AS2T: Arbitrary Source-To-Target Adversarial Attack on Speaker Recognition Systems

Guangke Chen, Zhe Zhao, Fu Song, Sen Chen, Lingling Fan, and Yang Liu *Senior Member, IEEE*

Abstract—Recent work has illuminated the vulnerability of speaker recognition systems (SRSs) against adversarial attacks, raising significant security concerns in deploying SRSs. However, they considered only a few settings (e.g., some combinations of source and target speakers), leaving many interesting and important settings in real-world attack scenarios alone. In this work, we present AS2T, the first attack in this domain which covers all the settings, thus allows the adversary to craft adversarial voices using arbitrary source and target speakers for any of three main recognition tasks. Since none of the existing loss functions can be applied to all the settings, we explore many candidate loss functions for each setting including the existing and newly designed ones. We thoroughly evaluate their efficacy and find that some existing loss functions are suboptimal. Then, to improve the robustness of AS2T towards practical over-the-air attack, we study the possible distortions occurred in over-the-air transmission, utilize different transformation functions with different parameters to model those distortions, and incorporate them into the generation of adversarial voices. Our simulated over-the-air evaluation validates the effectiveness of our solution in producing robust adversarial voices which remain effective under various hardware devices and various acoustic environments with different reverberation, ambient noises, and noise levels. Finally, we leverage AS2T to perform thus far the largest-scale evaluation to understand transferability among 14 diverse SRSs. The transferability analysis provides many interesting and useful insights which challenge several findings and conclusion drawn in previous works in the image domain. Our study also sheds light on future directions of adversarial attacks in the speaker recognition domain.

Index Terms—Adversarial examples, speaker recognition, speaker verification, over-the-air attack, transfer attack

1 Introduction

Speaker recognition (SR) is an automatic procedure of verifying or identifying individual speakers by extracting and interpreting their unique acoustic characteristics [1]. There are three main SR tasks: close-set identification (CSI), open-set identification (OSI), and speaker verification (SV). CSI recognizes unknown speakers from a group of enrolled speakers G. OSI is similar to CSI except that it may regard a speaker as an imposter (i.e., an unenrolled speaker). SV is similar to OSI except that only one speaker can be enrolled. SR has been adopted by open-source platforms (e.g., SpeechBrain [2]) and commercial products (e.g., Microsoft Azure [3]), and used in security-critical scenarios such as electrical appliances access control in smart home [4] and remote voice authentication in financial transaction [5].

Machine learning including deep learning is the dominant approach to implement the state-of-the-art SR systems (SRSs) [6], [7]. Unsurprisingly, SRSs have been shown to be fragile to adversarial attacks in some settings [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], namely, adding a tiny perturbation to a voice uttered by one source speaker is misclassified by the SRS, but still correctly recognized as the source speaker by ordinary users.

TABLE 1: Existing attacks

Attack	Task	Reference
$s \to \texttt{untar}$ for $s \in G$	CSI	[10], [15], [17], [18]
$s \to t$ for $s, t \in G$	CSI	[15], [16], [17]
$s \to t$ for $s \notin G$ and $t \in G$	SV	[8], [9], [11], [12], [13], [14], [16]
$s \to t$ for $s \notin G$ and $t \in G$	OSI	[16]

We will denote by $s \to \texttt{untar}$ and $s \to t$ ($s \neq t$) the untargeted and targeted attacks, respectively. When $s \in G$ (resp. $s \notin G$), the source speaker is one of the enrolled speakers (resp. an unenrolled speaker). When $t \in G$ (resp. $t = \texttt{imposter}$), the adversarial voice is recognized as the enrolled speaker t (resp. rejected as an imposter) by the SRS.

Prior works make remarkable progress in revealing the serve security implications of adversarial attacks on both open-source and commercial SRSs. However, they considered only a few settings as shown in TABLE 1, leaving many interesting and important settings alone, e.g., $s \to \texttt{imposter}$ for $s \in G$ on the OSI and SV tasks and $s \to \texttt{untar}$ for $s \notin G$ on the OSI and CSI tasks (cf. TABLE 3 for all the 10 settings). Yet these settings are interesting and important in real-world attack scenarios. For example, voice-controllable smart home offers for family members hands-free control of smart products, e.g., thermostat and light. Some ordinary products are controlled by all family members, while others are controlled exclusively by some members, such as parents. If an adversary attempts to acquire the permission of exclusively controlled products, he needs to perform a targeted attack $s \to t \in G$ for $s \notin G$ on the OSI task (considered in [16]) where the target speaker t owns the permission.

- *Guangke Chen is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China; Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China; and University of Chinese Academy of Sciences, Beijing, China.*
- *Zhe Zhao is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China.*
- *Fu Song (corresponding author) is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. Email: songfu@shanghaitech.edu.cn*
- *Sen Chen is with College of Intelligence and Computing, Tianjin University, Tianjin, China.*
- *Lingling Fan is with College of Cyber Science, Nankai University, Tianjin, China.*
- *Yang Liu is with School of Computer Science and Engineering, Nanyang Technological University, Singapore.*

In contrast, if he intends to obtain the privilege to control ordinary products, he can simply perform an untargeted attack $s \rightarrow \texttt{untar}$ for $s \notin G$, which is less difficult than targeted attack. Additionally, while the attack $s \rightarrow t \in G$ for $s \notin G$ on the SV task considered in previous works can be exploited to unlock the victim's smartphone, log into the victim's applications, etc., the attack $s \rightarrow \texttt{imposter}$ for $s \in G$ on the SV task can cause Denial-of-Service to the victim by disabling his authentication.

In this work, we aim to tackle the key limitation of prior works by proposing an attack, named **A**rbitrary **S**ource-**T**o-**T**arget adversarial attack (AS2T), covering all the above settings. Given a source speaker, either enrolled or unenrolled, a benign voice from the source speaker, and a target speaker that is either `untar` (for untargeted attack), or an enrolled or unenrolled speaker (for targeted attack), AS2T produces an adversarial voice with which the adversary can achieve his goal in the intended attack scenario. The design and implementation of AS2T is motivated by the following research questions: (RQ1) How to construct adversarial voices given arbitrary source and/or target speakers, considering that the adversary may own different levels of knowledge about SRSs? (RQ2) How to launch practical over-the-air attacks where distortions from both environment and hardware may disrupt adversarial voices? (RQ3) How about the attack capability of AS2T under totally black-box setting in which the adversary does not own any information about the model and cannot frequently query it?

To address RQ1, we formulate the crafting of adversarial voices as a constrained optimization problem and design appropriate source-/target-oriented loss functions for each setting, i.e., a combination of source and target speakers and a recognition task. Note that none of existing loss functions in the literature can be applied to all the considered settings. For example, the commonly adopted Cross Entropy Loss, defined as the negative logarithm of the predicted probability of the source class (untargeted attack) or target class (targeted attack), cannot be used for the attack $s \rightarrow \texttt{untar}$ with $s \notin G$ on both the OSI and CSI tasks, as no enrolled speaker of the SR model corresponds to the speaker s. Therefore, we explore various candidate loss functions including existing and newly designed ones, and conduct a thorough evaluation to compare their effectiveness and efficiency for each setting. We find that our loss functions outperform some loss functions adopted by prior works. According to our results, AS2T is designed to adaptively choose the optimal loss function for each attack setting. Furthermore, the adversary can freely integrate our loss functions into an optimization approach according to his prior knowledge about the SR model under attack, ranging from white-box access to model structure and parameters to black-box access to model's outputs (decision or score). Our experimental results with three representative white-box and one state-of-the-art black-box approaches (after necessary modifications) demonstrate the generic capability of AS2T to different optimization approaches.

To address RQ2, we first investigate the major sources of distortions occurred in the over-the-air transmission which may destruct the effectiveness of adversarial voices, i.e., reverberation, ambient noise, and equipment distortion. Based on their nature and properties, we utilize different proper transformation functions to model those distortions and incorporate them into AS2T, where the loss is computed on the transformed voice, thus it is expected that the crafted adversarial voices can survive from these distortions. To guarantee the attacker can use the adversarial voices to launch attack in various environments and even target multiple victims' devices simultaneously, we specify a variety of parameters for the transformation functions, e.g., different Signal-to-Noise ratio for the function simulating the effect of ambient noise. We empirically confirm that our solution to RQ2 significantly enhances the robustness of adversarial voices against these distortions under various attack conditions, thus improves the practicability of AS2T.

Under RQ3, the adversary cannot directly and frequently query the target model and thus has to leverage transfer attacks, i.e., crafting adversarial voices on a white-box source model and transferring them to the totally black-box target model [9], [10], [16]. Transfer attack is an obstacle to securely deploying machine learning models due to its ability to perform simple and practical black-box attacks [19], [20]. To thoroughly evaluate the capability of AS2Tunder this challenging scenario and study the factors that may influence the success rate of transfer attacks, we perform a large-scale transferability analysis among 14 SR models, covering five model architectures, five training datasets, four input types, and two scoring backends. We study the impact of both model-specific factors (architecture, training dataset, and input type) and attack-specific factors (number of iterations, step_size, and perturbation budget) on the transferability success rate of AS2T. Our analysis provides lots of useful insights and findings for better launching transfer attacks. For instance, model-specific factors are dominant factors over attack-specific ones; adversarial examples tend to, but not necessarily, transfer well to the target models with the same architecture as the source model; the transferability between two models may be asymmetric, which challenges the decision boundary similarity based explanation of transferable adversarial examples in the image domain [19], [20], since similarity is symmetric; iterative attack does not necessarily produce less transferable adversarial voices than single-step attack, contradicting the finding in the image domain [21].

In summary, we make the following main contributions.

- We propose AS2T, an arbitrary source-to-target adversarial attack on SRSs. It features source-/target-oriented and novel loss functions and enables the adversary to use arbitrary source and target speakers to craft adversarial voices to achieve attack goals in adversary-chosen attack scenarios.
- We successfully construct robust adversarial voices which remain effective when being played over-the-air in various scenes. Our solution is modeling and incorporating the possible distortions into the generation of adversarial voices using different parameterized transformation functions.
- To the best of our knowledge, we perform the largest-scale transferability analysis in the speaker recognition domain. We discover many valuable insights which can guide the future works in this domain.

TABLE 2: Different goals an adversary intends to achieve

Goal		Description	Setting	Source Speaker	Target Speaker	ID
Malicious	Unauthorized Access	Obtain privilege	Exclusive privilege owned by $G_1 \subset G$	$s \in G \setminus G1$	$t \in G_1$	S1-1
				$s \notin G$	$t \in G_1$	S1-2
			Low-level privilege shared by G	$s \notin G$	Untargeted	S1-3
	Denial-of-Service	Hinder accessing services	Common services shared by G	$s \in G$	$t \notin G$	S2-1
			Personalized services	$s \in G$	$t \notin G$	S2-2
				$s \in G$	Untargeted	S2.3
	Anonymous Access	Hides identity when accessing illegal service	—	$s \in G$	Untargeted	S3-1
			Cause reputation degrade to a specific speaker	$s \in G$	$t \in G \setminus \{s\}$	S3-2
	Evasion	Anomalous subject evades detection	Single anomalous subject	$s \in G$	Untargeted	S4-1
			Multiple anomalous subjects $G_2 \subset G$	$s \in G$	$t \in G \setminus G_2$	S4-2
Beneficial	Privacy Protection	Protect privacy against excessive surveillance	—	$s \in G$	Untargeted	S5-1

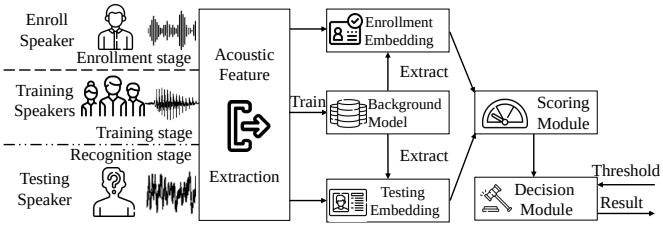


Fig. 1: A generic architecture of embedding-based SRSs

2 Speaker Recognition Systems

2.1 Overview of Speaker Recognition

Modern and cutting-edge SRSs represent characteristics of speakers as fixed-dimensional vectors, i.e., speaker embedding [22]. Fig. 1 depicts a generic architecture of embedding-based SRSs, consisting of training, enrollment, and recognition stages. In the training stage, a huge number of voices from thousands of training speakers are utilized to train a background model, which learns a mapping from voices to embeddings in the vector space. Classic background model utilizes Gaussian Mixture Model (GMM) [6], [23], which produces identity-vector (ivector) embeddings [24]. Recent promising background model utilizes deep neural networks, which produce deep embeddings, e.g., dvector [25] or xvector [26]. In the enrollment stage, the background model maps each enrolling speaker's voice to an *enrollment embedding*, regarded as a unique voice-identity of the enrolling speaker. In the recognition stage, given a voice of a testing speaker, its *testing embedding* is retrieved from the background model for scoring. The scoring module computes the similarity between the enrollment and testing embeddings based on which the result is produced by the decision module. There are two widely-used scoring approaches: Probabilistic Linear Discriminant Analysis (PLDA) [27] and COSine Similarity (COSS) [28]. The former one works well in most situations, but needs to be trained using the embeddings of training voices [22], while the latter one is a reasonable substitution of PLDA without training.

The acoustic feature extraction module converts raw speech signals to acoustic features carrying characteristics of the raw signals. Common feature extraction algorithms include speech spectrogram [29] fBank [30] [31], MFCC [32], and PLP [33]. Note that some end-to-end neural network-based SRSs, e.g., SincNet [34], extract features by the hidden neurons of neural networks instead of an explicit acoustic feature extraction module.

