# Parrot-Trained Adversarial Examples: Pushing the Practicality of Black-Box Audio Attacks against Speaker Recognition Models

Rui Duan
University of South Florida
ruiduan@usf.edu

Zhe Qu
Central South University
zhe_qu@csu.edu.cn

Leah Ding
American University
ding@american.edu

Yao Liu
University of South Florida
yliu@cse.usf.edu

Zhuo Lu
University of South Florida
zhuolu@usf.edu

***Abstract*—Audio adversarial examples (AEs) have posed significant security challenges to real-world speaker recognition systems. Most black-box attacks still require certain information from the speaker recognition model to be effective (e.g., keeping probing and requiring the knowledge of similarity scores). This work aims to push the practicality of the black-box attacks by minimizing the attacker's knowledge about a target speaker recognition model. Although it is not feasible for an attacker to succeed with completely zero knowledge, we assume that the attacker only knows a short (or a few seconds) speech sample of a target speaker. Without any probing to gain further knowledge about the target model, we propose a new mechanism, called parrot training, to generate AEs against the target model. Motivated by recent advancements in voice conversion (VC), we propose to use the one short sentence knowledge to generate more synthetic speech samples that sound like the target speaker, called parrot speech. Then, we use these parrot speech samples to train a parrot-trained (PT) surrogate model for the attacker. Under a joint transferability and perception framework, we investigate different ways to generate AEs on the PT model (called PT-AEs) to ensure the PT-AEs can be generated with high transferability to a black-box target model with good human perceptual quality. Real-world experiments show that the resultant PT-AEs achieve the attack success rates of 45.8%–80.8% against the open-source models in the digital-line scenario and 47.9%–58.3% against smart devices, including Apple HomePod (Siri), Amazon Echo, and Google Home, in the over-the-air scenario.**

## I. Introduction

Adversarial speech attacks against speech recognition [28], [114], [72], [101], [105], [32], [43], [118] and speaker recognition [43], [29], [118] have become one of the most active research areas of machine learning in computer audio security. These attacks craft audio adversarial examples (AEs) that can spoof the speech classifier in either white-box [28], [114], [72], [52] or black-box settings [105], [32], [43], [118], [29], [74], [17]. Compared with white-box attacks that require the full knowledge of a target audio classification model, black-box attacks do not assume the full knowledge and have been investigated in the literature under different attack scenarios [29], [118]. Despite the substantial progress in designing black-box attacks, they can still be challenging to launch in real-world scenarios in that the attacker is still required to gain information from the target model.

Generally, the attacker can use a query (or probing) process to gradually know the target model: repeatedly sending a speech signal to the target model, then measuring either the confidence level/prediction score [32], [43], [29] or the final output results [118], [113] of a classifier. The probing process usually requires a large number of interactions (e.g., over 1000 queries [113]), which can cost substantial labor and time. This may work in the digital line, such as interacting with local machine learning models (e.g., Kaldi toolkit [93]) or online commercial platforms (e.g., Microsoft Azure [7]). However, it can be even more cumbersome, if not possible, to probe physical devices because today's smart devices (e.g., Amazon Echo [3]) accept human speech over the air. Moreover, some internal knowledge of the target model still has to be assumed known to the attacker (e.g., the access to the similarity scores of the target model [29], [113]). Two recent studies further limited the attacker's knowledge to be (i) [118] only knowing the target speaker's one-sentence speech [118] and requiring probing to get the target model's hard-label (accept or reject) results (e.g., over 10,000 times) and (ii) [30] only knowing one-sentence speech for each speaker enrolled in the target model.

In this paper, we present a new, even more practical perspective for black-box attacks against speaker recognition. We first note that the most practical attack assumption is to let the attacker know nothing about the target model and never probe the model. However, such completely zero knowledge for the attacker unlikely leads to effective audio AEs. We have to assume some knowledge but keep it at the minimum level towards the attack practicality. Our work limits the attacker's knowledge to be only a one-sentence (or a few seconds) speech sample of her target speaker without knowing any other information about the target model. The attacker has neither knowledge of nor access to the internals of the target model. Moreover, she does not probe the classifier and needs no observation of the classification results (either soft or hard labels). To the best of our knowledge, our assumption of the

Network and Distributed System Security (NDSS) Symposium 2024
26 February - 1 March 2024, San Diego, CA, USA
ISBN 1-891562-93-2
https://dx.doi.org/10.14722/ndss.2024.24545
www.ndss-symposium.org

attacker's knowledge is the most restricted compared with prior work (in particular with the two recent attacks [118], [30]).

Centered around this one-sentence knowledge of the target speaker, our basic attack framework is to (i) propose a new training procedure, called parrot training, which generates a sufficient number of synthetic speech samples of the target speaker and uses them to construct a parrot-trained (PT) model for a further transfer attack, and (ii) systematically evaluate the transferability and perception of different AE generation mechanisms and create PT-model based AEs (PT-AEs) towards high attack success rates and good audio quality.

Our motivation behind parrot training is that the recent advancements in the voice conversion (VC) domain have shown that the one-shot speech methods [34], [77], [110], [31] are able to leverage the semantic human speech features to generate speech samples that sound like a target speaker's voice in different linguistic contents. Based on the attacker's one-sentence knowledge, we should be able to generate different synthetic speech samples of her target speaker and use them to build a PT model for speaker recognition. Our feasibility evaluations show that a PT model can perform similarly to a ground-truth trained (GT) model that uses the target speaker's actual speech samples.

The similarity between PT and GT models creates a new, interesting question of transferability: if we create a PT-AE from a PT model, can it perform similarly to an AE generated from the GT model (GT-AE) and transfer to a black-box target GT model? Transferability in adversarial machine learning is already an intriguing concept. It has been observed that the transferability depends on many aspects, such as model architecture, model parameters, training dataset, and attacking algorithms [79], [76]. Existing AE evaluations have been primarily focused on GT-AEs on GT models without involving synthetic data. As a result, we conduct a comprehensive study on PT-AEs in terms of their generation and quality.

