Deepfake Voice Command Attacks on Automatic Speaker Recognition Systems

Marco Micheletto
marco.micheletto@unica.it
Giulia Orru
giulia.orru@unica.it
Anna Setzu
a.setzu7@studenti.unica.it
Mattia Tronci
m.tronci21@studenti.unica.it
Matteo Trudu
m.trudu17@studenti.unica.it
Gian Luca Marcialis
marcialis@unica.it

Department of Electrical and
Electronic Engineering,
University of Cagliari, Italy

Abstract

Automatic speaker verification is increasingly deployed in security applications, including remote identity verification, internet banking, and access control systems. Although these systems have achieved strong performance under clean conditions, they remain vulnerable to logical access attacks, where synthetic speech is injected directly into the system to impersonate a legitimate user. This paper investigates the effectiveness of such attacks under a black-box threat model using a reproducible pipeline based on retrieval-based voice conversion. We evaluate spoofing success across multiple datasets, analyzing how pitch manipulation and gender pairing affect viability. A focused evaluation on command-based speech is performed using a new dataset of short utterances collected under controlled conditions. To complement biometric performance, we assess the intelligibility of cloned commands through automatic speech recognition, providing further insight into the risks posed by realistic voice cloning.

1 Introduction

Voice-based authentication is increasingly adopted in a wide range of security-critical applications, including smart assistants, home automation, and hands-free access control. In these scenarios, users typically interact with the system through short spoken commands, relying on automatic speaker verification (ASV) to authenticate their identity in a seamless and natural way [17]. While ASV systems have improved significantly in recent years, they remain vulnerable to spoofing attacks based on synthetic speech [6, 7]. These attacks are typically categorized into two types [23]: *presentation attacks* (PA), in which malicious inputs are physically presented to the sensor (e.g., via loudspeaker replay), and *logical access* *attacks* (LA), where synthetic or manipulated speech is injected directly into the system in digital form. This work focuses on the latter, which bypasses the sensor entirely and targets the system at the feature or signal level. Synthetic speech can be generated via *text-to-speech* (TTS) or *voice conversion* (VC). TTS maps text to speech, while VC adapts a source utterance to the vocal traits of a target speaker. In particular, Retrieval-based VC (RVC) enables high-fidelity mimicry with little data and no system access [27], heightening risks in command-driven scenarios where utterance content is predictable. Although ASV vulnerabilities have been studied under diverse models and conditions [10, 25], little work examines short fixed-phrase commands, intelligibility constraints, and composite ASV+ASR pipelines. We address this gap with a targeted evaluation of spoofing effectiveness, focusing on the interaction between voice similarity and command recognition. We present a reproducible black-box pipeline based on RVC, and perform a structured multi-dataset evaluation across two ASV systems. Additionally, we introduce the *Voice Command Identity Dataset (VocID)*[1], a command-oriented speech corpus acquired under controlled conditions, and use it to assess both biometric vulnerability and intelligibility under deepfake attacks. Our contributions are summarized as follows: (1) we replicate and contextualize the vulnerability of ASV systems under a realistic logical access threat model, focusing on short command-based utterances; (2) we analyze how pitch manipulation and base-target speaker gender pairing influence spoofing success across two ASV systems and three datasets; we assess the impact of intelligibility constraints by integrating an ASR component, and evaluate joint ASV+ASR acceptance rates to simulate realistic usage scenarios; we release the *VocID* dataset to support further research on command-level spoofing evaluation and intelligibility-aware authentication.

