mobile products. It has demonstrated the maturity of its AI technology through widespread adoption across various sectors in China. For example, in education, it supports 130 million students across 50,000 schools; in healthcare, it has assisted in over 820 million diagnoses across 600 districts; and in finance, it serves over 200 institutions. The enrollment and authentication procedures are the same as for the aforementioned three open-source VA systems. In our assessments, we directly use the voice authentication WebAPI provided by the open platform.

Real-World Applications. We consider two representative real-world VA applications: the popular social app *WeChat* and the financial mobile app of a national bank. WeChat allows users to log in using voice authentication, and the Bank App enables clients to make withdrawals and transactions via the voiceprint. During enrollment in WeChat, users are required to repeatedly read a fixed, randomly generated eight-digit sequence. For login, users read the same eight-digit sequence to pass the verification. In contrast, the Bank App requires clients to read five randomly generated eight-digit sequences during enrollment, but a new randomly generated sequence during the verification stage.

Commercial Smart Speakers. Two smart speakers are evaluated in this work, i.e., *Amazon Echo* and *TmallGenie*. Echo and TmallGenie are popular smart speakers globally and specifically in China, respectively. Both devices support voice authentication, allowing users to access their contacts and memos and make online purchases. For enrollment, Echo requires users to read several pre-specified sentences, while TmallGenie asks users to read a single sentence at various physical distances from the device. We use Amazon Echo Dot 5th Gen and TmallGenie X5 in our assessments.

Human Listeners. We recruit 30 volunteers (15 males and 15 females) as human spoof targets for our experiments. Specifically, we recruit volunteers on our campus forums. They age from 18 to 30 and consist of undergraduate students, graduate students, as well as some faculty members. Detailed information can be found in Section 4.6. Note that IRB approval is obtained in terms of our work involving human participants. Participants may voluntarily choose whether to grant authorization for our experiment.

3.4. Datasets

In our experiment, XTTS is trained on multi-dataset, FreeVC is trained on VCTK, VALL-E X and DDDM-VC are both trained on LibriTTS. And without transparency in commercial services, the datasets used by ElevenLabs are undisclosed. As mentioned in Section 3.1, in real-world scenarios, victims are likely unseen by any voice cloning techniques during training. To mitigate this issue, we build a large-scale dataset comprising over 7,000 speakers from our custom dataset and multiple public datasets that have various recording environments and devices. Most speakers were not used for training voice cloning techniques in our experiment to more accurately simulate real-world scenarios. We detail these datasets as follows:

TABLE 1: Summary of spoof targets involved in this work.

Type	SubType	Spoof Targets	Content of Enrollment Data
Voice Authentication Systems	Open-Source VA Systems	Ecapa SR [22] ResNet SR [57] Resemblyzer [58]	About 30*s* of arbitrary speeches.
	Closed Service Provider	iFLYTEK	About 30*s* of arbitrary speeches.
	Real-world Applications	WeChat A National-wide Bank	A fixed randomly generated 8-digit sequence. Five different randomly generated 8-digit sequences.
	Commercial Smart Speakers	Amazon Echo TmallGenie	Several sentences with pre-specified content. A single sentence repeatedly read at various distances.
Human Listeners	N/A	30 Participants	N/A

MCV. Mozilla Common Voice [63] is a large, crowd-sourced, and open-source multi-language dataset of voices. Most voice cloning techniques in our experiment do not use this dataset for training. We use Common Voice Corpus 16.1, which contains over 90,000 voices in English with various accents. Specifically, we randomly selected 4,495 speakers whose audio data have been validated and contain at least 60*s* of speeches per speaker.

FST. Free ST American English Corpus [64] were recorded in a silent indoor environment using cellphones. It has 10 speakers and each speaker has about 350 utterances. All utterances have been validated and checked by humans.

VCTK. This CSTR VCTK Corpus [65] includes speech data uttered by 110 English speakers with various accents. Each speaker reads out about 400 sentences, which were selected from a newspaper, the rainbow passage, and an elicitation paragraph used for the speech accent archive. Considering that two speakers have technical issues with the audio recording, we use the data from the remaining 108 speakers.

CSNED. This CrowdSourced high-quality Nigerian English speech Dataset [66] contains transcribed high-quality audio of Nigerian English sentences recorded by 31 volunteers. Each speaker contributed at least 2*min* of audio data and the data set has been manually quality-checked.

CSUKIED. This CrowdSourced high-quality UK and Ireland English Dialect speech dataset [67] contains transcribed high-quality audio of English sentences recorded by 120 volunteers speaking different dialects, including Irish, Midlands, Northern, Scottish, Southern, and Welsh English. Each volunteer provided about $10 \sim 15$*min* of audio data.

LibriSpeech. The LibriSpeech corpus [68] is derived from audiobooks that are part of the LibriVox project, and contains 1000*h* of speech. It contains audio data from 2,418 speakers across training, validation, and evaluation subsets, with each speaker contributing about 30*min* of audio data.

Custom. The real-world applications in our assessments require speakers to read specific phrases, but such audio data is missing from public datasets. To address this issue, we built a custom dataset containing recordings from 18 volunteers, each asked to record phrases in English and Mandarin. These phrases were designed to meet our experimental requirements, and thus allow us to evaluate real-world applications (WeChat and Bank App) and commercial smart speakers (TmallGenie and Amazon Echo). Detailed information can be found in Section 4.4 and 4.5.

4. RQ1: How Effective is Voice Cloning?

In this section, we explore the research question: *How effective are the state-of-the-art voice cloning techniques?* To address this, we assume an adversary that anyone could easily become in real-world scenarios, as described in Section 3.1. Based on the threat model, we generate 469,844 cloned voices for 7,200 victim speakers using out-of-the-box voice cloning techniques, each based on one stolen utterance of the respective speaker. We then conduct 1,344,484 evaluation trials to spoof voice authentication systems and deceive human listeners, allowing us to analyze the effectiveness of state-of-the-art voice cloning techniques.

4.1. Experimental Setup

4.1.1. Dataset. We conduct experiments on a large-scale dataset consisting of several public datasets and our custom dataset, as described in Section 3.4. Specifically, the dataset contains 7,182 speakers from public datasets and 18 volunteers from our custom dataset, across many accents and various recording conditions. The voice cloning and attack evaluation setup are detailed at the beginning of the following sections for each spoof target. In total, we generate 469,844 cloned voices using voice cloning techniques described in Section 3.2.

4.1.2. Running Environment. Given the scale of our dataset, we conduct experiments on our local server with 40 Intel Xeon Silver 4210R CPU, 256 GB RAM, and four 48 GB NVIDIA RTX A6000 GPU, running Ubuntu hirsute 21.04. Note that extensive computation resources are unnecessary because, in most real-world scenarios, the adversary only needs to generate a few cloned voices.

4.1.3. Evaluation Metrics. To evaluate the effectiveness of voice cloning techniques, we define the following two metrics:

- Cloned Voice **A**ttack **S**uccess **R**ate (ASR):

$$\text{ASR} = \frac{\text{\#Successful Cloned Voices}}{\text{\#Total Cloned Voices}}, \tag{1}$$

- **V**ictim Speaker **C**ompromise **R**ate (VCR_n):

$$\text{VCR}_n = \frac{\text{\#Successfully Attacked Victims}}{\text{\#Total Victim Speakers}}, \quad \text{s.t.} \quad \text{\#Attack Trials} \leq n, \quad \text{for each victim.} \tag{2}$$

4.2. Open-Source VA Systems

We first evaluate the effectiveness of voice cloning attacks on three open-source voice authentication (VA) systems. During the voice cloning process, a random utterance of less than 15*s* is selected as the reference audio for each victim speaker, with the desired text phrases listed in Table 16. For the two voice conversion techniques, we include two source speakers: one female and one male. Additionally, due to subscription constraints, cloned voices for a subset of 495 speakers were generated using ElevenLabs. In total, we produced 435,870 cloned voice samples for 7,182 speakers, including 10 samples per speaker using XTTS and VALL-E X, 20 samples per speaker using FreeVC and DDDMVC, and 10 samples per speaker for the 495 speakers using ElevenLabs.