- Generation: As an audio AE consists of the original signal and a perturbation signal. One essential difference in existing studies lies in finding the perturbation signal from different types of audio waveforms, which we call *carriers* in this paper. In particular, we summarize the carriers into the following major types: (i) noise carriers, which are the results of traditional methods [29], [118] during their search for the perturbation signals in the unrestricted $L_p$ space. (ii) feature-twisted carriers that are perturbation signals generated by only varying the auditory features of the original signal itself [113], [44], [17], [30], (iii) environmental sound carriers that are produced by environmental sounds [39]. Based on the built PT model, we create and evaluate PT-AEs based on these three types of carriers.
- Quality: We first need to define a quality metric to quantify whether a PT-AE is good or not. There are two important factors of PT-AEs: (i) transferability of PT-AEs to a black-box target model. We adopt the match rate, which has been comprehensively studied in the image domain [79], to measure the transferability. The match rate is defined as the percentage of PT-AEs that can still be misclassified as the same target label on a black-box GT model. (ii) The perception quality of audio AEs. We conduct a human study to let human participants rate the speech quality of AEs with different types of carriers in a unified scale of perception score from 1 (the worst) to 7 (the best) commonly used in speech evaluation studies [47], [108], [23], [19], [91], [36], and then build regression models to predict human scores of speech quality. However, these two factors are generally contradictory, as a high level of transferability likely results in poor perception quality. We then define a new metric called *transferability-perception ratio (TPR)* for PT-AEs generated using a specific type of carriers. This metric is based on their match rate and average perception score, and it quantifies the level of transferability a carrier type can achieve in degrading a unit score of human perception. A high TPR can be interpreted as high transferability achieved by a relatively small cost of perception degradation.

TABLE I: Summary of common attack strategies.

| Attack Strategy | Attack Scenario | Queries Needed | Knowledge Required | Human Perception |
|---|---|---|---|---|
| Carlini et al.[28] | White-box | $\sim$1000 | gradient info | ✗ |
| CommanderSong[114] | White-box | $\sim$100 | gradient info | ✗ |
| Psychoacoustic[95] | White-box | $\sim$5000 | gradient info | ✓ |
| AdvPulse[72] | White-box | $\sim$2000 | gradient info | ✗ |
| SpecPatch[52] | White-box | $\sim$1000 | gradient info | ✓ |
| Taori et al.[101] | Black-box | $\sim$300,000 | soft label | ✗ |
| SGEA[105] | Black-box | $\sim$300,000 | soft label | ✗ |
| Devil's Whisper[32] | Black-box | $\sim$1500 | soft label | ✗ |
| FakeBob[29] | Black-box | $\sim$5000 | soft label | ✗ |
| OCCAM[118] | Black-box | $\sim$10,000 | hard label | ✗ |
| TAINT[74] | Black-box | $\sim$1500 | hard label | ✓ |
| SMACK[113] | Black-box | $\sim$1000 | soft label | ✓ |
| QFA2SR [30] | Black-box | 0 | each speaker's sample | ✗ |
| PT-AE attack | Black-box | 0 | target speaker's sample | ✓ |

(i) Queries: indicating the typical number of probes need to interact with the black-box target model. (ii) Soft level: the confidence score [32] or prediction score [101], [105], [32], [29], [113] from the target model. (iii) Hard label: accept or reject result [118], [74] from the target model. (iv) QFA2SR [30] requires the speech sample of each enrolled speaker in the target model. (v) Human perception means integrating the human perception factor into the AE generation.

Under the TPR framework, we formulate a two-stage PT-AE attack that can be launched over the air against a black-box target model. In the first stage, we narrow down from a full set of carriers to a subset of candidates with high TPRs for the attacker's target speaker. In the second stage, we adopt an ensemble learning-based formulation [76] that selects the best carrier candidates from the first stage and manipulates their auditory features to minimize a joint loss objective of attack effectiveness and human perception. Real-world experiments show that the proposed PT-AE attack achieves the success rates of 45.8%–80.8% against open-source models in the digital-line scenario and 47.9%–58.3% against smart devices, including Apple HomePod (Siri), Amazon Echo, and Google Home, in the over-the-air scenario. Compared with two recent attack strategies Smack [113] and QFA2SR [30], our strategy achieves improvements of 263.7% (attack success) and 10.7% (human perception score) over Smack, and 95.9% (attack success) and 44.9% (human perception score) over QFA2SR. Table I provides a comparison of the required knowledge between the proposed PT-AE attack and existing strategies.

Our major contribution can be summarized as follows. (i) We propose a new concept of the PT model and investigate state-of-the-art VC methods to generate parrot speech samples to build a surrogate model for an attacker with the knowledge of only one sentence speech of the target speaker. (ii) We propose a new TPR framework to jointly evaluate the transferability and perceptual quality for PT-AE generations with different types of carriers. (iii) We create a two-stage PT-AE attack strategy that has been shown to be more effective than existing attacks strategies, while requiring the minimum level of the attack knowledge. Our attack demo can be found here[1].

## II. Background and Motivation

In this section, we first introduce the background of speaker recognition, then describe black-box adversarial attack formulations to create audio AEs against speaker recognition.

### A. Speaker Recognition

Speaker recognition becomes more and more popular in recent years. It brings machines the ability to identify a speaker via his/her personal speech characteristics, which can provide personalized services such as convenient login [5] and personalized experience [1] for calling and messaging. Commonly, the speaker recognition task includes three phases: training, enrollment, and recognition. It is important to highlight that speaker recognition tasks [29], [118], [113] can be either (i) multiple-speaker-based speaker identification (SI) or (ii) single-speaker-based speaker verification (SV). Specifically, SI can be divided into close-set identification (CSI) and open-set identification (OSI) [39], [29]. We provide detailed information in Appendix A.

### B. Adversarial Speech Attacks

Given a speaker recognition function $f$, which takes an input of the original speech signal $x$ and outputs a speaker's label $y$, an adversarial attacker aims to find a small perturbation signal $\delta \in \Omega$ to create an audio AE $x + \delta$ such that

$$f(x+\delta) = y_t, \quad D(x, x+\delta) \leq \epsilon, \tag{1}$$

where $y_t \neq y$ is the attacker's target label; $\Omega$ is the search space for $\delta$; $D(x, x+\delta)$ is a distance function that measures the difference between the original speech $x$ and the perturbed speech $x+\delta$ and can be the $L_p$ norm based distance [29], [118] or a measure of auditory feature difference (e.g., qDev [44] and NISQA [113]); and $\epsilon$ limits the change from $x$ to $x + \delta$.

A common white-box attack formulation [28], [72] to solve (1) can be written as

$$\arg\min_{\delta \in \Omega} \mathcal{J}(x+\delta, y_t) + c\, D(x, x+\delta), \tag{2}$$

where $\mathcal{J}(\cdot,\cdot)$ is the prediction loss in the classifier $f$ when associating the input $x + \delta$ to the target label $y_t$, which is assumed to be known by the attacker; and $c$ is a factor to balance attack effectiveness and change of the original speech.

A black-box attack has no knowledge of $\mathcal{J}(\cdot,\cdot)$ in (2) and thus has to adopt a different type of formulation depending on what other information it can obtain from the classifier $f$. If the attack can probe the classifier that gives a binary (accept or reject) result, the attack [118], [74] can be formulated as

$$\arg\min_{\delta \in \Omega} \mathcal{L}(x+\delta) = \begin{cases} D(x, x+\delta) & \text{if } f(x+\delta) = y_t, \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

Since (3) contains $f(x + \delta)$, the attacker has to create a probing strategy to continuously generate a different version of $\delta$ and measure the result of $f(x + \delta)$ until it succeeds. Accordingly, a large number of probes (e.g., over 10,000 [118]) are required, which makes real-world attacks less practical against commercial speaker recognition models that accept speech signals over the air.