2.2 Speaker Recognition Tasks

There are three main speaker recognition tasks: open-set identification (OSI), close-set identification (CSI), and speaker verification (SV). OSI allows multiple speakers to be enrolled during the enrollment stage, forming a speaker group G. Given a voice x of a testing speaker, it determines whether x is uttered by one of the enrolled speakers or none of them, according to the scores of all the enrolled speakers and a preset (score) threshold θ. Formally, suppose the speaker group G has n speakers $\{1, 2, \cdots, n\}$, 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 $[S(x)]_i$ for $i \in G$ denotes the score of the voice x that is likely uttered by the speaker i. Intuitively, the system classifies the input voice x as the speaker i if and only if the score $[S(x)]_i$ of the speaker i is the largest one among all the enrolled speakers, and no less than the threshold θ. If all the scores are less than θ, the system directly rejects the voice, namely, it is not uttered by any of the enrolled speakers.

CSI and SV accomplish similar tasks as OSI, but with some exceptions. A CSI system never rejects any input voices, i.e., an input will always be classified as one of the enrolled speakers. Whereas an SV system can have exactly *one* enrolled speaker and checks if an input voice is uttered by the enrolled speaker.

3 Our Attack: AS2T

In this section, we first highlight the motivation and threat model of our work and then present in detail our attack.

3.1 Motivation and Threat Model

We assume that the adversary intends to craft an adversarial example from a voice uttered by a source speaker, so that it is classified as one of the enrolled speakers (untargeted attack) or the target speaker or imposter (targeted attack) by the SRS under attack (effectiveness), but is still recognized as the source speaker by ordinary users (stealthiness). TABLE 2 summarizes possible goals of the adversary, including malicious goals, e.g., unauthorized access, Denial-of-Service (Dos), anonymous access, and evasion, and beneficial goals, e.g., privacy protection.

We emphasize that different goals often require different source and/or target speakers. Even for the same goal, the source and target speakers also differ with the settings. Consider the unauthorized access goal. When the privilege the adversary intends to obtain is exclusive to some specific enrolled speakers, e.g., unlocking the victim's smartphone, logging into the victim's applications, and conducting illegal financial transactions on behalf of the victim, the adversary should target an enrolled speaker who owns this privilege.

TABLE 3: Settings of source/target speakers

Task	ID	Source Speaker	Target	Goals
OSI	C1	enrolled speaker s	enrolled speaker $t \neq s$	S1-1, S3-2, S4-2
	C2	unenrolled speaker	enrolled speaker	S1-2
	C3	enrolled speaker	imposter	S2-1, S2-2
	C4	enrolled speaker	untargeted	S2-3, S3-1, S4-1, S5-1
	C5	unenrolled speaker	untargeted	S1-3
CSI	C6	enrolled speaker s	enrolled speaker $t \neq s$	S1-1, S3-2, S4-2
	C7	unenrolled speaker	enrolled speaker	S1-2
	C8	enrolled speaker	untargeted	S2-3, S3-1, S4-1, S5-1
SV	C9	enrolled speaker	imposter	S2-1, S2-2, S2-3 S3-1, S4-1, S5-1
	C10	unenrolled speaker	enrolled speaker	S1-2, S1-3

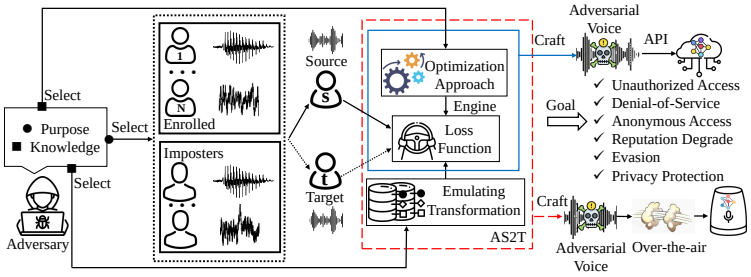


Fig. 2: The overview of our attack AS2T

In contrast, when the privilege is shared by all the enrolled speakers, e.g., all family members can remotely control some electrical appliances in the smart home, the adversary can simply perform untargeted attack.

Therefore, the adversarial attack with arbitrary source and target speakers is important and worthy to explore considering various attack scenarios in the physical world. As shown in TABLE 3, there are 10 combinations of source and target speakers on the three tasks, differing from the attacks in the image domain. Indeed, image recognition is a close-set multi-class problem and prior works on adversarial image attacks correspond to C6 and C8 in TABLE 3.

In this work, we propose an attack, named Arbitrary Source-To-Target adversarial attack (AS2T), covering all the possible combinations of source speaker and target speaker on all the three SRS tasks, as listed in TABLE 3. The source speaker could be an enrolled or unenrolled speaker and the target could be untargeted (untargeted attack), an enrolled speaker or imposter (targeted attack), leading to 2×3 combinations per task. Note that imposter represents all the unenrolled speakers. After excluding the meaningless combination unenrolled $\rightarrow$ imposter, there are 5 settings (C1-C5) for OSI task. Since CSI task never rejects, both enrolled $\rightarrow$ imposter and unenrolled $\rightarrow$ untargeted are excluded, leading to 3 settings (C6-C8) for CSI task. Note that unenrolled $\rightarrow$ untargeted is excluded since any input voice uttered by any unenrolled speaker will naturally be classified into one of enrolled speakers by CSI task. Since SV task is a binary classification, its setting is either enrolled $\rightarrow$ imposter (C9) or unenrolled $\rightarrow$ enrolled (C10). The overview of AS2T is depicted in Fig. 2. AS2T formulates the construction of adversarial voices as an optimization problem (cf. Section 3.2) and features source-/target-oriented loss functions (cf. Section 3.3). In addition, an adversary may own different levels of knowledge about the model under attack, from white-box access to model structure and parameters, to black-box access to model outputs (decisions or scores). The attack should be generic and can work under different levels of knowledge. We address this in the optimization approach (cf. Section 3.4). Furthermore, the crafted adversarial voices should also remain effective when being played over-the-air to launch practical attack in the physical world. The potential distortions occurred in this physical process are likely to destruct the effectiveness of adversarial examples. We address this challenge by modeling and incorporating the possible distortions into AS2T (cf. Section 3.5).

3.2 Problem Formulation

The problem of finding an adversarial voice from a voice x uttered by a source speaker s, is formalized as the following constrained optimization problem:

$$\text{argmin}_{x'} \; d(x', x) \quad \text{s.t. } D(x') = t \text{ and } x' \in [-1, 1]$$

where $d(x', x)$ is a distance metric quantifying the similarity between x' and x (stealthiness), and t is a target speaker (t is automatically selected in an untargeted attack).

The above minimization problem is difficult to solve due to the highly non-linear constraint $D(x') = t$. Therefore, we re-formulate it and turn to solve the following problem:

$$\text{argmin}_{x'} \quad \mathcal{L}(x', t) + \lambda \times d(x', x)$$

where $\mathcal{L}$ is the loss function indicating the effectiveness of the attack and the hyper-parameter λ is a trade-off between the effectiveness and stealthiness of the attack. In our attack, we utilize L_p norm as the distance metric, i.e., $d(x', x) = (\sum_i |x'_i - x_i|^p)^{\frac{1}{p}}$, which has been widely adopted in previous works, e.g., [16], [35], [36]. There are many choices of the loss function $\mathcal{L}$. We will explore various candidate loss functions in Section 3.3 and empirically evaluate them in Section 4.3.

3.3 Loss Function Design

As aforementioned, none of existing loss functions can be applied to all the settings listed in TABLE 3. In this subsection, we explore possible loss functions for each setting.

3.3.1 Loss functions for OSI task

We denote by $G = \{1, 2, \cdots, n\}$ the group of enrolled speakers. Then, the decision space of the OSI task is $\mathcal{D} = G \cup \{\texttt{imposter}\}$.

C1 and C2: Targeted attack ($t \in G$). Given a benign voice uttered by a source speaker s that is either an enrolled (i.e., $s \in G$, C1) or unenrolled (i.e., $s \notin G$, C2) speaker, we define the following four loss functions for targeted attack on the OSI task with a target speaker $t \in G$:

$$\mathcal{L}_{\text{CE}}(x, t) \triangleq -\log[\sigma(S(x))]_t \qquad \mathcal{L}_1(x, t) \triangleq -[S(x)]_t$$
$$\mathcal{L}_{\text{M}}(x, t) \triangleq \texttt{max}_{i \in G, i \neq t}[S(x)]_i - [S(x)]_t$$
$$\mathcal{L}_2(x, t) \triangleq \texttt{max}\{\theta, \texttt{max}_{i \in G, i \neq t}[S(x)]_i\} - [S(x)]_t$$

where σ denotes the softmax function, θ is a preset (score) threshold, and $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{M}}$ are the Cross Entropy Loss and the Margin Loss, respectively. $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{M}}$ are widely adopted to craft adversarial examples in the image domain, e.g., [35], [37]. Unlike $\mathcal{L}_{\text{M}}$ which aims to simultaneously increase the score of the target speaker t and reduce the scores of the other enrolled speakers, $\mathcal{L}_1$ is designed to increase the score of the target speaker t only, by which we can check the effectiveness of the term $\texttt{max}_{i \in G, i \neq t}[S(x)]_i$ in $\mathcal{L}_{\text{M}}$. $\mathcal{L}_2$ is designed such that $\mathcal{L}_2(x, t) \leq 0 \Leftrightarrow D(x) = t$. When $\mathcal{L}_2$

is minimized, the score $[S(x)]_t$ of the target speaker t is maximized to exceed the threshold θ and the scores of all the other enrolled speakers, indicating a successful attack.

C3: Targeted attack ($t \notin G$). Given a benign voice uttered by a source speaker s that is an enrolled speaker (i.e., $s \in G$), the following three loss functions are introduced to make its adversarial counterpart being rejected by the OSI task:

$$\mathcal{L}^s_{\text{CE}}(x,t) \triangleq \log[\sigma(S(x))]_s \qquad \mathcal{L}^s_1(x,t) \triangleq [S(x)]_s$$
$$\mathcal{L}_3(x,t) \triangleq \texttt{max}_{i\in G}[S(x)]_i - \theta$$

Note that the parameter t in the above loss functions is not necessary and is used to make notions consistent. $\mathcal{L}^s_{\text{CE}}$ (resp. $\mathcal{L}^s_1$) is the negation of $\mathcal{L}_{\text{CE}}$ (resp. $\mathcal{L}_1$) in which the target speaker t is replaced by the source speaker s, intended to minimize the score of the source speaker s (Note that $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_1$ are designed to maximize the score of the target speaker t). $\mathcal{L}_3$ is designed such that $\mathcal{L}_3(x,t) \leq 0 \Leftrightarrow D(x) = \texttt{imposter}$. Minimizing $\mathcal{L}_3$ makes the scores of all the enrolled speakers be less than the threshold θ, thus the adversarial voice is rejected, indicating a successful attack.

C4: Untargeted attack ($s \in G$). Given a benign voice uttered by a source speaker s such that $s \in G$, the untargeted attack may craft an adversarial voice such that: 1) it is rejected by the OSI task. This case is equivelent to "Targeted attack ($t \notin G$)", hence is omitted here; or 2) it is recognized as another enrolled speaker $t \in G$. We define the following five loss functions:

$$\mathcal{L}^s_{\text{CE}}(x,t) \triangleq \log[\sigma(S(x))]_s \qquad \mathcal{L}^s_1(x,t) \triangleq [S(x)]_s$$
$$\mathcal{L}^s_{\text{M}}(x,t) \triangleq [S(x)]_s - \texttt{max}_{i\in G, i\neq s}[S(x)]_i$$
$$\mathcal{L}^s_2(x,t) \triangleq \texttt{max}\{\theta, [S(x)]_s\} - \texttt{max}_{i\in G, i\neq s}[S(x)]_i$$
$$\mathcal{L}^s_4(x,t) \triangleq -\texttt{max}_{i\in G, i\neq s}[S(x)]_i$$

$\mathcal{L}^s_{\text{M}}$ is the negation of $\mathcal{L}_{\text{M}}$ in which the target speaker t is replaced by the source speaker s, intended to reduce the score of the source speaker s while increase the scores of other enrolled speakers (Remark that $\mathcal{L}_{\text{M}}$ is designed to increase the score of the target speaker t while reduce the scores of all the other enrolled speakers). $\mathcal{L}^s_1$ and $\mathcal{L}^s_4$ are the two terms of $\mathcal{L}^s_{\text{M}}$ used to check their effectiveness in $\mathcal{L}^s_{\text{M}}$. $\mathcal{L}^s_2$ is designed such that $\mathcal{L}^s_2(x,t) \leq 0 \Leftrightarrow D(x) \in G \setminus \{s\}$. When minimizing $\mathcal{L}^s_2$, we intend to find an enrolled speaker s' ($s' \neq s$) whose score is the largest one and exceeds the threshold θ, hence an adversarial voice is recognized as the speaker s', indicating a successful untargeted attack.

C5: Untargeted attack ($s \notin G$). Given a voice uttered by a source speaker s such that s is not an enrolled speaker (i.e., $s \notin G$), the adversary attempts to craft an adversarial voice such that it is accepted as an arbitrary enrolled speaker $t \in G$. The loss function to achieve this goal is defined as: $\mathcal{L}^-_3(x,t) \triangleq -\mathcal{L}_3(x,t) = \theta - \texttt{max}_{i\in G}[S(x)]_i$.