Results. As shown in Table 2, all open-source VA systems demonstrate significant vulnerability to certain voice cloning techniques. While voice conversion techniques achieve ASRs below 50%, text-to-speech techniques perform considerably better, with ASRs of 91.37%, 96.59%,

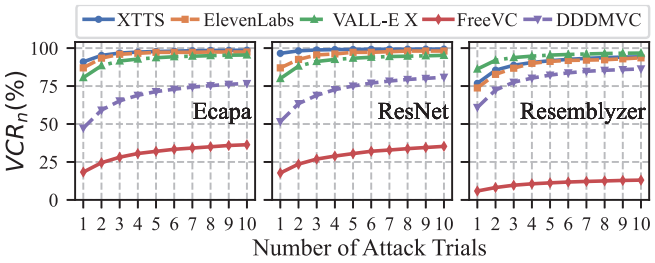


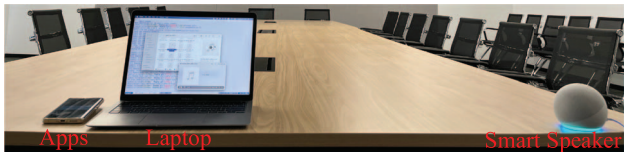
Figure 2: Victim compromise rate regarding a different number of attack trials.

and 86.39% on Ecapa SR, ResNet SR, and Resemblyzer, respectively. Figure 2 illustrates the Victim Compromise Rate (VCR) in relation to the number of attack trials. It reveals that text-to-speech techniques can achieve a VCR of at least 75% with just one attack trial, whereas DDDMVC requires five attack trials on Ecapa and ResNet, or three attack trials on Resemblyzer, to reach comparable performance levels. After 10 attack trials, the VCRs of voice cloning techniques exceed 96% across all three VA systems, indicating that most individuals are highly susceptible to voice cloning attacks.

4.3. Closed Service Provider

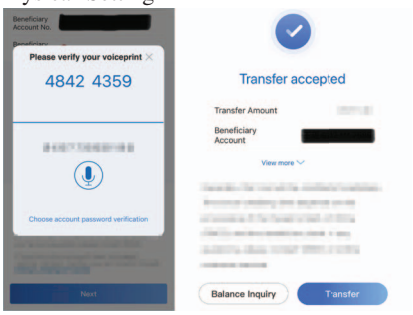
To further assess the effectiveness of voice cloning attacks in real-world scenarios, we conducted experiments on a closed-source VA system from the service provider iFLYTEK. Due to iFLYTEK's charging policy, we limited our evaluation to a subset of victim speakers from our dataset. Specifically, we included 496 speakers, comprising all speakers from the FST, VCTK, CSNED, and CSUKIED datasets, along with 100 speakers each from the MCV and LibriSpeech datasets. In total, we generated 32,830 cloned voice samples for 469 victim speakers, including 10 samples per speaker using XTTS, VALL-E X, and ElevenLabs, and 20 samples per speaker using FreeVC and DDDMVC.

Results. As summarized in Table 2, the results indicate that commercial VA systems are also highly vulnerable to voice cloning attacks, with ASRs exceeding 80%. Consistent with our findings on open-source systems, text-to-speech techniques outperformed voice conversion methods, showing an average ASR improvement of approximately 20%. Additionally, within ten attack trials, the VCRs for all voice cloning techniques, except for FreeVC, surpassed



(a) Physical Setting

(b) ATM Withdrawal



(c) App Transfer

Figure 3: Assessment on real-world VA applications.

88%. These results demonstrate that the commercial VA service provider is also vulnerable to voice cloning attacks, further emphasizing the susceptibility of commercial VA systems to modern voice cloning techniques.

4.4. Real-World Applications

In addition to evaluating open-source and closed-source VA systems, we assess two real-world applications that integrate VA systems in this section: the social app *WeChat* and the financial mobile app of a real-world national bank. Due to the lack of detailed information about these two VA systems deployed in real-world applications, we treat them as black-box VA systems. These evaluations are conducted using our custom dataset, which includes utterances from 18 volunteers (eight males and ten females), aged in their twenties to thirties, recruited through our campus forum. All participants are well-educated Chinese students fluent in both Mandarin and English. Before the experiment, each participant signed an informed consent form outlining the research purpose, procedures, and data usage. After completing a 10-minute recording session, each volunteer was compensated with $2.5.

Setup. As described in Section 3.3, both WeChat and the Bank App require users to read digit sequences for enrollment and verification. In this experiment, we first

TABLE 2: Attack success rate and victim compromise rate of voice cloning techniques on open-source voice authentication systems and closed-source service provider. The highest values across different voice cloning techniques are bolded.

Spoof Targets	ASR(%)					VCR_n(%)					
	XTTS	ElevenLabs	VALL-E X	FreeVC	DDDMVC	n	XTTS	ElevenLabs	VALL-E X	FreeVC	DDDMVC
Ecapa SR	**91.37**	86.46	80.89	11.23	32.81	10	**98.62**	97.44	95.48	36.35	76.69
ResNet SR	**96.59**	87.19	79.9	11.76	38.5	10	**99.36**	98.08	95.22	35.29	80.82
Resemblyzer	76.85	74.18	**86.39**	3.91	49.97	10	94.25	93.60	**96.73**	13.03	86.31
iFLYTEK	77.21	**81.05**	67.92	27.21	41.95	10	97.44	**98.51**	91.90	42.43	88.27

recorded ten-digit utterances for each speaker, which were then concatenated to form new voice samples for enrollment. Next, we generated cloned voices for each speaker to assess the voice authentication functionality of WeChat and the Bank App. Since both applications require digit sequences in Mandarin, we selected two text-to-speech techniques, XTTS and VALL-E X, which support multilingual functionality. For voice cloning, we employed two types of reference utterances from victim speakers: a paragraph of text and a sequence of digits from zero to ten. The impact of these reference types is analyzed in Section 5.2.2. For each reference type, we generated 10 cloned samples per speaker using two cloning techniques for the WeChat application, and 3 samples per speaker for the banking application. In total, we produced 936 cloned voice samples for 18 speakers. Figure 3a illustrates the physical setup of the experiment, where cloned voices were generated on a laptop and replayed through a phone to interact with the mobile applications.

Results. In this experiment, we successfully logged into the victim's WeChat account and, for the Bank App, completed both a mobile transaction and an ATM withdrawal. Table 3 presents the assessment results of XTTS and VALL-E X on WeChat and the Bank App. The ASRs of 30.00% on WeChat and 45.37% on the Bank App are notably lower than those observed on open-source VA systems, indicating that these real-world applications exhibit greater robustness and resilience to voice cloning attacks. Interestingly, despite being a financial application, the Bank App was more vulnerable to attacks than WeChat. However, the Bank App utilizes additional multi-factor authentication, with voice authentication being just one layer of security, providing relatively strong overall protection for users, and we did not consider overcoming other authentication factors in our experiment. Moreover, WeChat permits more than ten login attempts before blocking access, and under these conditions, over 80% of victims were compromised. In contrast, the Bank App enforces stricter limits on failed login attempts, yet within just three trials, approximately 69.45% of victims were still compromised.

Discussion. As noted in Section 2.4, consumers often trust that banking institutions provide robust security through voice authentication. However, our experimental results suggest that VA systems in banking environments perform significantly worse than expected. While some institutions, such as HSBC, have suspended their Voice ID systems, many others continue to rely on VA for authentication. Our results highlight the increasing threat posed by voice cloning attacks, especially in high-stakes environments like banking, where trust in authentication methods is paramount.

TABLE 3: Assessment results on WeChat and Bank App.

	ASR(%)		VCR_n(%)		
	XTTS	**VALL-E X**	**n**	**XTTS**	**VALL-E X**
WeChat	**30.00**	24.17	10	57.36	**80.56**
Bank App	**45.37**	31.48	3	**69.45**	66.67

Disclosure. We have reported our findings to WeChat and the Bank Institution following standard disclosure practices. The Bank Institution has acknowledged the vulnerability and requested a six-month delay after multiple meetings with our group.

4.5. Commercial Smart Speakers

In real-world scenarios, we further evaluated two commercial smart speakers: Amazon Echo Dot and Alibaba TmallGenie. This experiment was conducted using the custom dataset described in Section 4.4, which includes 18 volunteers. The physical setup is shown in Figure 3a, where the laptop and smart speaker are placed on a table with a distance of one meter between them.

