
# QFA2SR: Query-Free Adversarial Transfer Attacks to Speaker Recognition Systems

Guangke Chen, Yedi Zhang, and Zhe Zhao, *ShanghaiTech University;* Fu Song, *ShanghaiTech University; Automotive Software Innovation Center; Institute of Software, Chinese Academy of Sciences & University of Chinese Academy of Sciences*

https://www.usenix.org/conference/usenixsecurity23/presentation/chen-guangke

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

# QFA2SR: Query-Free Adversarial Transfer Attacks to Speaker Recognition Systems*

*Guangke Chen[1], Yedi Zhang[1], Zhe Zhao[1], Fu Song[1,2,3] (✉)*
*[1] ShanghaiTech University [2] Automotive Software Innovation Center*
*[3] Institute of Software, Chinese Academy of Sciences & University of Chinese Academy of Sciences*

## Abstract

Current adversarial attacks against speaker recognition systems (SRSs) require either white-box access or heavy black-box queries to the target SRS, thus still falling behind practical attacks against proprietary commercial APIs and voice-controlled devices. To fill this gap, we propose QFA2SR, an effective and imperceptible query-free black-box attack, by leveraging the transferability of adversarial voices. To improve transferability, we present three novel methods, tailored loss functions, SRS ensemble, and time-freq corrosion. The first one tailors loss functions to different attack scenarios. The latter two augment surrogate SRSs in two different ways. SRS ensemble combines diverse surrogate SRSs with new strategies, amenable to the unique scoring characteristics of SRSs. Time-freq corrosion augments surrogate SRSs by incorporating well-designed time-/frequency-domain modification functions, which simulate and approximate the decision boundary of the target SRS and distortions introduced during over-the-air attacks. QFA2SR boosts the targeted transferability by 20.9%-70.7% on four popular commercial APIs (Microsoft Azure, iFlytek, Jingdong, and TalentedSoft), significantly outperforming existing attacks in query-free setting, with negligible effect on the imperceptibility. QFA2SR is also highly effective when launched over the air against three wide-spread voice assistants (Google Assistant, Apple Siri, and TMall Genie) with 60%, 46%, and 70% targeted transferability, respectively.

## 1 Introduction

Speaker recognition (SR) is an automatic process recognizing the identity of a person with her voice. SR has versatile applications, such as authentication for financial transactions [10], access control for voice-controlled devices [12], and service personalization in voice assistants [11]. However, the popularity of SR has brought new security concerns. Recent studies have shown that SRSs are vulnerable to adversarial attacks as summarized in Table 1. Such attacks aims to craft an adversarial voice from a given voice uttered by a source speaker, so that it is misrecognized as another speaker by the target SRS, but does not sound like the misrecognized speaker from the perception of ordinary users. White-box attacks assume complete knowledge of the target SRS, which are powerful yet remarkably unpractical as it is *impossible* to acquire any internal information about protected proprietary systems. Black-box attacks do not rely on such information, but usually require a large number of queries to the target SRS to achieve considerable attack capabilities. Such black-box attacks suffer from two serious drawbacks: (1) they are cost-consuming because voice-controlled devices do not expose APIs thus voices have to be played over the air while commercial APIs require query-charges. Furthermore, both devices and APIs often pose limitations on the query frequency; (2) they are not very stealthy because a large bulk of queries to the target SRS leads to detectable abnormal traffics and behaviors.

Our work is motivated by the following research question: "*how to launch effective, stealthy, and practical adversarial attacks against black-box commercial APIs and voice-controlled devices without any queries to the target SRS when constructing adversarial voices (i.e., query-free)?*". A straightforward idea is to exploit the transferability of adversarial examples, i.e., crafting adversarial examples on a surrogate SRS (a local white-box SRS owned by the adversary) and then transferring them to the target SRS. However, until now, adversarial attacks in SR suffer from limited transferability since adversarial voices are easy to overfit the surrogate SRS and consequently become ineffective on the target SRS. This is because there are various aspects that the target SRS may differ in with the surrogate SRS (e.g., acoustic feature [15] and scoring method [39]) and a large number of updatable values of a seed voice due to a high audio sample rate. Indeed, we find that the transfer attack success rate (ASR) to most

*Voices and demo videos are available at our website [8] and the full version of this paper refers to [28]. We thank the reviewers and our shepherd for their constructive feedbacks. This work is supported by the National Key Research Program (2020AAA0107800), National Natural Science Foundation of China (62072309), CAS Project for Young Scientists in Basic Research (YSBR-040), and ISCAS New Cultivation Project (ISCAS-PYFX-202201).

Table 1: An overview of the state-of-the-art adversarial attacks to SRSs.

| Method | Threat Scenario | Knowledge | #Queries | Enrollment | Commercial | Attack Type | Attack Media | ASR (Digital) | ASR (Physical) |
|---|---|---|---|---|---|---|---|---|---|
| [49, 52] | White-box | Gradient | N/I | Same | ✗ | Untargeted | Digital | 41%-69% | N/A |
| [40, 44] | White-box | Gradient | N/I | N/A | ✗ | Untargeted | Digital | 40%-100% | N/A |
| AS2T [30] | White-/Black-box | Gradient/Scores | ∼ 5000 | Same | ✗ | Targeted, Untargeted | Digital, Physical | ∼ 100% | 89.4%-100% |
| FakeBob [27] | Black-box | Scores | ∼ 2500 | Same | TalentedSoft [9], Azure [18] | Targeted, Untargeted | Digital, Physical | 100% | 70% |
| SirenAttack [38] | Black-box | Scores | ∼ 7500 | N/A | ✗ | Targeted, Untargeted | Digital | ∼ 100% | N/A |
| Kenansville [22] | Black-box | Decision | ∼ 15 | Same | Azure | Untargeted | Digital | 5%-37% | N/A |
| Occam [84] | Black-box | Decision | ∼ 10000 | Same | Azure, Jingdong [17] | Targeted | Digital | 100% | N/A |
| QFA2SR (Ours) | Black-box | None | 0 | Different, Same | TalentedSoft, Azure, iFlytek [4], Jingdong, Google Assistant [11], Apple Siri [7], TMall Genie [14] | Targeted, Untargeted | Digital, Physical | 27.4%-99.5% | 46%-70% |

Note: (i)"While-box"/"Black-box": the adversary has complete/no knowledge of the target SRS. (ii) "Gradient"/"Scores"/"Decisions": the adversary requires the gradient information/the similarity scores/the identified speaker; "None": the adversary has no access to the target SRS when crafting adversarial voices. (iii) #Queries: the number of queries used for creating adversarial voices, which is not important (N/I) for white-box attacks, and "∼" denotes approximation. (iv) Same (Different): the enrollment voices used by the adversary are the same as (different from) the ones used for enrolling the target SRS, while N/A denotes that there is no enrollment voices. (v) "✗": commercial SRSs are not considered as a target SRS. (vi) "Untargeted" ("Targeted"): untargeted (resp. targeted) attack where the attack succeeds if the adversarial voice is misclassified as one of the enrolled speakers (resp. the target speaker). (vii) "Digital": adversarial voices are directly fed to SRSs in the form of audio file via exposed API; "Physical": adversarial voices are played and recorded by hardware and transmitted in the air. (viii) "ASR": attack success rate on commercial SRSs (if considered otherwise open-source SRSs).

target SRSs is less than 6% even the surrogate SRS shares the same architecture, training dataset, acoustic feature, and scoring method with the target SRSs (cf. Appendix G of [28]). Thus, the main problem is how to improve the transferability of adversarial voices without reducing imperceptibility.

In this work, we address the above problem by proposing an attack called **Q**uery-**F**ree Adversarial **A**ttack **to** **S**peaker **R**ecognition (QFA2SR). QFA2SR features three novel methods: Tailored Loss Functions, SRS Ensemble, and Time-Freq Corrosion, to improve the transferability of adversarial voices without reducing imperceptibility. The first one is proposed to find optimal loss functions, for which we design and empirically study various loss functions. Remarkably, we find that the commonly-adopted Cross Entropy Loss [41] and Margin Loss [26] for crafting adversarial images lead to less transferable adversarial voices than ours. The second one combines multiple surrogate SRSs via two novel strategies, so that adversarial voices crafted on the ensemble of surrogate SRSs can deceive as many surrogate SRSs as possible. The last one incorporates various well-designed time-/frequency-domain modification functions into surrogate SRSs to simulate and approximate unknown distribution of the target SRS and distortions introduced during over-the-air attacks [30].

We implement our approach in a tool and thoroughly evaluate the performance of QFA2SR on various open-source SRSs, commercial APIs, and voice assistants. The results confirm the effectiveness of our three novel methods and QFA2SR. For instance, QFA2SR on four commercial APIs, i.e., (Microsoft) Azure, Jingdong, iFlytek, and TalentedSoft, improves the targeted transfer ASR by 20.9%-70.7%, significantly outperforming the state-of-the-art attacks in the query-free setting, with negligible effect on the imperceptibility in terms of both perceptual objective metric and subjective human study. In particular, QFA2SR achieves 89.6%/99.6% targeted/untargeted transfer ASR to Azure, and 96% targeted transfer ASR to Jingdong (within 4 queries when launching QFA2SR). QFA2SR on three voice assistants, i.e., Google Assistant, Apple Siri, and Alibaba TMall Genie, achieves 46%-70% targeted transfer ASR when launched over the air.

In summary, the main contribution of our work includes:

- We study various loss functions and find better loss functions for transferability. We showcase that the promising Cross Entropy loss and Margin loss in the image domain are sub-optimal for the transfer attack in SR.
- We propose two novel strategies for the ensemble of the surrogate SRSs which outperforms the model ensemble for crafting adversarial images [56].
- We propose time-freq corrosion to enhance transferability, accompanied with diverse modification functions for simulating and approximating decision boundary of the target SRS and distortions introduced during over-the-air attacks.
- We propose QFA2SR, a query-free black-box adversarial attack against SRSs, by leveraging the transferability of adversarial voices, and aided by novel methods and strategies to boost the transferability, towards a truly usable transfer attack in the physical world.
- We extensively evaluate QFA2SR on 9 open-source SRSs, 4 commercial APIs, and 3 voice assistants, covering 3 attack scenarios, 2 recognition tasks, 2 attack types, 2 attack medias, and 3 settings of available voices to the adversary, with more than 144,800 adversarial voices in total. We find that QFA2SR can boost the transferability by a large margin with negligible effect on imperceptibility.

**Abbreviations and Acronyms.** For convenient reference, we summarize the abbreviations and acronyms in Table 2.

## 2 Ethical Considerations

We make the following ethical considerations:
**Strictly controlled experiments.** For commercial APIs, the target speakers in experiments are enrolled by us, so they

Table 2: Abbreviation and Acronym

| Acronym | Meaning | Acronym | Meaning |
|---|---|---|---|
| SR | Speaker recognition | SRS(s) | Speaker recognition system(s) |
| SV | Speaker verification | OSI | Open-set speaker identification |
| $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ | Targeted attack on OSI | TD-SV | Text-dependent speaker verification |
| $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ | Untargeted attack on OSI | $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ | Targeted attack on TD-SV |
| Same-enroll | Surrogate & target SRSs have the same enrollment voices | Differ-enroll | Surrogate & target SRSs have different enrollment voices |
| $\mathrm{ASR}_t$ | Targeted attack success rate | $\mathrm{ASR}_t$-s/$\mathrm{ASR}_t$-d | $\mathrm{ASR}_t$ under enroll-/differ-enroll |
| $\mathrm{ASR}_u$ | Untargeted attack success rate | $\mathrm{ASR}_u$-s/$\mathrm{ASR}_t$-d | $\mathrm{ASR}_u$ under enroll-/differ-enroll |
| RD | Reverberation-distortion | Sum-Global | Summation-based global score ranking |
| NF | Noise-flooding | Vote-Global | Voting-based global score ranking |
| SA | Speed-alteration | CD | Chunk-dropping |
| FD | Frequency-dropping | TW | Time-warping |
| TM | Time-masking | FM | Frequency-masking |

do not associate with any real-world financial or social accounts in the applications that exploit the APIs. For voice assistants, where target speakers associate with accounts, when launching our attack against them, we stopped once the attack bypassed the authentication of the target speaker. We did not take any further malicious actions, e.g., accessing the service exclusive to the target speaker. Additionally, all the used voice-controlled devices are our own facilities.

**Responsible disclosure.** We contacted the vendor TalentedSoft by email and other six vendors (Microsoft, iFlytek, Jingdong, Apple, Google, and Alibaba) with their official security vulnerability report websites, to report the vulnerabilities we found. We submitted reports with attack details, reproducibility of our attack using attached code, demonstration audios and videos, security risks brought by our attack, reason for the vulnerabilities, and suggested countermeasures. All vendors express their gratitude to our research and disclosure to keep their services, systems, and users secure. For instance, iFlytek has identified our reported vulnerability as a moderate risk level and awarded us a bounty of 1,000 RMB as recognition for our vulnerability report, and TalentedSoft replied that they will develop a plan to fix the vulnerability.