3.3.2 Loss functions for CSI task

C6 and C7: Targeted attack. The CSI task recognizes any input voice as one of the enrolled speakers, i.e., the decision space is G. Therefore, when launching targeted attack against SRSs performing the CSI task, an adversary can choose any enrolled speaker as the target speaker $t \in G$. The loss function can be derived from the ones defined for targeted attack with $t \in G$ on the OSI task (i.e., C1 and C2) by ignoring the threshold θ, i.e, the loss functions $\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{M}}$, and $\mathcal{L}_1$, no matter how the source speaker s is chosen, either one of the enrolled speakers (i.e., $s \in G$) or an unenrolled one (i.e., $s \notin G$).

C8: Untargeted attack. The loss functions can be derived from the ones defined for untargeted attack on the OSI task with $s \in G$ (i.e., C4) by ignoring the threshold θ, namely, the loss functions $\mathcal{L}^s_{\text{CE}}$, $\mathcal{L}^s_{\text{M}}$, $\mathcal{L}^s_1$, and $\mathcal{L}^s_4$.

3.3.3 Loss functions for SV task

The SV task involves only one enrolled speaker and determines if an input voice is uttered by the enrolled speaker or not. Hence, an adversary may potentially aim to achieve two opposite goals: 1) a voice uttered by the enrolled speaker is rejected as an imposter; 2) a voice uttered by an unenrolled speaker is recognized as the enrolled speaker.

C9: Enrolled speaker $\rightarrow$ imposter. Two loss functions which can be used to achieve this goal are formulated as:

$$\mathcal{L}_{\text{BCE}}(x,t) \triangleq -\log(1-\varphi(S(x))) \qquad \mathcal{L}_{3B}(x,t) \triangleq S(x) - \theta$$

where φ denotes the sigmoid function and $\mathcal{L}_{\text{BCE}}$ is the binary Cross Entropy Loss function. Note that $\mathcal{L}_{3B}$ is adapted from $\mathcal{L}_3$ by assuming the speaker group G is singleton. $\mathcal{L}_{\text{BCE}}$ (resp. $\mathcal{L}_{3B}$) is the special case of $\mathcal{L}_{\text{CE}}$ (resp. $\mathcal{L}_3$) for the binary classification task SV.

C10: Unenrolled speaker $\rightarrow$ enrolled speaker. Two loss functions for this goal can be derived from the above loss functions with minor modifications:

$$\mathcal{L}'_{\text{BCE}}(x,t) \triangleq -\log(\varphi(S(x))) \qquad \mathcal{L}^-_{3B}(x,t) \quad \triangleq \theta - S(x)$$

Remark that $\mathcal{L}^-_{3B}(x,t)$ is the special case of $\mathcal{L}^-_3(x,t)$ for the binary classification task SV.

3.4 Optimization Approachs

After formulating adversarial attack as an optimization problem with various source-/target-oriented loss functions, it remains to solve the optimization problem. In general, one can freely leverage any existing approaches to solve the optimization problem and craft adversarial examples. In this work, to increase the attack capability of AS2T and make it applicable to different levels of knowledge about the victim model, we consider the following solving approaches: FGSM [37], PGD [38], and CW [35] that were developed in the image domain, and FAKEBOB [16] that was tailored for SRSs. FGSM, PGD, and FAKEBOB use the standard gradient descent to solve the optimization problem, while CW utilizes Adam optimizer [39]. FGSM, PGD, and CW are representative white-box approaches, while FAKEBOB is the state-of-the-art black-box one. Below we integrate them into AS2T with necessary modifications. Since the ways that CW copes with the box-constraint and stealthiness differ from that of FGSM, PGD, and FAKEBOB, we first discuss how to adapt CW and then the others.

Adapting CW. CW deals with the box constraint $x' \in [-1,1]$ by introducing a new free variable z and optimizing over this new variable z instead of x'. Inspired by this trick, we define the variable $z = \text{arctanh}(x')$. Since $-1 \leq x' \leq 1$, we have $-\infty \leq z \leq \infty$, thus the box-constraint is removed

when we optimize over z. Note that our new free variable z is different from the one used in CW, due to the range difference between voices and images, i.e., $[-1, 1]$ vs. $[0, 1]$.

To achieve the stealthiness, CW minimizes the adversarial perturbation $\delta = x' - x$ by finding an optimal trade-off hyper-parameter λ using a binary search. Intuitively, a large λ will instruct the optimization to focus more on reducing the distance $d(x', x)$, leading to better stealthiness, but meanwhile pay less attention to the loss $\mathcal{L}$, undermining the effectiveness of the attack. We follow this practice and adopt L_2 norm [35] as the distance metric, i.e., $d(x', x) \triangleq \sqrt{\sum_i (x'_i - x_i)^2}$.

Since the objective function to be minimized is the combination of two terms, to focus only on reducing the perturbation after the attack succeeds (i.e., the loss $\mathcal{L}$ is small enough), CW adds a clamping operation to the loss function, i.e., $\texttt{max}(\mathcal{L} + \kappa, 0)$ where κ controls the strength of the adversarial examples. Due to the clamping operation, only part of our defined loss functions are suitable for CW approach, including $\mathcal{L}_2$, $\mathcal{L}_3$, $\mathcal{L}_2^s$, $\mathcal{L}_3^-$, $\mathcal{L}_{\text{M}}$ (on CSI), $\mathcal{L}_{\text{M}}^s$ (on CSI), $\mathcal{L}_{3B}$, and $\mathcal{L}_{3B}^-$. Other loss functions are excluded since they do not satisfy $\mathcal{L}(x, t) \leq 0 \Rightarrow D(x') = t$ (targeted attack) or $\mathcal{L}(x, t) \leq 0 \Rightarrow D(x') \neq s$ (untargeted attack).

Adapting FGSM, PGD, and FAKEBOB. Unlike the CW approach (i.e., $\lambda \neq 0$), FGSM, PGD, and FAKEBOB set $\lambda = 0$ and cope with the stealthiness and box-constraint by clipping the intermediate example at each iteration of gradient descent. Following this practice, we define the clipping function as

$$clip_{x,\varepsilon}(x') \triangleq \texttt{min}(x + \varepsilon, 1, \texttt{max}(x', x - \varepsilon, -1))$$

where ε is the upper bound of the perturbation magnitude measured in L_∞ norm. Note that our clipping function is different from the one exploited by FGSM and PGD, due to the range difference between images and voices.

FGSM is an one-iteration approach with the perturbation budget ε as the step_size, while PGD is an iterative version of FGSM with a small step size α. FAKEBOB is similar to PGD except that it is a black-box approach and estimates gradients via Natural Evolution Strategy [40]. Furthermore, FAKEBOB proposes the first algorithm to estimate the score threshold θ for the SV and OSI tasks.

Although FGSM, PGD, and CW were initially proposed as attack algorithms using loss functions $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{M}}$, we highlight that AS2T only utilizes them as optimization approaches since neither $\mathcal{L}_{\text{CE}}$ nor $\mathcal{L}_{\text{M}}$ fits to the purpose of AS2T: adversarial attack with arbitrary source/target speakers. We desgin suitable loss functions and integrate them into FGSM, PGD, and CW. This is the same for FAKEBOB, since our attack AS2T covers more cases of source and target speakers, i.e., C1, C3-C5, C7-C8, and C9. Even for the cases covered by FAKEBOB, AS2T provides more choices of loss functions. The common loss functions between our work and FAKEBOB are only $\mathcal{L}_2$, $\mathcal{L}_3^-$, $\mathcal{L}_{\text{M}}$ (for C6), $\mathcal{L}_{\text{M}}^s$ (for C8), and $\mathcal{L}_{3B}^-$. Furthermore, we empirically show that one of the loss functions adopted by FAKEBOB achieves inferior performance than ours (cf. Section 4.3).

3.5 Robust Over-the-Air AS2T

In practice, adversarial voices can be fed to SR models via different ways, including API and air channel. API attack directly feeds adversarial voices in the form of audio file to models via exposed API interfaces, thus it will not introduce any disruption to adversarial voices. In contrast, over-the-air attack, where adversarial voices are played/recorded by loudspeakers/microphones and transmitted over the air, is a lossy channel. The distortion occurred in the transmission could largely destruct the effectiveness of adversarial examples. Thus, an over-the-air attack is generally more practical and realistic yet more challenging than the API attack.

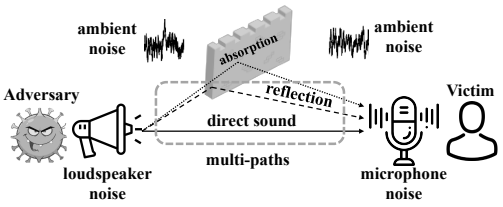


Fig. 3: The acoustic model of over-the-air attack

In this subsection, we first investigate the sources of distortions that may destruct the effectiveness of adversarial voices played over-the-air and then present our solution to enhance our attack AS2T towards robust over-the-air attack.

3.5.1 Sources of Distortions

Fig. 3 depicts the acoustic model for over-the-air attack. When an adversarial voice is played by a loudspeaker, it is transmitted over the air, and finally recorded by the victim's microphone. There are three sources of distortions: equipment distortion, ambient noise, and reverberation.

Equipment distortion. Equipment distortion is introduced by loudspeakers and microphones due to their frequency-selectivity feature. Specifically, their frequency response is non-uniform across the frequency band with amplification in some frequency ranges and attenuation in others, which may distort the adversarial voices and undermine the attack. In addition, different loudspeakers and microphones may exhibit different frequency responses, thus incur different equipment distortions. We can model the equipment distortion on adversarial voices in the time domain via loudspeakers and microphones' impulse response h. The adversarial voice disrupted by equipment distortion is formulated as $x^{adv} \otimes h$ where $\otimes$ denotes convolution operation.

Ambient noise. In practice, ambient noise includes ambient human voice, background music, traffic noise, and so on, depending on the specific attack scenario (e.g., living room, office, airport, and mall), but is inevitable in playback and recording environment. These noises, denoted by n, influence the effectiveness of adversarial voices by additively changing their magnitude, i.e., from x^{adv} to $x^{adv} + n$. However, the influence varies with the relative strength of ambient noise to adversarial voices, i.e., Signal-to-Noise Ratio (SNR) between adversarial voices and ambient noise. Ambient noise with low volume has limited impact since weak noises are easily overwhelmed by stronger voices [41].

Reverberation. When played by loudspeakers in indoor environment, a voice signal may transmit through multiple paths (i.e., direct path and other reflected paths) with various delays and absorption by many surfaces. When the direct sound and reflections blend and overlap with each other, reverberation is created. Reverberation will cause the

Algorithm 1 Robust AS2T

Input: Benign voice x^0; loss function $\mathcal{L}$; distance metric $d(\cdot)$; hyper-parameter λ; optimization approach $\mathcal{O}$; number of iterations #Iter; set of parameterized functions $\mathcal{F}$; sampling size K
Output: Adversarial voice x^{adv}

1: **for** $i = 1$ **to** #Iter **do**
2: $\quad L_i = 0$
3: $\quad$ **for** $j = 1$ **to** K **do**
4: $\quad\quad$ sampling a parameterized function $\mathcal{F}_p$ from $\mathcal{F}$
5: $\quad\quad L_i^j \leftarrow \mathcal{L}(\mathcal{F}_p(x^{i-1}))$
6: $\quad\quad L_i \leftarrow L_i + L_i^j$
7: $\quad L_i \leftarrow L_i/K + \lambda \times d(x^{i-1}, x^0)$
8: $\quad g_i \leftarrow \nabla_{x^{i-1}} L_i$
9: $\quad x^i \leftarrow \mathcal{O}(x^{i-1}, g_i)$
10: $x^{adv} \leftarrow x^{\#\text{Iter}}$
11: **return** x^{adv}

received voice by microphones to be largely different from the original voice sent out by loudspeakers. Room Impulse Response (RIR), represented by r, can well characterize the acoustic properties of a room regarding sound transmission and reflection. The adversarial voice with reverberation is created by convolving r with x^{adv}, i.e., $x^{adv} \otimes r$. RIR varies with the room configuration (e.g., room dimension, reverberation time, and absorption coefficient of reflective materials) as well as the location of loudspeakers and microphones. To obtain RIR for a specific room, two different methods are mostly used in practice: simulation approach and real-world measurement. Simulation approach leverages the well-known Image Source Method [42] which accepts room configuration and hardware's position as input and returns the simulated RIR. For real-world measurement, we can transmit a brief input signal (called impulse) by a loudspeaker in a room, then the response signal recorded by a microphone is the RIR of this room under the current position of loudspeaker and microphone. However, as the impulse signal also goes through the hardware during transmission, the RIR obtained by real-world measurement is indeed the composition of impulse responses of the room and hardware [41], [43], i.e., $r \circ h$.

3.5.2 *Robust* AS2T

To enhance the robustness of adversarial voices and enable physical over-the-air attack, we incorporate the aforementioned distortions into the generation process of adversarial voices. Our solution is described in Algorithm 1, based on a set $\mathcal{F}$ of parameterized functions modeling the distortions induced by the over-the-air transmission. In the i-th iteration, we randomly sample K functions from $\mathcal{F}$ (Lines 3-4). Each sampled function $\mathcal{F}_p$ is utilized to transform the intermediate voice x^{i-1} (Line 5), resulting in K transformed voices. Then we compute the average loss of those K voices based on which the gradient g_i is computed (Lines 7-8). Finally, the new voice x^i is created from x^{i-1} and g_i by invoking the optimization approach $\mathcal{O}$ (Line 9). In this work, we utilize the following parameterized functions $\mathcal{F}$.