For this experiment, we first recorded utterances containing specific phrases required by Echo and TmallGenie, as outlined in Table 15. Note that for enrollment, Echo requires users to read out various phrases, while TmallGenie mandates that users read the same phrase at varying distances from the device. We also assumed that the voiceprint authentication feature was enabled on both devices. To comprehensively evaluate the vulnerability of smart speakers to voice cloning attacks, we propose the following attack tasks:

- **(T1) Identity Theft** The attacker deceives the smart speaker into recognizing them as the legitimate user.
- **(T2) Privacy Breach:** The attacker gains access to the user's private information by querying the smart speaker.
- **(T3) Unauthorized Transaction:** The attacker makes online purchases through commands given to the smart speaker.

Based on these tasks, we designed corresponding phrases and generated cloned voices, as listed in Table 17. In total, we conducted 108 attack trials for 18 speakers on Echo Dot and TmallGenie under real-world conditions, using 3 cloned samples per speaker for each device.

Results. As shown in Table 4, XTTS achieved a 100% ASR in all tasks except T3 on Echo, while VALL-E X achieved an ASR above 94% on TmallGenie but a relatively lower ASR on Echo. In the Identity Theft task, we successfully used cloned voices to deceive both Echo and TmallGenie into recognizing the attacker as the enrolled user. For example, when we played a cloned voice saying "Alexa, who am I?" Echo Dot responded as it would to the legitimate user. In the Privacy Breach task, both Echo and TmallGenie revealed the legitimate user's private information when queried with cloned voices. Furthermore,

TABLE 4: Attack success rate on smart speakers. T1, T2, and T3 stand for Identity Theft, Privacy Breach, and Unauthorized Transaction, respectively.

ASR(%)	Echo			TmallGenie		
	T1	**T2**	**T3**	**T1**	**T2**	**T3**
XTTS	100	100	83.33	100	100	100
VALL-E X	77.78	55.56	38.89	100	100	94.4

in the Unauthorized Transaction task, we were able to make Echo and TmallGenie complete online purchases and finalize payments using cloned voices.

Disclosure. We have reported our findings to Alibaba and Amazon following standard disclosure practices. Both Amazon and Alibaba acknowledged the vulnerability, and Alibaba sought more collaboration with our group to develop countermeasures.

4.6. Human Listeners

In this section, we conduct an experiment to evaluate the attack success rate of voice cloning attacks on human listeners. As described in Section 3.3, we recruited 30 volunteers (15 males and 15 females) with IRB approval from our university. Participants were screened for hearing issues as part of the recruitment process, and informed consent was obtained from all volunteers.

The recruited participants, aged 18 to 30, included undergraduate and graduate students, as well as faculty members. The experiment took approximately 30*min* per participant, with each compensated $5 for their time. Before the experiment began, participants were informed that any audio they heard could either be a genuine or cloned voice. Note that the awareness of volunteers regarding the purpose of our study may lead to different detection behaviors compared to real-world scenarios, which may affects the results. For each participant, we played 100 pairs of audio, with one being a genuine voice and the other a cloned voice. They were asked to answer two questions for each pair: 1) label each played audio as either genuine or synthetic, and 2) determine whether the two audios came from the same speaker. Based on these responses, we defined the following metrics to assess the effectiveness of voice cloning attacks on human listeners:

- **A**ttack **S**uccess **R**ate on naturalness and intelligibility:

$$\text{ASR}_{genuine} = \frac{\text{\#Cloned Voices Labeled as } \textbf{Genuine}}{\text{\#Total Cloned Audio Samples}}, \quad (3)$$

- **A**ttack **S**uccess **R**ate on speaker similarity:

$$\text{ASR}_{same} = \frac{\text{\#Cloned Voices Labeled As Same}}{\text{\#Total Cloned Audio Samples}}. \quad (4)$$

Results. Table 5 presents the attack success rates of different voice cloning techniques on human listeners, evaluated in terms of both naturalness and speaker similarity. The results show that ElevenLabs achieved the highest attack success rates in both dimensions, with 68.17% for naturalness and 67% for speaker similarity. For naturalness, all three text-to-speech techniques achieved attack success rates above 50%, while the two voice conversion techniques had lower success rates, both below 38%. In terms of speaker similarity, all voice cloning techniques, except for DDDMVC, achieved success rates above 56%. Additionally, we observed an interesting trend: when volunteers were aware that the audio could be either genuine or cloned, their accuracy in identifying genuine voices dropped to just 69%.

TABLE 5: Attack success rate on human listeners.

	XTTS	ElevenLabs	VALL-E X	FreeVC	DDDMVC
$\text{ASR}_{genuine}(\%)$	51.83	**68.17**	62.17	35.17	37.67
$\text{ASR}_{same}(\%)$	61.50	**67.00**	56.17	56.67	43.00

> ***Finding 1.*** State-of-the-art voice cloning techniques pose more severe threats than reported in previous studies, even under more constrained conditions in real-world scenarios. With only a single utterance from the victim speaker as reference and without any fine-tuning of the voice cloning models, an attacker can easily and successfully impersonate any victim speaker across both open-source and commercial VA systems, real-world applications, and human listeners, resulting in privacy breaches and, in more severe cases, economic losses.

5. RQ2: How Various Factors Impact Attacks?

In this section, we explore the research question: *How do various factors from the victim's, attacker's, and human listener's perspective impact the success rate of voice cloning attacks?* To address this, we first comprehensively analyze the impact of different characteristics of the victim to the attack success rate, including their ages and genders. To better understand voice cloning attacks, from the attacker's perspective, we analyze the impact of different voice cloning techniques and various characteristics of the reference utterance used in the voice cloning process, including the quality, length, and phoneme coverage. Furthermore, to better understand the circumstances of human listeners faced with voice cloning attacks, we analyze the attack success rate on human listeners regarding their inborn conditions like gender or nurture knowledge and sentiment towards voice cloning. Note that in all our statistical analyses, significance levels are denoted as follows: $p < .05^{*}$, $p < .01^{**}$, and $p < .001^{***}$.

5.1. Victim's Perspective

5.1.1. Impact of Victim's Gender. A previous study [14] reported that female victim speakers exhibited over a 20% higher attack success rate than male victim speakers, based on experiments involving fewer than 100 participants. In contrast, our study expands this analysis by conducting experiments with a significantly larger dataset, comprising 7,047 speakers, as detailed in Table 6. Note that in our experiment, we only consider the gender as the sex assigned at birth, and excluded data from speakers who have non-binary identity or do not want to share gender information.

To rigorously examine the influence of victim gender on attack success rates (ASR), we employed a linear mixed-effects model (LMM) to account for system-level variability.

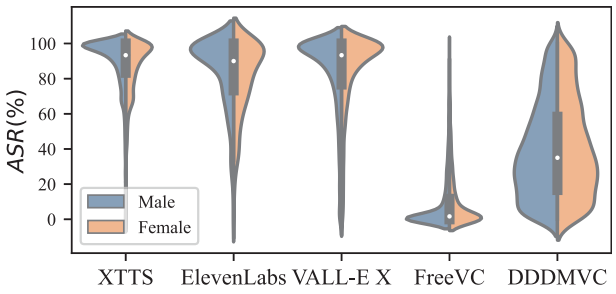


Figure 4: Probability density distribution of attack success rate by gender.

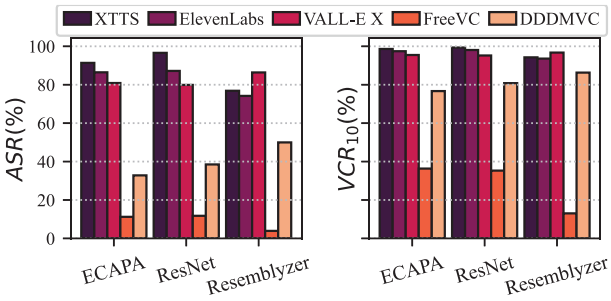


Figure 5: Attack success rate of different voice cloning techniques.

Specifically, gender was treated as a fixed effect, while voice cloning techniques and voice authentication systems were also included as fixed effects to control for their substantial influence on ASR. Victim identity was modeled as a random effect, allowing us to capture individual-level variation across repeated measurements.

In the LMM specification, male was set as the reference category. The resulting coefficient for female was $\beta = 0.259$ with a p-value of 0.475 and a 95% confidence interval of $[-0.451,\ 0.969]$. Although this coefficient suggests that female victims had a slightly higher average ASR than males, the difference was not statistically significant ($p > 0.05$). To further illustrate this finding, we visualized the distribution of attack success rates for male and female victims using violin plots, as shown in Figure 4. The distributions are nearly symmetric across genders, with only minor differences in density and spread. Based on these observations, we conclude that gender has a negligible effect on the attack success rate of voice cloning techniques.