## 3 Background & Related Work

### 3.1 Speaker Recognition System (SRS)

**Speaker recognition**. Speaker recognition (SR) is the task of automatically recognizing individual speakers from their voices, typically representing acoustic characteristics as fixed-dimensional vectors via speaker embedding [75]. An architecture of generic speaker recognition systems (SRSs) is shown in Fig. 1, comprising three stages: *training*, *enrollment*, and *recognition*. All of them extract acoustic features from raw speech signals via an acoustic feature extraction module, yielding the acoustic characteristics. Common acoustic features include speech spectrogram [42], fBank [64], and MFCC [60]. The training stage trains a background model which learns a mapping from training voices to embeddings. Classic background model utilizes Gaussian Mixture Model (GMM) [68], to produce identity-vector (ivector) embeddings [34]. Recent promising background model utilizes deep

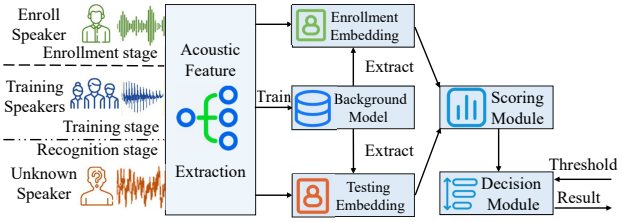



Fig. 1: Framework of SRSs.

neural networks (DNNs) to produce deep embeddings, e.g., xvector [71]. The enrollment stage maps a voice uttered by an enrolling speaker to an *enrollment embedding* using the background model. The recognition stage first retrieves the *testing embedding* of a given voice $x$ from the background model and then measures the similarity between the enrollment and testing embeddings via the scoring module. The scoring module produces a score vector $S(x)$ based on which the decision module produces the result. Probabilistic Linear Discriminant Analysis (PLDA) [62] and COSine Similarity (COSS) [33] are two widely-adopted scoring methods.

**SR task**. The SR can be classified into two major tasks: speaker identification and speaker verification (SV), where the former can be further classified into open-set identification (OSI) and close-set identification (CSI) both allowing multiple speakers to be enrolled forming a speaker group $G$. OSI determines if a given voice is uttered by either one of the enrolled speakers or imposter (i.e., an unenrolled speaker), according to the scores of all the enrolled speakers and a predefined score threshold $\theta$. Formally, assuming $G = \{1, \cdots, n\}$, given a voice $x$, the decision module outputs $D(x)$:

$$D(x) = \begin{cases} \arg\max\limits_{i \in G} [S(x)]_i, & \text{if } \max\limits_{i \in G} [S(x)]_i \geq \theta; \\ \texttt{imposter}, & \text{otherwise.} \end{cases}$$

where $[S(x)]_i$ denotes the $i$-th entry of the score vector $S(x)$, namely, the score of the voice $x$ that is likely uttered by the enrolled speaker $i$. Intuitively, the speaker $i$ that gives the maximal score is assigned as the speaker of the voice $x$, if $[S(x)]_i$ is no less than the threshold $\theta$. Otherwise, the voice $x$ is rejected, regarding it as being uttered by an imposter. In contrast, CSI always identifies the speaker that gives the maximal score as the speaker of the voice $x$, i.e., the decision module outputs $D(x) = \arg\max_{i \in G} [S(x)]_i$, and SV is a restricted case of OSI, which has exactly *one* enrolled speaker.

**Text dependency**. SR can be text-dependent (TD) and text-independent (TID). TD requires speakers to utter some predefined phrases or words during both the enrollment and recognition stages while TID does not pose any such constraints. TD can achieve good performance on short voices, but needs a large number of training voices with the same phrases or words, thus it is only used in the SV task, called TD-SV. TID needs longer voices to achieve good performance, but is more convenient and can be used in all tasks.

## 3.2 Attacks on SRS

**Adversarial attack**. An adversarial attack on SRS aims to craft an adversarial voice from a given voice uttered by a *source* speaker, so that the SRS under attack misclassifies it as one of the enrolled speakers (untargeted attack) or the target speaker (targeted attack), but ordinary users do not determine it as the recognized speakers by the SRS (imperceptibility).

The problem of finding such an adversarial voice $x'$ from a voice $x$ can be formalized as the optimization problem:

$$\text{argmin}_{x'} f(x') \text{ subject to } d(x',x) \leq \varepsilon \text{ and } x' \in [-1,1]$$

where $f$ is a loss function measuring the effectiveness of the attack, $d(x',x)$ is a distance metric quantifying the similarity between $x'$ and $x$ (imperceptibility), and $\varepsilon$ is the budget of added adversarial perturbation to ensure imperceptibility. The most widely adopted distance metric is $L_p$ norm [26], i.e., $d(x',x) = \sqrt[p]{\sum_i |x'_i - x_i|^p}$. Under the white-box setting where the adversary has full knowledge of the target SRS, the optimization problem can be solved by gradient descent using the exact gradient obtained by backpropagation [30, 40, 44, 49, 52, 53]. Under the black-box setting where the exact gradient is not available, the attack either estimates the gradient (e.g., FakeBob [27] and AS2T [30]) or utilizes gradient-free optimization approaches (e.g., SirenAttack [38], Kenansville [22], and Occam [84]). All these black-box attacks access the target SRS as an oracle, i.e., providing a series of carefully crafted inputs to the model and observing its outputs (either scores [27, 30, 38] or decisions [22, 84]).

**Hidden voice and spoofing attacks**. Hidden voice attack [20] perturbs given a voice uttered by a target speaker so that the resulting voice is treated as mere noise by humans, but still correctly recognized as the target speaker by the SRS. The spoofing attack [79] (e.g., replay attack [70] and voice cloning attack [78]) aims to obtain a voice that is correctly classified as the target speaker by the SRS and also sound like the target speaker listened to by ordinary users. Specifically, a replay attack aims to bypass the SRS using pre-recorded voices surreptitiously captured from the target speaker, and is usually used for attacking *text-independent* SV, as the collected voices usually do not contain the required text by *text-dependent* SV. In contrast, given a few voices of a speaker and the desired text, voice cloning attack creates a voice that sounds like the speaker and contains the specified speech content, thus can be exploited to attack text-dependent SV.

Hidden voice and spoofing attacks have different attack purposes and scenarios from adversarial attacks [27, 31, 78]. The perception of human listeners is inconsistent with that of the SRS under adversarial and hidden voice attacks, while it is consistent under the spoofing attack (cf. § 6.3). Furthermore, we will show that our adversarial attack QFA2SR achieves a higher attack success rate than hidden voice and spoofing attacks in the query-free setting (cf. § 6.3).

# 4 Methodology of QFA2SR

## 4.1 Threat model

We consider so far the most practical threat model concerning the knowledge of the target SRS and attack capability in the adversarial speaker recognition domain.

**Target SRS.** Regarding the target SRS, we assume that the adversary neither has white-box access to any of its internal information (e.g., architecture, parameters, training algorithm, and dataset), nor perform queries to the SRS during the generation of adversarial voices, so-called *query-free black-box* setting. First, it is almost *impossible* for the adversary to acquire internal information of a strictly protected proprietary SRS in the real life, e.g., commercial service APIs and voice controlled devices, thus preventing from white-box attacks [30, 40, 44, 49, 52]. Second, query-free is necessary and significant for achieving truly practical attacks in the real world considering that: (1) Voice assistants can *only* be interacted via the air channel, while the generation of adversarial voices via air channels would be difficult and time-consuming as the generation is an iterative process, and at each iteration, intermediate voices have to be played by loudspeakers. (2) Commercial APIs usually pose a limit on the query frequency, e.g., Jingdong SRS resticts 2 queries per second with a maximum of 500 queries per day. The limit can be solved by using time slots between queries, but making attacks time-consuming. (3) Commercial APIs may charge on the query, e.g., JingDong SRS charges 500 RMB for 1,000 queries, making attacks expensive. (4) Voice assistants and some commercial APIs only return final decision without any scores, thus stopping all the score-based black-box attacks [27, 30, 38]. Query-free attacks overcome all the above limitations.

**Voice resources.** Regarding voice resources, we assume that the adversary: (1) has a large number of voices for training the *surrogate* SRS but could be different from those used for training the *target* SRS and (2) knows all the enrolled speakers of the *target* SRS and has some voices for each of them which are used to enroll *surrogate* SRSs but also could be different from those used for enrolling the *target* SRS. The first assumption is reasonable thanks to many large-scale open-source speech corpora, e.g, Librispeech [63] and VoxCeleb1 [61]. The second assumption is also reasonable as the adversary can either use enrolled speakers' public videos on social media or record their speeches via social engineering. In § 6.5, we will relax the second assumption by considering that the adversary only has the target speaker's voice instead of all the enrolled speakers of the *target* SRS. In contrast, prior works [22, 27, 30, 38, 40, 44, 49, 52, 84] are either white-box or query-based black-box attacks, thus require neither voice datasets to train *surrogate* SRSs nor voices of enrolled speakers of the *target* SRS to enroll *surrogate* SRSs, but used the same enrollment speakers and the same voices between the surrogate and target SRSs when launching transfer attacks.

**Attack scenarios and risks.** Regarding attack scenarios, different combinations of source/target speaker, and recognition task enables the adversary to achieve different goals, e.g., unauthorized access, denial-of-service, anonymous access, evasion, and privacy protection [30]. In this work, we consider three combinations, denoted by $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$, all of which attempt to craft an adversarial voice from a given benign voice uttered by an *imposter* such that the adversarial voice is accepted by the target SRS. Both $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ and $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ focus on the OSI task, but $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ is a targeted attack that specifies an enrolled speaker as the target speaker, while $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ is an untargeted attack that succeeds when the adversarial voice is accepted as any enrolled speaker. $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ focuses on the text-dependent SV task (i.e., TD-SV), where the adversary has target speakers' voices *not* containing the desired text but knows the text in advance. It is practical as systems should inform customers of the text, e.g., "Hey Siri". Adversary can use voices of imposters with such text to craft adversarial voices. We found our attack rarely alters the text as it focuses on identity instead of speech content. Since SV is a binary classification problem with only one enrolled speaker, the target speaker of $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ is the unique enrolled speaker. We do not consider the CSI task since the OSI task is more difficult to attack than the CSI task [30], and to the best of our knowledge, no commercial SRSs use the CSI task.

Our attack exposes the following risks. (1) SR has been used for access control in smart home [12], smartphones [3], and mobile applications [74]. Our attack may enable unauthorized access, e.g., controlling over critical appliances, unlocking and logging target speakers' smartphones and applications. (2) Speaker recognition has been used for identity verification in banks' telephone-communication [2, 10] and password-free payment [13], so our attack may lead to property damage. (3) Speaker recognition has been used in key-word detection of voice assistants [11], so our attack can activate assistants and then issue malicious instructions (e.g., reading messages, deleting reminders, circumventing the confidentiality and integrity of data), or launch follow-up attacks targeting speech-to-text, e.g., Dolphin-attack [83] and CommanderSong [82]. Readers are recommended to watch recorded videos on our website [8]. However, our attack cannot achieve certain objectives, e.g., (i) denial-of-service to the target speaker, or (ii) actively hiding the identity of the target speaker to achieve anonymous access to illegal services, protect personal privacy, or evade being detected [30]. Realizing these purposes requires crafting adversarial voices from the *target* speaker's benign voices such that they are rejected or recognized as other speakers by the target SRS, which is beyond the scope of $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$.

## 4.2 Technical Challenges

Under the query-free black-box setting, all the prior attacks cannot be directly mounted, as they are either white-box or query-based black-box attacks. To tackle this issue, one has to exploit the intriguing property of adversarial examples, i.e., transferability – an adversarial example crafted with respect to one model is often found effective against other models as well. Thus, the adversary can first craft an adversarial voice on a local surrogate SRS and then transfer it to the target SRS. While advanced transfer attacks against computer vision systems have been extensively studied in the literature (e.g., [37, 54, 56, 58]), current transfer attacks on SRSs are considerably limited (e.g., the targeted/untargeted transfer attack success rate to most target SRSs is less than 6% (cf. Appendix G of [28]) due to the following technical challenges.

**Challenge CH-I**. The target SRS may be different from the surrogate SRS in various aspects, such as dataset and hyper-parameters for training background model, architecture (e.g., GMM and DNN), acoustic feature (e.g., fBank and MFCC), scoring method (e.g., PLDA and COSS), and input pre-processing, all of which can largely affect the transferability [27, 30]. More specifically, different datasets may obey different voice distributions due to different recording environments, hardware, and subjects, while different voice pre-processing can change the voice distributions in different ways. Thus, SRSs trained with different datasets and input pre-processing may learn different voice distributions. As a result, an adversarial voice crafted from a surrogate SRS would be highly sensitive to the voice distribution of the surrogate SRS, leading to low transferability. A piece of evidence is that adversarial voices are more likely to be destroyed by some input transformation [29, 31]. Similarly, SRSs with different training hyper-parameters, architectures, acoustic features and scoring methods may learn different voice distributions and decision boundaries. For instance, removing an MFCC acoustic feature extraction module from the surrogate system improves the transferability in the speech recognition domain [21]. We highlight that in the audio domain, the surrogate system may differ from the target one in more aspects than in the image domain because audio systems are usually more complicated and own several unique components and pipelines, e.g., acoustic feature extraction module and scoring method, making the transfer attacks more challenging [23, 38, 84].

**Challenge CH-II**. The iterative generation process of adversarial examples can be seen as the "training" of the input data with a fixed model, in contrast to the standard training where the model is trained with a fixed input dataset. Due to the high audio sampling rate (e.g., 16 khz), an audio has a large number of trainable variables, leading to the curse of dimensionality. For instance, a 1-second audio with 16k Hz sampling rate has totally 16,000 updatable variables, much larger than 784 ($28 \times 28$) and 3072 ($32 \times 32 \times 3$) variables of an image from MNIST and CIFAR-10, respectively. As a result, similar to significant overfitting and poor generalization of training DNNs with a larger number of parameters [51, 72], the crafted adversarial voices are easy to over-fit to the surrogate SRS,

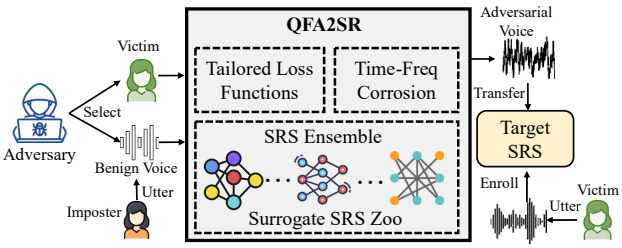



Fig. 2: The overview of our attack.

resulting in ineffective transfer attacks [37, 76]. This phenomenon has been reported in the image domain [56], where targeted transfer attacks that are effective on low-resolution images (e.g., MNIST or CIFAR-10) become significantly less effective on high-resolution images (e.g., ImageNet).