Addition with random noise. We use random noise n (e.g., uniform noise and white Gaussian noise) to model ambient noise in the physical attack. Formally, we define the functions $\mathcal{F} = \{\mathcal{F}_{\text{SNR},n} | \mathcal{F}_{\text{SNR},n}(x) = x + \Phi(\text{SNR}, x, n), n \sim \mathcal{Z}, \text{SNR}_l \leq \text{SNR} \leq \text{SNR}_u\}$, where $\mathcal{Z}$ denotes the distribution of the random noise, SNR is the Signal-to-Noise Ratio between the adversarial voice x and the random noise n (i.e., $\text{SNR} = 10 \times \log_{10} \frac{P_x}{P_n}$, P_x and P_n are powers of x and n respectively), SNR_l and SNR_u are the lower and upper bounds of SNR, and $\Phi(\text{SNR}, x, n)$ scales the magnitude of the random noise n such that the SNR requirement is satisfied. We denote by AS2T+RN our attack AS2T with random noise.

TABLE 4: Details of the 14 SR models, where Arch and Trans represent architecture and Transformer, respectively

Arch	Name	#Params	Input type	Training dataset	Scoring Backend
GMM	**Ivector** [44]	80.37M	MFCC	VoxCeleb1&2	PLDA
TDNN	**ECAPA** [45]	20.77M	fBank	VoxCeleb1	COSS
	Xvector-P [46]	5.79M	MFCC	VoxCeleb1&2	PLDA
	Xvector-C [2]	4.21M	fBank	VoxCeleb1	COSS
CNN	**AudioNet** [47]	0.21M	fBank	LibriSpeech	COSS
	SincNet [34]	21.84M	wavform	TIMIT	COSS
	Res18-I [48]	11.17M	spectrogram	VoxCeleb1	COSS
	Res18-V [48]	11.17M	spectrogram	VoxCeleb1	COSS
	Res34-I [49]	21.28M	spectrogram	VoxCeleb1	COSS
	Res34-V [49]	21.28M	spectrogram	VoxCeleb1	COSS
	Auto-I [50]	15.11M	spectrogram	VoxCeleb1	COSS
	Auto-V [50]	15.11M	spectrogram	VoxCeleb1	COSS
LSTM	**GE2E** [51]	12.13M	fBank	TIMIT	COSS
Trans	**Hubert** [52]	316.61M	wavform	LibriLight	COSS

Convolution with room impulse response. To improve the robustness of adversarial voices against reverberation in the physical attack, we define the functions $\mathcal{F} = \{\mathcal{F}_r | \mathcal{F}_r(x) = r \otimes x, r \sim \mathcal{R}\}$, where $\mathcal{R}$ denotes the distribution of the RIR. We denote by AS2T+RIR our attack AS2T with RIR. Furthermore, we denote by AS2T+RN+RIR our attack with both random noise and RIR, in which $\mathcal{F} = \{\mathcal{F}_{r,\text{SNR},n} | \mathcal{F}_{r,\text{SNR},n}(x) = r \otimes x + \Phi(\text{SNR}, x, n)\}$.

Remark that we do not explicitly model the equipment distortion in this work. Our experimental results show that the effect of such distortion is not substantial and can be coped with by AS2T+RN, AS2T+RIR, and AS2T+RN+RIR. We highlight that in Algorithm 1 we randomly sample K parameterized functions from $\mathcal{F}$ instead of using a single function across all the iterations. In this way, it is expected that the obtained adversarial voices can work in various scenarios and attack numerous victim devices simultaneously.

4 Evaluation

In this section, we first present the common evaluation setup and metrics of our evaluation, then evaluate effectiveness of AS2T with arbitrary source/target speakers over API and over-the-air, and finally conduct a thorough transferability study of AS2T among 14 SR models (cf. TABLE 4).

4.1 Common Evaluation Setup

Datasets. Our evaluation is based on three datasets derived from Librispeech [53] and released in [36]. namely, Spk_{10}-enroll, Spk_{10}-test, and Spk_{10}-imposter. Spk_{10}-enroll and Spk_{10}-test consist of 10 and 100 distinct voices per speaker from the same ten speakers (five male and five female), while Spk_{10}-imposter consists of 100 voices per speaker from another ten speakers (five male and five female). We also build another three datasets from LibriSpeech with more speakers, where Spk_{100}-enroll and Spk_{100}-test respectively consist of 10 and 100 distinct voices per speaker from

TABLE 5: Performance of the SR models on three tasks

model	CSI	SV		OSI		
	Acc (%)	EER (%)	θ	EER (%)	IER (%)	θ
Ivector	100	1.05	9.74	4.8	0	12.77
ECAPA	99.9	0.97	0.42	2.2	0	0.49
Xvector-P	100	0.8	13.78	6	0	18.72
Xvector-C	96.4	4.22	0.63	9.85	0	0.7
AudioNet	99.9	4.2	0.82	14.01	0	0.87
SincNet	96.3	5	0.59	16.9	0	0.74
Res18-I	100	0.84	0.51	5.5	0	0.61
Res18-V	99.9	1.3	0.49	6.61	0	0.59
Res34-I	100	0.9	0.52	6.7	0	0.61
Res34-V	99.8	1.2	0.51	4.3	0	0.58
Auto-I	58.8	19.95	0.23	17.01	0	0.58
Auto-V	99.3	1.6	0.29	3.6	0	0.39
GE2E	78.7	10.65	0.67	23.2	0	0.87
Hubert	95.2	9.13	0.57	17.23	0	0.62

the same 100 speakers (52 female and 48 male) and Spk_{100}-imposter consists of 100 voices per speaker from another 100 speakers (49 female and 51 female).

Models. In this work, we select 14 SR models for our experiments. These models cover five architectures (GMM, TDNN, CNN, LSTM, and Transformer), four input types (waveform, spectrogram, fBank, and MFCC), and two scoring methods (PLDA and COSS). More details of these models are shown in TABLE 4, such as their training dataset, the number of parameters, and etc.

We measure the performance of these models on three tasks using the above datasets and the results are shown in TABLE 5, where the best cases are highlighted in blue color. Column (Acc) shows accuracy. Column (EER) shows the equal error rate, i.e., when False Acceptance Rate (FAR) equals False Rejection Rate (FRR), where FAR is the proportion of voices that are uttered by unenrolled speakers but accepted by the model, and FRR is the proportion of voices that are uttered by enrolled speakers but rejected (i.e., classified as `imposter`). Column (IER) shows Identification Error Rate, i.e., the rate of voices uttered by enrolled speakers which are not rejected but incorrectly classified [16]. We tune the threshold θ for the SV and OSI tasks based on EER.

Attacks. We implement four optimization approaches for AS2T, namely, FGSM, PGD, CW_2, and FAKEBOB (cf. Section 3.4). We limit the perturbation budget ϵ to 0.002 for L_∞ attacks, the same as [10], [16], unless explicitly stated. Note that the CW_2 attack minimizes adversarial perturbations in the loss function, and hence does not have any limitations.

We conduct experiments on a machine with Ubuntu 18.04, an Intel Xeon E5-2697 v2 2.70GHz CPU, 376GiB memory, and a GeForce RTX 2080Ti GPU.

4.2 Evaluation Metrics

We mainly use the following evaluation metrics.

Effectiveness. To evaluate the effectiveness of an attack, we adopt untargeted (resp. targeted) attack success rate ASR_u (resp. ASR_t), which refers to the proportion of adversarial voices that are misclassified (resp. classified as the target result). Formally, ASR_u=100%-Acc for the CSI task. For the SV/OSI task, ASR_u=FAR when the benign voices are uttered by unenrolled speakers and ASR_u=FRR when the benign voices are uttered by enrolled speakers.

Stealthiness. To measure the stealthiness of adversarial voices, we use the standard L_2 norm, SNR (cf. Section 3.5.2), and Perceptual Evaluation of Speech Quality (PESQ) [54]. PESQ is one of the objective perceptual measures, which simulates the human auditory system [55]. The calculation of PESQ is involved. Intuitively, PESQ first applies an auditory transform to obtain the loudness spectra of the original and adversarial voices, and then compares these two loudness spectra to obtain a metric score whose value is in the range of -0.5 to 4.5. We refer readers to [54] for more details. Smaller L_2, larger SNR, and higher PESQ indicate better stealthiness.

4.3 Evaluation of AS2T over API

In this section, we evaluate the effectiveness of AS2T with arbitrary source/target speakers over API, i.e., directly feeding adversarial voices in the form of audio file to models.

4.3.1 Evaluation Setup

We target the ECAPA model in TABLE 4 since this model achieves the best overall performance over the three tasks. We enroll ten speakers using the dataset Spk_{10}-enroll, forming a speaker group G for the CSI and OSI tasks and ten speaker models for the SV task. The thresholds θ for the SV and OSI tasks are set to the ones shown in TABLE 5.

We mount attacks using two categories of benign voices, namely, voices uttered by enrolled speakers in the dataset Spk_{10}-test and voices uttered by unenrolled speakers in the dataset Spk_{10}-imposter. For target speakers, since the decision space of the OSI task is $\mathcal{D} = G \cup \{\texttt{imposter}\}$, we consider two categories of target speakers, i.e., the enrolled speakers in G and `imposter`. In contrast, since the decision space of the CSI task is G, the target speakers are limited to the enrolled speakers. We adopt two different ways to set a target speaker, i.e., randomly choosing among the enrolled speakers except the source speaker (Targeted Random) and choosing the enrolled speaker whose score is the least one (Targeted Least Likely). Note that we use the same target speakers for the CSI and OSI tasks. Since the SV task makes binary decision, its target speaker is `imposter` (resp. the enrolled speaker) when the source speaker is the enrolled speaker (resp. unenrolled speaker).

We implemented the loss functions defined in Section 3.3 and the approaches FGSM, PGD, CW_2, and FAKEBOB. The number of iterations for PGD is 2, 3, 4, and 5 with step size $\alpha = \frac{\varepsilon}{5} = 0.0004$. The step size of FGSM is $\varepsilon = 0.002$. For CW_2, we set the initial trade-off constant $\lambda = 0.1$, use 9 binary search steps to minimize adversarial perturbations, run 900-9000 iterations to converge, and set the parameter $\kappa = 0$. For FAKEBOB, the maximum number of iterations is set to 1000 with samples_per_draw of NES $m = 50,\ 100$.

4.3.2 Results

The results of AS2T with FGSM and PGD as the optimization approaches on the OSI task and the CSI/SV tasks are showed in TABLE 6 and TABLE 7, respectively.

Loss function comparison. In general, with the increase of the number of iterations, the attack success rate of all the loss functions approaches 100%. However, some loss functions are more effective and efficient, i.e., achieve higher attack

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2022.3189397

TABLE 6: The attack success rate (%) of AS2T on the OSI task with FGSM and PGD as the optimization approaches

			OSI																									
			Targeted ($t \in G$) Random								Targeted ($t \in G$) Least Likely								Targeted ($t =$ imposter)						Untargeted			
			ASR_t (%)				ASR_u (%)				ASR_t (%)				ASR_u (%)				ASR_t (%)			ASR_u (%)			ASR_u (%)			
			$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}^s_{CE}$	$\mathcal{L}_3$	$\mathcal{L}^s_1$	$\mathcal{L}^s_{CE}$	$\mathcal{L}_3$	$\mathcal{L}^s_1$	$\mathcal{L}^s_{CE}$	$\mathcal{L}^s_M$	$\mathcal{L}^s_4$	$\mathcal{L}^s_2$	$\mathcal{L}^s_1$	$\mathcal{L}_3^-$
$s \in G$	FGSM		9.7	9.2	9.4	9.4	30.6	62.8	22.3	62.8	0.1	0	0.3	0	27.4	61.8	14.8	61.8	80.1	83.9	83.9	80.1	83.9	83.9	0	41.2	37.6	41.6	0	N/A
	PGD	2	65.7	49.6	67.9	62.1	77	92.6	71	91.3	46.6	21.5	54.2	39.8	74.2	92.7	64	91.5	96.8	98.2	98.2	97	98.2	98.2	0.2	82.4	81.4	84.6	0	
		3	92.1	75.5	92.7	91.8	95.7	99.4	93.4	98.5	85.8	53.8	88.3	85.1	93	99.5	89.4	98.1	100	100	100	100	100	100	0	95.5	97.2	98.7	0	
		4	99.1	90.7	98.9	98.8	99.8	100	99	99.9	97.1	82.1	97.6	97	98.8	100	97.9	100	100	100	100	100	100	100	0	98.6	99.8	100	0	
		5	99.8	96	100	99.8	100	100	100	100	99.4	92	99.8	99.7	100	100	99.8	100	99.9	100	100	100	100	100	0.1	98.8	100	100	0	
$s \notin G$	FGSM		26	10.8	34.3	33.9	26.1	10.8	35.4	34.9	0.4	0.1	2.2	2.1	0.4	0.1	2.9	2.4	N/A			N/A			N/A				95	
	PGD	2	86.5	43.8	93.4	92.9	86.6	43.8	93.8	93.2	64.7	6.9	79.2	78.4	64.7	6.9	80.8	79.5											100	
		3	99.1	80.9	99.7	99.6	99.1	80.9	99.7	99.6	98.2	52.3	99.3	98.9	98.2	52.3	99.3	98.9											100	
		4	99.7	96.5	99.9	99.9	99.7	96.5	99.9	99.9	99.6	87.7	99.9	99.9	99.6	87.7	99.9	99.9											100	
		5	99.9	98.9	99.9	99.9	99.9	98.9	99.9	99.9	100	97.6	99.9	99.9	100	97.6	99.9	99.9											100	

TABLE 7: The attack success rate (%) of AS2T on the CSI and SV tasks with FGSM and PGD as the optimization approaches

			CSI													SV	
			Targeted Random						Targeted Least Likely						Untargeted	Targeted	
			ASR_t (%)			ASR_u (%)			ASR_t (%)			ASR_u (%)			ASR_u (%)	ASR_t (%)	
			$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}_{CE}$	$\mathcal{L}_M$	$\mathcal{L}_1$	$\mathcal{L}^s_{CE}$	$\mathcal{L}_{BCE}/\mathcal{L}'_{BCE}$	$\mathcal{L}_{3B}/\mathcal{L}^-_{3B}$
$s \in G$	FGSM		13.6	23.3	11.7	13.6	25.7	11.8	1.3	4.9	0.7	1.6	9.9	1.1	28.2	68.7	68.7
	PGD	2	73.6	81.9	70	73.6	82.5	70	64.4	75.2	60.1	64.4	77.4	60.1	82.5	95.2	95.2
		3	95	97.8	93.2	95	97.8	93.2	90.4	94.8	88.9	90.4	95.1	88.9	96.2	100	100
		4	99.7	99.9	99	99.7	99.9	99	98.5	99.8	97.7	98.5	99.8	97.7	99.6	100	100
		5	100	100	100	100	100	100	99.9	99.9	99.8	99.9	99.9	99.8	100	100	100
$s \notin G$	FGSM		81.6	70.7	76	100	100	100	57	36.6	47.7	100	100	100	N/A	56.9	56.9
	PGD	2	99.3	98.3	98.4	100	100	100	97.5	92.3	94.9	100	100	100		99.1	99.1
		3	100	99.9	100	100	100	100	100	99.9	100	100	100	100		100	100
		4	100	100	100	100	100	100	100	100	100	100	100	100		100	100
		5	100	100	100	100	100	100	100	100	100	100	100	100		100	100

success rate within the same number of iterations and obtain 100% attack success rate with fewer iterations.