2 Related works

The rise of generative speech models has significantly impacted biometric authentication, particularly in speaker verification. Text-to-speech systems have evolved from rule-based pipelines to neural networks capable of generating highly natural, speaker-specific audio. Models like Tacotron [26] and FastSpeech [21] laid the foundation for expressive end-to-end synthesis via spectrogram prediction. More recent approaches integrate waveform generation into unified architectures, often combining linguistic encoding with adversarial training to improve fidelity [12, 13]. Voice conversion, by contrast, aims to alter speaker identity while preserving linguistic content. A major advance is the shift to non-parallel training, which removes the need for aligned recordings [4, 11]. These capabilities pose a growing threat to speaker verification, especially in black-box settings where attackers have no access to the system internals but can collect short audio samples of the target [9, 22]. Recent work has shown that even brief samples suffice to build effective impersonators when high-capacity VC techniques are used [3, 16]. Among them, retrieval-based voice conversion (RVC) stands out for its ability to disentangle speaker and linguistic information, re-synthesizing high-fidelity speech with minimal training data. In response to these threats, considerable effort has been devoted to the development of spoofing detection methods, with the ASVspoof challenge playing a key role in benchmarking countermeasures under standardized conditions [6, 23]. The SpoofCeleb benchmark [10] further expanded this line of research by introducing a large-scale evaluation suite spanning 23 spoofing techniques and multiple ASV systems, including spoofing-aware ASV (SASV) pipelines trained on VoxCeleb. Similarly, recent analyses examined ASV robustness under TTS/VC attacks [9] and

[1] https://github.com/PRALabBiometrics/VocID/

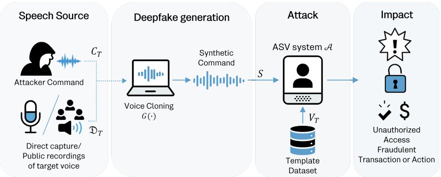

Figure 1: Attack scenario investigated in this study.

multimodal deepfakes [14]. Practical studies also tested commercial assistants (Alexa, Siri, Bixby) against synthetic speech [1, 25], exposing challenges from device-specific latency and ASR integration.

While prior work mainly considers long-form or replay scenarios, limited research has focused on short, fixed-phrase commands typical of authentication pipelines. Our contribution is a targeted, multi-dataset evaluation of deepfake spoofing under pitch and gender variation within a composite ASV+ASR framework.

3 Methodology

In this section, we describe the overall pipeline of our attack simulation, aligned with the scenario (Figure 1), which is organized into four main stages: *Speech Source*, *Deepfake Generation*, *Attack*, and *Impact*. In the following subsections, we formalize the attack path and introduce the datasets, generation models, and speaker recognition systems.

3.1 Threat Model

We consider a logical access attack scenario in which a malicious actor attempts to impersonate a target speaker by injecting synthetic speech signals generated through voice cloning techniques into an ASV system. The attacker's objective is to produce a synthetic speech signal that is accepted by the system as if it originated from the legitimate enrolled speaker. The attack is carried out in a black-box setting, that is, without knowledge of the architecture, embeddings, or thresholds of the ASV system. However, access to the following resources is considered available to the adversary:

- a small set of speech recordings belonging to the target speaker, denoted as $\mathcal{D}_T$;
- a generative function $G(\cdot)$ capable of synthesizing speech conditioned on an input utterance and the speaker identity inferred from $\mathcal{D}_T$;
- knowledge of the intended voice command to be issued, denoted C_T.

The attacker speaks the desired command C_T using their own voice, and the generative model produces a synthetic version S of the utterance in the target speaker's voice:

$$S = G(C_T \mid \mathcal{D}_T) \tag{1}$$

The attacker then presents S to the ASV system. Let $\mathcal{A}$ denote the ASV system, and T_V the enrolled template corresponding to speaker V_T. The attack is considered successful if the similarity score $\mathcal{A}(S, T_V)$ exceeds the decision threshold τ, leading to a false acceptance:

$$\mathcal{A}(S, T_V) \geq \tau \tag{2}$$

Such attacks are especially dangerous in voice-controlled systems that rely on short commands for secure operations, including banking interfaces, smart home systems and item and voice-based authentication.

3.2 Speech Source

We use three datasets to simulate enrollment data and to provide speaker-specific material for voice cloning and verification. Two of them (VoxCeleb1 and VoxCeleb2) are publicly available datasets widely adopted in the speaker recognition community. The third is a proprietary collection designed explicitly for command-driven speaker verification under controlled conditions: the Voice Command Identity Dataset (VocID).