**Challenge CH-III**. To attack voice-controlled devices, adversarial voices should be played by loudspeakers, transmitted over the air, and recorded by microphones, during which loudspeakers and microphones will induce distortions to voices due to their non-uniform frequency selectivity [53]. Even worse, different loudspeakers and microphones may exhibit distinct frequency responses, thus incurring distinct distortions [53]. Moreover, both ambient noise and reverberation could distort adversarial voices and undermine the attack as well, and their impacts depend on the specific attack environments [30]. Therefore, over-the-air transfer attacks undergo additional challenges, compared to pure API transfer attacks.

## 4.3 Overview of QFA2SR

A straightforward idea to improve the transferability is to enlarge the perturbation budget or increase the confidence of adversarial voices [27, 30]. However, it not only makes the adversarial voices less imperceptible, thus much easier to increase the awareness of human, but only is almost ineffective when there is a large gap between the surrogate and the target SRSs, as SRS-specific factors (e.g., architecture) are dominant factors over attack-specific ones (e.g., perturbation budget and confidence) [30]. We also note that confidence is not a good tool to increase the transferability in commercial computer vision platforms (cf. [58, Observation 9]).

In this work, we propose an effective and imperceptible adversarial transfer attack on SRSs, named QFA2SR, addressing all the above three challenges. The overview of QFA2SR is depicted in Fig. 2, which consists of three key components: tailored loss functions, time-frequency (time-freq) corrosion and SRS ensemble, designed to increase the transferability without sacrificing imperceptibility, where the latter two are proposed to address the above three challenges.

**Tailored loss functions**. We study and evaluate various loss functions for achieving the optimal transferability for each attack scenario (i.e., $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$) (cf. § 5.1). It is essential to explore different loss functions for improving the transferability, as their effectiveness may vary in different attack scenarios (cf. Appendix E of [28]). The evaluation leads to the best tailored loss function for each attack scenario.

**SRS ensemble**. Inspired by the ensemble-based approach for improving the transferability in the image domain [56], we propose SRS ensemble (cf. § 5.2), which builds a surrogate SRS zoo with multiple surrogate SRSs. To alleviate the overfitting problem of adversarial voices to a single surrogate SRS, adversarial voices are crafted to fool as many as surrogate SRSs simultaneously, so they will be more transferable to an unknown target SRS. We emphasize that our SRS ensemble differs from the one in the image domain [56] (cf. § 5.2).

**Time-freq corrosion**. We propose time-frequency corrosion (cf. § 5.3), which randomly manipulates voice signals in the time domain and acoustic features in the frequency domain using well-designed modification functions. These functions are inserted into proper positions of the surrogate SRSs (before the acoustic feature extraction for time-domain modification functions and after the acoustic feature extraction for frequency-domain modification functions). During the generation of adversarial voices, intermediate voices are randomly modified in both the time and frequency domains. Each modification function is intentionally designed to be random and changes the distribution of the surrogate SRS in a different way, thus we can simulate and approximate as many distributions as possible. The adversarial voices crafted in this way will be more robust against different distributions (e.g., the unknown distribution of the target SRS) and the distortions introduced during over-the-air attacks, thus more transferable to an unknown target SRS even being played over the air.

# 5 Design of QFA2SR

In this section, we present the details of our attack QFA2SR.

## 5.1 Tailored Loss Functions

We study various loss functions for the attack scenarios $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ whose effectiveness will be thoroughly evaluated to choose the best one for better transferability.

**Attack scenario** $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$. Given a benign voice uttered by an imposter $s$, the adversarial attack in $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ aims to craft a voice such that the OSI SRS recognizes it being uttered by a given target speaker $t \in G$. We define the following loss functions:

$$f_{\mathrm{CE}}(x) = -\log[\mathtt{Softmax}(S(x))]_t \qquad f_1(x) = -[S(x)]_t$$
$$f_{\mathrm{M}}(x) = \max_{i\in G, i\neq t}[S(x)]_i - [S(x)]_t$$
$$f_2(x) = \max\{\theta, \max_{i\in G, i\neq t}[S(x)]_i\} - [S(x)]_t$$

where $\theta$ is a preset score threshold, $f_{\mathrm{CE}}$ and $f_{\mathrm{M}}$ are respectively the Cross Entropy Loss [41] and the Margin Loss [26] that have been widely used to craft adversarial images. $f_1$ is designed to increase the score of the target speaker $t$ *only*, in

contrast to the loss function $f_{\mathrm{M}}$ which is designed to simultaneously increase the score of the target speaker $t$ and reduce the scores of the other enrolled speakers. $f_2$ is designed such that $f_2(x) \leq 0 \Leftrightarrow D(x) = t$, when minimized, the score $[S(x)]_t$ of the target speaker $t$ is maximized to exceed $\theta$ and the scores of all the other enrolled speakers. Note that $\theta$ is the threshold of the surrogate SRS, which is known to the adversary.

**Attack scenario** $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$. Given a benign voice $x_0$ uttered by an imposter $s'$, the adversarial attack in $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ aims to craft a voice such that it is accepted as an arbitrary enrolled speaker $t \in G$ by the OSI SRS. We define the following loss functions:

$$f^s_{\mathrm{CE}}(x) = -\log[\mathtt{Softmax}(S(x))]_s$$
$$f^s_2(x) = \mathtt{max}\{\theta, \mathtt{max}_{i\in G, i\neq s}[S(x)]_i\} - [S(x)]_s$$
$$f^s_{\mathrm{M}}(x) = \mathtt{max}_{i\in G, i\neq s}[S(x)]_i - [S(x)]_s$$
$$f^s_1(x) = -[S(x)]_s \qquad f_3(x) = \theta - \mathtt{max}_{i\in G}[S(x)]_i$$

where $s = \arg\max_i [S(x_0)]_i$ and $x_0$ is the input voice. The loss functions $f^s_{\mathrm{CE}}(x)$, $f^s_1(x)$, $f^s_{\mathrm{M}}(x)$ and $f^s_2(x)$ are defined the same as $f_{\mathrm{CE}}(x)$, $f_1(x)$, $f_{\mathrm{M}}(x)$ and $f_2(x)$ except that the enrolled speaker $s$ giving the maximal score on the *input* voice $x_0$ is used as the target speaker. $f_3$ is designed such that $f_3(x) \leq 0 \Leftrightarrow D(x) =$ any enrolled speaker. Minimizing $f_3$ makes the maximal score among all the enrolled speakers exceed the threshold $\theta$, thus the adversarial voice is accepted. Unlike the others which always optimize towards the speaker $s$ that gives the maximal score on the *input* voice $x_0$ throughout the optimization, $f_3$ dynamically adjusts the target speaker based on the scores of each *intermediate* voice.

**Attack scenario** $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$. Given a benign voice uttered by an imposter, the adversarial attack in $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ aims to craft a voice that is recognized as the enrolled speaker by the SV SRS. We consider the following two loss functions for this goal:

$$f_{\mathrm{BCE}}(x) = -\log(\varphi(S(x))) \qquad f_{3B}(x) = \theta - S(x)$$

where $\varphi$ denotes the sigmoid function. Intuitively, $f_{\mathrm{BCE}}$ is the binary Cross Entropy Loss function, and $f_{3B}(x)$ is the special case of $f_3(x)$ for the binary classification task SV. We note that $f_{3B}(x)$ is also equivalent to the loss functions $f_1$, $f_M$ and $f_2$ when only one speaker is enrolled.

## 5.2 SRS Ensemble

**Ensemble of multiple SRSs**. To combine multiple SRSs, a straightforward idea is to adopt *loss-level fusion* [56], originally proposed for the ensemble of image classification models. The loss-level fusion computes the loss of the ensemble model using the weighted sum of losses of multiple SRSs. Formally, the loss function of the ensemble model is defined as $f_{\mathtt{ens}} = \sum_{k=1}^{K} w_k \times f(x; R_k)$, where $K$ is the number of surrogate models, $f(x;R_k)$ is the loss function of the $k$-th surrogate model $R_k$ with the ensemble weight $w_k$.

We tried uniform weights, i.e., $w_k = \frac{1}{K}$ for $k = 1, \cdots, K$, which works well for the ensemble of multiple image classification models [56]. However, it has limited effectiveness and sometimes even reduces the transferability compared to the attack using a single surrogate SRS (cf. Appendix G of [28]), probably because different SRSs produce scores with different ranges and scales. For example, the scoring method PLDA produces unconstrained scores, while COSS outputs scores within the range $[-1, 1]$. The loss function also varies with SRSs in the range and scale, due to its dependency on the scores. Thus, uniform weights cause the optimization to concentrate more on the SRSs with large losses than the SRSs with small losses, definitely reducing the effect of SRS ensemble. An intuitive way to address this issue is to treat the weights as hyper-parameters and manually tune them. But, searching for (approximately) optimal weights is prohibitively expensive and difficult with the increase of surrogate SRSs [46]. Moreover, the weights obtained via tuning depend on both the surrogate and subjunctive target SRSs, thus may have to be re-tuned when either the surrogate SRSs or target SRSs change.

We propose to craft adversarial voices using multiple surrogate SRSs as multi-task learning [46] and use the following method to automatically and adaptively choose appropriate weights (called dynamic weighting) for balancing different loss terms. During each iteration of crafting adversarial voices, we normalize the loss of the $k$-th SRS $f_k = f(x; R_k)$ by its mean $\mu_k$ and standard derivative $\sigma_k$, i.e., $f'_k = \frac{f_k - \mu_k}{\sqrt{\sigma_k}}$. Remark that both $\mu_k$ and $\sigma_k$ are SRS-specific and are iteratively updated via $\mu_k = \mu_k + \frac{f_k - \mu_k}{n}$ and $\sigma_k = \sigma_k + \frac{1}{n}((f_k - \mu_k)^2 - \sigma_k)$, where $n$ is the current iteration. Finally, the loss function of the ensemble model is defined as $f_{\mathtt{ens}} = \sum_{k=1}^{K} f'_k$.

**Global ranking for untargeted attack**. We now face another problem when combining the surrogate SRSs for untargeted attack (i.e., scenario $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$). Recall that the loss functions for $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ (i.e., $f^s_1$, $f^s_{\mathrm{M}}$, $f^s_2$ and $f_3$) depend on the maximal score among the enrolled speakers. Due to the diversity, the ranking of the enrolled speakers according to their scores on each surrogate SRS (called *local rank*) may differ from that of the others. If we solely use the maximal scores based on the local ranks in the loss functions, the optimization directions on the surrogate SRSs may differ, definitely reducing the effect of SRS ensemble. This is in contrast to the targeted attack (i.e., $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$) where the target speaker is the same in all the surrogate SRSs. To solve this problem, instead of using local ranks, we utilize the global rank which aggregates the local ranks of all the surrogate SRSs. We define the following two different global ranks, i.e., summation and voting.

Consider the surrogate SRS zoo $\{R_1, \cdots, R_K\}$ that has the same group $G$ of enrolled speakers. Let $\mathtt{rnk}_{k,x}$ be the *local rank* of the SRS $R_k$ on a voice $x$ that maps enrolled speakers to ranks according to their scores, i.e., speaker $i \in G$ has the $\mathtt{rnk}_{k,x}(i)$-th maximal score in the score vector $S(x)$ of the SRS $R_k$. We define the *summation-based global*

*ranking* (Sum-Global) as $\texttt{rnk}_x(i) = \sum_{k=1}^{K} \texttt{rnk}_{k,x}(i)$, where the global rank of the speaker $i$ is $\texttt{rnk}_x(i)$. We define the *voting-based global ranking* (Vote-Global) as $\texttt{rnk}_x(i) = \arg\max_{j \in G \setminus \{\texttt{rnk}_x(1), \cdots, \texttt{rnk}_x(i-1)\}} \sum_{1 \le k \le K} \mathbb{I}(\texttt{rnk}_{k,x}(j) \le i)$, where $\mathbb{I}(\texttt{rnk}_{k,x}(j) \le i)$ is 1 if $\texttt{rnk}_{k,x}(j) \le i$ otherwise 0.

The loss functions $f^s_{\text{CE}}(x)$, $f^s_1(x)$, $f^s_{\text{M}}(x)$, $f^s_2(x)$ and $f_3$ for untargeted attacks against the ensemble of the surrogate SRSs are now generalized as follows:

$$f^s_{\text{CE}}(x) = -\log[\texttt{Softmax}(S(x))]_s \qquad f^s_1(x) = -[S(x)]_s$$
$$f^s_2(x) = \texttt{max}\{\theta, \texttt{max}_{i \in G, i \neq s}[S(x)]_i\} - [S(x)]_s$$
$$f^s_{\text{M}}(x) = \texttt{max}_{i \in G, i \neq s}[S(x)]_i - [S(x)]_s \quad f_3(x) = \theta - [S(x)]_{s'}$$

where $x_0$ is the input voice, $s = \arg\min_i[\texttt{rnk}_{x_0}(i)]$ and $s' = \arg\min_i[\texttt{rnk}_x(i)]$ for Sum-Global, and $s = \texttt{rnk}_{x_0}(1)$ and $s' = \texttt{rnk}_x(1)$ for Vote-Global. Finally, the loss function of the ensemble model $f_{\texttt{ens}} = \sum_{k=1}^{K} f'_k$ is defined the same as above. Remark that the above loss functions, adapted by replacing the local rank with a global rank, only differ from the original loss functions for $\mathcal{A}^{\texttt{UT}}_{\texttt{OSI}}$ in the enrolled speaker (i.e., $s$ or $s'$).