5.1.2. Impact of Victim's Age. We further examined whether age influences the susceptibility to voice cloning attacks. To the best of our knowledge, no existing study has examined the attack success rate in relation to the age of victim speakers. To address this gap, we conducted experiments on speakers from the MCV dataset, analyzing how the age of the victim impacts the attack success rate.

Similar to previous analysis on victim gender, we employed a linear mixed-effects model (LMM), treating ASR as the dependent variable. The model included age group, voice cloning techniques, and voice authentication systems as fixed effects, and incorporated victim identity as a random intercept to account for repeated measurements. The age variable was treated as a categorical factor, with "teens" used as the reference group. The modeling results are summarized in Table 7, which reports the estimated coefficients, p-values, and 95% confidence intervals. Compared to the reference group, several age groups exhibited significantly lower ASR values. Specifically, the twenties group showed a difference of $\beta = -3.112$ ($p < 0.001$), while other groups such as forties ($\beta = -2.071$, $p = 0.017$), fifties ($\beta = -2.913$, $p = 0.003$), sixties ($\beta = -3.362$, $p = 0.004$), and seventies ($\beta = -5.621$, $p = 0.001$) also demonstrated statistically significant effects. The thirties group showed a non-significant trend ($\beta = -1.145$, $p = 0.116$).

These results suggest that age is associated with variation in attack success rates. The overall pattern indicates that ASR tends to decrease with increasing age across many groups, although the effect is not strictly monotonic. To further evaluate whether age as a factor contributes significantly to model fit, we conducted a likelihood ratio test (LRT) comparing the full model (including age group) to a reduced model without age. The result showed a significant improvement in model fit ($\chi^2(6) = 43.70$, $p < 0.001$), indicating that age group, as a whole, has a meaningful effect on ASR.

TABLE 6: Number of speakers by gender.

	MCV	FST	VCTK	CSNED	CSUKIED	LibriSpeech	Total
M	3287	5	47	12	71	1250	4672
F	1073	5	61	19	49	1168	2375

TABLE 7: Estimated effects of age group on ASR using a linear mixed-effects model (reference group: Teens).

Age Group	#Speakers	Coefficient	p-value	95% CI
Teens	**402**	-	-	-
Twenties	**1913**	-3.112	0.001***	[−4.394, −1.830]
Thirties	**957**	-1.145	0.116	[−2.571, 0.281]
Forties	**453**	-2.071	0.017*	[−3.774, −0.368]
Fifties	**309**	-2.913	0.003**	[−4.828, −0.998]
Sixties	**191**	-3.362	0.004**	[−5.633, −1.092]
Seventies	**70**	-5.621	0.001***	[−9.072, −2.171]

CI = Confidence Interval.

> ***Finding 2.*** Victim gender has no statistically significant effect on attack success rates, in contrast to a previous study reporting a 20% higher success rate for female victims. Age shows a statistically significant effect, but its practical impact is modest, with less than 6% variation across age groups.

5.2. Attacker's Perspective

5.2.1. Impact of Voice Cloning Technique. Figure 5 presents the attack success rate and victim compromise rate

within 10 attack trials, comparing different voice cloning techniques across three open-source voice authentication (VA) systems, based on the setup described in Section 4.1. We observe that the three text-to-speech techniques consistently achieve significantly higher ASR and VCR compared to the two voice conversion techniques across all VA systems, indicating that text-to-speech systems demonstrate a stronger capability for cloning voices. Given the ease of use of text-to-speech systems, this may allow more non-professional attackers to launch such attacks.

5.2.2. Impact of Reference Utterance.

Utterance Quality. Table 8 presents the results of the attack success rate on utterances with varying levels of quality, as measured by the Signal-to-Noise Ratio (SNR), on the LibriSpeech dataset including 2,418 speakers. In this experiment, we systematically varied only the amount of noise added to the original reference utterances, keeping all other conditions constant to isolate the effect of utterance quality. A Chi-square test showed a significant difference in ASR across SNR levels of reference utterance with $p < .0001$. We observe that for utterances with extremely low SNR, the ASR drops significantly to 3.64%. However, as the SNR improves, the ASR increases steadily, eventually stabilizing at around 84%. These results indicate that higher-quality reference utterances substantially improve the likelihood of a successful voice cloning attack while using utterances with an SNR above 30 is sufficient to achieve a satisfactory attack success rate.

Utterance Length. To explore the impact of utterance length, we clipped audio segments of varying durations from the LibriSpeech dataset, ensuring that all other experimental conditions were held constant. A Chi-square test also showed a significant difference in ASR across the length of reference utterance with $p < .0001$. As shown in Table 8, even with a short utterance of only 5*s*, state-of-the-art voice cloning techniques can achieve an ASR above 60%. As the length of the reference utterance increases, the ASR also improves, stabilizing at 80% with utterances of around 15 to 20*s*. These results indicate that longer reference utterances indeed improve the attack success rate while using utterances about 15*s* is sufficient to achieve a satisfactory success rate exceeding 80%.

Phoneme Coverage. Intuitively, in the voice cloning process, using reference utterances containing more similar phonemes as desired content may improve attack success rates. This indicates that attackers may improve their success rate by using reference audio with content that closely matches the target speech in terms of phoneme coverage. In this experiment, we examine how phoneme coverage (i.e., the proportion of desired phonemes present in the reference utterance) impacts the effectiveness of voice cloning attacks, while keeping other conditions such as utterance quality and background noise constant. As discussed in Section 4.4, an attacker needs to generate a cloned voice that replicates a sequence of digits to bypass the WeChat voice authentication (VA) system. To this end, we used two types of reference utterances: one with normal text and another consisting of ten digits. For both types, we calculated the phoneme coverage, with results shown in Table 9. Our results demonstrate that, for both voice cloning techniques, higher phoneme coverage consistently leads to a higher attack success rate and a greater victim compromise rate. This suggests that attackers can improve their success rate by using reference audio whose phoneme content closely matches target speeches.

> ***Finding 3.*** For voice cloning attacks, text-to-speech techniques are generally more effective than voice conversion techniques, and a single 15*s* utterance with a signal-to-noise ratio similar to a quiet office is sufficient to achieve a successful attack. Moreover, greater phoneme coverage in the reference utterance can further enhance these attacks.

TABLE 8: Attack success rate across different SNR (top) and utterance length (bottom) used in the cloning process. Both factors significantly affect ASR ($\chi^2(5) = 5979.97$ for SNR; $\chi^2(5) = 410.98$ for length, $p < .0001$).

SNR (dB)	**0**	**10**	**20**	**30**	**40**	**50**
ASR(%)	3.64	22.08	58.06	79.94	84.66	84.62
Length (s)	**5**	**10**	**15**	**20**	**25**	**30**
ASR(%)	64.64	75.64	80.11	83.62	82.92	84.62

TABLE 9: Attack success rate and victim compromise rate on WeChat regarding phoneme coverage.

Evaluation Metric	**ASR(%)**				**VCR_{10}(%)**	
Phoneme Coverage	**50**	**100**	χ^2	*p*-value	**50**	**100**
XTTS	23.89	**36.11**	5.83	0.0157*	50.00	**64.71**
VALL-E X	16.67	**31.67**	10.25	0.0014***	72.22	**88.89**

5.3. Listener's Perspective

5.3.1. Impact of Demographic Characteristics.

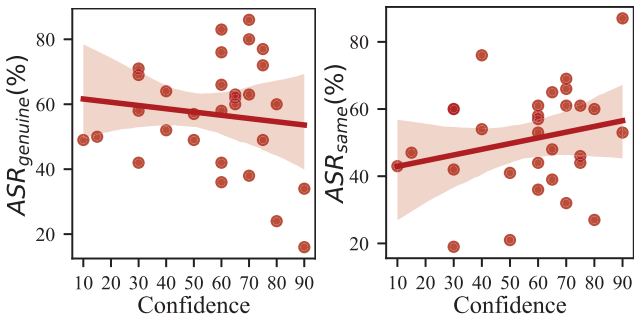
Impact of Age. Hearing ability changes with age, potentially influencing individuals' ability to accurately identify cloned voices. According to an audiology and hearing aid

TABLE 10: Attack success rate on human listeners with 95% confidence intervals by age group.