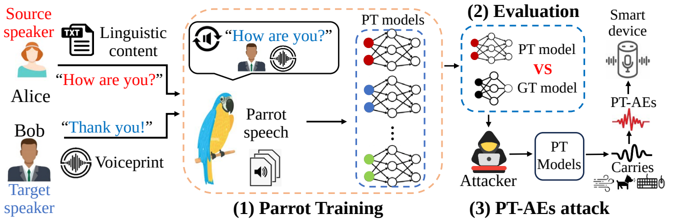

Fig. 1: The procedure of parrot training based black-box attack.

### C. Design Motivation

To overcome the cumbersome probing process of a black-box attack, we aim to find an alternative way to create practical black-box attacks. Given the fact that a black-box attack is not possible without probing or knowing any knowledge of a classifier, we adopt an assumption of prior knowledge used in [118] that the attacker possesses a very short audio sample of the target speaker (note that [118] has to probe the target model in addition to this knowledge). This assumption is more practical than letting the attacker know the classifier's internals. Given this limited knowledge, we aim to remove the probing process and create effective AEs.

To this end, we go back to the white-box attack formulation in (2) and try to build a local function $\mathcal{J}^*$ similar to the loss prediction function $\mathcal{J}$ in (2), then replace $\mathcal{J}$ with $\mathcal{J}^*$ to create an audio AE. This may look like a traditional transfer attack strategy [32]. But the key difference is that the traditional transfer attack still needs to keep probing the classifier (e.g., 1500 queries [32]) to build the local model $\mathcal{J}^*$; in contrast, the attacker here only has a very short sample of the target speaker to construct $\mathcal{J}^*$ without probing.

As a result, the first challenge we need to solve is how to build $\mathcal{J}^*$ based on a very short audio sample. As human speech is semantic, the recent advancements in the VC domain have shown that the one-shot speech methods [34], [77], [110], [31], commonly taking a source speaker's audio sample and a target speaker's sample as two inputs, are able to output a speech sample that sounds like the target speaker's voice in the source speaker's linguistic content. Hence, we are motivated to explore the feasibility of using the one-shot speech methods to create synthetic audio data of the attacker's target speaker. As this process is similar to training a parrot to reproduce more speech samples that can mimic the target speaker, we call them *parrot speech samples*, based on which we train the local model $\mathcal{J}^*$ to create audio AEs. We call this method *parrot training*, in contrast to the *ground-truth training* that uses a speaker's real audio samples to train.

Existing studies have focused on a wide range of aspects regarding ground-truth trained AEs (GT-AEs). The concepts of parrot speech and parrot training create a new type of AEs, parrot-trained AEs (PT-AEs), and also raise three major questions of the feasibility and effectiveness of PT-AEs towards a practical black-box attack: (i) Can a PT model approximate a GT model? (ii) Are PT-AEs built upon a PT model as transferable as GT-AEs against a black-box GT model? (iii)

[1] https://sites.google.com/view/pt-attack-demo

How to optimize the generation of PT-AEs towards an effective black-box attack? Fig. 1 shows the overall procedure for us to address these questions towards a new, practical and non-probing black-box attack: (1) we propose a two-step one-shot conversion method to create parrot speech for parrot training in Section III; (2) we study different types of PT-AE generations from a PT model regarding their transferability and perception quality in Section IV; and (3) we formulate an optimized black-box attack based on PT-AEs in Section V. Then, we perform comprehensive evaluations to understand the impact of the proposed attack on commercial audio systems in Section VI.

### *D. Threat Model*

In this paper, we consider an attacker that attempts to create an audio AE to fool a speaker recognition model such that the model recognizes the AE as a target speaker's voice. We adopt a black-box attack assumption that the attacker has no knowledge about the architecture, parameters, and training data used in the speech recognition model. We assume that the attacker has a very short speech sample (a few seconds in our evaluations) of the target speaker, which can be collected in public settings [118], but the sample is not necessarily used for training in the target model. We focus on a more realistic scenario where the attacker does not probe the model, which is different from most black-box attack studies [113], [29], [118] that require many probes. We assume that the attacker needs to launch the over-the-air injection against the model (e.g., Amazon Echo, Apple HomePod, and Google Assistant).

## III. Parrot Training: Feasibility and Evaluation

In this section, we study the feasibility of creating parrot speech for parrot training. As the parrot speech is the one-shot speech synthesized by a VC method, we first introduce the state-of-the-art of VC, then propose a two-step method to generate parrot speech, and finally evaluate how a PT model can approximate a GT model.

### *A. One-shot Voice Conversion*

**Data synthesis:** Generating data with certain properties is commonly used in the image domain, including transforming the existing data via data augmentation [92], [98], [80], [98], generating similar training data via Generative Adversarial Networks (GAN) [50], [35], [18], and generating new variations of the existing data by Variational Autoencoders (VAE) [62], [48], [53], [24]. These approaches can also be found in the audio domain, such as speech augmentation [63], [90], [64], [69], GAN-based speech synthesis [33], [65], [59], [21], and VAE-based speech synthesis [62], [55], [117]. Specifically, VC [94], [75], [70], [104], [31] is a specific data synthesis approach that can utilize a source speaker's speech to generate more voice samples that sound like a target speaker. Recent studies [107], [40] have revealed that it can be difficult for humans to distinguish whether the speech generated by a VC method is real or fake.

**One-shot voice conversion:** Recent VC has been developed by only using one-shot speech [34], [77], [110], [31] (i.e., the methods only knowing one sentence spoken by the target speaker) to convert the source speaker's voice to the target speaker's. This limited knowledge assumption well fits the black-box scenario considered in this paper and motivates us to use one-shot speech data to train a local model for the black-box attacker. As shown in the left-hand side of Fig. 1, a VC model takes the source speaker's and the target speaker's speech samples as two inputs and yields a parrot speech sample as the output. The attacker can pair the only speech sample, obtained from the target speaker, with different speech samples from public speech datasets as different pairs of inputs to the VC model to generate different parrot speech samples, which are expected to sound like the target speaker's voice to build parrot training.

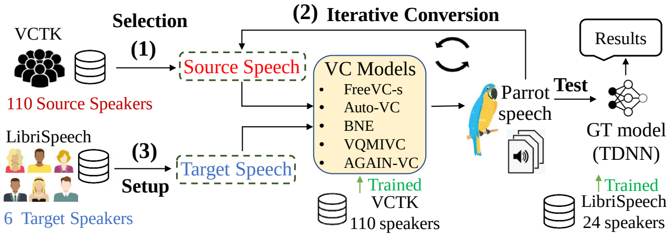

Fig. 2: Parrot speech generation: setups and evaluations.