**Attack with multiple SRSs**. An adversarial attack with SRS ensemble is shown in Alg. 1. It first initializes the means $\mu$ and derivatives $\sigma$ for each surrogate SRS (Line 1). The second for-loop (Lines 2–12) iteratively searches for an adversarial voice starting from a seed voice $x_0$. In each iteration, the third for-loop (Lines 4–10) iteratively computes the loss of each surrogate SRS, normalizes them by their individual means and standard derivatives, based on which we compute SRS ensemble loss $f_{\texttt{ens}}$ as the sum of these normalized losses (Line 10). Since the surrogate SRSs may introduce some randomness (e.g., the randomized pre-processing), we independently draw a randomness $r$ for each surrogate SRS $\beta$ times (Lines 6–7) and obtain the loss using the average of the $\beta$ losses (Line 8). We found that this leads to better transferability. Next, the intermediate voice $x_{n-1}$ is updated according to the gradient sign of $f_{\texttt{ens}}$ w.r.t. $x_{n-1}$ and the step size $\alpha$ (Line 11), which is further clipped into the $L_\infty$ $\varepsilon$-neighbourhood of the seed $x_0$ and the valid range of voices $[-1, 1]$ (Line 12).

**Algorithm 1:** SRS Ensemble

**Input:** seed voice $x_0$; $L_\infty$ perturbation budget $\varepsilon$; number of steps $N$; step size $\alpha$; surrogate SRS zoo $\{R_1, \cdots, R_K\}$; sampling size $\beta$; loss function $f(\cdot)$
**Output:** adversarial voice

1 **for** $k$ *from* 1 *to* $K$ **do** $\mu_k \leftarrow 0$; $\sigma_k \leftarrow 1$;
2 **for** $n$ *from* 1 *to* $N$ **do**
3  $f_{\texttt{ens}} \leftarrow 0$;
4  **for** $k$ *from* 1 *to* $K$ **do**
5  $f_k \leftarrow 0$;
6  **for** $r$ *from* 1 *to* $\beta$ **do** ▷ $R^r_k$ denotes the SRS $R_k$ with
7  $f_k \leftarrow f_k + f(x_{n-1}; R^r_k)$; ▷ the sampled randomness $r$
8  $f_k \leftarrow \frac{f_k}{\beta}$;
9  $\mu_k \leftarrow \mu_k + \frac{f_k - \mu_k}{n}$; $\sigma_k \leftarrow \sigma_k + \frac{1}{n}((f_k - \mu_k)^2 - \sigma_k)$;
10  $f_{\texttt{ens}} \leftarrow f_{\texttt{ens}} + \frac{f_k - \mu_k}{\sqrt{\sigma_k}}$; ▷ SRS ensemble loss
11  $x_n \leftarrow x_{n-1} - \alpha \times \texttt{sign}(\nabla_{x_{n-1}} f_{\texttt{ens}})$;
12  $x_n \leftarrow \max\{\min\{x_n, 1, x_0 + \varepsilon\}, -1, x_0 - \varepsilon\}$;
13 **return** $x_N$;

## 5.3 Time-Freq Corrosion

Due to the time-varying non-stationary property, voices are not resilient enough to noises and other variations, and waveform signals themselves cannot effectively represent speaker characteristics [66]. Thus, to achieve better performance [81], a raw input voice is often transformed into a two-dimensional time-frequency representation via an acoustic feature extraction (cf. Fig. 1). This motivates us to design functions for manipulating voices in both the time and frequency domains.

### 5.3.1 Time Domain Modification Functions

We consider the following five modification functions for manipulating voice signals in the time domain.

**Reverberation-distortion (RD)** [48]. Reverberation occurs when a signal propagates through multiple paths (direct and reflected paths) in a room, where the direct sound and reflections overlap with each other. Room Impulse Response (RIR), denoted by $r$, can characterize the acoustic properties of a room regarding sound transmission and reflection. Given an input voice $x$, the reverberant voice is created by convolving $r$ with $x$. Given a list of RIRs ($\mathcal{R}$), each of which models a room configuration, RD randomly applies one RIR each step.

**Noise-flooding (NF)** [43]. NF modifies a voice by superimposing it with a random white Gaussian noise. The magnitude of the noise is controlled via the signal-to-noise ratio (SNR) $10\log_{10}\frac{P_v}{P_n}$, where $P_v$ and $P_n$ are the power of the input voice and the noise, respectively. The SNR is randomly chosen from $[\text{SNR}_l, \text{SNR}_h]$ during each step, where $\text{SNR}_l$ and $\text{SNR}_h$ denote the lower bound and upper bound of the SNR.

**Speed-alteration (SA)** [47]. Given a voice $x(t)$ and the speed ratio $\alpha$ between the new and original speeds, SA produces the time-scaled voice $x(\alpha t)$, which sounds faster (resp. slower) when $\alpha > 1$ (resp. $\alpha < 1$). SA changes the duration of the utterance, thus affects the number of frames of the voice. The randomized version of SA selects a one-speed ratio from a candidate list of speed ratios ($\mathcal{A}$) each step.

**Chunk-dropping (CD)** [67]. Given a voice with $T$ sample points, CD sets the magnitudes of the sample points within $[t_0, t_0 + t)$ to zero, where $t$ and $t_0$ are randomly chosen from $[T_l, T_u]$ and $[0, T - t]$, respectively. $T_l$ and $T_u$ are the lower and upper bounds of the chunk lengths to be dropped. In addition, given the lower $C_l$ and upper $C_u$ bounds of the number of the chunks to be dropped, the process is independently repeated $c$ times where $c$ is randomly selected from $[C_l, C_u]$.

**Frequency-dropping (FD)** [67]. A voice signal can be decomposed into multiple pure tones with different frequencies. Given the lower $F_l$ and upper $F_u$ bounds of the frequencies to be dropped, FD applies a notch filter to remove the pure tone with frequency $f$ which is randomly chosen from $[F_l, F_u]$.

This process can also be independently repeated $c$ times for a randomly chosen number $c$ between the lower $C_l$ and upper $C_u$ bounds of the number of the frequencies to be dropped.

#### 5.3.2 Frequency Domain Modification Functions

We consider three modification functions for manipulating voice signals in the frequency domain which was used for voice data augmentation in [65]. We denote by $M \in \mathbb{R}^{T \times F}$ the acoustic feature matrix, where $T$ and $F$ are the number of time frames and frequency channels, respectively.

**Time-warping (TW)**. TW introduces deformations in the time frame dimension of $M$. First, an entry $p$ of $M$ is selected such that its horizontal coordinate is the center and the vertical coordinate $P$ is randomly chosen from $[W, T-W]$ where $W$ is the time warping parameter. Then, the sub-region above the horizontal line passing $p$ with the size $P \times F$ is scaled to the size $w \times F$, while the sub-region below the horizontal line passing $p$ with the size $(T-P) \times F$ is scaled to the size $(T-w) \times F$, where $w$ is randomly chosen from $[P-W, P+W]$.

**Time-masking (TM)**. TM introduces deformations in the time frame dimension of $M$ by applying zero masking to $t$ consecutive time frames $[t_0, t_0+t)$ where $t$ is randomly chosen from $[0, t']$ for a given TM parameter $t'$ and $t_0$ is randomly chosen from $[0, T-t]$.

**Frequency-masking (FM)**. FM introduces deformations in the frequency channel dimension of $M$ by replacing the coefficients of $f$ consecutive frequency channels $[f_0, f_0+f)$ with 0 where $f$ is randomly chosen from $[0, f']$ for a given FM parameter $f'$, and $f_0$ is randomly chosen from $[0, F-f]$.

#### 5.3.3 Combination of Modification Functions

We explore the combinations of the above modification functions to improve transferability. Denoting the individual modification functions by $m_1, \cdots, m_K$, we consider two combination strategies: *serial* and *parallel*. Consider the indices $i_1, \cdots, i_K \in \{1, \cdots, K\}$ such that if $m_{i_j}$ is a time domain modification function, then $m_{i_1}, \cdots, m_{i_{j-1}}$ are all time domain modification functions. The *serial composite modification function* $M_s(\cdot) = m_{i_K}(m_{i_{K-1}}(\cdots m_{i_1}(\cdot)))$ sequentially applies the functions $m_{i_1}, \cdots, m_{i_K}$ either at signal-level or feature-level depending on the function. $M_s$ is achieved by building the *simulated* SRS $R_{M_s}$ from a given surrogate SRS $R$ where all the modification functions $m_i$ of $M_s$ are inserted at proper positions, i.e., before (resp. after) the acoustic extraction module if $m_i$ is a time (resp. frequency) domain function. The *parallel composite modification function* $M_p(\cdot) = M_1||\cdots||M_K$ modifies an input voice by applying the functions $M_1, \cdots, M_K$ in parallel, leading to $K$ different modified voices and $K$ loss values $v_1, \cdots, v_K$. Note that $M_i$ in $M_p$ could be a serial composite modification function. $M_p$ is achieved by building $K$ *simulated* SRSs $\{R_{M_1}, \cdots, R_{M_K}\}$ from a given surrogate SRS $R$ where $R_{M_i}$ is the simulated SRS of the surrogate SRS $R$ for the modification functions $M_i$. The $K$ *simulated* SRSs $\{R_{M_1}, \cdots, R_{M_K}\}$ can be combined using our ensemble method (cf. § 5.2). In this work, we consider three serial composite modification functions: RD+NF, SA+FD+CD, and TW+TM+FM. For *parallel* combination, we consider the combination of these three serial composite functions. We leave other composite functions as future work.

**Algorithm 2:** QFA2SR

**Input:** seed voice $x_0$; modification functions $\mathcal{M} = \{\cdots, M_j, \cdots\}$; sampling size $\beta$; surrogate SRS zoo $\mathcal{R} = \{\cdots, R_i, \cdots\}$; number of steps $N$; the step size $\alpha$; $L_\infty$ perturbation budget $\varepsilon$; the optimal loss function for the attack scenario $f_{\texttt{opt}}(\cdot)$

**Output:** adversarial voice $x^{adv}$

1 $\mathcal{Z} \leftarrow \{R_M \mid R \in \mathcal{R}, M \in \mathcal{M}\}$ if $\mathcal{M} \neq \emptyset$ else $\mathcal{R}$;

2 $x^{adv} \leftarrow$ invoke Alg. 1 with the surrogate SRS zoo $\mathcal{Z}$ and parameters ($x_0$, $\beta$, $N$, $\alpha$, $\varepsilon$, and $f_{\texttt{opt}}(\cdot)$);

3 **return** $x^{adv}$;

### 5.4 QFA2SR: Our Final Attack

QFA2SR for one seed voice is shown in Alg. 2, where $\mathcal{M}$ is a set of (serial composite) modification functions $\{M_1, \cdots, M_K\}$ and used as a parallel composite modification function $M_p(\cdot) = M_1||\cdots||M_K$ if $k \geq 2$. Alg. 2 first builds a *simulated* surrogate SRS zoo $\mathcal{Z}$ by combining each surrogate SRS $R \in \mathcal{R}$ with each modification function $M \in \mathcal{M}$, to get rid of modification functions. Then, it invokes Alg. 1 with $\mathcal{Z}$ as the surrogate SRS zoo and other necessary input parameters to craft an adversarial voice. We note that the surrogate SRSs $\{R_M \mid R \in \mathcal{R}, M \in \mathcal{M}\}$ are combined using our ensemble method (cf. § 5.2) when crafting adversarial voices.

In practice, the adversary can generate many adversarial voices but can only query the target SRS few times during transfer attack. Thus, we propose a method to select the adversarial voices which are the most likely transferable to the target SRS, thus largely improves the success rate of QFA2SR with few allowed queries. Details refer to Appendix A of [28].

## 6 Evaluation of Attack

### 6.1 Experimental Setting and Design

**Enrollment settings.** The enrollment voices used in the target SRS may be the same as (resp. different from) that used by the adversary to enroll surrogate SRSs, called *same-enroll* (resp. *differ-enroll*). Note that all the prior works consider same-enroll *only* which is less realistic than differ-enroll. Here we consider both same-enroll and differ-enroll except for text-dependent verification in the scenario $\mathcal{A}^{\texttt{T}}_{\texttt{TD-SV}}$ for which we consider differ-enroll *only* where the target speaker's voices available to the adversary do *not* contain the desired text.

**Datasets.** Our evaluation is mainly based on eight datasets, the details of which are shown in Table 3.

Table 3: Details of Datasets.

| Name | #Voices | Voice Source & Attack Scenario | Description |
|---|---|---|---|
| **Spk$_{10}$-enroll-P$_1$** | 10 × 5 | Provided by SEC4SR [29] that are derived from LibriSpeech [63], for $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ and $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ | Used to enroll OSI surrogate and target SRSs for same-enroll |
| **Spk$_{10}$-enroll-P$_2$** | 10 × 5 | | Same speakers but different voices as Spk$_{10}$-enroll-P$_1$, used to enroll OSI target SRSs for differ-enroll |
| **Spk$_{10}$-test** | 10 × 100 | | Same speakers but different voices as Spk$_{10}$-enroll-P$_1$&P$_2$, used to test the performance of SRSs |
| **Spk$_{10}$-imposter** | 10 × 100 | | Speakers different from Spk$_{10}$-test, used to test the performance of SRSs and as seeds for crafting adversarial voices |
| **Spk$_5$-TD-P$_1$** | 5 × 10 × 4 | Recruiting volunteers to record voices for $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ | Ten different sentences in Appendix A, used to enroll TD-SV target SRSs and as seeds for crafting adversarial voices |
| **Spk$_5$-TD-P$_2$** | 5 × 10 × 1 | | Ten sentences from [78], same speakers but different texts as Spk$_5$-TD-P$_1$, used to enroll surrogate SRSs |
| **Spk$_{1000}$-enroll-P$_1$** | 1000 × 5 | Derived from LibriSpeech, for $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ and $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ | Expand Spk$_{10}$-enroll-P$_1$ with 990 speakers, used to enroll OSI surrogate and target SRSs for same-enroll |
| **Spk$_{1000}$-enroll-P$_2$** | 1000 × 5 | | Same speakers but different voices as Spk$_{1000}$-enroll-P$_1$, used to enroll OSI target SRSs for differ-enroll |

Note: In the #Voices column, x × y denotes x speakers and y voices per speaker, and x × y × z denotes x speakers, y different texts, and z voices per text and per speaker.