On the OSI task, for targeted attack ($t \in G$), the loss function $\mathcal{L}_1$ often achieves better ASR_t than $\mathcal{L}_{CE}$, $\mathcal{L}_M$ and $\mathcal{L}_2$ (Note that $\mathcal{L}_2$ was adopted by FAKEBOB in [16]). In contrast, the most effective loss functions for the untargeted attack and targeted attack with $t =$ imposter are $\mathcal{L}^s_2$ and $\mathcal{L}_3/\mathcal{L}^s_1$, respectively. Interestingly, $\mathcal{L}^s_{CE}$ and $\mathcal{L}^s_1$ are rather effective for launching targeted attack ($t =$ imposter), but achieves extremely limited attack success rate for untargeted attack.

On the CSI task, when the source speaker is an enrolled speaker ($s \in G$), the loss function $\mathcal{L}_M/\mathcal{L}^s_M$ performs better than the others in terms of both ASR_t and ASR_u for targeted/untargeted attacks. A similar result has been reported in the image domain [35]. In contrast, when the source speaker is an unenrolled speaker ($s \notin G$), the loss function $\mathcal{L}^s_{CE}$ outperforms the others. On the SV task, we find that the loss functions have the same performance. This is because FGSM and PGD optimize them on the signs of the gradients instead of the actual gradients, and the signs are the same.

Remark 1. The effectiveness of loss functions varies with setting (e.g., task, source/target speakers). On the OSI task, the best loss functions for untargeted, targeted ($t \in G$), and targeted ($t =$ imposter) are $\mathcal{L}^s_2$, $\mathcal{L}_1$, and $\mathcal{L}_3$, respectively. On the CSI task, the best loss functions for untargeted/targeted attack are $\mathcal{L}_M/\mathcal{L}^s_M$ for $s \in G$ and $\mathcal{L}^s_{CE}$ for $s \notin G$.

Different source/target speakers. We observe that the targeted attack with $t =$ imposter (i.e., C3) often outperforms the ones with $t \in G$ (i.e., C1 and C2) on the OSI task. This is because the former only concentrates on the relative magnitude of the scores between the enrolled speakers and the threshold θ, while the latter additionally has to consider the relative magnitude of the scores between the source and target speakers. Another observation is that the attack with $s \notin G$ tends to be easier than the one with $s \in G$, e.g., C2 vs. C1, C5 vs. C4, and C7 vs. C6. An in-depth analysis reveals that the score of an enrolled speaker is much higher than the others if the benign voice is uttered by himself (i.e., $s \in G$), while scores of all the enrolled speakers are similar if the benign voice is uttered by an unenrolled speaker (i.e., $s \notin G$). Therefore, it is much easier to increase the score of the target speaker if the source speaker is an unenrolled one.

Remark 2. Targeted attack with $t =$ imposter is often easier than the targeted attack with $t \in G$ and attacks using unenrolled speaker as the source speaker (i.e., $s \notin G$) is often easier than the attacks with $s \in G$.

Different tasks. We notice that the same loss function often achieves lower targeted attack success rate (ASR_t) on the OSI task than on the CSI task. Recall that we specify the same target speakers for the two tasks. This indicates that targeted attack on the OSI task is more difficult than that on the CSI task. It is because the OSI task rejects an input if the scores are less than the threshold θ, thus a successful attack must simultaneously guarantee that the score of the target speaker is the maximal one and larger than the threshold θ.

Remark 3. The OSI task is more difficult to attack than the CSI task.

To be exhaustive, TABLE 8 reports the results of AS2T on the CSI task with $s \in G$ using the most effective loss functions $\mathcal{L}_M/\mathcal{L}^s_M$ and FAKEBOB and CW_2 approaches. The

TABLE 8: The attack success rate (%) and stealthiness of AS2T on the CSI task with CW_2 and FAKEBOB as optimization approaches

			CSI													
			Targeted Random					Targeted Least Likely					Untargeted			
			Success rate		Stealthiness			Success rate		Stealthiness			Success rate	Stealthiness		
			ASR_t (%)	ASR_u(%)	L2	SNR (dB)	PESQ	ASR_t (%)	ASR_u(%)	L2	SNR (dB)	PESQ	ASR_u (%)	L2	SNR (dB)	PESQ
			$\mathcal{L}_{\mathrm{M}}$					$\mathcal{L}_{\mathrm{M}}$					$\mathcal{L}_{\mathrm{M}}^s$			
$s \in G$	CW2		100	100	0.081	47.01	4.00	100	100	0.102	44.54	3.85	100	0.057	50.47	4.17
	FAKEBOB	m=50	86.5	86.5	0.357	31.14	2.74	78.3	79.5	0.340	31.21	2.76	93.5	0.382	30.92	2.70
		m=100	99.3	99.3	0.390	31.17	2.69	98.4	98.4	0.376	31.38	2.72	100	0.412	30.76	2.64

TABLE 9: The room configuration and hardware setting of over-the-air attack. RWCP, REVERB, and AIRD are different real RIR datasets.

		Room configuration						Hardware position		Hardware	
		Room type	Length (m)	Width (m)	Height (m)	Absorption coefficient	Reverberation time (s)	Distance (m)	Angle (°)	Loudspeaker	Microphone
Simulated RIR		small	1-10	1-10	2-5	0.2-0.8	-	1.06-11.33		-	-
		medium	10-30	10-30			-	1.96-35.07		-	-
		large	30-50	30-50			-	2.63-59.5		-	-
Real RIR	RWCP	variable reverberant room	6.66	4.18	-	-	0.3-1.3	2	10-170	1) Diatone DS-7 loud speaker 2) B&K Type 4128 Head-Torso	1) 54ch Spherical array 2) 14ch Linear array (2.83cm spacing) 3) 16ch Circle array
		anechoic room	-	-	-	-	0.01-2	-	-		
		office	-	-	-	-	0.01-2	-	-		
	REVERB	small	5.57	3.77	-	-	0.25	0.5, 2	45, 135	BOSE 101MM	SONY ECM-77B
		medium	6.27	4.89	2.59	-	0.5	0.5, 2	45, 135	Genelec 1029A	AKG capsules CE20 V17
		large	6.67	6.14	-	-	0.7	0.5, 2	45, 135	BOSE 101MM	SONY ECM-77B
		reverberant meeting room	-	-	-	-	0.7	1, 2.5	-	single stationary speaker	8-ch circular array with omni-directional microphones
	AIRD	studio booth	3	1.8	2.2	-	0.08-0.18	0.5, 1.0, 1.5	-	2-way active studio monitor Genelec 8130	two Beyerdynamic MM1 omnidirectional condenser measurement
		office room	5	6.4	2.9	-	0.37-0.48	1.0, 2.0, 3.0	-		
		meeting room	8	5	3.1	-	0.21-0.25	1.45, 1.7, 1.9, 2.25, 2.8	-		
		lecture room	10.8	10.9	3.15	-	0.70-0.83	4.0, 5.56, 7.1, 8.68, 10.2	-		

results show that FAKEBOB and CW_2 are also effective, although the black-box approach FAKEBOB performs slightly worse than the white-box approach CW_2 in terms of attack success rate and stealthiness. We remark that at the same perturbation budget and comparable attack success rate, the single-step approach FGSM produces the least stealthy adversarial voices in terms of L_2, SNR, and PESQ, while PGD is better than FAKEBOB but worse than CW_2.

We also evaluate AS2T on the datasets with more speakers, i.e., Spk_{100}-enroll, Spk_{100}-test, and Spk_{100}-imposter, from which we can draw the same conclusions as Remarks 1-3. Therefore, the results are omitted here and reported in our technical report [56].

4.4 Evaluation of AS2T Over-the-Air

4.4.1 Evaluation Setup

It is non-trivial to conduct a large-scale and thorough evaluation of adversarial attacks over-the-air in the physical world. Thus, we simulate over-the-air attacks, as done in the speech community [57], [58], [59], [60], using the untargeted attack of AS2T on the CSI task, where the source speakers are enrolled speakers (i.e., $s \in G$) and the loss function is $\mathcal{L}_{\mathrm{M}}^s$ as suggested in Remark 1.

To simulate reverberation, we use 2,000 randomly simulated RIR and all the 325 publicly available real-world RIR in [57], and convolve them with the adversarial voices. The simulated RIR is generated by Image Source Method [42], covering various room dimension and the position of devices. The real-world RIR is collected in different physical rooms with various loudspeakers and microphones, thus also reflects various equipment distortions. The room configurations, positions of loudspeakers and microphones (e.g., distance and angle), and brands of hardware devices are given in TABLE 9. To simulate the ambient noise, we use both the widely-spread white noise and three types of representative noise occurred in real-world scenarios provided by the MUSAN dataset [61]: point-source, musical, and speech babble noise. Point-source noise includes technical sounds (e.g., cellphone noises and dialtones) and non-technical sounds (e.g., hunder, car horns, and animal sounds). Musical noise consists of several music genres, e.g., Country, Hip-Hop, and Jazz. Speech babble noise is introduced when multiple speakers utter at the same time and part or even the whole speech is mixed with others. Different types of ambient noise simulate different environments where the hardware is placed. For ambient noise, we set the SNR between adversarial voices and the noise to 0, 5, 10, 15, and 20 to imitate different volume of loudspeakers and the environments with different levels of noise.

For AS2T+RN, we use additive white Gaussian noise, i.e., $\mathcal{Z} = \mathcal{N}(0,1)$ where $\mathcal{N}(0,1)$ is standard normal distribution, and set $\mathrm{SNR}_l = 0$ and $\mathrm{SNR}_h = 20$. For AS2T+RIR, we approximate the distribution of RIR $\mathcal{R}$ using 1,000 simulated RIR, and the other disjoint 1,000 simulated RIR and 325 real-world RIR are used for testing. The sampling size K in Algorithm 1 is set to 10. We exploit PGD as the optimization approach except that the standard gradient descent (SGD) is replaced with Adam [39], because

TABLE 10: The attack success rate (%) of over-the-air AS2T, where RN and RIR denote random noise and reverberation impulse response, respectively

ε	Evaluation / Training	Reverberation		Ambient Noise (SNR=?dB)																				
				White Noise					Point Source Noise					Music Noise					Speech Babble Noise				
		Sim.	Real	20	15	10	5	0	20	15	10	5	0	20	15	10	5	0	20	15	10	5	0
0.008	AS2T	85.3	80.0	69.7	31.0	7.5	1.4	1.3	95.7	87.8	75.7	60.7	47.5	98.8	94.5	85.4	71.7	59.6	99.8	97.3	90.9	83.5	73.4
	AS2T+RN	100.0	99.8	100.0	100.0	99.9	96.0	89.6	100.0	100.0	99.9	99.6	97.9	100.0	100.0	100.0	100.0	99.4	100.0	100.0	100.0	100.0	99.7
	AS2T+RIR	100.0	99.9	100.0	95.1	73.4	38.6	16.7	100.0	99.7	98.0	93.7	86.3	100.0	100.0	99.5	98.5	95.6	100.0	100.0	100.0	99.9	99.0
	AS2T+RN+RIR	100.0	100.0	100.0	100.0	99.8	95.5	89.4	100.0	100.0	99.8	99.2	98.0	100.0	100.0	100.0	100.0	99.7	100.0	100.0	100.0	100.0	99.9
0.01	AS2T	88.5	83.1	78.1	40.8	12.3	1.8	1.9	97.2	90.7	79.8	66.6	51.5	99.6	96.1	89.1	77.8	64.1	99.9	98.3	93.1	87.2	76.8
	AS2T+RN	100.0	99.9	100.0	100.0	100.0	98.7	92.6	100.0	100.0	100.0	99.8	98.9	100.0	100.0	100.0	100.0	99.6	100.0	100.0	100.0	100.0	99.9
	AS2T+RIR	100.0	100.0	100.0	98.5	84.9	47.0	23.3	99.9	99.7	98.6	95.8	87.9	100.0	100.0	99.9	98.7	97.0	100.0	100.0	100.0	99.9	99.3
	AS2T+RN+RIR	100.0	100.0	100.0	100.0	100.0	98.1	92.3	100.0	100.0	100.0	99.9	99.2	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
0.02	AS2T	96.1	89.9	94.4	78.6	39.0	11.3	2.3	99.3	96.8	90.2	80.0	67.4	99.8	99.4	96.7	89.2	81.4	100.0	100.0	98.7	94.9	89.7
	AS2T+RN	100.0	100.0	100.0	100.0	100.0	100.0	99.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	AS2T+RIR	100.0	100.0	100.0	100.0	99.3	83.3	42.9	100.0	100.0	99.8	98.6	94.8	100.0	100.0	100.0	99.9	99.5	100.0	100.0	100.0	100.0	99.8
	AS2T+RN+RIR	100.0	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	100.0	99.9	99.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
0.03	AS2T	97.4	95.1	99.2	91.1	63.5	27.3	6.8	99.8	98.4	94.1	87.7	77.4	100.0	99.8	98.2	95.2	88.9	100.0	100.0	99.8	96.8	92.8
	AS2T+RN	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	AS2T+RIR	100.0	100.0	100.0	100.0	100.0	96.3	64.8	100.0	100.0	100.0	99.0	97.3	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0
	AS2T+RN+RIR	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Adam is more efficient for crafting robust adversarial voices against over-the-air distortions. We set the perturbation budget $\varepsilon = 0.008, 0.01, 0.02, 0.03$, the number of iterations #Iter=400 and step_size $\alpha = \frac{5\times\varepsilon}{\#\text{Iter}}$. Note that for each crafted adversarial voice, we only simulate the over-the-air transmission *one* time.