### *B. Parrot Speech Sample Generation and Performance*

We first propose our method to generate parrot speech samples and then use them to build and evaluate a PT model. To generate parrot speech, we propose two design components, motivated by existing results based on one-shot VC methods [60], [61], [40].

1) Initial selection of the source speaker. Existing VC studies [60], [61] have shown that intra-gender VC (e.g., female to female) appears to have better performance than inter-gender one (e.g., female to male). As a major difference between male and female voices is the pitch feature [70], [104], [75], which represents the basic frequency information of an audio signal, our intuition is that selecting a source speaker whose voice has the pitch feature similar to the target speaker may improve the VC performance. Therefore, for an attacker that knows a short speech sample of the target speaker to generate more parrot speech samples, the first step in our design is to find the best source speaker in a speech dataset (which can be a public dataset or the attacker's own dataset) such that the source speaker has the minimum average pitch distance to the target speaker.
2) Iterative conversions. After selecting the initial source speaker, we can adopt an existing one-shot VC method to output a speech sample given a pair of the initial source speaker's and target speaker's samples. As the output sample, under the VC mechanism, is expected to feature the target speaker's audio characteristics better than the initial source speaker, we use this output as the input of a new source speaker's sample and run the VC method again to get the second output sample. We run this process iteratively to eventually get a parrot speech sample. Iterative VC conversions have been investigated in a recent audio forensic study [40], which found that changing the target speakers during iterative conversions can help the source speaker hide his/her voiceprints, i.e., obtaining more features from other speakers to make the voice features of the original

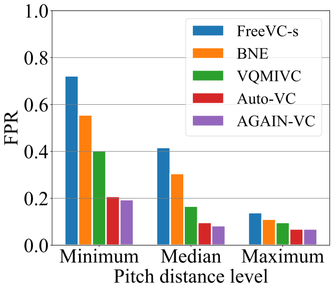

Fig. 3: FPRs under different initial source speakers.

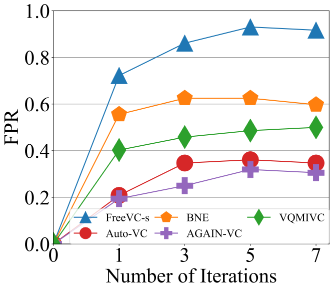

Fig. 4: FPRs under different numbers of iterations.

source speaker less evident. Compared with this feature-hiding method, our iterative conversions can be considered as a way of amplifying the audio features of the same target speaker to generate parrot speech.

We set up source speaker selection and iterative conversions with one-shot VC models to generate and evaluate the performance of parrot speech samples in Fig. 2.

**Experimental setup:** There are a wide range of one-shot VC methods recently available for parrot speech generation. We consider and compare the performance of AutoVC [10], BNE [13], VQMIVC [15], FreeVC-s [11], and AGAIN-VC [9]. As shown in Fig. 2, we use the VCTK dataset [103] to train each VC model. The dataset includes 109 English speakers with around 20 minutes of speech. We also select the source speakers from this dataset. We select 6 target speakers from the LibriSpeech dataset [87], which is different from the VCTK dataset, such that the VC training does not have any prior knowledge of the target speaker. Only one short sample (around 4 seconds with 10 English words) of a target speaker is supplied to each VC model to generate different parrot speech samples. We build a time delay neural network (TDNN) as the GT model for a CSI task to evaluate how parrot samples can be accurately classified as the target speaker's voice. The GT model is trained with 24 (12 male and 12 female) speakers from LibriSpeech (including the 6 target speakers and 18 randomly selected speakers). The model trains 120 speech samples (4 to 15 seconds) for each speaker and yields a test accuracy of 99.3%.

**Evaluation metrics:** We use the False Positive Rate (FPR) [56], [29] to evaluate the effectiveness of parrot speech, i.e., the percentage of parrot speech samples that are classified by the TDNN classifier as the target speaker's voice. Specifically, $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$, where False Positives (FP) indicates the number of cases that the classifier wrongly identifies parrot speech samples as target speaker's label; True Negatives (TN) represents the number of cases that the classifier correctly rejects parrot speech samples as any other label except for the target speaker.

**Evaluation results:** We first evaluate the impact of the initial source speaker selection on different VC models. We set the number of iterative conversions to be one, and the target speaker's speech sample is around 4.0 seconds (10 English words), which is the same for all VC models. We use the pitch distance between the source and target speakers as the evaluation standard. Specifically, we first sort all 110 source speakers in the VCTK dataset with respect to their average pitch distances to the target speaker. We use *minimum*, *median*, *maximum* to denote the source speakers who have the smallest, median, and largest pitch distances out of all the source speakers, respectively. We use each VC method to generate 12 different parrot speech samples for each target speaker (i.e., a total of 72 samples for 6 target speakers under each VC method). Fig. 3 shows that the pitch distance of the source speaker can substantially affect the FPR. For the most effective VC model, Free-VCs, we can observe that the FPR can reach 0.7222 when the source speaker is chosen to have the minimum distance to the target speaker, indicating that 72.22% parrot speech samples can fool the GT TDNN model in Fig. 2. Even for the worst-performing AGAIN-VC model, we can still observe that the minimum-distance FPR (0.1944) is nearly 3 times the maximum-distance FPR (0.0694). As a result, the source speaker with the less pitch distance is more effective to improve the VC performance (i.e., leading to a higher FPR).

TABLE II: VC Performance under different knowledge levels.

| Knowledge Level | FreeVC-s | AutoVC | BNE | VQMIVC | AGAIN-VC |
|---|---|---|---|---|---|
| 2-second | 0.5416 | 0.0972 | 0.3194 | 0.1667 | 0.0833 |
| 4-second | 0.8750 | 0.4028 | 0.5139 | 0.4583 | 0.2639 |
| 8-second | 0.9167 | 0.5417 | 0.7083 | 0.5833 | 0.3750 |
| 12-second | 0.9305 | 0.5556 | 0.7222 | 0.5972 | 0.3889 |

Next, we evaluate the impact of iterative conversions on the FPR. Fig. 4 shows the FPRs with different numbers of iterations for each VC model (with zero iteration meaning no conversion and directly using the TDNN to classify each source speaker's speech). It is noted from the figure that with increasing the number of iterations, the FPR initially gains and then stays within a relatively stable range. For example, the FPR of FreeVC-s achieves the highest value of 0.9305 after 5 iterations and then drops slightly to 0.9167 after 7 iterations. Based on the results in Fig. 4, we set 5 iterations for parrot speech generation.