Table 4: Results on commercial APIs in $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$.

| | Microsoft Azure | | | | TalentedSoft | | | | IFlytek | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR$_t$-s | ASR$_t$-d | SNR | PESQ | ASR$_t$-s | ASR$_t$-d | SNR | PESQ | ASR$_t$-s | ASR$_t$-d | SNR | PESQ |
| **SirenAttack** | 1 | 2.1 | 8.02 | 1.12 | 1.4 | 1.3 | 10.07 | 1.18 | 0 | 0 | 8 | 1.12 |
| **Kenansville** | 0 | 0 | **16.23** | **1.75** | 0 | 0 | **16.23** | **1.75** | 0 | 0 | **16.23** | **1.75** |
| **FakeBob** | 4.2 | 3.1 | 12.23 | 1.22 | 5.0 | 2.4 | 12.50 | 1.23 | 0 | 0 | 12.16 | 1.24 |
| **FakeBob** + ① | 6.2 | 4.1 | 12.23 | 1.23 | 5.6 | 2.7 | 12.51 | 1.24 | 1.9 | 1.9 | 12.16 | 1.23 |
| **FakeBob** + ①② | 17.5 | 17.2 | 12.22 | 1.24 | 9.3 | 4.7 | 12.22 | 1.24 | 9.1 | 8.8 | 12.22 | 1.24 |
| **FakeBob** + ①②③ | 3.8 | 2.7 | 12.71 | 1.28 | 4.0 | 2.5 | 12.71 | 1.28 | 0.6 | 0.6 | 12.71 | 1.28 |
| **BIM** | 18.9 | 12.7 | 11.49 | 1.18 | 8.9 | 6.5 | 11.28 | 1.19 | 16 | 15.5 | 11.50 | 1.18 |
| **BIM** + ① | 27.2 | 21.8 | 11.50 | 1.18 | 9.3 | 6.6 | 11.28 | 1.19 | 24 | 17.5 | 11.52 | 1.19 |
| **BIM** + ①② | 42.8 | 34.2 | 11.29 | 1.18 | 16.9 | 12.5 | 11.29 | 1.18 | 25.9 | 21.6 | 11.29 | 1.18 |
| **BIM** + ①②③ (**QFA2SR**) | **89.6** ↑ 70.7 | **82.8** ↑ 70.1 | 10.85 | 1.18 | **40.1** ↑ 31.2 | **27.4** ↑ 20.9 | 10.85 | 1.18 | **46.1** ↑ 30.1 | **39.5** ↑ 24 | 10.85 | 1.18 |

Table 5: Results of QFA2SR on commercial APIs in $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$.

| | Microsoft Azure | | | | TalentedSoft | | | | IFlytek | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ASR$_u$-s | ASR$_u$-d | SNR | PESQ | ASR$_u$-s | ASR$_u$-d | SNR | PESQ | ASR$_u$-s | ASR$_u$-d | SNR | PESQ |
| **SirenAttack** | 16.67 | 8.25 | 8.16 | 1.12 | 23.9 | 18.7 | 10.07 | 1.18 | 0 | 0 | 8.07 | 1.12 |
| **Kenansville** | 0 | 0 | **16.97** | **1.8** | 7 | 4 | **17.58** | **1.84** | 0 | 0 | **16.66** | **1.77** |
| **Hidden** | 21.4 | 23 | -2.84 | 1.14 | 22.9 | 21.9 | -2.9 | 1.18 | 0 | 0 | -2.95 | 1.15 |
| **FakeBob** | 33.33 | 15.46 | 12.24 | 1.23 | 26.8 | 24 | 12.41 | 1.24 | 11.5 | 5.8 | 12.12 | 1.23 |
| **FakeBob** + ① | 33.33 | 15.46 | 12.24 | 1.23 | 26.8 | 24 | 12.41 | 1.24 | 11.5 | 5.8 | 12.12 | 1.23 |
| **FakeBob** + ①② | 47.92 | 37.11 | 12.22 | 1.22 | 31 | 26.7 | 12.22 | 1.22 | 19.2 | 13.5 | 12.22 | 1.22 |
| **FakeBob** + ①②③ | 15.42 | 6.41 | 12.55 | 1.27 | 11.7 | 7.2 | 12.55 | 1.27 | 5.0 | 2.7 | 12.55 | 1.27 |
| **BIM** | 61.22 | 47.21 | 11.55 | 1.18 | 17.8 | 16.2 | 11.37 | 1.18 | 60 | 58 | 11.53 | 1.17 |
| **BIM** + ① | 68.4 | 50.8 | 11.54 | 1.18 | 22.7 | 19.9 | 11.37 | 1.19 | 64 | 61.9 | 11.54 | 1.18 |
| **BIM** + ①② | 80.62 | 66.53 | 11.37 | 1.19 | 30.1 | 23.5 | 11.37 | 1.19 | 69 | 62.9 | 11.37 | 1.19 |
| **BIM** + ①②③ (**QFA2SR**) | **99.49** ↑ 38.27 | **92.39** ↑ 45.18 | 11.01 | 1.19 | **55** ↑ 28.2 | **39.6** ↑ 15.6 | 11.01 | 1.19 | **70** ↑ 10 | **68** ↑ 10 | 11.01 | 1.19 |

Note: ①, ②, ③ denote Tailored Loss Functions, SRS Ensemble, and Time-Freq Corrosion, respectively. ↑ is the improvement of QFA2SR over the most effective baseline.

**SRSs**. We use 9 open-source SRSs: Ivector-PLDA (IV) [16], ECAPA-TDNN (ECAPA) [35], Xvector-PLDA (XV-P) [19], Xvector-COSS (XV-C) [67], Resnet18 trained for OSI (Res18-I) and SV (Res18-V) [25], Resnet34 trained for OSI (Res34-I) and SV (Res34-V) [32], and AutoSpeech (Auto) [36]. We also include four commercial APIs: (Microsoft) Azure [18], TalentedSoft [9], iFlytek [4], and Jingdong [17], and three voice assistants: Google Assistants [11], Apple Siri [7], and TMall Genie [14]. Details of these SRSs, and their threshold θ and performance are given in Appendix B.

**Metrics**. We use transfer attack success rate (ASR) to measure attack effectiveness, and denote by ASR$_u$ (resp. ASR$_t$) the untargeted (resp. targeted) ASR, which refers to the proportion of adversarial voices that are misrecognized as any enrolled speakers (resp. target speaker) by the target SRS. Let ASR$_u$-s (resp. ASR$_u$-d) denote ASR$_u$ under the same-enroll (resp. differ-enroll) setting. ASR$_t$-s and ASR$_t$-d are defined similarly. The ASR improvement $x$% by our attack compared over a baseline is calculated as $x = b - a$, where b% (resp. a%) is the ASR of our attack (resp. baseline). To quantify imperceptibility, we use Signal-to-Noise Ratio (SNR) [31] and Perceptual Evaluation of Speech Quality (PESQ) [69]. SNR is defined as $10\log_{10}\frac{P_x}{P_\delta}$, where $P_x$ and $P_\delta$ are the power of the original voice and the perturbation. PESQ is an objective perceptual measure that simulates the human auditory system [80]. Higher SNR/PESQ indicates better imperceptibility.

**Experimental design.** We first summarize the results of tuning parameters of QFA2SR (§ 6.2). We then evaluate QFA2SR on commercial APIs where adversarial voices are directly fed to the exposed APIs as audio files (§ 6.3) and voice assistants where adversarial voices are played over the air to attack voice assistants (§ 6.4). We finally study the effect of adversarial knowledge on the enrolled speakers of target SRSs (§ 6.5), and the attack scalability of QFA2SR (§ 6.6).

## 6.2 Tuning Parameters of QFA2SR

We tune the parameters of QFA2SR on open-source SRSs, simulating a real-world adversary who tunes parameters within the surrogate SRS zoo, and attacks commercial SRSs in § 6.3 and § 6.4 using the resulting parameters. Due to space limit, here we only summarize the results of parameter tuning. Details are given in Appendixes E, F, and G of [28].

**Tailored loss functions.** The losses $f_1$ and $f_2$ are comparable and outperform the others for $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $f_3$ in general performs better than the others for $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, and $f_{\mathrm{BCE}}$ and $f_{3B}$ have the same performance for $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$. The comparison results among loss functions keep consistent across different surrogate and target SRSs, and between a single surrogate SRS and the ensemble of multiple SRSs with adapted losses (cf. § 5.2). Thus, we will use $f_1$ for $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $f_3$ for $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, and $f_{3B}$ for $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ for all systems and settings, rather than performing the comparison and selection repeatedly. Note that $f_1$ is preferable than $f_2$ since $f_1$ depends *only* on the score of the targeted speaker.

**SRS ensemble.** Our dynamic weighting outperforms the uniform weighting used in the image domain [56]. Summation-based global ranking (Sum-Global) and voting-based global ranking (Vote-Global) are comparable, and both of them perform better than the local ranking. Hence, we will use our dynamic weighting and Sum-Global for SRS ensemble.

**Time-freq corrosion.** All individual modification functions can improve transferability. The serial composite functions (RD+NF, SA+FD+CD, and TW+TM+FM) achieve higher

transferability than individual functions. Their parallel combination yields the best transferability, hence will be utilized as the default modification function for time-freq corrosion.

### 6.3 QFA2SR against Commercial APIs

**Setting.** For OSI task (i.e., $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ and $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$), we attack 3 commercial SRSs: Azure, TalentedSoft, and iFlytek. For TD-SV task (i.e., $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$), we attack 2 commercial SRSs: Azure and Jingdong. Note that Jingdong does not support OSI while TalentedSoft and iFlytek do not support TD-SV. For surrogate SRSs, we only consider IV, ECAPA, XV-P, and XV-C since they yield the best transferability in general according to the results in Appendix G of [28]. We compare QFA2SR with baselines: Basic-Iterative-Method (BIM) [50], FakeBob, SirenAttack, and Kenansville. Occam is not available and AS2T is based on BIM and FakeBob, thus are not compared. In $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, we also compare with the hidden voice attack [20], where 100 voices are randomly selected from $\mathrm{Spk}_{10}$-test as the seed voices. Note that the hidden voice attack can neither launch targeted attack (cf. § 3) nor change the speech content, so is not applicable to $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$ and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$. In $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$, we also compare with the voice cloning attack using the few-shot voice cloning toolkit Real-Time-Voice-Cloning [5, 45]. It produces a voice with the desired speech content given a set of the target speaker's voice samples and a speech content. We use the voices in $\mathrm{Spk}_5$-TD-P2 as voice samples and the ten sentences from Azure (cf. Appendix A) as the desired contents.

We set $L_\infty$ perturbation budget $\varepsilon = 0.02$, step size $\alpha = \frac{\varepsilon}{5} = 0.004$, number of steps $N = 300$, and sampling size $\beta = 5$ for QFA2SR, and detailed setups of the baselines refer to Appendix C. As we focus on query-free attacks (i.e., no query to target SRSs during adversarial voice generation), all the baselines are used to launch transfer attacks. We report the best transferability among different surrogate SRSs for them.

**Results of scenario $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$.** The results are shown in Table 4. QFA2SR achieves 20.9%-70.7% higher $\mathrm{ASR}_t$ than BIM which is generally the most effective one among the baselines. QFA2SR can achieve more than 82% $\mathrm{ASR}_t$ on Azure.

**Results of scenario $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$.** The results for $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$ are shown in Table 5. Compared with the most effective baseline, QFA2SR improves the $\mathrm{ASR}_u$ by 10%-45.1%, achieving more than 92% $\mathrm{ASR}_u$ on Azure. It also achieves much higher $\mathrm{ASR}_u$ than the hidden voice attack, probably because the least incomprehensible hidden voices crafted with respect to the source SRS are difficult for the target to correctly recognize.

**Results of scenario $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$.** The results for $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ are shown in Table 6. Compared to the best baseline, QFA2SR improves the $\mathrm{ASR}_t$ by 48.85% and 54% on Azure and Jingdong, respectively. QFA2SR also achieves 26%-51.86% higher $\mathrm{ASR}_t$ than the voice cloning attack. It is because the voice cloning attack generates artificially fake voices, which usually contain artifacts and suffer from low quality, e.g., the characteristic prosody is lost [45]. As a result, the cloned voice does not have sufficient acoustic similarity with the genuine enrollment voice of the target speaker and thus fails to bypass the SRS. In contrast, QFA2SR starts from genuine voices of an imposter and only add to them imperceptible perturbations that sound like background noise to improve the score of the target speaker, thus the adversarial voices crafted by QFA2SR have sufficient acoustic similarity to bypass the target SRS.

**Imperceptibility.** In $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$, QFA2SR has higher SNR and PESQ than SirenAttack and hidden voice attack. Kenansville and FakeBob have better imperceptibility than QFA2SR, but their transferability is too low to effectively mislead the target SRS and thus far from being practical. The SNR of QFA2SR is slightly lower than BIM, but PESQ is the same or even larger in most cases. Note that PESQ is an objective perceptual measure that simulates human auditory system, but SNR is not, we believe PESQ can better characterize the imperceptibility.