4.4.2 Results

The results are shown in TABLE 10.

Different sources of distortions. We observe that the success rate of AS2T increases with the budget ε, indicating a trade-off between the robust and stealthiness of adversarial voices. The success rate of AS2T is positively correlated with the SNR between adversarial voices and ambient noise. This is not surprising, as the adversarial voices with larger magnitude can overwhelm the weaker ambient noise.

We notice that the attack under the real-world RIR is slightly less effective than that under simulated RIR. This is because the real-world RIR contains additional distortion from the loudspeakers and microphones. Nevertheless, the difference is minor, indicating that the equipment distortion is not substantial compared to the other distortions.

AS2T vs. AS2T+X. With the budget $\varepsilon = 0.008$, AS2T+RIR achieves 19% higher success rate than AS2T under the real-world reverberation. This indicates that incorporating the simulated RIR into AS2T improves the robustness of adversarial voices against real-world reverberation. Similarly, modeling white-noise in AS2T enhances the robustness of adversarial voices against white, point-source, musical, and speech babble noises, with at least 26% attack success rate improvement when $\varepsilon = 0.008$ and SNR=0 dB. Unsurprisingly, the combination of AS2T+RIR and AS2T+RN, i.e., AS2T+RN+RIR, improves the practicability of AS2T under both real-world reverberation and various ambient noises.

Remark 4. The impact of equipment distortion is minor compared to that of reverberation and ambient noise. By incorporating the simulated reverberation and white noise into the generation of adversarial voices, AS2T can be improved towards robust over-ther-air attack against real-world reverberation and different types of ambient noises.

4.5 Transferability Analysis

In this section, we conduct a thorough transferability study of AS2T among 14 SR models (cf. TABLE 4) along two major axes: model-specific factors (e.g., model architecture, training dataset, and input types) and attack-specific factors (e.g., number of iterations, step_size, and perturbation budget).

4.5.1 Model-specific Factors

Evaluation setup. We mainly evaluate the transferability of AS2T with FGSM and PGD optimization approaches, where the perturbation budget ε is 0.002, step_size α of FGSM (resp. PGD) is ε (resp. $\frac{\varepsilon}{5} = 0.0004$). To avoid the bias introduced by attack-specific factors, we vary the number of iterations (i.e., #Iter) of PGD from 2 to 30 with step 1 and report the best one among FGSM and PGD of all the #Iter.

Results. TABLE 11 reports the results of AS2T with untargeted attack on the CSI task, where the source speakers are enrolled speakers, i.e., C8 in TABLE 3.

We find that Ivector and all the TDNN models transfer to each other quite well, especially for the transfer attacks *Xvector-C* $\rightarrow$ *Ivector* and *Xvector-C* $\rightarrow$ *Xvector-P*, with 51.5% and 58.6% accuracy drop, respectively.

Among CNN-based models, ResNet (i.e., Res18/34-I/V) and AutoSpeech (i.e., AutoI/V) transfer to each other quite well with at least 9.4% accuracy drop, but they are much less transferable to AudioNet and SincNet, and the vice versa (except for *AudioNet* $\rightarrow$ *Auto-I* and *SincNet* $\rightarrow$ *Auto-I*). This gap is possibly due to: 1) difference between training datasets: ResNet and AutoSpeech are trained on VoxCeleb1, while AudioNet and SincNet are on LibriSpeech and TIMIT, respectively; 2) difference between input types: the input type of ResNet and AutoSpeech is spectrogram while the input types of AudioNet and SincNet are fBank and raw waveform, respectively. Indeed, the most transferable model to AudioNet is GE2E which has the same input type.

We find that the transferability from CNN models to TDNN models is rather limited, indicating the adversarial examples often do not transfer well to the target models with different architectures from the source model.

Remarkably, the transferability between two models are not necessarily symmetric, even they have the same architecture. For instance, the accuracy drop of *SincNet* $\rightarrow$

TABLE 11: Transferability of AS2T with the FGSM and PGD optimization approaches. For each pair of source and target systems, the result indicates the accuracy drop of the target system. CSI task, untargeted attack, $s \in G$.

⇓ Accuracy (%) Source \ Target		Ivector	ECAPA	Xvector-P	Xvector-C	AudioNet	SincNet	Res18-I	Res18-V	Res34-I	Res34-V	Auto-I	Auto-V	GE2E	Hubert
GMM	Ivector	100.0	7.3	30.5	26.7	21.2	10.3	0.5	0.6	3.2	2.9	11	6.8	23	33.8
TDNN	ECAPA	8.6	99.9	14.3	11.4	10.9	8.6	1.1	1.0	7.1	1.4	12.5	12.2	14.5	28.8
	Xvector-P	44.4	22.5	100.0	34.4	18.3	9.4	1.3	1.2	4.2	2.5	12.8	8.2	21.1	34.9
	Xvector-C	51.5	42.7	58.6	96.4	13.5	7.7	1.9	2.1	7.0	7.5	15.8	14.2	18.9	25.6
CNN	AudioNet	4.3	0.1	0.5	-1.9	99.9	8.1	0.9	0.7	4	2.8	8.2	4	18.5	26.9
	SincNet	4.5	0.5	0.7	4.2	20.8	67.6	2.2	3.6	6.3	5.9	14.0	8.7	36.6	10.2
	Res18-I	1.4	0.1	0.4	-2.9	15.2	8.6	100.0	34.2	41.8	41.4	23.9	27.3	18.3	23.5
	Res18-V	2	0.2	0.3	-2.9	14.8	8.2	36.6	99.9	47.2	48.7	28.5	34.8	18.4	23.3
	Res34-I	1.2	0.1	0.3	-3.1	15.3	9.5	15.2	20	100.0	27.1	18.2	22.6	17.6	23.9
	Res34-V	1.6	0.2	0.2	-3.1	15.4	8	20	24.9	38.9	99.8	23.4	28.4	18.7	22.6
	Auto-I	2.6	0.1	0.4	-3.1	12.2	7.3	9.4	10.2	18	23.5	58.8	47.3	16.5	22.4
	Auto-V	2.6	0.6	1.1	-2.2	13.9	8.9	12.1	17.8	25.2	31.8	44.2	99.3	18.5	20.2
LSTM	GE2E	3.5	0.1	0.3	-2.7	25.2	4.3	1.1	3.2	2.7	5.2	12.1	9.8	68.7	28.2
Trans	Hubert	1.7	0.3	0.2	-3.2	3.9	4.3	0.1	-0.1	0.7	0.7	9.7	8.0	7.6	95.2

TABLE 12: The input gradient size of different SRS models. CSI task, untargeted attack, $s \in G$.

Model	IV	ECAPA	XV-P	XV-C	AudioNet	SincNet	Res18-I	Res18-V	Res34-I	Res34-V	Auto-I	Auto-V	GE2E	Hubert
$\|\nabla_x \mathcal{L}_t^x\|_1$	1.75e+09	0.173	8.54e+08	0.110	0.156	0.007	0.187	0.151	0.285	0.231	0.131	0.165	0.026	0.248

TABLE 13: Transferability of AS2T using FGSM and PGD. For each pair of source and target systems, the result indicates the increase of ASR_t. (Note that some target models misclassify some benign voices). CSI task, targeted attack, $s \in G$.

⇑ ASR_t (%) Source \ Target		Ivector	ECAPA	Xvector-P	Xvector-C	AudioNet	SincNet	Res18-I	Res18-V	Res34-I	Res34-V	Auto-I	Auto-V	GE2E	Hubert
GMM	Ivector	100.0	2.5	6.4	6.4	2.2	1.5	0.1	0.1	0.3	0.3	1.4	0.9	3.8	-0.3
TDNN	ECAPA	2.0	100.0	4.5	3.3	1.5	1	0.1	0.3	0.9	0.3	1.9	1.9	1.8	-0.2
	Xvector-P	13.3	7.1	100.0	8.9	1.8	1.5	0.0	0.1	0.4	0.2	2.3	1.2	2.7	0.0
	Xvector-C	16.0	15.5	17.8	99.7	2	1.2	0.3	0.2	0.6	0.5	3	2	5.7	0.1
CNN	AudioNet	0.4	0.1	0.1	-0.2	99.9	1.4	0.1	0.2	0.1	0.3	0.6	0.5	3.3	0.3
	SincNet	0.8	0.2	0.2	0.9	5.6	34.4	0.3	0.5	1.2	0.9	1.4	1.9	13.6	0.5
	Res18-I	0.2	0.0	0.0	-0.3	1.7	1.3	100.0	6.4	9.0	10.1	7.7	6	1.9	0.2
	Res18-V	0.3	0.1	0.1	-0.3	2.2	1.1	11.2	99.9	14.3	13.7	9.2	9	1.9	0.2
	Res34-I	0.3	0.0	0.1	-0.3	1.9	1	3	3	100.0	7.3	5.7	4.7	1.7	0.2
	Res34-V	0.3	0.0	0.1	-0.3	2.1	1.1	4.6	4.7	8.6	100.0	7.6	6.1	1.7	0.1
	Auto-I	0.2	0.1	0.2	-0.3	1.5	1.2	1.9	1.7	3.8	4.5	73.2	11.6	1.3	0.1
	Auto-V	0.4	0.1	0.4	-0.2	2.2	1.1	3.9	4.5	7.2	8.4	15.6	99.9	2.5	0.2
LSTM	GE2E	0.5	0.1	0.1	-0.1	5.3	1.2	0.0	0.2	0.4	0.7	0.6	1.2	72.6	0.4
Trans	Hubert	0.2	0.0	0.1	-0.3	0.7	0.5	0.0	0.1	0.1	0.1	0.5	0.8	1.2	88.1

AudioNet (same architecture) is over 20%, while the accuracy drop of *AudioNet* → *SincNet* is merely 8.1%; the accuracy drop of *SincNet* → *GE2E* (different architecture) is 36.6%, while the accuracy drop of *GE2E* → *SincNet* is less than 5%.

To understand asymmetry of transferability, consider the source SRS model with parameter θ_s, the target SRS model with parameter θ_t, the original voice x, and the adversarial perturbation δ crafted against the source model, the loss attained by the target model on the adversarial voice $x' = x + \delta$, denoted by $\mathcal{L}_t^{x'}$, can be simplified through a linear approximation as: $\mathcal{L}_t^{x'} = \mathcal{L}_t^x + \delta^T \nabla_x \mathcal{L}_t^x$.

For L_∞ attack, $\delta = \varepsilon \frac{\nabla_x \mathcal{L}_s^x}{\|\nabla_x \mathcal{L}_s^x\|_\infty}$, where $\mathcal{L}_s^x$ is the loss attained by the source model on the original voice. Then, according to Cauchy-Schwartz inequality, we can derive the upper bound of the change of loss as follows:

$$\delta^T \nabla_x \mathcal{L}_t^x = \varepsilon \frac{(\nabla_x \mathcal{L}_s^x)^T}{\|\nabla_x \mathcal{L}_s^x\|_\infty} \nabla_x \mathcal{L}_t^x \leq \varepsilon \|\nabla_x \mathcal{L}_t^x\|_1$$

Obviously, the upper bound of the change of the loss is positively co-related with $\|\nabla_x \mathcal{L}_t^x\|_1$, called the input gradient size in [62]. TABLE 12 shows the input gradient size of all the models. We can observe that the input gradient size of AudioNet and GE2E is much larger than that of SincNet, which results in asymmetry of transferability.

We also report the results of AS2T with randomly chosen target speakers on the CSI task in TABLE 13, where the source speakers are enrolled speakers, i.e., C6 in TABLE 3. We can draw a similar conclusion. However, the best attack *Xvector-C* → *Xvector-P* only achieves 17.8% success rate for targeted attack, compared to 58.6% for untargeted attack. This indicates that targeted transfer attack is much more difficult than untargeted transfer attack.

Remark 5. Transfer attack is effective in general when models have the same architecture, but the same architecture is neither sufficient nor necessary, as the transferability is also affected by the training dataset and input type.

We also evaluated the transferability of AS2T on the OSI and SV tasks, namely, C5 and C10, from which a similar conclusion can be drawn. Therefore, the results are omitted here and reported in our technical report [56].