We are also interested in how much knowledge of the target speaker is needed for each VC model to generate effective parrot speech. We set the knowledge level based on the length of the target speaker's speech given to the VC. Specifically, we crop the target speaker's speech into four levels: i) 2-second length level (around 5 words), ii) 4-second level (10 words), iii) 8-second level (15 words), and iv) 12-second level: (22 words). For each VC model, we generate 288 parrot speech samples (12 for each target speaker with each different knowledge level) to interact with the GT model. All samples are generated by choosing the initial source speaker with the minimum pitch distance and setting the number of iterations to be 5.

Table II evaluates the FPRs under different knowledge levels of the target speaker. It can be seen that the length of the target speaker's speech substantially affects the effectiveness of parrot speech samples. For example, AutoVC achieves the FPRs of 0.0972 and 0.5417 given 2- and 4-second speech samples of the target speaker, and finally increases to 0.5556 with the 12-second knowledge. It is also observed that FreeVC-s performs the best in all VC methods for each knowledge level (e.g., 0.9167 for the 8-second knowledge level). We can also find that the increase in FPR becomes slight from 8-second to 12-second speech knowledge. For example, FreeVC-s increases from 0.9167 (8-second) to 0.9305 (12-second), and VQMIVC increases from 0.5833 (8-second) to 0.5972 (12-second). Overall, the results of Table II reveal that even based on a very limited amount (i.e., a few seconds) of the target speaker's speech, parrot speech samples can still be efficiently generated to mimic the speaker's voice features and fool a speaker classifier to a great extent.

### C. *Parrot Training Compared with Ground-Truth Training*

We have shown that parrot speech samples can be effective in misleading a GT-trained speaker classification model. Additionally, we use experiments to further evaluate how a PT model trained by parrot speech samples is compared with a GT model. We compare the classification performance of PT and GT models. Based on our findings, PT models exhibit classification performance that is comparable to, and can approximate, GT models. We include experimental setups and results in Appendix B.

## IV. PT-AE Generation: A Joint Transferability and Perception Perspective

In this section, we aim to evaluate whether the PT-AEs are as effective as GT-AEs against a black-box GT model. We first summarize AE generation methods that use different types of audio waveforms (i.e., carriers). Next, we quantify the human perceptual quality of AEs with different carriers, then use the match rate to measure the transferability of PT-AEs to GT models. Finally, we define the unified metric, transferability-perception ratio (TPR), to evaluate PT-AEs.

### A. *Carriers in Audio AE Generation*

Recent audio attack studies have considered different audio perturbation carriers to generate AEs via specific generation algorithms. We summarize three main types of carriers.

**Noise carriers:** Traditional methods [29], [74] usually adopt a gradient estimation method to generate audio AEs in the unrestricted $L_p$ space with the initial perturbation signal set commonly as a Gaussian noise. This leads to a noisy sound despite some psychoacoustic methods [95], [52], [74] that can be used to alleviate the noisy effect.

**Feature-twisted carriers:** Directly manipulating the auditory feature of a speech signal could make a classifier sensitive but stealthy to the human ears. Existing works [17], [113] have found that modifying the phonemes or changing the prosody of the speech can also spoof the audio classifier while preserving the perception quality.

**Environmental sound carriers:** The enrollment phase attack [39] employed environmental sounds (e.g., traffic) to create the perturbation signal to poison a speaker recognition model.

### B. *Quantifying Perceptual Quality of Speech AEs*

We first need to find an appropriate perception metric to accurately measure the human perceptual quality of AEs based on different carriers. Recent studies [44], [113] have pointed out that traditional metrics, such as signal-to-noise ratio (SNR) [32] and the $L_p$ norm [114], [29], [118], cannot directly reflect the human perception. They have used different human study based metrics to measure the perceptual quality of AEs with certain types of carriers (i.e., qDev for music AEs in [44] and NISQA for feature-twisted AEs [113]). In addition, we also notice that the harmonics-to-noise ratio (HNR) [115] is a common metric adopted in speech science to measure the quality of a speech signal. Given these potential perception metrics, we aim at conducting a human study to find out the best metric to measure the perceptual quality across a diversity of AE carriers that we are interested in.

**Dataset generation for human study:** We create the human study dataset with noise carriers [28], [95], [52], [29], [118], [74], feature-twisted carriers [113], and environmental carriers [39]. We choose 30 original speech signals (with length from 5 to 15 seconds) from the existing speech dataset [82]. We modify these original signals by adding different types of carriers to form perturbed speech signals for the human study. We use the signal-to-carrier ratio (SCR) to control the energy of a perturbation carrier added to an original signal. For example, an SCR of 0dB means that the carrier and the original signal have the same energy level. We consider the following carriers to be added to the original signals.

i) Noise carriers: The dataset [82] provides a wide range of noisy speech signals. The noise is Gaussian-distributed and can be generated with different SNRs. We generated 30 speech samples whose SNRs are uniformly distributed in 0-30 dB. Note that the metric SCR is equivalent to the metric of SNR in the case of noise carriers.

ii) Feature-twisted carriers: For feature-twisted speech signals, we shift the tone (i.e., the pitch) [113] to generate pitch-twisted carriers. Specifically, we shift up/down by 25 semitones[2] of the original speech to craft the pitch-twisted carriers, and add these carriers to the original speech with different SCR levels. For twisting the rhythm, we speed up and slow down the speech ranging from 0.5 to 2 times of its speech rate.

iii) Environmental sound carriers: Environmental sound carriers are selected from the large-scale human-labeled environmental sound datasets [47] with categories including natural sounds (e.g., wind and sea waves), sounds of things (e.g., vehicle and engine), human sounds (e.g., whistling), animal sounds (e.g., pets), and music (e.g., musical instruments). For each category, we randomly selected 6 audio clips.

We have created a total of 90 perturbed speech samples, 30 samples for each carrier set at different SCR levels.

**Human participant involvement:** We have recruited 30 volunteers, who are college students with no hearing issues (self-reported). Our study procedure was approved by our Institutional Review Board (IRB). Each volunteer is asked to rate the similarity between a pair of original and carrier-perturbed speech clips using a scale from 1 to 7 commonly adopted in speech evaluation studies [47], [108], [23], [19], [91], [36], where 1 indicates the least similarity (i.e., speakers sound very different between the two clips) and 7 represents the most similarity (i.e., speakers sound very similar).

**Perceptual quality of different carriers:** Fig. 5 compares the average human scores at varying SCR levels for different

[2] 1 semitone = $12\log_2(f'/f)$, where $f$ and $f'$ are the original and perturbed speech frequencies, respectively [14].