VocID – Voice Command Identity Dataset

As part of this study, we collected a dataset tailored for speaker verification based on short spoken commands. The corpus comprises recordings from 30 Italian-native participants (18 male, 12 female), acquired under supervised and noise-controlled conditions. The participants span a range of ages, although no specific restrictions on regional accent were imposed.

Each speaker completed a fixed protocol comprising:

- 8 single-command sessions: 15 repetitions each of `conferma`, `confirm`, `accetta`, `accept`, `rifiuta`, `reject`, `elimina`, `delete`;
- 2 multi-command sessions with 2 repetitions per command, grouped by language;
- 4 reading tasks: two in Italian (short and long) and two in English (short and long).

Recordings were captured using two smartphones: Google Pixel 3a and iPhone 14. Each speaker completed a single session in a quiet environment. All recordings are stored in 16-bit PCM WAV format at 16 kHz. Available metadata include speaker ID, gender, device, and language. Given the command-oriented nature of the utterances, the corpus is particularly suited for studying voice-driven security applications such as transaction approval, smart home access, and digital authorization. This results in a total of 50.400 audio samples, composed of 43.200 spoofed and 7.200 bona fide utterances All participants were fully informed about the scope, objectives, and intended use of the data collected. Participation was voluntary with explicit written consent.

VoxCeleb1 and VoxCeleb2

We used pre-trained ASV back-end models that were originally trained on the VoxCeleb1 (*Vox1*) and VoxCeleb2 (*Vox2*) datasets. These corpora are widely adopted benchmarks in the literature, offering large-scale, unconstrained audio data collected from speakers in diverse conditions. *Vox1* [18] and *Vox2* [6] are large-scale corpora of unconstrained speech collected from available online interview videos. They feature thousands of speakers recorded under various acoustic conditions and are standard benchmarks for speaker recognition research. For these reasons, in the experimental evaluation, we selected these datasets to simulate a baseline *intra-dataset* protocol. To maintain consistency with the trial structure used in the VocID dataset, we selected 30 speakers from the development subset (15 male, 15 female). For each speaker, one enrollment segment of about 20 seconds was extracted. The remaining speech material was segmented into 60 non-overlapping 2-second clips, for a total of 1.800 probe utterances. All probes were drawn from regions disjoint from the enrollment audio to avoid temporal overlap and ensure a clean separation between enrollment and test material.

3.3 Deepfake Generation

To synthesize spoofed utterances for each target speaker, we employ the Retrieval-based Voice Conversion (RVC) framework[2], a non-parallel speech-to-speech synthesis system that enables high-fidelity voice cloning using limited training data. RVC operates by disentangling the linguistic content from speaker identity in a source utterance and reconstructing

[2] Implementation of RVC can be found at: https://github.com/RVC-Project/Retrieval-based-Voice-Conversio n-WebUI. Last accessed: 5 August 2025

the same linguistic content using the vocal characteristics of a target speaker. The generative process $G(\cdot)$ used in this work follows a four-stage pipeline:
1. Feature extraction: the input utterance C_T is processed with a pre-trained HuBERT model [8], which converts the waveform into a sequence of discrete linguistic units $\mathbf{V}$ that are invariant to speaker identity. These features encode the phonetic content of the utterance.
2. Speaker adaptation: for each target speaker, a dedicated voice conversion model is fine-tuned using a small set of enrollment utterances $\mathcal{D}_T$. This allows the model to learn the speaker-specific representation $\mathbf{e}_T$ required for accurate identity transfer.
3. Voice conversion: the features $\mathbf{V}$ and the speaker representation $\mathbf{e}_T$ are passed to a neural vocoder based on VITS [12], which synthesizes a waveform $\hat{S}$ in the target speaker's voice:

$$\hat{S} = G(C_T \mid \mathcal{D}_T) = \text{Vocoder}(\mathbf{V}, \mathbf{S}) \tag{3}$$

4. Pitch adjustment: A pitch shift T is applied to the waveform $\hat{S}$ to simulate cross-gender adaptation.

3.4 Automatic Speaker Verification Systems

We evaluate the effectiveness of the spoofed commands against two representative automatic speaker verification systems. Both systems operate in a verification setting, comparing a probe signal S against an enrollment template T_V to produce a similarity score $\mathcal{A}(S, T_V)$.