Age Group	**(18, 22)**	**(23, 26)**	**(27, 30)**	*p*-value
#Participants	8	17	5	–
$ASR_{genuine}$(%)	59.75 ± 10.60	57.41 ± 10.26	50.40 ± 15.87	0.633
ASR_{same}(%)	52.88 ± 9.97	51.88 ± 8.62	45.00 ± 11.34	0.867

TABLE 11: Attack success rate on human listeners with 95% confidence intervals by gender.

Gender	**Male**	**Female**	*p*-value
#Participants	15	15	-
$ASR_{genuine}$(%)	56.33 ± 10.28	57.40 ± 8.89	0.867
ASR_{same}(%)	52.87 ± 10.70	49.13 ± 5.93	0.518



(a) Naturalness (b) Similarity

Figure 6: Attack success rate on human listeners regarding their confidence.

center [69], hearing acuity peaks at around 18 years of age, remains optimal through the early 20s, and typically declines after approximately 25 years of age. Based on this, we hypothesize that younger listeners (aged between 18 and 22 years) would exhibit higher resistance to voice cloning attacks due to superior hearing acuity, whereas older listeners (aged between 27 and 30 years) might demonstrate increased vulnerability as hearing acuity diminishes. Table 10 summarizes our experimental results. The ASR indeed declines with increasing age, dropping from approximately 60% among listeners aged 18 to 22 years to around 50% among those aged 27 to 30 years. However, a one-way ANOVA test yielded non-significant p-values (0.633 for $ASR_{genuine}$, 0.645 for ASR_{same}), indicating that the overall effect of age on vulnerability to voice cloning is subtle.

Impact of Gender. Table 11 shows the average attack success rates for male and female listeners. The results indicate comparable susceptibility levels between genders, with differences of approximately 1% in $ASR_{genuine}$ and around 3.7% in ASR_{same}. Further statistical analysis using independent t-tests produced p-values of 0.867 for $ASR_{genuine}$ and 0.518 for ASR_{same} confirming that gender differences do not have statistically significant impacts on human vulnerability to voice cloning attacks.

5.3.2. Impact of Perceptual Factors.

Impact of Subjective Opinions and Background. In addition to inherent factors such as age and gender, we also investigate how non-inherent factors influence the attack success rate. Before conducting the human listening experiment in Section 4.6, participants were asked a series of questions regarding their background: whether they had a computer science background, prior experience with AI-related tools, familiarity with the concept of voice cloning, and whether they held negative attitudes toward voice cloning technology. Based on their responses, we categorized participants and compared the average attack success rate, as shown in Table 12. Interestingly, the attack success rate among participants without a computer science background was lower than those with such a background, which contradicts our initial expectations. Furthermore, individuals who had previously used AIGC tools were found to be more susceptible to voice cloning attacks, with an increased success rate of 6.25%. Regarding the naturalness of cloned voices, participants familiar with the voice cloning concept exhibited a lower attack success rate. However, when it came to speaker similarity, the attack success rate was paradoxically higher among these individuals. This suggests that while familiarity with voice cloning may enhance the ability to differentiate cloned voices from authentic ones, it does not necessarily aid in identifying the original speaker. Moreover, we found that participants with a negative attitude toward voice cloning technology demonstrated a significantly lower attack success rate when evaluating naturalness, showing a decrease of more than 11% compared to those with a positive attitude.

Impact of Identification Confidence. After the human listening experiment, participants were asked to rate their confidence in their identification results. As shown in Figure 6a, higher average confidence levels were generally associated with lower attack success rates in naturalness evaluations, indicating greater accuracy. In contrast, for speaker similarity, the attack success rate increased with confidence, suggesting that higher confidence in this context was linked to lower identification accuracy. These findings suggest that while greater confidence may enhance the detection of unnatural speech, it does not necessarily improve recognition of the original speaker.

> ***Finding 4.*** Differences across genders and ages of human listeners have a subtle impact on their susceptibility to voice cloning attacks. In contrast, factors such as individuals' subjective opinions and their level of caution towards voice cloning appear to play a more significant role. Furthermore, human listeners who are more confident in the accuracy of their identifications on cloned voice samples tend to be more accurate when assessing the naturalness of the samples, but relatively less accurate when evaluating speaker similarity.

TABLE 12: Attack success rate on human listeners regarding subject opinions and background. VC stands for Voice Clone.

Criteria	Comp. Science Background			AIGC Tools Usage			Familiar with VC Concept			Negative Attitude Toward VC		
	Yes	No	Δ	Yes	No	Δ	Yes	No	Δ	Yes	No	Δ
#Participants	16	14	-	24	6	-	14	16	-	20	10	-
$ASR_{genuine}$(%)	60.00	53.29	**6.71**	56.75	57.33	0.58	55.21	58.31	3.10	53.10	64.40	**11.3**
ASR_{same}(%)	53.56	48.07	**5.49**	52.25	46.00	**6.25**	54.14	48.25	**5.89**	50.90	51.20	0.30

6. RQ3: How Do Detection Methods Perform?

In this section, we address the research question: *How do existing detection methods perform when challenged by state-of-the-art voice cloning techniques?* Specifically, we evaluate several cutting-edge detection methods: two passive approaches (AASIST [37] and SSL-AS [70]), and a proactive approach, AudioSeal [42]. Passive approaches aim to determine whether suspect speech is machine-generated, while proactive approaches actively mark machine-generated content to enable identification once it is released. AASIST achieved the best performance with an equal error rate (EER) of 1.13% on ASVSpoof 2019 dataset, while SSL-AS demonstrated a state-of-the-art performance with an EER of 2.85% on ASVSpoof 2021 Deepfake database. Also, Meta's AudioSeal achieves the best results in proactive voice cloning detection through localized watermarking. Using the available open-source implementations, we apply these methods to assess all the cloned voices described in Section 4.1. In total, we conducted over two million trials across more than 7,000 speakers.

6.1. Passive Detection

To assess the effectiveness of two passive detection methods, we directly applied the pre-trained detector to our generated cloned voices along with their corresponding source audios. For each audio sample, both AASIST and SSL-AS output a score representing the probability that the sample is a cloned voice. Based on these scores, we computed the equal error rate (EER).

Results. As shown in Table 13, AASIST achieves poor performance with an EER of 46.14%, while SSL-AS achieves an average EER of 16.24%, which represents a significant performance decline compared to results reported on the ASVSpoof 2019 and 2021 database. This drop in performance highlights the limitations of passive detection approaches, primarily due to the poor generalization capability of existing methods [71]. Although SSL-AS incorporates a self-supervised wav2vec model [72] and data augmentation to enhance generalization across both unseen speakers and generation algorithms outside the training set, our results suggest that achieving reliable and robust detection performance for real-world applications remains a significant challenge.

6.2. Proactive Detection

To proactively detect cloned voices using watermarks, AudioSeal utilizes two primary modules: a generator and a detector. The generator embeds a watermark into an input audio sample, while the detector identifies whether an audio sample has been watermarked. In real-world scenarios, attackers might use voice cloning tools protected by different watermark generators or even tools without watermark protection. To simulate this, we conducted experiments under three conditions: cloned voices protected by 1) *NOMark:* no watermark, 2) *ASMark:* the AudioSeal watermark generator, and 3) *WMMark:* a mismatched watermark generator (WavMark [73]). During the detection phase, the AudioSeal detector evaluated these altered voices, outputting scores indicating the probability that each audio sample was watermarked. We then calculated the EER based on these scores.

Results. As shown in Table 14, AudioSeal consistently failed to identify cloned voices when they were unwatermarked, yielding an average EER of 84.49%. In contrast, AudioSeal performed nearly perfectly, achieving an EER below 0.001%, when cloned voices were protected by the designated generator. However, when protected by other watermark generators, AudioSeal's performance dropped sharply once more, with an average EER of 86.21%. These results demonstrate that the proactive detection method yields excellent performance only when strict compliance with watermarking protocols is enforced, while detection effectiveness remains limited without such restrictions.

> ***Finding 5.*** Existing methods still struggle with detecting voice cloning samples in real-world scenarios. In particular, passive detection methods are prone to missed and false detections due to limited generalization capabilities, while proactive methods perform poorly because they rely on enforcing watermarking protocols for online tools.