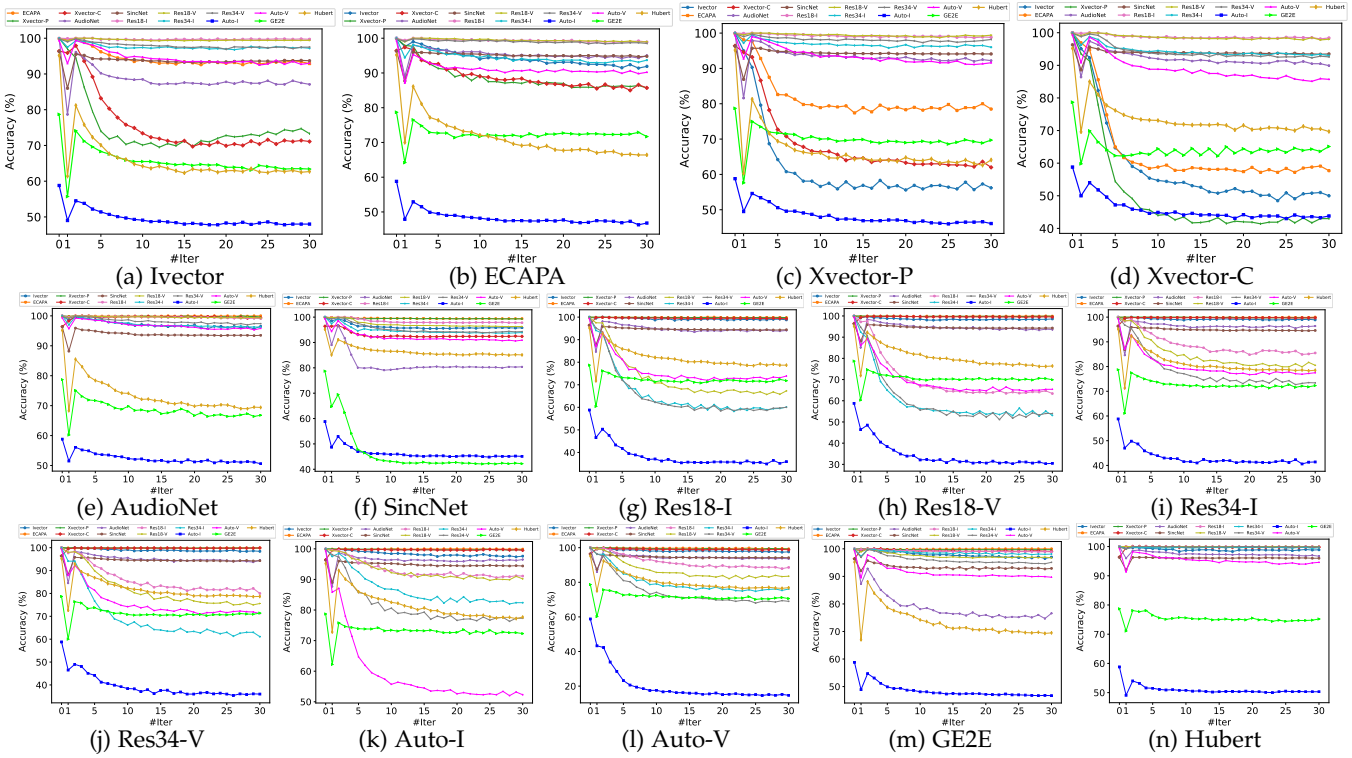


Fig. 4: Transferability of AS2T using FGSM and PGD w.r.t. #Iter. The subfigures' captions and the legends indicate the source and target models, respectively. #Iter=0 and #Iter=1 are no attack and AS2T attack with FGSM, respectively.

4.5.2 Attack-specific Factors

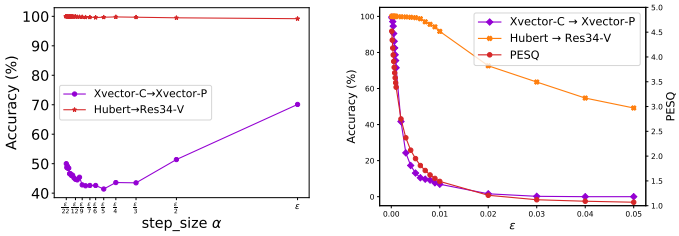
According to the above results, we study the impact of the number of iterations, step_size, and perturbation budget on the transferability of AS2T with the untargeted attack on CSI task, where the source speakers are enrolled speakers, i.e., C8 in TABLE 3.

The number of iterations (#Iter). We craft adversarial examples using AS2T with the PGD and FGSM optimization approaches, where ε=0.002, α=$\varepsilon/5$ for PGD, and α=0.002 for FGSM. We vary #Iter from 2 to 30 for PGD with step 1.

The results are plotted in Fig. 4. In general, the transferability rate increases with #Iter. This is possibly because large #Iter makes adversarial examples far from the decision boundary, hence can fool the other models with high probability. However, for some pairs of source and target models, especially when they have different architectures, increasing #Iter fails to improve the transferability.

We also find that the transferability of PGD with small #Iter is lower than that of FGSM. With the increase of #Iter, the transferability of PGD is higher than that of FGSM on some pairs of source and target models (e.g., *Ivector* $\rightarrow$ *Xvector-C*), but not on some pairs (e.g., *Xvector-P* $\rightarrow$ *Hubert*). The latter is reasonable since FGSM tends to craft adversarial examples with larger perturbation. The former is interesting since it differs from the conclusion in the image domain that multiple iterations attack is less transferable than single iteration attack [21].

The step_size (α). To explore the effect of step_size of PGD on the transferability, we choose two pairs of source and target models according to the results in TABLE 11: *Xvector-C* $\rightarrow$ *Xvector-P* and *Hubert* $\rightarrow$ *Res34-V*, where the former (resp. latter) achieves the best (worst) transferability among all the pairs in TABLE 11. We set $\varepsilon = 0.002$ and vary α from $\frac{\varepsilon}{\#\text{Iter}_b}$ to ε where $\#\text{Iter}_b$ is the number of iterations yielding the best transferability in Fig. 4. $\#\text{Iter}_b = 22$ for *Xvector-C* $\rightarrow$ *Xvector-P* and $\#\text{Iter}_b = 23$ for *Hubert* $\rightarrow$ *Res34-V*.



(a) Accuracy w.r.t. α (b) Accuracy and stealthiness (PESQ) w.r.t. ε

Fig. 5: Transferability of AS2T w.r.t. α and ϵ

The results are plotted in Fig. 5a. We find that the step_size is a crucial parameter for *Xvector-C* $\rightarrow$ *Xvector-P*, where too small or too large α will harm the transferability of *Xvector-C* $\rightarrow$ *Xvector-P*. However, it seems that α has negligible effect on the transferability of *Hubert* $\rightarrow$ *Res34-V*.

The perturbation budget (ε). We study the impact of the budget ε on the transfer attacks *Xvector-C* $\rightarrow$ *Xvector-P* and *Hubert* $\rightarrow$ *Res34-V*, with #Iter $= 22$ and #Iter $= 23$ respectively, $\alpha = \frac{\varepsilon}{5}$, and ε ranging from 1e-4 to 0.05, according to the results in Fig. 5a.

The results are plotted in Fig. 5b. The accuracy decreases with the increase of the perturbation budget, though the first attack drops more quickly. However, the PESQ also decreases with the increase of ε. To validate whether PESQ is consistent with human perception, we conduct a human study (cf. Section 4.6) and the results confirmed this. Therefore, we can conclude that transferability rate can be

improved by allowing large perturbation budget, but at the cost of sacrificing the stealthiness.

4.6 Human Study

To demonstrate that the perceptual objective metric PESQ is consistent with human perception in quantifying the stealthiness of adversarial voices, we conduct a human study on MTurk [63] under the approval of the Institutional Review Board (IRB) of our institutes.

Setup of human study. We recruit participants from MTurk and ask them to tell whether the voices in a pair are uttered by the same speaker (The three options are *same*, *different*, and *not sure*). Specifically, we randomly select 4 speakers (2 male and 2 female), and randomly choose 1 normal voice per speaker (called reference voice). Then for each speaker, we randomly select 1 normal voice with different text from reference voice, 1 distinct adversarial voice per perturbation budget (we set $\varepsilon = 0.05, 0.01, 0.001$) that are crafted from other normal voices of the same speaker, and 1 normal voice from other speakers with the same gender. Together, we build 24 pairs of voices: 4 pairs are *normal pairs* (one reference voice and one normal voice from the same speaker), 4 pairs are *other pairs* (one reference voice and one normal voice from another speaker) and 16 pairs are *adversarial pairs* (one reference voice and one adversarial voice from the same speaker; 4 pairs per perturbation budget).

To guarantee the quality of our questionnaire and validity of the results, we filter out the questionnaires that are randomly chosen by participants. In particular, we insert three pairs of voices, where each pair contains one male voice and one female voice as a concentration test. Only when all of them are correctly answered (i.e, the answer *different* is selected), we regard it as a valid questionnaire, otherwise, we exclude it.

Results of human study. We finally received 100 questionnaires where 13 questionnaires are filtered out as they failed to pass our concentration tests. Therefore, there are 77 valid questionnaires. The results of the human study are shown in Fig. 6. 76.1% of participants believe that voices in each other pair are uttered by different speakers, indicating the quality of collected questionnaires. For the adversarial pairs with $\varepsilon = 0.001$, 42.1% of participants believe that voices in each pair are uttered by the same speaker, very close to the baseline 45.2% of normal pairs. However, with the increase of ε, more participants think that voices in each pair are uttered by different speakers. We note that the PESQ also decreases with the increase of ε, indicating that PESQ is consistent with human perception to some extent in quantifying the stealthiness.

Remark 6. Increasing the number of iterations and perturbation budget improves transferability at the cost of sacrificing stealthiness. However, for some pairs of source and target models, strengthening the attack-specific factors is ineffective in improving transferability, indicating that model-specific factors are dominant factors over attack-specific ones.

5 Related Work

Adversarial attacks recently have attracted intensive attention in various domains, e.g., [64], [65], [66], [67], [68], [69].

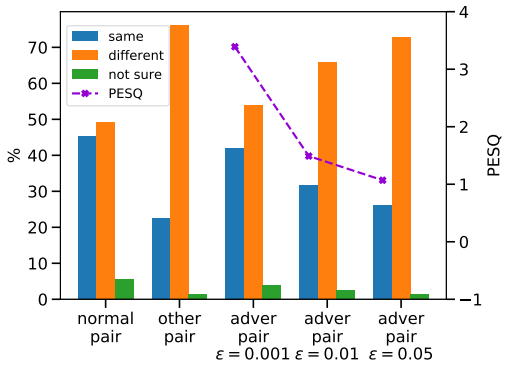


Fig. 6: Results of human study, adver is short for adversarial.

In this section, we mainly discuss related works in the speaker recognition (SR) domain.

Adversarial attacks. There exist both white-box and black-box attacks in the SR domain. Existing white-box attacks utilize FGSM and PGD from the image domain to create adversarial voices. The works in [10], [15], [18] demonstrated their attacks for C6/C8 (cf. TABLE 3), while other white-box attacks [8], [9], [11], [12], [13], [14] focused on the SV task with C10 only.

FAKEBOB [16] and SirenAttack [17] are two black-box attacks targeting SRSs. SirenAttack [17] only considers C6/C8 and is less effective than FAKEBOB [16], thus we do not integrate it into AS2T. FAKEBOB [16] covers more settings than the above works including C2, C6, and C10.

Compared over those works, we highlight the following three contributions. (i) To our knowledge, AS2T is the first attack that covers all the combinations of the source and target speakers on the three tasks (i.e., C1-C10 in TABLE 3) and is capable to launch arbitrary source-to-target speaker adversarial attack by utilizing various loss functions. In contrast, prior works consider only a few settings. We remark that each setting (i.e., C1-C10) is meaningful since it can be used to achieve at least one goal by the adversary. (ii) AS2T is applicable to different levels of knowledge about the victim model by integrating our loss functions into any optimization approaches, e.g., white-box FGSM, PGD, and CW, and black-box FAKEBOB, while prior works are limited to either white-box or black-box attacks. (iii) We explore different choices of loss functions and conduct a thorough evaluation by which we find the optimal loss functions leading to the most effective and efficient attacks. We remark that the loss function for C2 (resp. C8) adopted by FAKEBOB (resp. [10] and [18]) achieved inferior performance than our optimal loss functions.

Some previous works also demonstrate the practicability of transfer attacks, where the source and target models differ in at least one of the following factors: architecture, input type, training dataset, and scoring method. All the existing works consider only two architectures, namely, GMM and TDNN in [9], [16], CNN and TDNN in [10], and LSTM and TDNN in [13]. In this work, we consider all these four architectures and an additional Transformer. The works in [8], [16] train the SR models using two datasets, while our work involve five training datasets. The works in [8], [9], [16] cover two input types, e.g., MFCC and PLP in [16], MFCC and fBank in [8], and MFCC and spectrogram in

[9], while this work covers more input types, including waveform, spectrogram, fBank, and MFCC. In addition, our work covers both PLDA and COSS as the scoring method, while previous works merely consider one. In summary, we performed thus far the largest-scale transferability analysis among 14 SR models. We also evaluated the impact of attack-specific factors on the transferability, which has not been considered in the previous works.

Robust over-the-air adversarial attacks. To enhance the robustness of adversarial voices and enable physical over-the-air attack, prior works either simply craft adversarial voices with high-confidence [16] or integrate the distortions occurred in the over-the-air transmission into the generation process of adversarial voices [43], [70], [71], [72].