WeSpeaker (denoted as WS) [24] is an open-source speaker verification framework based on deep discriminative embeddings. In our experiments, we adopt the English pre-trained model provided by the toolkit, built on a ResNet-based architecture trained on *Vox2* data. The system operates in a text-independent setting and outputs fixed-length embeddings from speech utterances, which are compared via cosine similarity for verification.

SpeechBrain (denoted as SB) [20] is an open-source toolkit that includes several speaker verification recipes. In our experiments, we use the ECAPA-TDNN-based speaker recognition model trained on *Vox1* and *Vox2*.Similar to WeSpeaker, the system produces fixed-length embeddings and performs cosine-based verification.

4 Experimental Protocol

This Section presents the evaluation setup used to assess the vulnerability of automatic speaker verification systems to cloned voice commands. The protocol, spoofing pipeline, and training configuration are detailed below, along with the evaluation metrics.
To operationalize the threat model defined in Section 3.1, we implement a controlled pipeline for the generation of deepfake probes. For each enrolled speaker V_T, a set of synthetic utterances is generated by conditioning the voice conversion function $G(\cdot)$ on a spoken command C_T issued by the attacker. The output is a deepfake waveform $S = G(C_T \mid \mathcal{D}_T)$ designed to imitate the voice characteristics of V_T while preserving the linguistic content of C_T.
The dataset $\mathcal{D}_T$ used to fine-tune G consists of bona fide recordings from the target speaker. In our setup, $\mathcal{D}_T$ includes approximately 90 seconds of speech for VocID and 120 seconds for VoxCeleb speakers, with slight per-speaker variability. The longer duration for VoxCeleb reflects its more fragmented recording structure and helps maintain balance across datasets.
Three experimental parameters drive the cloning process: (i) the **base voice** V_B, e.i. the identity of the cloning source, either male or female; (ii) the **transpose factor** T, a fixed pitch shift applied to the synthesized signal, with values $T \in \{-8, 0, +8\}$ and (iii) the **command** C_T, short phrases related to voice-activated actions.
Each target-specific cloning model is fine-tuned for 1000 epochs with a batch size of 4. The model operates at a sampling rate of 40 kHz, consistent with the architecture and RVC

pre-trained checkpoints. Training leverages pre-initialized weights for the generator and discriminator. The resulting deepfake commands S are then submitted to the ASV systems for verification against the enrolled templates, forming the basis of logical attack trials.
We adopt a closed-set verification setup. Each speaker is enrolled using a 20-second bona fide segment. Probe comparisons fall into three categories: (i) *mated comparisons*, where the probe and reference belong to the same speaker; (ii) *non-mated comparisons*, involving different speakers; and (iii) *deepfake attacks*, consisting of deepfake utterances generated to mimic the enrolled identity. Then, we assess the system performance using standard verification metrics, including Genuine Acceptance Rate (GAR), False Match Rate (FMR), Receiver Operating Characteristic (ROC), and Area Under the Curve (AUC). To assess spoofing vulnerability, we include the Spoof False Acceptance Rate (SFAR), which quantifies the acceptance rate of deepfake trials. Unlike the ISO-defined IAPAR [2], SFAR reflects logical access attacks with no physical presentation. For spoof detection, we report Attack Presentation Classification Error Rate (APCER) and Bona Fide Presentation Classification Error Rate (BPCER) [2], measuring misclassification of spoofed and bona fide inputs, respectively, at a fixed decision threshold of 0.5.

Finally, we additionally integrate an automatic speech recognition (ASR) module into the protocol. We adopt Whisper [19], a multilingual Transformer-based ASR model, to assess the intelligibility of both bona fide and synthetic commands. Given an audio input, Whisper outputs a transcription $\hat{C_T}$, which is compared to the ground-truth command C_T using the Levenshtein distance [15], denoted as $d_L(\hat{C_T}, C_T)$. A prediction is considered correct if $d_L(\hat{C_T}, C_T) \leq d_{\max}$, where $d_{\max}$ is a tunable threshold controlling the tolerance to transcription errors. This ASR-based intelligibility measure is later combined with ASV scores to analyze the behavior of a composite verification pipeline under logical access attacks.