7. Discussion

In this section, we discuss threat scenarios enabled by voice cloning, implications for defense and authentication services, and key challenges in detecting cloned voices.

Abuse Scenarios. Our findings demonstrate that modern voice cloning techniques pose realistic and immediate threats across a range of misuse scenarios. In addition to bypassing voice-based authentication systems, cloned voices may be exploited in social engineering attacks, including impersonation of family members or supervisors in fraudulent calls, manipulation of public opinion through fabricated statements attributed to celebrities or politicians, and unauthorized activation of voice-controlled systems such as smart assistants. The widespread availability of high-fidelity voice

TABLE 13: Results of passive detection methods.

EER(%)	XTTS	ElevenLabs	VALL-E X	FreeVC	DDDMVC	Average
AASIST	31.80	53.60	52.28	58.41	37.86	46.14
SSL-AS	13.35	25.70	20.28	24.89	4.40	16.24

TABLE 14: Results of proactive detection methods.

EER(%)	XTTS	ElevenLabs	VALL-E X	FreeVC	DDDMVC	Average
NOMark	76.94	82.70	77.51	87.62	87.48	84.49
ASMark	0.0	0.0	0.0	0.0	0.0	0.0
WMMark	77.27	86.70	82.27	89.04	87.76	86.21

cloning tools significantly lowers the barrier for executing such attacks.

Defense Implications. Our results show that even a single short utterance is sufficient for adversaries to generate convincing voice clones. Individuals are therefore encouraged to limit the public sharing of voice samples, including those found in social media content, podcasts, or video recordings. Notably, even low-quality audio may be exploitable after enhancement using denoising or source separation techniques. While fully preventing voice exposure is challenging, service providers should avoid relying solely on voice biometrics. Instead, voice-based authentication should be paired with multi-factor authentication or supplementary behavioral verification to ensure security in sensitive applications.

Detection Challenges. Although we evaluated both passive and proactive detection techniques, reliably identifying cloned voices in the wild remains an open research challenge. Existing methods either fail to generalize to unseen conditions or require strict adherence to generation protocols. Future research should focus on developing detection systems that are robust, adaptable, and feasible for real-world deployment.

8. Limitations

While this study offers a comprehensive assessment of voice cloning risks, it has several limitations that warrant further attention.

Controlled Reference Conditions. Our experiments on real-world custom datasets relied on relatively clean reference samples recorded under controlled conditions. In practice, adversaries may obtain voice data from less ideal sources such as social media or phone conversations, where background noise, compression artifacts, and spontaneous speech patterns may reduce the effectiveness of cloning attacks.

Informed Human Evaluation. Our human evaluation was conducted in quiet environments, and participants were informed that their task was to identify cloned voices. This awareness may have introduced bias and does not fully reflect real-world situations where listeners are typically unaware of the presence of synthetic audio.

Partial Authentication Modeling. Our analysis of real-world banking apps focused solely on the voice authentication layer. We did not evaluate other security components such as multi-factor authentication (e.g., SMS verification), liveness detection, or behavioral monitoring, which may mitigate the attack success in practice.

Future Directions. Future research may expand the evaluation to include real-world noisy speech, spontaneous dialogue, and more diverse application contexts. In addition, further exploration of adaptive attack strategies, the robustness of liveness detection, and the integration of watermarking-based defenses would provide a more comprehensive understanding of evolving threats.

9. Conclusion

In this paper, we investigate the threats posed by current advanced voice cloning techniques by presenting a large-scale study and exploring several key research questions. Our results demonstrate that with only a single utterance of the victim speaker, an attacker can achieve a successful attack, and existing detection methods fail to detect voice cloning samples in real-world scenarios. The detailed analysis reveals that human subjective opinions and perceptions of voice cloning also impact our ability to identify a cloned voice. Our work highlights the severity of risks posed by state-of-the-art voice cloning to everyone in the current era, provides a further understanding of how demographic characteristics and human subjective perception impact voice cloning attacks, and emphasizes the deficiency of existing countermeasures.

Acknowledgements

We appreciate the shepherd's efforts and all the reviewers' valuable comments. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China (LY24F020007, LD24F020010); in part by the Key R&D Programme of Zhejiang Province (2025C02264, 2024C01012); in part by the National Natural Science Foundation of China (62102354, 62032021, 62172359); National Key Research and Development Program of China (2023YFB3107402, 2023YFB2904000, 2023YFB2904001).

References

[1] Tencent Holdings Limited, "Wechat." [Online]. Available: https://www.wechat.com

[2] JP Morgan Chase, "Chase voice id." [Online]. Available: https://www.chase.com/personal/voice-biometrics

[3] HSBC, "Hsbc voice id." [Online]. Available: https://www.us.hsbc.com/customer-service/voice/

[4] Amazon, "Alexa ai." [Online]. Available: https://developer.amazon.com/en-US/alexa

[5] Alibaba AI, "Tmall genie." [Online]. Available: https://www.alizila.com/video/what-is-tmall-genie/

[6] Apple, "Apple siri." [Online]. Available: https://www.apple.com/siri/

[7] Samsung, "Samsung bixby." [Online]. Available: https://www.samsung.com/us/apps/bixby/

[8] Catherine Stupp, "Fraudsters used ai to mimic ceo's voice in unusual cybercrime case," Wall Street Journal, 2019.

[9] Dennislaw News, "Fraudsters cloned company director's voice in $35 million bank heist, police find," Dennislaw News, 2021.

[10] B. I. Report, "Audio deepfakes: Cutting-edge tech with cutting-edge risks," https://www.bradley.com/insights/publications/2024/01/audio-deepfakes-cutting-edge-tech-with-cutting-edge-risks, 2023.

[11] C. Mui, "A mysterious phone call cloned biden's voice. can the next one be stopped?" https://www.politico.com/news/2024/01/29/phone-call-biden-voice-clone-00108941, 2024.

[12] Carnegie Mellon University, "Festvox," http://festvox.org/.

[13] M. Shirvanian, S. Vo, and N. Saxena, "Quantifying the breakability of voice assistants." in *Proc. of IEEE PERCOM*, Kyoto, Japan, 2019, pp. 1–11.

[14] E. Wenger, M. Bronckers, C. Cianfarani, J. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, ""hello, it's me": Deep learning-based speech synthesis attacks in the real world." in *Proc. of ACM CCS*, Virtual Event, Korea, 2021, pp. 235–251.

[15] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss." in *Proc. of ACM ICML*, Long Beach, California, USA, 2019, pp. 5210–5219.

[16] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. López-Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." in *Proc. of MIT Press NeurIPS*, Montréal, Canada, 2018, pp. 4485–4495.

[17] The ElevenLabs Team, "Ai voice cloning," 2024. [Online]. Available: https://elevenlabs.io/voice-cloning

[18] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers." *arXiv preprint*, 2023.

[19] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, "Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models." *arXiv preprint*, 2024.

[20] G. R. Doddington, "Speaker recognition–identifying people by their voices." *Proc. IEEE*, vol. 73, no. 11, pp. 1651–1664, 1985.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition." in *Proc. of INTERSPEECH*, Hyderabad, India, 2018, pp. 1086–1090.

[22] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification." in *Proc. of INTERSPEECH*, Shanghai, China, 2020, pp. 3830–3834.

[23] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis." in *Proc. of IEEE ICASSP*, Istanbul, Turkey, 2000, pp. 1315–1318.

[24] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis." in *Proc. of IEEE ICASSP*, Seattle, Washington, USA, 1998, pp. 285–288.

[25] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." in *Proc. of ACM ICML*, Virtual Event, 2021, pp. 5530–5540.

[26] C. Donahue, J. J. McAuley, and M. S. Puckette, "Adversarial audio synthesis." in *Proc. of ICLR*, New Orleans, LA, USA, 2019.

[27] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network." in *Proc. of AAAI AAAI*, Honolulu, Hawaii, USA, 2019, pp. 6706–6713.

[28] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme." in *Proc. of ICLR*, Virtual Event, 2022.

[29] S. Chen, S. Liu, L. Zhou, Y. Liu, X. Tan, J. Li, S. Zhao, Y. Qian, and F. Wei, "Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers." *arXiv preprint*, 2024.

[30] A. Ogihara, H. Unno, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification." *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. 88, no. 1, pp. 280–286, 2005.

[31] P. L. D. Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis." in *Proc. of INTERSPEECH*, Portland, Oregon, USA, 2012, pp. 370–373.