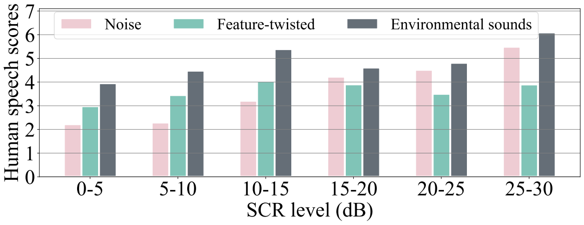

Fig. 5: Human scores for carrier-perturbed speech signals.

carriers. We can clearly see that the perception quality for noise carriers improves gradually with increasing the SCR, which indicates the less loudness of the noise carrier, the better perception of the perturbed speech. Interestingly, the human scores of the feature-twisted and environmental sound carriers are not closely correlated with the SCR. Both of them can indeed get better human scores at lower SCR levels (e.g., 10-15 dB vs 15-20 dB). Fig. 5 also shows that overall, environmental sound carriers yield the better human scores than the feature-twisted carriers and noise carriers.

**Evaluation of speech quality metrics:** Next, we evaluate the accuracy of existing metrics to characterize the speech quality based on our human study results. We compare the metrics of $L_2$ and $L_\infty$ norms [114], [29], [118], SCR (equivalent to SNR [32]), HNR [115], audio-feature-regression-based qDev [44], and DNN-based NISQA [113], [82]. Note that the qDev model [44] was originally trained using music instead of speech. We follow the procedure in [44] to train a random forest regression model using our speech samples. We call the resultant metric speech-regression score (SRS).

To evaluate how well a speech quality metric matches the human score from the human study, we use two correlation coefficients, Pearson's and Spearman's coefficients [54], to measure the correlation between the metric and the human score. Table III computes all correlation coefficients from our human study. It is observed from the table that SRS has the best accuracy across almost all carriers, except for noise carriers, where $L_2$-norm achieves the highest Spearman's coefficient. The DNN-based NISQA has high coefficients for noise carriers, but has degraded accuracy for feature-twisted carriers. One potential reason is that NISQA is trained with the noise carrier and environmental sound carrier dataset [82], which may not be effective for feature-twisted speech as the diversity of training data is important to the prediction performance [44]. Based on Table III, we use the metric of SRS to measure the perpetual quality of an audio AE.

### C. Measuring Transferability of PT-AEs

We then move to evaluate the transferability of different carriers for PT-based AEs.

*1) Building Target and Surrogate Models:* The first step in evaluating the transferability is to build i) target models, which refer to the models to be attacked by the attacker using PT-AEs, and ii) surrogate models, which are used by the attacker to generate PT-AEs against the target models. It is known that the difference between the target and surrogate models can affect the transferability of AEs [76].

**Building target models:** We consider building a diversity of target models with 4 DNN-based speaker recognition models including 2 CNN [58] and 2 TDNN models [99], [100]. These 4 target models are trained with the same 6 target speakers (3 males and 3 females). We randomly select them from LibriSpeech, and use 120 speech samples for each speaker for training. As the 4 target models have varying architectures and parameters (i.e., number of layers and weights), we denote them as CNN-A, CNN-B, TDNN-A, and TDNN-B. Their accuracies are 100.0%, 96.5%, 99.3%, and 97.2%, respectively.

TABLE III: Evaluation of different metrics.

| Carrier Type | Metrics | SRS | HNR | $L_2$ | $L_\infty$ | SCR | NISQA |
|---|---|---|---|---|---|---|---|
| Noise | Pearson | **0.9387** | 0.6339 | -0.7699 | -0.6680 | 0.2524 | 0.9279 |
| | Spearman | 0.7882 | 0.7303 | **-0.9349** | -0.9229 | 0.3956 | 0.8409 |
| Environ. Sounds | Pearson | **0.9647** | 0.4265 | 0.0923 | -0.5426 | 0.2348 | 0.6657 |
| | Spearman | **0.9566** | 0.5355 | -0.2843 | -0.4761 | 0.4152 | 0.7280 |
| Feature-twisted | Pearson | **0.9234** | 0.1099 | -0.1959 | 0.0744 | -0.097 | 0.3859 |
| | Spearman | **0.9139** | 0.1173 | -0.0985 | -0.0097 | 0.0397 | 0.2978 |
| Overall | Pearson | **0.9299** | 0.0855 | -0.3108 | -0.4068 | 0.0438 | 0.2372 |
| | Spearman | **0.9187** | 0.0785 | -0.3691 | -0.4603 | 0.1331 | 0.1434 |

**Building surrogate models:** We also aim to build a diversity of surrogate models for the attacker. As the attacker, without the knowledge of target models, is free to use any architecture for parrot training, we build two CNN-based and two TDNN-based surrogate architectures with different parameters, denoted by PT-CNN-C, PT-CNN-D, PT-TDNN-C, and PT-TDNN-D. Since there are 6 speakers trained in a target model, we consider each of them to be the attacker's target under each of the four surrogate architectures. For example, when the attacker uses the PT-CNN-C architecture and she targets speaker $i \in [1, 6]$ in the target models, the attacker is assumed to only know speaker $i$'s 8-second speech, and uses it to generate parrot speech samples, together with speech samples from 3 to 8 speakers randomly selected from the VCTK dataset (none is in the target models that use the LibriSpeech dataset), to build her surrogate model, denoted by PT-CNN-C-$i$. As a result, we construct a set of 6 surrogate models under each surrogate architecture (totally 24 models), denoted by $\{$PT-CNN-C-$i\}_{i\in[1,6]}$, $\{$PT-CNN-D-$i\}_{i\in[1,6]}$, $\{$PT-TDNN-C-$i\}_{i\in[1,6]}$, and $\{$PT-TDNN-D-$i\}_{i\in[1,6]}$.

**Compare PT with benchmark GT models.** To better understand the transferability of the PT-AEs in comparison with GT-AEs, we also use the target speaker $i$'s ground-truth speech instead of the parrot speech to build the attacker's surrogate models under the four surrogate architectures, denoted by $\{$GT-CNN-C-$i\}_{i\in[1,6]}$, $\{$GT-CNN-D-$i\}_{i\in[1,6]}$, $\{$GT-TDNN-C-$i\}_{i\in[1,6]}$, and $\{$GT-TDNN-D-$i\}_{i\in[1,6]}$. We will also generate GT-AEs based on these GT-surrogate models to attack the target models. They will serve as the benchmark for comparison with their PT counterparts.

*2) AE generations via different carriers:* After building the surrogate and target models, we generate AEs from the surrogate models using the three types of carriers based on existing studies.

i) For the noise carrier, we solve the white-box problem (2) via projected gradient descent (PGD) [49], and we choose $L_\infty$ norm as the distance metric, which shows a good performance in Table III. We set $\epsilon = 0.05$ to control the $L_\infty$ norm.

ii) For the feature-twisted carrier, we twist the pitch and rhythm of the original speech [113], [44] using the perception metric SRS as the distance measurement. As the random-forest-based SRS is non-differentiable, we use grid search to solve 2.