5 Experimental Results

In this Section, we present a detailed evaluation of the logical access attacks. The results are organized to highlight (i) the general behavior of the ASV systems across different datasets and spoofing conditions, (ii) the influence of pitch and speaker identity on attack effectiveness, and (iii) the specific properties of deepfake speech in the VocID dataset.

Intra- and Cross-Domain Evaluation Figure 2 reports the ROC curves for the ASV systems WS and SB, evaluated on the *VocID*, *Vox1*, and *Vox2* datasets. Two verification conditions are reported: standard biometric verification (GAR vs. FMR) and logical access attack evaluation (GAR vs. SFAR). The systems achieve high performance in the standard verification setting, with ROC curves showing clear separation between mated and non-mated comparisons. Despite being a cross-dataset evaluation, VocID yields results comparable to the intra-dataset scenarios of Vox1 and Vox2, indicating that both ASV systems generalize well to unseen domains when tested on bona fide speech. However, a different picture emerges under deepfake attack conditions. Although Vox1 and Vox2 maintain relatively stable behavior, VocID exhibits marked degradation, with significantly higher SFAR values. The increased vulnerability might be related to the quality of the speech material used during the fine-tuning of the RVC models. In the VocID case, the attacker has access to clean, carefully recorded samples acquired under controlled conditions, potentially enabling more effective speaker adaptation and higher-quality synthesis. In contrast, the deepfakes generated for *Vox1* and *Vox2* are based on in-the-wild recordings, often affected by background noise, acoustic variability, and overlapping speech, factors that reduce the fidelity of the synthesized utterances and limit their effectiveness as attack probes. Moreover, while

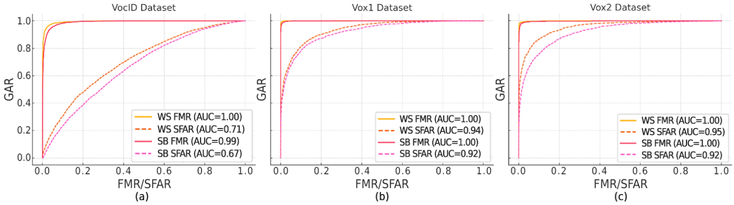

Figure 2: ROC curves for WS and SB on VocID (a), Vox1 (b), and Vox2 (c).

performance differences between ASV systems are generally small in the standard setting, WS consistently yields lower SFAR values under attack, particularly in the most challenging scenario (e.i., VocID). Both systems, however, remain vulnerable to high-quality cloned commands.

Intra- and Cross-Gender Evaluation To better understand how signal-level and speaker-level properties influence spoofing effectiveness, we perform a fine-grained analysis conditioned on the parameters introduced in Section 4. Each attack is defined by a base voice identity $V_B \in \{\text{male}, \text{female}\}$, a target speaker identity $V_T \in \{\text{male}, \text{female}\}$, and a transpose factor $T \in \{-8, 0, +8\}$ controlling the pitch of the generated speech. Figure 3 reports the SFAR in all configurations of V_B, V_T, and T, across the three datasets and the two systems. Several trends emerge from this analysis. We first consider the case where the base voice and target voice belong to the same gender ($V_B = V_T$). In this setting, the transpose factor T plays a crucial role. Attacks with $T = 0$ consistently achieve the highest SFAR, indicating that neutral pitch produces synthetic signals that closely resemble the target speaker, and are therefore more likely to be accepted by the verification system. In contrast, both $T = -8$ and $T = +8$ significantly reduce the effectiveness of the spoofing attempt, particularly in the *Vox1* and *Vox2* datasets, suggesting that unnatural pitch modifications can impair the similarity between S and T_V. When base and target genders differ, the role of T becomes even more critical. Results show a clear asymmetry depending on the direction of the attack. In female-to-male attacks, higher transpose values such as $T = +8$ lead to substantially lower SFAR, likely because the resulting signal is excessively high-pitched. In contrast, $T = -8$ partially mitigates the mismatch by lowering the pitch and improving the success rate. Conversely, in male-to-female attacks, upward transposition results in better performance as it helps approximate the expected spectral characteristics of the female target. In general, no consistent advantage is observed for any gender as a base voice, suggesting that the effectiveness of spoofing depends more on the interaction between V_B, V_T, and T than on the intrinsic properties of the source speaker alone. A mismatch between the natural pitch of V_B and the expected pitch of V_T, if not corrected appropriately through T, degrades the quality of the attack. These dynamics become more evident when comparing datasets. For VocID dataset, even under non-neutral pitch transformations ($T \neq 0$), SFAR values remain high across both intra- and cross-gender conditions. In contrast, Vox1 and Vox2 datasets show greater sensitivity to T, with attacks becoming noticeably less effective as pitch diverges from the natural target range. Finally, the gap between systems remains consistent: SB exhibits higher SFAR than WS across nearly all configurations.