As SNR and PESQ may not fully measure human imperceptibility, we conduct a human study on MTurk [1] with approval from the Institutional Review Board (IRB) of our institute. The participants are presented with a pair of voices and asked to tell after listening whether they are uttered by the same speaker, provided with three options: *same*, *different*, and *not sure*. We compare the imperceptibility of QFA2SR with BIM and voice cloning attack, while other attacks are excluded since their transfer success rates are too low to be practical. Furthermore, we conduct the human study in $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$ because voice cloning attack is text-dependent. Specifically, we build 24 pairs: 4 normal pairs (two clean voices from distinct speakers), 10 adversarial pairs (one adversarial voice from an imposter and one clean voice from the target speaker; 5 pairs are from QFA2SR and 5 pairs are from BIM), and 10 cloning pairs (one voice generated by voice cloning and one clean voice from the target speaker). To guarantee the quality of the answers and validity of the results, we filter out the answers that are randomly chosen by participants. In particular, we insert 6 special voice pairs (two clean voices from different speakers with opposite gender) as the concentration test. All the submitted answers from a participant will be excluded as long as she/he does not choose the *different* option for any one of the special pairs.

After excluding 14 participants who failed to pass our concentration tests, we finally received the answers from 126 participants. The results of human study is shown in Fig. 3. 76.7% of participants think that the adversarial voices crafted by QFA2SR do not sound like the target speaker, merely 6% lower than that of the normal pairs and even 4.6% higher than that of BIM. This demonstrates that QFA2SR enhances the transferability without harming the human imperceptibility. Interestingly, 39.3% of participants choose the *same* option for the cloning pairs, very close to the $\mathrm{ASR}_t$ against Jingdong SRS in Table 6 and much higher than that for adversarial pairs. In contrast, only 20% of participants choose the *same* option for QFA2SR, although QFA2SR achieves more than 60%

Table 6: Results on commercial APIs in $\mathcal{A}^{\mathrm{T}}_{\text{TD-SV}}$.

| | Microsoft Azure | | | Jingdong | | |
|---|---|---|---|---|---|---|
| | differ-enroll $\text{ASR}_t$ | SNR (dB) | PESQ | differ-enroll $\text{ASR}_t$ | SNR (dB) | PESQ |
| **SirenAttack** | 0.49 | 8.97 | 1.15 | 0 | 10.15 | 1.18 |
| **Kenansville** | 0 | **20.64** | **2.11** | 0 | **20.64** | **2.11** |
| **Voice Cloning** | 10 | - | - | 40 | - | - |
| **FakeBob** | 0.52 | 13.16 | 1.28 | 8 | 13.32 | 1.28 |
| **FakeBob + ①** | 0.52 | 13.16 | 1.28 | 8 | 13.32 | 1.28 |
| **FakeBob + ① ②** | 16.67 | 13.14 | 1.28 | 11 | 13.14 | 1.28 |
| **FakeBob + ① ② ③** | 0.1 | 13.45 | 1.30 | 3 | 13.45 | 1.30 |
| **BIM** | 13.01 | 12.40 | 1.24 | 12 | 12.21 | 1.23 |
| **BIM + ①** | 13.01 | 12.40 | 1.24 | 12 | 12.21 | 1.23 |
| **BIM + ① ②** | 27.78 | 12.21 | 1.23 | 23.5 | 12.21 | 1.23 |
| **BIM + ① ② ③ (QFA2SR)** | **61.86** ↑ 48.85 | 11.84 | 1.24 | **66** ↑ 26 | 11.84 | 1.24 |

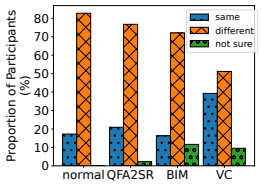



Fig. 3: Results of human study. VC=voice cloning.

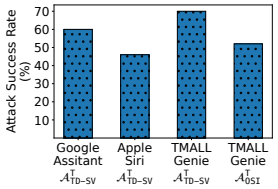
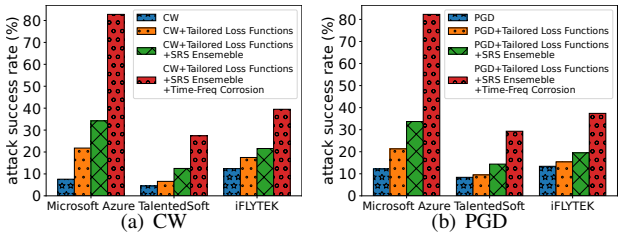
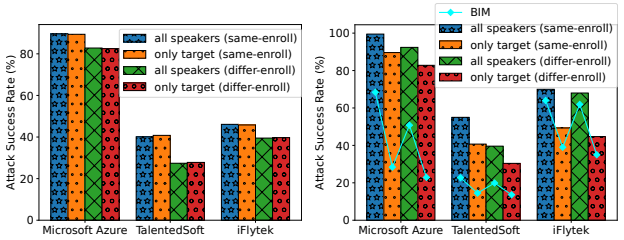



Fig. 4: Results on voice assistants.




(a) CW (b) PGD

Fig. 5: Ablation study using the CW and PGD attacks in $\mathcal{A}^{\mathrm{T}}_{\text{OSI}}$ under different-enrollment setting.




(a) $\text{ASR}_t$ of QFA2SR in $\mathcal{A}^{\mathrm{T}}_{\text{OSI}}$ (b) $\text{ASR}_u$ of QFA2SR/BIM in $\mathcal{A}^{\mathrm{UT}}_{\text{OSI}}$

Fig. 6: Effect of knowledge on enrolled speakers of target SRSs.

$\text{ASR}_t$ against target SRSs in Table 6. This confirms the difference between adversarial and voice cloning attacks regrading the human-machine perception consistency.

Summary: QFA2SR *significantly improves transferability under all the three attack scenarios, with negligible effect on imperceptibility in terms of both perceptual objective metric and subjective human study, compared to the best baseline.*

**Ablation Study.** To understand the contributions of tailored loss functions, SRS ensemble, and time-freq corrosion, we perform ablation study by gradually incorporating them into BIM and FakeBob, which are in general the most effective baselines. Note that QFA2SR bases on BIM. From Tables 4–6, we observe that: *all the three methods improve transferability, and in general, time-freq corrosion contributes the most, while tailored loss functions contribute the least, regardless of attacks, scenarios, and enrollment settings, with the following two exceptions.*

First, the tailored loss function $f_{3B}$ (resp. $f_3$) does not enhance the transferability of BIM and FakeBob on $\mathcal{A}^{\mathrm{T}}_{\text{TD-SV}}$ (resp. FakeBob on $\mathcal{A}^{\mathrm{UT}}_{\text{OSI}}$). This is because FakeBob uses the same loss $f_{3B}$ and $f_3$ for $\mathcal{A}^{\mathrm{T}}_{\text{TD-SV}}$ and $\mathcal{A}^{\mathrm{UT}}_{\text{OSI}}$, respectively, and BIM uses the loss function $f_{\text{BCE}}$ that has the same performance as $f_{3B}$ for $\mathcal{A}^{\mathrm{T}}_{\text{TD-SV}}$ (see § 6.2). Second, time-freq corrosion does not improve or even worsens the transferability of FakeBob. It is because the black-box attack FakeBob estimates gradients instead of using exact gradients as BIM, and the randomness introduced by time-freq corrosion makes the estimated gradients uninformative and hence the optimization direction unreliable, consistent with the finding in [31]. We try to address this by enlarging the parameter of FakeBob that is positively correlated with the precision of estimated gradients from 50 to 1,000, but the improvement is rather limited, and the computation cost is totally unacceptable (1000 × 4 surrogate SRSs × 50 steps = $2e^5$ queries for a *single* adversarial voice). These suggest that time-freq corrosion is more compatible with white-box attacks that utilize exact gradients. To confirm this, we perform additional ablation study using two white-box attacks: Carlini and Wagner's attack (CW) [26] and Projected Gradient Descent (PGD) attack [57] (cf. Appendix C). The results are shown in Fig. 5. *All the three methods enhance the transferability, demonstrating their generalizability for incorporating into white-box attacks.* Note that we also perform the ablation study on open-source SRSs in Appendix I of [28], where we can draw the same conclusion on the contributions of individual methods of QFA2SR.

## 6.4 QFA2SR against Voice Assistants

**Settings.** For $\mathcal{A}^{\mathrm{T}}_{\text{TD-SV}}$, we consider three voice assistants supporting speaker recognition, i.e., Google Assistant in Google Pixel 5 [11], Siri in Apple iPad Pro 10.5 [7], and TMall Genie in smart speaker X5 [14]. For $\mathcal{A}^{\mathrm{T}}_{\text{OSI}}$ (when the adversary only has voices of target speakers), we only consider TMall Genie as the others do not support speaker identification. $\mathcal{A}^{\mathrm{UT}}_{\text{OSI}}$ is omitted since it is easier than $\mathcal{A}^{\mathrm{T}}_{\text{OSI}}$ and $\mathcal{A}^{\mathrm{T}}_{\text{TD-SV}}$. To be diverse, we adopt JBL clip3 portable loudspeaker [6], TMall smart speaker X5, and iPad Pro 10.5 as the loudspeaker to play adversarial voices when attacking Google Assistant, Apple Siri, and TMall Genie, respectively. We conduct experiments in a meeting room with air-conditioner noise and the distance

between voice assistants and loudspeakers is set to 1.5 meters.

The enrollment and test voices of these voice assistants are text-dependent, so we recruited six volunteers (four male and two female) to utter the desired phrases. To cover the differ-enroll setting, we also ask them to utter the ten English sentences used in $\text{Spk}_5$-TD-$\text{P}_2$ for Google Assistant and Apple Siri, and five Chinese quotes for TMall Genie, which are used to enroll surrogate SRSs. Details refer to Appendix D.

The results are depicted in Fig. 4. QFA2SR achieves 60%, 46%, and 70% $\text{ASR}_t$ in $\mathcal{A}^{\text{T}}_{\text{TD-SV}}$ on Google Assistant, Apple Siri, and TMall Genie, respectively, indicating that different voice assistants have different frangibility to adversarial attacks. For TMall Genie, the $\text{ASR}_t$ for $\mathcal{A}^{\text{T}}_{\text{OSI}}$ is lower than that for $\mathcal{A}^{\text{T}}_{\text{TD-SV}}$, because without the voices of other enrolled speakers in $\mathcal{A}^{\text{T}}_{\text{OSI}}$, the crafted adversarial voices may be recognized as another enrolled speaker whose voiceprint is similar to the target speaker. *These results demonstrate the effectiveness of* QFA2SR *in crafting transferable adversarial voices which can be played over the air against popular voice assistants.*

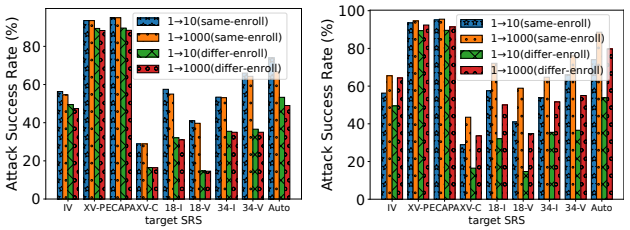
## 6.5 Effect of Knowledge on Enrolled Speakers

We show that QFA2SR is still effective when the surrogate SRS is *only* enrolled with the target speaker, which relaxes the assumption that the adversary knows and has some voices of *all* the enrolled speakers of a target SRS. $\mathcal{A}^{\text{T}}_{\text{TD-SV}}$ is not considered as only one speaker is enrolled for speaker verification.

We conduct experiments on commercial APIs in the same settings as § 6.3. The results are depicted in Fig. 6. We observe that *whether knowing and having voices of the other enrolled speakers have almost no effect for* $\mathcal{A}^{\text{T}}_{\text{OSI}}$, and the minor difference in $\text{ASR}_t$ is due to the randomness in crafting adversarial voices. It is no surprising as the optimal loss function ($f_1(x) = -S[(x)]_t$) *only* depends on the score of the target speaker which are independent on the scores of the other enrolled speakers. *For* $\mathcal{A}^{\text{UT}}_{\text{OSI}}$, *the* $\text{ASR}_u$ *decreases moderately if the adversary only knows the target speaker.* It is because the optimal loss function of $\mathcal{A}^{\text{UT}}_{\text{OSI}}$ ($f_3(x) = \theta - \max_{i \in G}[S(x)]_i$) can dynamically select the most transferable enrolled speaker as the optimization direction when the enrolled speakers of surrogate and target SRSs are the same, but when only the target speaker is enrolled in the surrogate SRS, $f_3$ becomes $\theta - [S(x)]_t$ that always optimizes towards the "target speaker" which may not be the most transferable one. *This problem also occurs in the most effective baseline attack (BIM), and* QFA2SR *still improves its transferability by a large margin.*

## 6.6 Scalability of QFA2SR

We have shown that QFA2SR is effective in attacking target SRSs with no more than 10 enrolled speakers. Now we evaluate attack scalability by increasing the enrolled speakers to 1,000, while the surrogate SRS is *only* enrolled with the target speaker. We use all the nine open-source SRSs as target SRSs and $\text{Spk}_{1000}$-enroll-P1&P2 as enrollment voices. Fig. 7 compares the ASR of QFA2SR between 10 and 1,000 enrolled speakers. *With the increase of enrolled speakers, the* $\text{ASR}_t$ *of* $\mathcal{A}^{\text{T}}_{\text{OSI}}$ *decreases slightly on some target SRSs, while the* $\text{ASR}_u$ *of* $\mathcal{A}^{\text{UT}}_{\text{OSI}}$ *increases, indicating the scalability of* QFA2SR. It is because those adversarial voices, optimized towards the target speaker and successfully transferring to the target SRS, often have higher scores on the target speaker than on other enrolled speakers, thus rarely get recognized as other enrolled speakers when increasing enrolled speakers. Thus, the $\text{ASR}_t$ of $\mathcal{A}^{\text{T}}_{\text{OSI}}$ does not decrease too much. On the other hand, those that fail to transfer to the target SRS are more likely to be recognized as other enrolled speakers by the target SRS when increasing enrolled speakers, thus, the $\text{ASR}_u$ of $\mathcal{A}^{\text{UT}}_{\text{OSI}}$ increases.