TABLE IV: Match rates between surrogate and target models.

| AE Carrier Type: | Noise | | | | | Feature-twisted | | | | | Environmental sound | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Target Model: | CNN-A | CNN-B | TDNN-A | TDNN-B | **Average** | CNN-A | CNN-B | TDNN-A | TDNN-B | **Average** | CNN-A | CNN-B | TDNN-A | TDNN-B | **Average** |
| GT-CNN-C | 0.2167 | 0.1500 | 0.1167 | 0.1417 | **0.1563** | 0.2333 | 0.2083 | 0.1583 | 0.1750 | **0.1937** | 0.3500 | 0.3250 | 0.2417 | 0.2250 | **0.2854** |
| PT-CNN-C | 0.1917 | 0.1417 | 0.0917 | 0.1250 | **0.1375** | 0.2083 | 0.1750 | 0.1083 | 0.1583 | **0.1625** | 0.3083 | 0.2583 | 0.2000 | 0.1750 | **0.2353** |
| GT-CNN-D | 0.0917 | 0.2167 | 0.0833 | 0.1917 | **0.1458** | 0.1667 | 0.1917 | 0.1500 | 0.1833 | **0.1729** | 0.1833 | 0.3250 | 0.2417 | 0.2917 | **0.2604** |
| PT-CNN-D | 0.0417 | 0.1667 | 0.0583 | 0.1583 | **0.1063** | 0.1417 | 0.1500 | 0.1417 | 0.1583 | **0.1479** | 0.1583 | 0.2167 | 0.2750 | 0.2583 | **0.2271** |
| GT-TDNN-C | 0.1000 | 0.1500 | 0.1750 | 0.1583 | **0.1458** | 0.1500 | 0.1833 | 0.2583 | 0.1417 | **0.1833** | 0.3500 | 0.1833 | 0.3583 | 0.3417 | **0.3083** |
| PT-TDNN-C | 0.0917 | 0.1417 | 0.1667 | 0.1333 | **0.1333** | 0.1167 | 0.1750 | 0.2500 | 0.1333 | **0.1688** | 0.3167 | 0.1750 | 0.2833 | 0.3083 | **0.2708** |
| GT-TDNN-D | 0.1333 | 0.1000 | 0.2083 | 0.2083 | **0.1625** | 0.1583 | 0.2750 | 0.2833 | 0.2917 | **0.2520** | 0.1417 | 0.3083 | 0.3917 | 0.4083 | **0.3125** |
| PT-TDNN-D | 0.1250 | 0.0833 | 0.1750 | 0.1667 | **0.1375** | 0.1417 | 0.2500 | 0.2500 | 0.2583 | **0.2225** | 0.1250 | 0.2667 | 0.3417 | 0.3333 | **0.2667** |

Specifically, we shift up/down for 25 semitones of the pitch, and the minimal shift-pitch step $\Delta_p = 1$ semitone. We speed up and slow down the speech ranging from 0.2 to 2.0 its speech rate with the minimal rhythm-changed step $\Delta_r$ to be 0.2.

iii) For the environmental sound carrier, we choose 30 environmental sounds from [47] which includes natural sounds, sounds of things, human sounds, animal sounds, and music. Based on the SRS to represent the distance $D$ in (2), we solve (2) via finding the best linear weights [44] of different environmental sounds using grid search with the minimal search step to be $0.1\epsilon$ with threshold $\epsilon$ set to be 0.05 (the same as the noise carrier's threshold).

For each carrier type, we generate 20 PT-AEs from each PT-surrogate model (a total of 480 PT-AEs). In addition, we generate 20 GT-AEs from each GT-surrogate model for the comparison purpose (also a total of 480 GT-AEs).

*3) Evaluation metric for transferability:* The transferability has been extensively studied in the image domain [88], [76], [89], [79]. One important evaluation metric in the transfer attacks [79], [76] is the match rate, which measures the percentage of AEs that can make both a surrogate model and a target model predict the same wrong label. We use the metric of the match rate to measure the transferability of PT-AEs in this work. Specifically, we can test a generated PT-AE: $x+\delta$ with both surrogate model $f(\cdot)$ and target model $f'(\cdot)$. If $f(x+\delta)=f'(x+\delta) \neq f(x)$, we can say $x+\delta$ is a matched AE for both $f(\cdot)$ and $f'(\cdot)$. The match rate is the ratio between the number of matched AEs and the total number of AEs.

*4) Results analysis:* It would be tedious to show the match rate of each pair in the 24 surrogate models and 6 target models that we have built. We average the match rates of the surrogate models under the same surrogate architecture (i.e., PT-CNN-C, PT-CNN-D, PT-TDNN-C, and PT-TDNN-D). For example, we compute the match rate of the PT-CNN-C based surrogate architecture by averaging the six match rates of $\{$PT-CNN-C-$i\}_{i\in[1,6]}$ models against a target model.

Table IV shows the match rates between different surrogate and target models under the 3 types of AE carriers. We can see that the environmental sound carrier achieves better AE transferability than the noise and feature-twisted carriers in terms of the average match rate over the 4 target models. In particular, PT-AEs based on environmental sounds have match rates from 0.23 to 0.27, compared with 0.10 to 0.14 (noise carrier) and 0.15 to 0.22 (feature-twisted carrier). The results demonstrate that using environmental sounds as the carrier achieves the best transferability of PT-AEs from a PT-surrogate model to a target model.

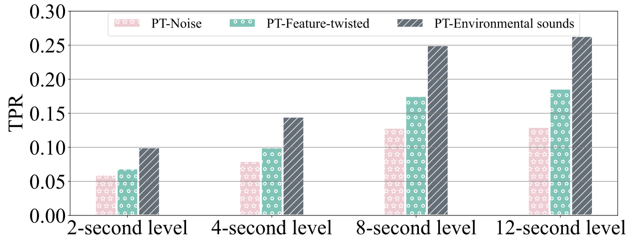

Fig. 6: TPRs of carriers with different attack knowledge levels.

Table IV also compares the match rates of PT-AEs generated from PT models in comparison with GT-AEs generated from GT models. We can observe that the match rate of PT-AEs is slightly lower than their GT counterparts. For example, using the noise carrier, GT-AEs based on GT-TDNN-D achieve the best average match rate of 0.1625; in contrast, PT-AEs based on PT-TDNN-D obtain a slightly lower average match rate of 0.1375. Overall, we can see that PT-AEs are slightly less transferable than GT-AEs, but still effective against target models, especially using the environmental sound carrier.