Deepfake Robustness in Command-Oriented Scenarios To further explore the behavior of ASV systems under LAs, we analyze how similarity scores vary depending on recording devices, command content, and recognition accuracy. Figure 4 shows the similarity scores

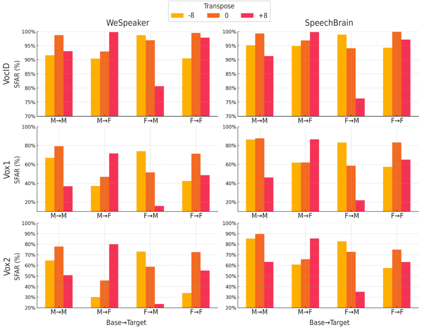

Figure 3: SFAR across datasets and gender configurations. Each bar group shows the effect of pitch manipulation ($T = -8, 0, +8$). Gender pairs on the x-axis indicate the base → target voice.

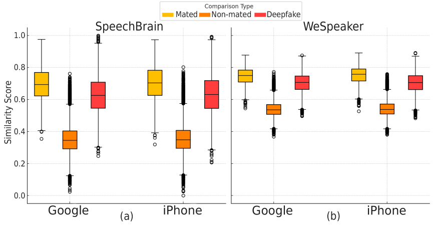

Figure 4: Box plots of similarity scores on VocID, grouped by device and comparison type (mated, non-mated, deepfake) for (a) SB and (b) WS.

between enrollment and probe recordings, recorded using *Google* and *iPhone* devices, evaluated using the two systems SB and WS. Overall, the distribution patterns are consistent with previous findings. Some minor variations across devices can be observed, particularly in the spread of deepfake scores, but no substantial or systematic device effect emerges. Both systems follow the same trend, with SB showing slightly higher scores across all probe types.

A complementary view is offered in Figure 5, which reports the similarity score distributions grouped by target command. Across both systems, non-mated comparisons remain consistently well-separated from the other classes, showing narrow distributions centred at lower values. The distribution of deepfake scores shows some variability across commands, with certain keywords such as *Accept* or *Reject* exhibiting slightly higher overlap with mated probes. This variation may reflect differences in phonetic content, which could influence how well speaker-specific characteristics are retained in the synthesized audio. Overall, WS exhibits more regular and compact score distributions, suggesting a more stable behavior.

Real-world voice interfaces, however, do not rely on similarity scores alone. As outlined in Section 4, our evaluation integrates the Whisper ASR module to account for intelligibility in command-oriented scenarios. Figure 6 reports the cumulative ASR accuracy as a function of the maximum allowed distance $d_{\max}$, separately for bona fide and deepfake utterances. While bona fide speech consistently achieves high accuracy, deepfakes perform noticeably worse at stricter thresholds. Nevertheless, accuracy already approaches 60% at $d_{\max} = 2$, highlighting the potential threat cloned commands pose in voice-controlled security-sensitive scenarios.