(a) $\text{ASR}_t$ of QFA2SR in $\mathcal{A}^{\text{T}}_{\text{OSI}}$ (b) $\text{ASR}_u$ of QFA2SR/BIM in $\mathcal{A}^{\text{UT}}_{\text{OSI}}$

Fig. 7: The scalability of QFA2SR.

# 7 Countermeasures

We discuss and evaluate possible countermeasures by considering transformation-based defenses and liveness detection.

## 7.1 Transformation-based Defenses

Transformation-based defenses apply some transformations to input voices to disrupt adversarial perturbations. We consider the seven most efficient such defenses in [31], i.e., Quantization (QT), Audio Turbulence (AT), Average Smoothing (AS), Median Smoothing (MS), Down Sampling (DS), Low Pass Filter (LPF), and Band Pass Filter (BPF), some of which are reported promising for mitigating existing adversarial attacks. We incorporate each of them to target SRSs and check the accuracy of normal voices of enrolled speakers ($\text{Spk}_{10}$-test) and imposters ($\text{Spk}_{10}$-imposter), and the accuracy of adversarial voices crafted from $\text{Spk}_{10}$-imposter. We use XV-P and Res18-V as target SRSs which have different architectures, and all nine open-source SRSs except for the target SRS are used as surrogate SRSs. We conduct the evaluation in $\mathcal{A}^{\text{T}}_{\text{OSI}}$ and set perturbation budget $\varepsilon = 0.02$, step size $\alpha = \varepsilon/5$, number of steps $N = 300$, and sampling size $\beta = 5$. The results are shown in Table 7. We find that *they are either effective in defending* QFA2SR *but significantly sacrificing the accuracy on the normal voices of enrolled speakers, or ineffective, thus are not suitable for mitigating* QFA2SR *regarding the trade-off*

Table 7: The accuracy (%) of normal voices and adversarial voices crafted by QFA2SR on target SRSs with defenses.

| | | Baseline | QT | AT | AS | MS | DS | LPF | BPF |
|---|---|---|---|---|---|---|---|---|---|
| XV-P | Enrolled | 97.2 | 75.3 | 71.4 | 91.5 | 29.6 | 44.8 | 62.8 | 70.6 |
| | Imposter | 97.4 | 99.6 | 99.5 | 96.4 | 100 | 100 | 99.8 | 99.6 |
| | QFA2SR | 10.4 | 16.5 | 26.8 | 12.1 | 84.6 | 88.6 | 47.2 | 42 |
| Res18-V | Enrolled | 92.5 | 46 | 6.4 | 7.1 | 3.9 | 8.1 | 10.6 | 0 |
| | Imposter | 92.6 | 92.2 | 100 | 99.9 | 98.4 | 100 | 100 | 100 |
| | QFA2SR | 85.3 | 91.5 | 97.9 | 99.8 | 100 | 100 | 100 | 100 |

(1) Baseline: target SRS without any defense. (2) Enrolled/imposter: normal voices from enrolled speakers/imposters. (3) Only differ-enroll is considered, an easier setting for defenses.

Table 8: Results of liveness detection.

| Detector | Benign voices | | Adversarial voices | | | |
|---|---|---|---|---|---|---|
| | TNR | FPR | Physical | | Digital | |
| | | | TPR | FNR | TPR | FNR |
| Void | **79.1%** | **20.9%** | **80.2%** | **19.8%** | 11.0% | 89.0% |
| LFCC-LCNN | 59.2% | 40.8% | 59.3% | 40.7% | 15.5% | 84.5% |
| LFCC-GMM | 61.8% | 38.2% | 61.6% | 38.4% | **25.0%** | **75.0%** |

*between normal and adversarial accuracy.* This is because they mitigate QFA2SR by lowering the scores of adversarial voices to fall below the threshold $\theta$ of target SRSs, which also incurs the same side-effect on normal voices.

## 7.2 Liveness Detection

By exploiting the different characteristics of the voices generated by human vocal tract and electronic loudspeakers, liveness detection predicts whether or not input voices are uttered by humans. Such defense can be used to defend against QFA2SR when launched over the air, e.g., when attacking voice assistants deployed in voice-controlled devices.

We use three recent liveness detectors that are open sourced and reported promising in the ASVspoof challenge [24, 55]: Void [24], LFCC-LCNN [77], and LFCC-GMM [55]. These detectors are trained using the physical access dataset of ASVspoof. Following [55], we compute True/False Positive/Negative Rate (i.e., TPR, TNR, FPR, and FNR) on the adversarial and benign voices used in § 6.4 (i.e., experiments on voice assistants). To void confusing, we use *Physical* to refer the adversarial voices that are played and recorded with 3 loudspeakers (JBL clip3 portable loudspeaker, TMall Genie smart speaker X5, and DELL laptop) and 3 microphones (Google Pixel 5 and iPhone 6 Plus, and iPad Pro 10.5), leading to 9 different hardware setups, and use *Digital* to refer the adversarial voices that are directly fed to the detector using the audio files. The average results are shown in Table 8. *These detectors can detect adversarial voices in the physical world (i.e., played over the air) at the cost of falsely rejecting many benign voices (more than 20%). Unsurprisingly, they have a remarkably high FNR (at least 75%) on adversarial voices in the digital world, indicating that liveness detection cannot defeat our attack when adversarial voices are launched via APIs.* This is no surprising since these adversarial voices do not contain the characteristics of loudspeakers.

# 8 Discussion

We discuss the generalizability of our methods for improving transferability and interesting future works.

**Generalizability of the three methods.** The optimal loss functions we selected are scenario-dependent, so they may not be optimal for other scenarios other than $\mathcal{A}^{\mathtt{T}}_{\mathtt{OSI}}$, $\mathcal{A}^{\mathtt{UT}}_{\mathtt{OSI}}$, and $\mathcal{A}^{\mathtt{T}}_{\mathtt{TD\text{-}SV}}$. It is interesting to consider other scenarios and design specific loss functions for them in future. SRS ensemble and time-freq corrosion are scenario-independent, but their effectiveness should still be evaluated in other scenarios.

**How to further improve QFA2SR?** While QFA2SR significantly improves the transferability, there is still space for improvement. Possible directions include using more advanced optimization methods (e.g., momentum-based gradient [37, 73] and Nesterov accelerated gradient [54]) and adopting more effective loss balancing strategies for SRS ensemble, (e.g., uncertainty-based balancing [46]).

**How to launch effective transfer attack without voices of the target speaker?** It is challenging to craft adversarial voices on surrogate SRSs when the adversary has no voices of the target speaker, due to the lack of optimization guidance by the embedding of the target speaker. One potential solution is dictionary attack [59], which creates a master voice that matches the identity of a large population such that it is likely to bypass the authentication of the target speaker. However, this attack is extremely limited in the query-free black-box setting. Future works can address this by incorporating the methods of QFA2SR into dictionary attack.

# 9 Conclusion

We proposed QFA2SR, so far the most effective query-free black-box adversarial attacks against SRSs. It leverages the transferability of adversarial voices and features three novel methods, i.e., tailored loss functions, SRS ensemble, and time-freq corrosion, which significantly improves the transferability. From the adversary perspective, our work unveils the feasibility of launching realistic and practical adversarial attacks against strictly protected proprietary commercial SRS APIs and voice-controlled devices in a complete black-box manner without queries them when crafting adversarial voices, thus enabling lots of follow-up attacks, e.g., those targeting speech recognition systems. From the perspective of SRSs maintainers and inspectors, our attack can serve as a strong baseline for measuring adversarial robustness under a realistic setting.

# References


[1] Amazon Mechanical Turk Platform. https://www.mturk.com.

[2] Citi Uses Voice Prints To Authenticate Customers Quickly And Effortlessly. https://www.citibank.com.hk/english/

info/pdf/VoiceBiometrics_PressRelease_Eng_final_online.pdf.

[3] Dragon ID from Nuance Uses Voiceprint to Unlock Phones. https://www.phonescoop.com/articles/article.php?a=10547.

[4] iflytek. http://www.iflytek.com/en.

[5] Implementation of SV2TTS. https://github.com/CorentinJ/Real-Time-Voice-Cloning.

[6] JBL clip3 portable speaker. https://www.jbl.com/bluetooth-speakers/JBL+CLIP+3.html.

[7] Personalized Hey Siri. https://machinelearning.apple.com/research/personalized-hey-siri.

[8] QFA2SR. https://sites.google.com/view/qfa2sr.

[9] Talentedsoft. http://www.talentedsoft.com.

[10] TD Bank Uses Voice Prints To Authenticate Customers during phone calls. https://tdbank.intelliresponse.com/index.jsp?requestType=NormalRequest&question=What+is+TD+VoicePrint+and+how+do+I+enroll).

[11] Teach google assistant to recognize your voice with voice match. https://support.google.com/assistant/answer/9071681?hl=en&co=GENIE.Platform%3DAndroid.

[12] The applications of Junlin's voiceprint recognition solutions on Smart Household Appliances can realize user permission management to distinguish different family members' permission to different appliances. https://www.junlinpro.com/en/solution/index.html?id=11&sid=42.

[13] TMall Genie has voiceprint recognition to ensure that only authorized users can place orders. http://thearf-org-unified-admin.s3.amazonaws.com/MSI/2021/04/MSI_Report_21-114.pdf.

[14] TMall Genie smart speaker X5. https://www.dxomark.com/speakers/TMall/Genie-X5.

[15] The most popular acoustic features. http://speech.ee.ntu.edu.tw/~tlkagk/courses/DLHLP20/ASR%20(v12).pdf, 2020.

[16] Ivector-plda model released by kaldi. https://kaldi-asr.org/models/m7, 2022.

[17] Jingdong Speaker Recognition. https://neuhub.jd.com/market/api/325, 2022.

[18] Microsoft azure speaker recognition. https://docs.microsoft.com/en-us/rest/api/speakerrecognition/, 2022.

[19] Xvector-plda model released by kaldi. https://kaldi-asr.org/models/m8, 2022.

[20] Hadi Abdullah, Washington Garcia, Christian Peeters, Patrick Traynor, Kevin R. B. Butler, and Joseph Wilson. Practical hidden voice attacks against speech and speaker recognition systems. In *NDSS*, 2019.

[21] Hadi Abdullah, Aditya Karlekar, Vincent Bindschaedler, and Patrick Traynor. Demystifying limited adversarial transferability in automatic speech recognition systems. In *ICLR*, 2021.

[22] Hadi Abdullah, Muhammad Sajidur Rahman, Washington Garcia, Logan Blue, Kevin Warren, Anurag Swarnim Yadav, Tom Shrimpton, and Patrick Traynor. Hear "no evil", see "kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems. In *IEEE S&P*, 2021.

[23] Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, and Patrick Traynor. Sok: The faults in our asrs: An overview of attacks against automatic speech recognition and speaker identification systems. In *S&P*, 2021.

[24] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. Void: A fast and light voice liveness detection system. In *USENIX Security Symposium*, 2020.

[25] Gautam Bhattacharya, Md. Jahangir Alam, and Patrick Kenny. Deep speaker recognition: Modular or monolithic? In *Interspeech*, 2019.

[26] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *S&P*, 2017.

[27] Guangke Chen, Sen Chen, Lingling Fan, Xiaoning Du, Zhe Zhao, Fu Song, and Yang Liu. Who is real Bob? adversarial attacks on speaker recognition systems. In *S&P*, 2021.

[28] Guangke Chen, Yedi Zhang, Zhe Zhao, and Fu Song. QFA2SR: Query-free adversarial transfer attacks to speaker recognition systems. https://arxiv.org/abs/2305.14097, 2023.

[29] Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, and Yang Liu. SEC4SR: A security analysis platform for speaker recognition. *CoRR*, abs/2109.01766, 2021.

[30] Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, and Yang Liu. AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems. *IEEE Transactions on Dependable and Secure Computing*, pages 1–17, 2022.

[31] Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, Feng Wang, and Jiashui Wang. Towards understanding and mitigating audio adversarial examples for speaker recognition. *IEEE Transactions on Dependable and Secure Computing*, pages 1–17, 2022.

[32] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, 2018.

[33] Najim Dehak, Reda Dehak, James R Glass, Douglas A Reynolds, Patrick Kenny, et al. Cosine similarity scoring without score normalization techniques. In *Odyssey*, 2010.

[34] Najim Dehak, Réda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet, and Pierre Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *INTERSPEECH*, 2009.

[35] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech*, 2020.

[36] Shaojin Ding, Tianlong Chen, Xinyu Gong, Weiwei Zha, and Zhangyang Wang. Autospeech: Neural architecture search for speaker recognition. In *Interspeech*, 2020.

[37] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018*, 2018.

[38] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *ASIACCS*, 2020.

[39] Luciana Ferrer, Mitchell McLaren, and Niko Brümmer. A speaker verification backend with robust performance across conditions. *Comput. Speech Lang.*, 2022.

[40] Yuan Gong and Christian Poellabauer. Crafting adversarial examples for speech paralinguistics applications. *CoRR*, abs/1711.03280, 2017.

[41] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[42] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014.

[43] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014.

[44] Arindam Jati, Chin-Cheng Hsu, Monisankha Pal, Raghuveer Peri, Wael AbdAlmageed, and Shrikanth Narayanan. Adversarial attack and defense strategies for deep speaker recognition systems. *Computer Speech & Language*, 68:101199, 2021.

[45] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *NeurIPS*, 2018.

[46] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018.

[47] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *INTERSPEECH*, 2015.

[48] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 5220–5224. IEEE, 2017.

[49] Felix Kreuk, Yossi Adi, Moustapha Cissé, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. In *ICASSP*, 2018.

[50] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *ICLR 2017*, 2017.