Chen et al. [16] argue that the high sensitivity of adversarial voices to the distortions incurred by over-the-air transmission is attributed to their closeness from the decision boundary. Motivated by this, they craft high-confidence adversarial voices using FAKEBOB and show that these adversarial voices are robust against over-the-air distortion, and are device-/environment-independent to some extent.

To improve the robustness of adversarial voices against reverberation in physical environment, pre-coding RIR is coordinated into the loss function using simulated RIR [70], [71] and real-world RIR [43], [72]. [72] uses isotropic noise and band-pass filter to handle ambient noise and equipment distortion and incorporate them into the loss function.

Our solution towards robust adversarial voices under the over-the-air setting is similar to [43], [70], [71], [72] except that: (i) although we do not model the equipment distortion as in [72], we find that this distortion is not substantial and modeling the reverberation and ambient noise is sufficient. (ii) we incorporate random noise with different levels of SNR to simulate different acoustic environments in the real-life scenarios, which enables our attack to remain effective under various physical environments.

Regrading the evaluation of over-the-air attack, [16], [43], [72] adopt real-world evaluation, i.e. repeatedly playing and recording voices by hand, while [70], [71] and our work use a simulated manner. Compared to real-world evaluation, simulated evaluation makes it possible to perform a large-scale and thorough evaluation, covering different scenarios such as attacking scenes, acoustic environments with various ambient noises, hardware devices, and positions of the adversary and the victim. Compared to [70], [71] which only consider reverberation in the simulated evaluation, our work additionally considers equipment distortions and four types of real-life representative ambient noises with various SNR, making our simulated evaluation more realizable in the real-world evaluation. We hope that the simulated evaluation manner in this work will serve as a benchmark for evaluating over-the-air attack for future works.

6 Conclusion and Future Work

We proposed AS2T, the first attack in speaker recognition domain that covers all the combinations of source and target speakers on all the three tasks. It features novel source-/target-oriented loss functions and enables the adversary to create adversarial voices using arbitrary source and target speakers to achieve various goals in diverse attack scenarios. The loss functions can be freely composed with both white-box and black-box optimization approaches to adapt to the adversary's knowledge about the target model.

We improved the robustness of AS2T for launching over-the-air attacks in the physical world by utilizing various parameterized transformation functions to model diverse distortions and incorporate them into the generation of adversarial voices. The effectiveness of our approach is confirmed by a thorough evaluation.

We leveraged AS2T to conduct thus far the larger-scale transferability analysis among 14 SR models, covering 5 architectures, 5 training datasets, 4 input types, and 2 scoring backends. The transferability analysis reveals how model-specific and attack-specific factors impact the transferability of AS2T and results in many useful findings and insights.

Our work motivates the following future works. (i) The transferability rate is limited in some cases, probably because the loss functions tailored for non-transfer attack are not optimal for transfer attack. One future work is to explore better loss functions for transfer attack. (ii) The transferability between models may be asymmetric, contradicting the decision boundary similarity based explanation of transferable adversarial examples in the image domain, because similarity is a systematical metric. An interesting future work is to study explanation that is consistent with this asymmetric phenomenon. (iii) Another future work is to investigate defense solutions against adversarial attacks in the speaker recognition domains under various settings.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62072309, Ant Group, and the National Key Research and Development Program under Grant No. 2020AAA0107800.

References

[1] H. Beigi, *Fundamentals of Speaker Recognition*. Springer, 12 2011.
[2] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J. Chou, S. Yeh, S. Fu, C. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," *CoRR*, vol. abs/2106.04624, 2021.
[3] "Microsoft azure speaker recognition," https://azure.microsoft.com/en-us/services/cognitive-services/speaker-recognition.
[4] H. Ren, Y. Song, S. Yang, and F. Situ, "Secure smart home: A voiceprint and internet based authentication system for remote accessing," in *ICCSE*, 2016.
[5] TD Bank voiceprint, https://www.tdbank.com/bank/tdvoiceprint.html.
[6] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digit. Signal Process.*, 2000.
[7] S. Sremath Tirumala and S. R. Shahamiri, "A review on deep learning approaches in speaker identification," in *ICSPS*, 2016.
[8] F. Kreuk, Y. Adi, M. Cissé, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *ICASSP*, 2018.
[9] X. Li, J. Zhong, X. Wu, J. Yu, X. Liu, and H. Meng, "Adversarial attacks on gmm i-vector based speaker verification systems," in *ICASSP*, 2020.
[10] A. Jati, C.-C. Hsu, M. Pal, R. Peri, W. AbdAlmageed, and S. Narayanan, "Adversarial attack and defense strategies for deep speaker recognition systems," *Computer Speech & Language*, vol. 68, p. 101199, 2021.
[11] W. Zhang, S. Zhao, L. Liu, J. Li, X. Cheng, T. F. Zheng, and X. Hu, "Attack on practical speaker verification system using universal adversarial perturbations," in *ICASSP*, 2021.

[12] J. Li, X. Zhang, C. Jia, J. Xu, L. Zhang, Y. Wang, S. Ma, and W. Gao, "Universal adversarial perturbations generative network for speaker recognition," in *ICME*, 2020.

[13] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Enabling fast and universal audio adversarial attack using generative model," in *AAAI*, 2021.

[14] Q. Wang, P. Guo, and L. Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *INTERSPEECH*, 2020.

[15] A. S. Shamsabadi, F. S. Teixeira, A. Abad, B. Raj, A. Cavallaro, and I. Trancoso, "Foolhd: Fooling speaker identification by highly imperceptible adversarial disturbances," in *ICASSP*, 2021.

[16] G. Chen, S. Chen, L. Fan, X. Du, Z. Zhao, F. Song, and Y. Liu, "Who is real Bob? adversarial attacks on speaker recognition systems," in *S&P*, 2021.

[17] T. Du, S. Ji, J. Li, Q. Gu, T. Wang, and R. Beyah, "Sirenattack: Generating adversarial audio for end-to-end acoustic systems," in *ASIACCS*, 2020.

[18] Y. Gong and C. Poellabauer, "Crafting adversarial examples for speech paralinguistics applications," *CoRR*, vol. abs/1711.03280, 2017.

[19] F. Tramèr, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "The space of transferable adversarial examples," *CoRR*, vol. abs/1704.03453, 2017.

[20] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *ICLR*, 2017.

[21] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," in *ICLR*, 2017.

[22] D. Wang, "A simulation study on optimal scores for speaker recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, no. 1, pp. 1–23, 2020.

[23] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[24] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *INTERSPEECH*, 2009.

[25] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *ICASSP*, 2014.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *ICASSP*, 2018.

[27] M. K. Nandwana, L. Ferrer, M. McLaren, D. Castan, and A. Lawson, "Analysis of critical metadata factors for the calibration of speaker recognition systems," in *INTERSPEECH*, 2019.

[28] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, P. Kenny *et al.*, "Cosine similarity scoring without score normalization techniques," in *Odyssey*, 2010.

[29] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.

[30] 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.

[31] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *CoRR*, vol. abs/1705.02304, 2017.

[32] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *CoRR*, vol. abs/1003.4083, 2010.

[33] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[34] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *SLT*, 2018.

[35] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *S&P*, 2017.

[36] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, and Y. Liu, "SEC4SR: A security analysis platform for speaker recognition," *CoRR*, vol. abs/2109.01766, 2021.

[37] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[38] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[40] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *ICML*, 2018, pp. 2142–2151.

[41] T. Chen, L. Shangguan, Z. Li, and K. Jamieson, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *NDSS*, 2020.

[42] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[43] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *HotMobile*, 2020.

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

[45] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Interspeech*, 2020.

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

[47] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, 2018.

[48] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker recognition: Modular or monolithic?" in *Interspeech*, 2019.

[49] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[50] S. Ding, T. Chen, X. Gong, W. Zha, and Z. Wang, "Autospeech: Neural architecture search for speaker recognition," in *Interspeech*, 2020.

[51] L. Wan, Q. Wang, A. Papir, and I. Lopez-Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[52] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2021.

[53] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[54] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001.

[55] Y. Xiang, G. Hua, and B. Yan, *Digital audio watermarking: fundamentals, techniques and challenges*. Springer, 2017.

[56] G. Chen, Z. Zhao, F. Song, S. Chen, L. Fan, and Y. Liu, "AS2T: Arbitrary source-to-target adversarial attack on speaker recognition systems," ShanghaiTech University, Tech. Rep., 2022, https://faculty.sist.shanghaitech.edu.cn/faculty/songfu/publications/AS2T.pdf.

[57] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017.

[58] M. Harper, "The automatic speech recogition in reverberant environments (aspire) challenge," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, 2015.

[59] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU aspire system: Robust LVCSR with tdnns, ivector adaptation and RNN-LMS," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, 2015.

[60] R. Hsiao, J. Z. Ma, W. Hartmann, M. Karafiát, F. Grézl, L. Burget, I. Szöke, J. Cernocký, S. Watanabe, Z. Chen, S. H. R. Mallidi, H. Hermansky, S. Tsakalidis, and R. M. Schwartz, "Robust speech recognition in unknown reverberant and noisy conditions," in *IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, 2015.

[61] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[62] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli, "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks," in *28th USENIX Security Symposium*, 2019.

[63] Amazon Mechanical Turk Platform. https://azure.microsoft.com.

[64] W. Liu, F. Song, T. Zhang, and J. Wang, "Verifying relu neural networks from a model checking perspective," *J. Comput. Sci. Technol.*, vol. 35, no. 6, pp. 1365–1381, 2020.
[65] X. Guo, W. Wan, Z. Zhang, M. Zhang, F. Song, and X. Wen, "Eager falsification for accelerating robustness verification of deep neural networks," in *ISSRE*, 2021, pp. 345–356.
[66] Y. Zhang, Z. Zhao, G. Chen, F. Song, and T. Chen, "BDD4BNN: A bdd-based quantitative analysis framework for binarized neural networks," in *CAV*, 2021, pp. 175–200.
[67] L. Bu, Z. Zhao, Y. Duan, and F. Song, "Taking care of the discretization problem: A comprehensive study of the discretization problem and a black-box adversarial attack in discrete integer domain," *IEEE TDSC*, 2021.
[68] Z. Zhao, G. Chen, J. Wang, Y. Yang, F. Song, and J. Sun, "Attack as defense: Characterizing adversarial examples using robustness," in *ISSTA*, 2021.
[69] F. Song, Y. Lei, S. Chen, L. Fan, and Y. Liu, "Advanced evasion attacks and mitigations on practical ml-based phishing website classifiers," *Int. J. Intell. Syst.*, vol. 36, no. 9, pp. 5210–5240, 2021.
[70] Y. Xie, Z. Li, C. Shi, J. Liu, Y. Chen, and B. Yuan, "Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems," *Journal of Signal Processing Systems*, pp. 1–14, 2021.
[71] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *ICASSP*, 2020.
[72] Z. Li, Y. Wu, J. Liu, Y. Chen, and B. Yuan, "Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations," in *CCS*, 2020.

Guangke Chen received his BEng degree from South China University of Technology, Guangzhou, China, in 2019. He is currently pursuing the Ph.D. degree with ShanghaiTech University, advised by Dr. Song. His research interest lies in the area of multimedia and machine learning security and privacy. He is currently doing research on the security issues of speaker and speech recognition systems. More information is available at http://guangkechen.site/.

Zhe Zhao received his B.S. degree from Ocean University of China, Tsingtao, China, in 2016. From 2016 to 2018, he was a software engineer at Hewlett-Packard Company. Now he is a Ph.D. student at School of Information Science and Technology, ShanghaiTech University. His research interest lies in the area of software engineering and testing. He is currently doing research in trusted artificial intelligence. His supervisor is Dr. Song.

Fu Song received the B.S. degree from Ningbo University, Ningbo, China, in 2006, the M.S. degree from East China Normal University, Shanghai, China, in 2009, and the Ph.D. degree in computer science from University Paris-Diderot, Paris, France, in 2013. From 2013 to 2016, he was a Lecturer and Associate Research Professor at East China Normal University. From August 2016 to July 2021, he is an Assistant Professor with ShanghaiTech University, Shanghai, China. Since July 2021, he is an Associate Professor with ShanghaiTech University. His research interests include formal methods and computer/AI security.

Sen Chen (Member, IEEE) is an Associate Professor at Tianjin University, China. Before that, he was a Research Assistant Professor at Nanyang Technological University (NTU), Singapore, and a Research Assistant of NTU from 2016 to 2019 and a Research Fellow from 2019-2020. He received his Ph.D. degree in Computer Science East China Normal University, China, in 2019. His research focuses on Security and Software Engineering. More information is available on https://sen-chen.github.io/.

Lingling Fan is an Associate Professor at Nankai University, China. She received her Ph.D and BEng degrees in computer science from East China Normal University, Shanghai, China in June 2019 and June 2014, respectively. In 2017, she joined Nanyang Technological University (NTU), Singapore as a Research Assistant and then had been as a Research Fellow of NTU since 2019. Her research focuses on program analysis and testing, software security. She got an ACM SIGSOFT Distinguished Paper Award at ICSE 2018.

Liu Yang (Senior Member, IEEE) graduated in 2005 with a Bachelor of Computing (Honours) in the National University of Singapore (NUS). In 2010, he obtained his PhD and started his post doctoral work in NUS, MIT and SUTD. In 2011, In 2012 fall, he joined Nanyang Technological University (NTU) as a Nanyang Assistant Professor. He is currently a full professor and the director of the cybersecurity lab in NTU. He specializes in software verification, security and software engineering. His research has bridged the gap between the theory and practical usage of formal methods and program analysis to evaluate the design and implementation of software for high assurance and security. He has more than 300 publications and 6 best paper awards in top tier conferences and journals.