To assess the role of intelligibility on spoofing effectiveness, we evaluate a composite ASR+ASV pipeline in which a trial is accepted only if both the ASV score exceeds a biometric threshold and the ASR transcription distance satisfies $d_L \leq d_{\max}$. We select the ASV threshold that yields 99% GAR on the test set (in the absence of ASR) to examine the residual attack surface under strong verification performance. Table 5 reports GAR and SFAR under varying ASR constraints. While stricter ASR filters significantly reduce SFAR, they also lower GAR, as genuine commands may be rejected due to transcription errors. Importantly, even under realistic constraints (e.g., $d_{\max} = 1$), over 50% of deepfake trials are still accepted, underscoring the persistent risk posed by high-fidelity cloned speech.

To complement the evaluation of the spoofing pipeline, a standard deepfake detector

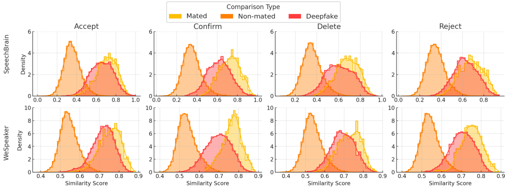

Figure 5: Similarity score distributions per target command, separated by comparison type (Mated, Non-Mated, Deepfake).

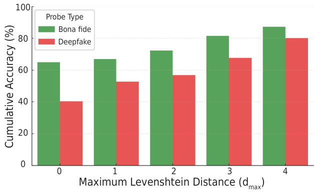

Figure 6: Cumulative ASR accuracy on bona fide and deepfake utterances as a function of the maximum allowed Levenshtein distance $d_{\max}$.

VC Condition	APCER (%)	BPCER (%)
Male voice, transpose 0	4.24	40.11
Male voice, transpose +8	7.03	
Male voice, transpose –8	6.33	
Female voice, transpose 0	24.04	
Female voice, transpose +8	34.75	
Female voice, transpose –8	16.72	

Table 2: APCER and BPCER results of RawNet2 on the VocID dataset.

d_{max}	*no ASR*	*0*	*1*	*2*	*3*	*4*
GAR SB (%)	99.00	64.67	66.56	71.83	80.94	86.42
SFAR SB (%)	94.39	38.13	50.06	53.88	64.01	75.60
GAR WS (%)	99.00	64.64	66.56	71.86	81.00	86.56
SFAR WS (%)	94.15	37.98	49.94	53.69	63.80	75.35

Table 3: GAR and SFAR at varying ASR thresholds d, with fixed biometric threshold ensuring GAR = 99% without ASR.

based on RawNet2 [6], the official baseline from the ASVspoof 2021 challenge, was evaluated on the VocID dataset. The model was used without fine-tuning, simulating a zero-shot scenario to assess cross-domain robustness. As shown in Table 2, the detector achieves low APCER in male voice conditions (4.24–7.03%), but degrades sharply for female voices, reaching 34.75% under pitch-shifted attacks. This gap may stem from gender imbalance in training data or greater pitch variability in female voices, which may obscure synthetic artifacts. Further investigation is required to isolate the acoustic factors responsible for this vulnerability. Despite this, the most critical limitation lies in the high BPCER, which reaches 40.11%. Such performance highlights the unsuitability of out-of-domain detectors in interactive scenarios, where high rejection rates of genuine users undermine system usability.

6 Conclusions

This study investigates vulnerabilities of speaker verification in realistic voice command scenarios, focusing on short utterances and constrained prompts. We developed a reproducible black-box attack pipeline using RVC and evaluated its impact across multiple systems and datasets, including *VocID*, a corpus tailored to this setting. Analysis of pitch manipulation and speaker pairing shows that aligning these parameters with source–target gender relations increases spoofing success. Integrating ASV and ASR into a joint decision pipeline, we showed that deepfakes remain effective even under intelligibility constraints, with over half the attacks succeeding. Finally, evaluation of a state-of-the-art detector in a zero-shot setting reveals critical weaknesses, particularly against pitch-shifted female voice attacks. Overall, our findings reinforce the need for countermeasures that address both biometric and linguistic aspects in voice-based authentication, especially under constrained input and operational conditions.

Acknowledgment

This work is supported by SERICS (PE00000014) under the Italian Ministry of University (MUR) and Research National Recovery and Resilience Plan by the EU - NGEU.