[32] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting ai-synthesized speech using bispectral analysis." in *Proc. of IEEE CVPR*, Long Beach, CA, USA, 2019, pp. 104–109.

[33] L. Attorresi, D. Salvi, C. Borrelli, P. Bestagini, and S. Tubaro, "Combining automatic speaker verification and prosody analysis for synthetic speech detection." in *Proc. of IEEE ICPR*, Montreal, QC, Canada, 2022, pp. 247–263.

[34] L. Blue, K. Warren, H. Abdullah, C. Gibson, L. Vargas, J. O'Dell, K. R. B. Butler, and P. Traynor, "Who are you (i really wanna know)? detecting audio deepfakes through vocal tract reconstruction." in *Proc. of USENIX Security*, Boston, MA, USA, 2022, pp. 2691–2708.

[35] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A comparative study on physical and perceptual features for deepfake audio detection." in *Proc. of ACM ACM MM*, Lisboa, Portugal, 2022, pp. 35–41.

[36] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. W. D. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2." in *Proc. of IEEE ICASSP*, Virtual Event / Toronto, ON, Canada, 2021, pp. 6369–6373.

[37] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *arXiv preprint arXiv:2110.01200*, 2021.

[38] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models." in *Proc. of ACM ICML*, Honolulu, HI, USA, 2023, pp. 17 061–17 084.

[39] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models." in *Proc. of IEEE ICCV*, Paris, France, 2023, pp. 22 409–22 420.

[40] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "Wavmark: Watermarking for audio generation." *arXiv preprint*, 2023.

[41] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, "Detecting voice cloning attacks via timbre watermarking." in *Proc. of ISOC NDSS*, San Diego, CA, USA, 2024.

[42] R. S. Roman, P. Fernandez, A. Défossez, T. Furon, T. Tran, and H. Elsahar, "Proactive detection of voice cloning with localized watermarking." in *Proc. of ACM ICML*, Vienna, Austria, 2024.

[43] B. Kinsella and A. Herndon, "Deepfake and voice clone consumer sentiment report." voicebot.ai, Synthedia, Pindrop, Tech. Rep., 2023.

[44] Joseph Cox, "How i broke into a bank account with an ai-generated voice," https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice, 2023.

[45] P. L. D. Leon, V. R. Apsingekar, M. Pucher, and J. Yamagishi, "Revisiting the security of speaker verification systems against imposture using synthetic speech." in *Proc. of IEEE ICASSP*, Dallas, Texas, USA, 2010, pp. 1798–1801.

[46] T. Kinnunen, Z. Wu, K.-A. Lee, F. Sedlak, E. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech." in *Proc. of IEEE ICASSP*, Kyoto, Japan, 2012, pp. 4401–4404.

[47] Z. Wu, A. Larcher, K.-A. Lee, E. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints." in *Proc. of INTERSPEECH*, Lyon, France, 2013, pp. 950–954.

[48] D. Mukhopadhyay, M. Shirvanian, and N. Saxena, "All your voices are belong to us: Stealing voices to fool humans and machines." in *Proc. of Springer ESORICS*, Vienna, Austria, 2015, pp. 599–621.

[49] E. Khoury, L. E. Shafey, and S. Marcel, "Spear: An open source toolbox for speaker recognition based on bob." in *Proc. of IEEE ICASSP*, Florence, Italy, 2014, pp. 1655–1659.

[50] J. Betker, "Better speech synthesis through scaling." *arXiv preprint*, 2023.

[51] Kanetkar, C. Burroughs, and Riddhi, "The fomo is real for venture capitalists paying big premiums to invest in ai startups right now," *Business Insider*, 2023.

[52] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling." *arXiv preprint*, 2023.

[53] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint*, 2023.

[54] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[55] G. Eren and The Coqui TTS Team, "Coqui TTS," 2021. [Online]. Available: https://github.com/coqui-ai/TTS

[56] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion." in *Proc. of AAAI AAAI*, Vancouver, Canada, 2024, pp. 17 862–17 870.

[57] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P. A. Torres-Carrasquillo, and N. Dehak, "State-of-the-art speaker recognition with neural network embeddings in nist sre18 and speakers in the wild evaluations." *Comput. Speech Lang.*, vol. 60, 2020.

[58] Resemblyzer, "Resemblyzer." [Online]. Available: https://github.com/resemble-ai/Resemblyzer

[59] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset." in *Proc. of INTERSPEECH*, Stockholm, Sweden, 2017, pp. 2616–2620.

[60] X. Xiang, S. Wang, H. Huang, Y. Qian, and K. Yu, "Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition." in *Proc. of APSIPA*, Lanzhou, China, 2019, pp. 1652–1656.

[61] L. Wan, Q. Wang, A. Papir, and I. López-Moreno, "Generalized end-to-end loss for speaker verification." in *Proc. of IEEE ICASSP*, Calgary, AB, Canada, 2018, pp. 4879–4883.

[62] M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, Y. Wang, P. Mousavi, L. D. Libera, A. Ploujnikov, F. Paissan, D. Borra, S. Zaiem, Z. Zhao, S. Zhang, G. Karakasidis, S.-L. Yeh, P. Champion, A. Rouhe, R. Braun, F. Mai, J. Zuluaga-Gomez, S. M. Mousavi, A. Nautsch, X. Liu, S. Sagar, J. Duret, S. Mdhaffar, G. Laperriere, M. Rouvier, R. D. Mori, and Y. Esteve, "Open-source conversational ai with speechbrain 1.0," 2024. [Online]. Available: https://arxiv.org/abs/2407.00463

[63] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus." in *Proc. of LREC*, Marseille, France, 2020, pp. 4218–4222.

[64] Surfing.AI, "ST-AEDS-20180100_1, Free ST American English Corpus," https://www.openslr.org/45/.

[65] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for cstr voice cloning toolkit." 2019.

[66] Google, "Crowdsourced high-quality nigerian english speech data set," https://openslr.elda.org/70/.

[67] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source Multi-speaker Corpora of the English Accents in the British Isles," in *Proc. of LREC*, Marseille, France, 2020, pp. 6532–6541.

[68] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books." in *Proc. of IEEE ICASSP*, South Brisbane, Queensland, Australia, 2015, pp. 5206–5210.

[69] Clark, "At what age is your hearing best?" https://www.betterhearinghealth.com/blog/249738-at-what-age-is-your-hearing-best, 2020.

[70] H. Tak, M. Todisco, X. Wang, J. weon Jung, J. Yamagishi, and N. W. D. Evans, "Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation." in *Proc. of Odyssey*, Beijing, China, 2022, pp. 112–119.

[71] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Proc. of INTERSPEECH*, Incheon, Korea, 2022, pp. 2783–2787.

[72] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations." in *Proc. of MIT Press NeurIPS*, Virtual Event, 2020.

[73] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, "Wavmark: Watermarking for audio generation," *arXiv preprint*, 2023.

Appendix A. Phrases for Enrollment

For all voice authentication systems, user enrollment is required initially. In this study, we evaluated several real-world voice authentication systems, each with different enrollment requirements. Below, we provide details on the phrases used during enrollment.

Specifically, for the two smart speakers, Echo requires the user to pronounce five fixed sentences (see Table 15), while TmallGenie requires the user to pronounce a single sentence from varying physical distances. For WeChat and the bank app, both require the user to pronounce sequences of eight digits. The difference is that during the verification stage, WeChat uses the same digit sequence as in enrollment, while the bank app randomly generates new digit sequences for enhanced security.

TABLE 15: Phrases for enrollment.

Phrases Used for Echo Dot.
Alexa
Alexa, What's the temperature outside?
Alexa, play music
Alexa, turn off the light
Alexa, add milk to my shopping list
Phrases Used for TmallGenie.
Tian mao jing ling.
Phrases Used for WeChat/Bank.
Sequences of eight digits.

Appendix B. Phrases for Voice Cloning

To simulate attacks on various real-world voice authentication systems, we designed different malicious phrases for voice cloning. For attacks on open-source and real-world voice authentication systems in Sections 4.2 and 4.3, we use the phrases presented in Table 16, which cover multiple possible real-world attack scenarios.

For the two smart speakers, we developed specific phrases for different attack tasks, as described in Section 4.5, to simulate three types of attacks: Identity Theft, Privacy Breach, and Unauthorized Transaction (see Table 17).