[51] Steve Lawrence, C. Lee Giles, and Ah Chung Tsoi. Lessons in neural network training: Overfitting may be harder than expected. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference*, pages 540–545, 1997.

[52] Xu Li, Jinghua Zhong, Xixin Wu, Jianwei Yu, Xunying Liu, and Helen Meng. Adversarial attacks on gmm i-vector based speaker verification systems. In *ICASSP*, 2020.

[53] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In *CCS*, 2020.

[54] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *ICLR*, 2020.

[55] Xuechen Liu, Xin Wang, Md. Sahidullah, Jose Patino, Héctor Delgado, Tomi Kinnunen, Massimiliano Todisco, Junichi Yamagishi, Nicholas W. D. Evans, Andreas Nautsch, and Kong Aik Lee. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *CoRR*, abs/2210.02437, 2022.

[56] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[57] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[58] Yuhao Mao, Chong Fu, Saizhuo Wang, Shouling Ji, Xuhong Zhang, Zhenguang Liu, Jun Zhou, Alex X. Liu, Raheem Beyah, and Ting Wang. Transfer attacks revisited: A large-scale empirical study in real computer vision settings. In *Proceedings of the 43rd IEEE Symposium on Security and Privacy*, pages 1423–1439, 2022.

[59] Mirko Marras, Pawel Korus, Nasir D. Memon, and Gianni Fenu. Adversarial optimization for dictionary attacks on speaker verification. In *Interspeech*, 2019.

[60] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *CoRR*, abs/1003.4083, 2010.

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

[62] Mahesh Kumar Nandwana, Luciana Ferrer, Mitchell McLaren, Diego Castan, and Aaron Lawson. Analysis of critical metadata factors for the calibration of speaker recognition systems. In *INTERSPEECH*, 2019.

[63] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *ICASSP*, 2015.

[64] H. F. Pardede, V. Zilvan, D. Krisnandi, A. Heryana, and R. B. S. Kusumo. Generalized filter-bank features for robust speech recognition against reverberation. In *IC3INA*, 2019.

[65] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech*. ISCA, 2019.

[66] D. Prabakaran and R. Shyamala. A review on performance of voice feature extraction techniques. In *Proceedings of the 3rd International Conference on Computing and Communications Technologies*, 2019.

[67] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. Speechbrain: A general-purpose speech toolkit. *CoRR*, abs/2106.04624, 2021.

[68] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digit. Signal Process.*, 2000.

[69] Antony W Rix, John G Beerends, Michael P Hollier, and An-

dries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001.

[70] Maliheh Shirvanian, Summer Vo, and Nitesh Saxena. Quantifying the breakability of voice assistants. In *PerCom*, 2019.

[71] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *ICASSP*, 2018.

[72] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

[73] Hao Tan, Zhaoquan Gu, Le Wang, Huan Zhang, Brij B. Gupta, and Zhihong Tian. Improving adversarial transferability by temporal and spatial momentum in urban speaker recognition systems. *Comput. Electr. Eng.*, 104:108446, 2022.

[74] Voiceprint: The New WeChat Password. https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password/.

[75] Dong Wang. A simulation study on optimal scores for speaker recognition. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1):1–23, 2020.

[76] Guoqiu Wang, Huanqian Yan, Ying Guo, and Xingxing Wei. Improving adversarial transferability with gradient refining. *CoRR*, abs/2105.04834, 2021.

[77] Xin Wang and Junichi Yamagishi. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. In *Interspeech*, 2021.

[78] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. "hello, it's me": Deep learning-based speech synthesis attacks in the real world. In *CCS*, 2021.

[79] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *Speech Commun.*, 2015.

[80] Yong Xiang, Guang Hua, and Bin Yan. *Digital audio watermarking: fundamentals, techniques and challenges*. Springer, 2017.

[81] Xiong Xiao, Shengkui Zhao, Duc Hoang Ha Nguyen, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li. Speech dereverberation for enhancement and recognition using dynamic features constrained deep neural networks and feature adaptation. *EURASIP Journal on Advances in Signal Processing*, 2016(1):4, 2016.

[82] Xuejing Yuan, Yuxuan Chen, Yue Zhao, Yunhui Long, Xiaokang Liu, Kai Chen, Shengzhi Zhang, Heqing Huang, Xiaofeng Wang, and Carl A. Gunter. Commandersong: A systematic approach for practical adversarial voice recognition. In *USENIX Security*, 2018.

[83] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *CCS*, 2017.

[84] Baolin Zheng, Peipei Jiang, Qian Wang, Qi Li, Chao Shen, Cong Wang, Yunjie Ge, Qingyang Teng, and Shenyi Zhang. Black-box adversarial attacks on commercial speech platforms with minimal information. In *CCS*, 2021.

# A Phrases for Text-Dependent SV

We use the following ten phrases supported by Microsoft Azure [18]:

- I am going to make him an offer he cannot refuse.
- Houston we have had a problem.
- My voice is my passport verify me.
- Apple juice tastes funny after toothpaste.
- You can get in without your password.
- You can activate security system now.
- My voice is stronger than passwords.
- My password is not your business.
- My name is unknown to you.
- Be yourself everyone else is already taken.

# B More Detail about SRSs

The details of the nine adopted open-source SRSs are shown in Table 9. They cover three architectures, i.e., the typical GMM [34] and the state-of-the-art deep neural networks (TDNN [35] and CNN [25]). GMM is a generative model, while the others are discriminative models. Auto is an automatically searched architecture by [36] while the others are manually designed by the existing works. They also cover three most popular acoustic features [15] (i.e., spectrogram [42], fBank [64], and MFCC [60]), and two commonly-used scoring methods (i.e., PLDA [62] and COSS [33]). They are trained using two datasets, i.e., VoxCeleb1 [61] and VoxCeleb2 [32], which have different number of speakers, utterances, and subjects background (e.g., ethnicities, accents, age, and profession).

We tune the threshold θ of the open-source SRSs listed in Table 9 based on the Equal Error Rate (EER) meaning the same FAR and FRR, where False Acceptance Rate (FAR) is the proportion of voices that are uttered by unenrolled speakers but accepted by the SRS, and False Rejection Rate (FRR is the proportion of voices that are uttered by enrolled speakers but rejected. The tuned threshold and the performance of SRSs are shown in Table 11 where column (IER) denotes Identification Error Rate, i.e., the proportion of voices uttered by enrolled speakers which should not be rejected but incorrectly classified by the SRS [27].

For the commercial SRSs, the responses from Microsoft Azure, TalentedSoft, and iFltek only contain the scores given by the enrolled speakers without the final decision results, which should be determined by the developers to adapt to the specific applications. Therefore, we tune the threshold θ of these commercial SRSs the same as for open-source SRSs. In contrast, Jingdong, Google Assistant, Apple Siri, and TMall Genie only provide the decision result without any scores, so there is no need to tune the threshold θ.

Table 9: Details of the 9 open-source SRSs where Arch denotes architecture.

| Arch | Name | #Params | Acoustic feature | Training dataset | Scoring Backend |
|---|---|---|---|---|---|
| GMM | Ivector-PLDA (IV) [16] | 80.37M | MFCC | VoxCeleb1&2 | PLDA |
| TDNN | ECAPA-TDNN (ECAPA) [35] | 20.77M | fBank | VoxCeleb1 | COSS |
| | Xvector-PLDA (XV-P) [19] | 5.79M | MFCC | VoxCeleb1&2 | PLDA |
| | Xvector-COSS (XV-C) [67] | 4.21M | fBank | VoxCeleb1 | COSS |
| CNN | Res18-Identification (Res18-I) [25] | 11.17M | spectrogram | VoxCeleb1 | COSS |
| | Res18-verification (Res18-V) [25] | 11.17M | spectrogram | VoxCeleb1 | COSS |
| | Res34-Identification (Res34-I) [32] | 21.28M | spectrogram | VoxCeleb1 | COSS |
| | Res34-Verification (Res34-V) [32] | 21.28M | spectrogram | VoxCeleb1 | COSS |
| | AutoSpeech (Auto) [36] | 15.11M | spectrogram | VoxCeleb1 | COSS |

Table 10: The datasets used for enrolling and attacking voice assistants.

| Voice Assitant | Activation Phrase | Number |
|---|---|---|
| Google Assitant | Ok Google | 5 |
| Apple Siri | Hey Siri | 1 |
| | Hey Siri, send a message | 1 |
| | Hey Siri, how's the weather today | 1 |
| | Hey Siri, set a timer for three minutes | 1 |
| | Hey Siri, play some music | 1 |
| TMall Genie | TMall Genie (Chinese) | 3 |
| | TMall Genie, who am I (Chinese) | 5 |

Table 11: The threshold and performance of SRSs.

| SRS | SV | | OSI | | |
|---|---|---|---|---|---|
| | EER (%) | θ | EER (%) | IER (%) | θ |
| IV | 1.40 | 10.41 | 6.50 | 0 | 12.90 |
| ECAPA | 1.43 | 0.40 | 3.01 | 0 | 0.48 |
| XV-P | 1.12 | 12.64 | 3.02 | 0 | 16.23 |
| XV-C | 6.10 | 0.60 | 11.57 | 0 | 0.69 |
| Res18-I | 1.92 | 0.45 | 6.91 | 0 | 0.57 |
| Res18-V | 2.83 | 0.41 | 6.80 | 0 | 0.55 |
| Res34-I | 1.50 | 0.46 | 9.60 | 0 | 0.57 |
| Res34-V | 2.80 | 0.43 | 5.83 | 0 | 0.56 |
| Auto | 1.52 | 0.29 | 5.61 | 0 | 0.38 |
| Microsoft Azure | 0.72 | 0.49 | 1.6 | 0 | 0.53 |
| Talentedsoft | - | - | 5.1 | 0 | 0.19 |
| iFlytek | - | - | 14 | 0 | 0.64 |
| Jingdong | $0.5^{\sharp}$ | $0^{\natural}$ | - | - | - |
| Google Assistant | $0.8^{\sharp}$ | $0^{\natural}$ | - | - | - |
| Apple Siri | $1.2^{\sharp}$ | $0^{\natural}$ | - | - | - |
| TMall Genie | $0.4^{\sharp}$ | $0^{\natural}$ | $0.5^{\sharp}$ | 0 | $0^{\natural}$ |

Note: the number with "♯" and "♮" superscript denote FAR and FRR, respectively. "-" means unsupported, i.e., Jingdong, Google Assistant, Apple Siri do not support OSI, and TalentedSoft and iFltek do not support TD-SV.

## C Details of the Compared Attacks

**Adversarial attacks.** We set $L_\infty$ perturbation budget $\varepsilon = 0.02$, step size $\alpha = \frac{\varepsilon}{5} = 0.004$, number of steps $N = 300$, and sampling size $\beta = 5$ for QFA2SR, and discard those seed voices that are falsely accepted by the target commercial SRSs. BIM is implemented as a special case of QFA2SR with only one surrogate SRS but without time-freq corrosion. For FakeBob (resp. SirenAttack), we set the number of iterations (resp. maximum number of epochs) to 1500 (resp. 100), which is sufficient for the attacks to converge according to our experiments, while other parameters are the same as the original work[27, 38]. Additionally, we set the confidence value $\kappa = 5 \times \theta$ in FakeBob and SirenAttack where $\theta$ is the threshold of the surrogate SRS. This enables the attacks to continue searching for high-confident adversarial voices instead of early-stopping, which may benefit the transferability [26, 27]. For Kenansville, we use the Fast Fourier Transform (FFT) method to perturb a voice with 15 maximal number of iterations (the same as the original work [22]), while the Singular Spectrum Analysis method is not considered since it is comparable with FFT method regarding the transferability, but is much less efficient. PGD is the same as the BIM attack except that it starts from a randomly perturbed example, which may help the attack find a better local optimum. We run the random start 10 times and select the adversarial voice with the minimal loss that is more likely to transfer, and other settings are the same as BIM. For CW attack, we adopt its $L_\infty$ version, set the confidence value $\kappa = 5 \times \theta$, and adopt the efficient implementation in [57].

**Hidden voice attack.** We exploit Time Domain Inversion (TDI) to perturb a voice since it is one of the most effective method [20]. TDI features the parameter window size $w$, where the smaller $w$, the less comprehensible the voices for human and the harder the voices to be correctly recognized by the SRS. To produce the least understandable voices for human when ensuring the correct recognition of the SRS, we start from $w = 1$ milliseconds (ms), gradually increase to 10 ms with step of 0.5 ms,

## D More Details of Experimental Setting on Attacking Voice Assistants

**Datasets.** The activation phrase as well as the recording number is shown in Table 10. For Google Assistant and Apple Siri, these activation phrases are used for both the enrollment voices and the seed voices for the attack. For TMall Genie, "TMall Genie" is used for enrolling and "TMall Genie, who am I" is used as the attack seed voices. The reason is that the activation of TMall Genie by "TMall Genie" is speaker-independent, and we have to ask the TMall Genie "who am I" to determine the identity of the speaker.

**Attack success rate.** For Google Assistant and Apple Siri, we count a successful attack only when the voice assistants are activated within the number of allowed queries to the target SRS. For TMall Genie, there are three kinds of response to "TMall Genie, who am I", each reflecting the confidence that TMall Genie recognizes the voice as from the speaker `SPK_ID`, namely, "Hello, `SPK_ID`, happy to serve you." (high-confidence), "I think you are `SPK_ID`, am I right?" (medium confidence), and "I am unfamiliar with your voice." (low confidence). We regard an attack as a successful attack when one of the following conditions holds: (1) The seed voice receives the low-confidence response or the other two responses where `SPK_ID` is different from the target speaker, and the adversarial voice receives the medium or high confidence response where `SPK_ID` is identical to the target speaker. (2) The seed voice receives the medium confidence response, the adversarial voice receives the high confidence response, and both of their `SPK_ID` are identical to the target speaker.