References

[1] Domna Bilika, Nikoletta Michopoulou, Efthimios Alepis, and Constantinos Patsakis. Hello me, meet the real me: Voice synthesis attacks on voice assistants. *Computers & Security*, 137:103617, 2024.

[2] ISO/IEC JTC1 SC37 Biometrics. *ISO/IEC 30107-3. Information technology - biometric presentation attack detection - part 3: testing and reporting*. International Organization for Standardization, 2023.

[3] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pages 2709–2720. PMLR, 2022.

[4] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv preprint arXiv:1904.05742*, 2019.

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, pages 1086–1090, 2018.

[6] Héctor Delgado, Nicholas Evans, Tomi Kinnunen, Kong Aik Lee, Xuechen Liu, Andreas Nautsch, Jose Patino, Md Sahidullah, Massimiliano Todisco, Xin Wang, et al. Asvspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. *arXiv preprint arXiv:2109.00535*, 2021.

[7] Priyanka Gupta, Hemant A Patil, and Rodrigo Capobianco Guido. Vulnerability issues in automatic speaker verification (asv) systems. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):10, 2024.

[8] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[9] Jee-weon Jung, Xin Wang, Nicholas Evans, Shinji Watanabe, Hye-jin Shim, Hemlata Tak, Sidhhant Arora, Junichi Yamagishi, and Joon Son Chung. To what extent can asv systems naturally defend against spoofing attacks? *arXiv preprint arXiv:2406.05339*, 2024.

[10] Jee-weon Jung, Yihan Wu, Xin Wang, Ji-Hoon Kim, Soumi Maiti, Yuta Matsunaga, Hye-jin Shim, Jinchuan Tian, Nicholas Evans, Joon Son Chung, et al. Spoofceleb: Speech deepfake detection and sasv in the wild. *IEEE Open Journal of Signal Processing*, 2025.

[11] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273. IEEE, 2018.

[12] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[13] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[14] Pavel Korshunov, Haolin Chen, Philip N Garner, and Sébastien Marcel. Vulnerability of automatic identity recognition to audio-visual deepfakes. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[15] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[16] Junjie Li, Yiwei Guo, Xie Chen, and Kai Yu. Sef-vc: Speaker embedding free zero-shot voice conversion with cross attention. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12296–12300. IEEE, 2024.

[17] Aakshi Mittal and Mohit Dua. Automatic speaker verification systems and spoof detection techniques: review and analysis. *International Journal of Speech Technology*, 25(1):105–134, 2022.

[18] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *INTERSPEECH*, pages 2616–2620, 2017.

[19] Alec Radford et al. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2023.

[20] Mirco Ravanelli, Titouan Parcollet, Adel Moumen, Sylvain de Langen, Cem Subakan, Peter Plantinga, Yingzhi Wang, Pooneh Mousavi, Luca Della Libera, Artem Ploujnikov, et al. Open-source conversational ai with speechbrain 1.0. *Journal of Machine Learning Research*, 25(333):1–11, 2024.

[21] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

[22] Xiaohai Tian, Rohan Kumar Das, and Haizhou Li. Black-box attacks on automatic speaker verification using feedback-controlled voice conversion. *arXiv preprint arXiv:1909.07655*, 2019.

[23] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[24] Shuai Wang, Zhengyang Chen, Bing Han, Hongji Wang, Chengdong Liang, Binbin Zhang, Xu Xiang, Wen Ding, Johan Rohdin, Anna Silnova, et al. Advancing speaker embedding learning: Wespeaker toolkit for research and production. *Speech Communication*, 162:103104, 2024.

[25] Yuanda Wang, Qiben Yan, Nikolay Ivanov, and Xun Chen. A practical survey on emerging threats from ai-driven voice attacks: How vulnerable are commercial voice control systems? *arXiv preprint arXiv:2312.06010*, 2023.

[26] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

[27] Bowen Zhang, Hui Cui, Van Nguyen, and Monica Whitty. Audio deepfake detection: What has been achieved and what lies ahead. *Sensors (Basel, Switzerland)*, 25(7):1989, 2025.