Appendix C. Phoneme Coverage

In Section 5.2, we examine the impact of phoneme coverage on voice cloning attacks. Phoneme coverage refers to the proportion of desired phonemes that are present in the reference utterance. Here, we detail the phonemes of the two types of reference utterances used in Section 5.2.2 (see Table 18) and explain how we calculated phoneme coverage. Since the target voice samples in this experiment are all 8-digit sequences, the phoneme coverage of the digits reference is 100%. For the text reference utterance, 8 out of the 16 phonemes match those in a typical 10-digit sequence. Therefore, for any 8-digit sequence, the phoneme coverage for the text reference utterance is at most 50%.

Appendix D. Detailed Results across Datasets

Here, we present the results of voice cloning attacks using different techniques across multiple datasets and various voice authentication systems. Note that the average attack success rate (ASR) is a weighted average calculated as follows:

$$Average = \sum_{D \in Datasets} W_D * ASR_D$$

where W_D represents the proportion of speakers from dataset D relative to the total number of speakers, and ASR_D is the attack success rate for dataset D. The calculation procedure for VCR (Victim Compromise Rate) follows a similar approach.

TABLE 16: Desired phrases targeting on open-source and real-world voice authentication systems.

Phrases Used for Voice Cloning
Open the door and turn on the light.
Activate the access control system.
Transfer ten thousands dollars to Bob.
Confirm that the recipient of the transfer is Bob.
This is Alice, confirm the transaction.
Play my recently recorded voice memos.
Display all the family contacts in my address book.
Please provide me with the location history of my phone for the past week.
I am Barack Obama, and I am running for this year's election.
I am Donald Trump, and I am preparing to propose a new immigration policy.

TABLE 17: Desired phrases used for smart speakers. T1, T2, and T3 stands for Identity Theft, Privacy Breach, and Unauthorized Transaction, respectively.

Task	**Phrases Used for Echo Dot.**
T1	Alexa, who am I?
T2	Alexa, what is my latest event?
T3	Alexa, add a pen to my cart. Buy it now.
Task	**Phrases Used for TmallGenie.**
T1	Tianmao Jingling, wo shi shui?
T2	Tianmao Jingling, zui xin kuai di xiao xi.
T3	Tianmao Jingling, wo yao mai niu nai. Xia dan bing zhi fu.

TABLE 18: Phonemes of two types of reference utterances (normal text and digits sequence) in Section 5.2.2.

Text	Tiānmāo Jīnglíng shì Ālǐbābā qíxià zhìnéng chǎnpǐn pǐnpái, tā kěyǐ tōngguò Tiānmāo Jīnglíng ruǎnjiàn liánjiē hé kòngzhì nǐ de Tiānmāo Jīnglíng xìliè yīnxiāng.
Phonemes	ian2, n, iang1, j, an3, b, q, l, r, h, i3, a1, ai2, in1, ong1, ia4, in3, t, ch, zh, ian1, ie4, uan3, uo4, i4, e, k, sh, i2, ing1, ian4, ing2, g, d, ao1, x, m, ie1, ong4, eng2, e2, e3, y, p
Digits	líng yī èr sān sì wǔ liù qī bā jiǔ
Phonemes	w, iu4, q, y, iu3, i4, b,an1, l, er4, ing2, j, u3, a1, i1, s

TABLE 19: Detailed results of different voice cloning techniques across datasets based on Ecapa SR system.

Metric	ASR(%)							VCR(%)						
Dataset	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**
XTTS	58.71	92.33	100.00	89.74	92.43	91.57	91.37	96.77	100	100	98.01	98.91	99.07	98.62
FreeVC	0.00	16.88	77.50	9.67	10.40	71.53	11.23	0	52.5	100	36.68	34.59	88.89	36.35
DDDMVC	0.32	23.79	72.00	44.06	27.55	15.56	32.81	3.23	84.17	100	83.87	73.33	66.67	76.69
ElevenLabs	61.29	99.67	100.00	79.70	81.40	88.70	86.46	93.55	100	100	95	96	99.07	97.44
vallex	32.26	80.75	100.00	82.45	80.22	85.89	80.89	58.06	96.67	100	95.94	95.36	99.07	95.48

TABLE 20: Detailed results of different voice cloning techniques across datasets based on ResNet SR system.

Metric	ASR(%)							VCR(%)						
Dataset	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**
XTTS	96.45	98.42	100.00	96.24	96.72	96.48	96.59	100	100	100	99.34	99.33	100	99.36
FreeVC	0.00	14.33	74.00	9.07	11.69	69.72	11.76	0	43.33	90	31.6	36	84.26	35.29
DDDMVC	2.74	23.79	78.00	50.41	33.37	8.06	38.50	19.35	81.67	90	85.4	79.56	46.3	80.82
ElevenLabs	88.06	99.50	100.00	79.70	83.00	82.87	87.19	96.77	100	100	98	97	97.22	98.08
vallex	49.68	85.67	95.00	78.30	80.91	74.26	79.90	74.19	97.5	100	94.54	95.66	95.37	95.22

TABLE 21: Detailed results of different voice cloning techniques across datasets based on Resemblyzer system.

Metric	ASR(%)							VCR(%)						
Dataset	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**
XTTS	19.03	57.67	17.00	77.46	78.92	20.74	76.85	58.06	85	60	95.33	95.19	54.63	94.25
FreeVC	0.16	0.25	0.00	8.07	1.71	7.73	3.91	0	0.83	0	25.64	6.38	25.93	13.03
DDDMVC	4.84	24.46	0.00	55.98	45.68	7.13	47.97	25.81	70	0	91.69	85.67	36.11	86.31
ElevenLabs	43.55	91.83	17.00	88.60	80.10	49.81	74.18	83.87	100	50	98	96	87.04	93.60
vallex	62.90	78.83	55.00	88.95	85.79	72.04	86.39	83.87	96.67	90	96.65	96.86	97.22	96.73

TABLE 22: Detailed results of different voice cloning techniques across datasets based on iFlytek voice authentication service.

Metric	ASR(%)							VCR(%)						
Dataset	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**	**CSNED**	**CSUKIED**	**FST**	**LS**	**MCV**	**VCTK**	**Average**
XTTS	94.84	69.83	68.00	69.80	68.10	96.48	77.21	100	98.33	90	94	97	100	97.44
FreeVC	29.52	0.62	3.50	14.50	5.30	91.25	27.42	67.74	7.5	20	47	12	100	42.43
DDDMVC	53.87	25.17	19.50	58.75	49.60	36.62	41.95	93.55	78.33	70	97	92	87.96	88.27
ElevenLabs	92.58	84.00	80.00	70.80	71.60	92.78	81.05	100	100	100	95	98	100	98.51
vallex	68.20	52.81	63.00	65.18	66.01	89.38	67.92	87.1	87.5	100	90	92	99.07	91.90

Appendix E. Meta-Review

The following meta-review was prepared by the program committee for the 2025 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

E.1. Summary

This work performs a large-scale evaluation of 5 VC systems against 8 VA systems and human listeners, using voice data from over 7,000 speakers. It explores their ability to spoof identities against voice recognition systems, the factors of the listener/victim that affect this success, and whether detection methods can identify these spoofs. It demonstrates that modern VC techniques can successfully spoof both automated systems and human listeners with high success rates, even with just one utterance. Overall, the work finds that several commercial systems and people are vulnerable.

E.2. Scientific Contributions

- Identifies an Impactful Vulnerability
- Provides a Valuable Step Forward in an Established Field
- Independent Confirmation of Important Results with Limited Prior Research

E.3. Reasons for Acceptance

1) The work evaluates diverse voices and accents and uses several real-world voice cloning and authentication systems.
2) This work collects numerous public datasets, curates a custom dataset by recruiting participants, develops custom evaluation metrics, and provides a detailed overview of voice cloning.

E.4. Noteworthy Concerns

1) The creators of VALL-E did not release an official version of its implementation. However, an open-source implementation is used in this work to demonstrate the risks of voice cloning. The Conqui platform doesn't exist anymore. However, this work reflects the risk due to FreeVC and XTTS (which are open source).
2) There is limited information about WeChat's or the Bank's VA system. Since this work treats these systems as black-box VA systems, there is limited transparency on how these VA systems are affected by Voice Cloning attacks.
3) This work assumes that the six datasets weren't used to train voice cloning techniques. Furthermore, there is limited visibility into the datasets used in commercial VC services.