### *D. Defining Transferability-Perception Ratio for Evaluation*

Now, given an AE carrier type $C \in \{$noise, feature-twisted, environmental sounds$\}$, we have the metrics of SRS$(C)$ and match rate $m(C)$ to measure the perceptual quality and transferability of PT-AEs of type $C$, respectively. We define a joint metric, named Transferability-Perception Ratio (TPR), as

$$\text{TPR}(C) = m(C)/(8 - \text{SRS}(C)), \tag{4}$$

where $8 - \text{SRS}(C)$ ranges from 1 to 7, denoting the score loss to the best human perceptual quality. The resultant value of TPR$(C)$ is in $[0, 1]$ and quantifies, on average, how much transferability (in terms of the match rate) we can obtain by degrading one unit of human perceptual quality (in terms of the SRS). A higher TPR indicates a better AE quality from a joint perspective of transferability and perception.

As the attacker only knows one-sentence speech of her target speaker, the length of the speech (measured by seconds) is an important factor for the attacker to build the PT model and determines the effectiveness of PT-AEs. Fig. 6 shows the TPRs of PT-AEs using the 3 types of carriers under different attack knowledge levels (2, 4, 8, and 12 seconds). It is observed in

Fig. 6 that the TPRs of all AE carriers increase by giving more knowledge about the target speaker's speech. For example, the TPR of the environmental sound carrier increases substantially from 0.14 (4-second level) to 0.25 (8-second level), and then slightly to 0.259 (12-second level).

Note that the environmental sound carrier in all three types has the highest TPR at each knowledge level, which is consistent with the findings in Fig. 5 and Table IV. We also see that the feature-twisted carrier achieves the second-highest TPR, while the noise carrier has the lowest TPR. In summary, our TPR results show that we can base environment sounds to generate PT-AEs to improve their transferability to a black-box target model.

## V. Optimized Black-box PT-AE Attacks

In this section, we propose an optimized PT-AE generation mechanism to attack a black-box target model. We first investigate the TPRs of PT-AEs generated from combined carriers, then formulate a two-stage attack to generate PT-AEs against the target model.

### A. *Combining Carriers for Optimized PT-AEs*

The findings in Fig 6 reveal that the environmental sound carrier achieves the highest TPR and should be a good choice to generate PT-AEs. But using the environmental sound carrier does not exclude us to further twist the auditory feature of the carrier or adding additional noise to it (e.g., an enrollment-phase attack [39] used both environmental sounds and noise). In other words, there is a potential way to combine the environmental sound carrier with feature-twisting or noise-adding method to further improve the TPR.

We consider two additional types of carriers: (i) Feature-twisted environmental sounds, and manipulating the pitch [113] or the rhythm [44] is a straightforward way to twist the features of environmental sounds. We follow the same feature-twisting procedure in Section IV-C2 to twist the pitch and rhythm features of environmental sounds to generate PT-AEs. (ii) Noise-based environmental sounds. We first add environmental sounds to the original speech and then use the noise attack procedure in Section IV-C2 to generate PT-AEs.

Fig. 7 shows the TPRs of various PT-AEs generated based on (i) adding noise to, (ii) twisting the rhythm, and (iii) twisting the pitch of a type of environmental sounds. We can find that the TPR is sensitive to the choice of environmental sounds. For example, the music sounds do not seem very effective to increase the TPRs even with twisted features. It is noted that natural sounds have overall higher TPRs than other types of carriers. For example, using the brook sounds can achieve 0.29 TPR compared with alarm (0.25), rooster (0.26), and Rock2 (0.16) in the existing dataset [47]. Moreover, Fig. 7 illustrates the uniform advantage of twisting the pitch of environmental sound over twisting the rhythm and adding noise. For example, built upon the hail sounds, twisting the pitch feature obtains a TPR of 0.26, substantially higher than twisting the rhythm (0.18) and adding noise (0.05). In addition, Fig. 7 shows that adding noise is the least effective way to improve the TPR. Based on the results in Fig. 7, we consider generating PT-AEs against a black-box target model via twisting the pitch feature of environmental sounds.

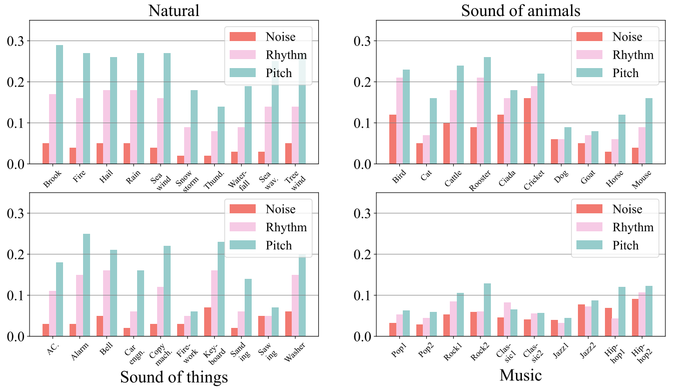

Fig. 7: TPR of different optimized carriers.

### B. *Two-stage Black-box Attack Formulation*

We now formulate the black-box PT-AE attack strategy against a target speaker in a target speaker recognition model. The attack strategy consists of two stages.

In the first stage, the attacker needs to determine a set of candidate environmental sounds as there are a wide range of environmental sounds available and not all of them can be effective against the target speaker (as shown in Figure. 7). To this end, we first build a PT-surrogate model for the attacker, evaluate the TPR of each type of environmental sounds based on the surrogate model, and choose $K$ sounds with the best TPRs to form the candidate set. Then, we pre-process each environmental sound in the candidate set by shifting its pitch to obtain its best TPR, and obtain a new candidate set of $K$ pitch-shifted sounds, denoted by $\{\delta_k\}_{k\in[1,K]}$.

In the second stage, we build additional PT-surrogate models for the attacker. We use the same parrot speech samples generated for the target speaker and speech samples of different other speakers to build each PT model. Denote all $N$ PT-surrogate models as $\{\mathcal{J}_n\}_{n\in[1,N]}$. We employ an ensemble-based method [42], [46], [73], [76], [106], [111], which linearly combines the loss functions of all the surrogate models (i.e., the ensemble loss), to further improve the transferability of PT-AEs. The attack can be formulated as finding the optimal carrier weights $\gamma_k$ for the pitch-twisted candidate set $\{\delta_k\}_{k\in[1,K]}$ to minimize the ensemble loss:

$$\text{Objective:} \quad \arg\min_{\gamma_k} \Sigma_{n=1}^{N} w_n \mathcal{J}_n \left(x + \Sigma_{k=1}^{K}\gamma_k\delta_k, y_t\right) + c\,\text{SRS}\big(x, x + \Sigma_{k=1}^{K}\gamma_k\delta_k\big) \tag{5}$$

$$\text{Subject to:} \quad \Sigma_{k=1}^{K}\gamma_k \le \epsilon \tag{6}$$

where $x$ is the original speech to be perturbed to generate the attack speech; $y_t$ is the target speaker's label; (6) limits the total energy of the AE carrier within the threshold $\epsilon$; and we uniformly set the model weights $w_n = 1/N$. The optimization (5) is a problem to find multiple carrier weights $\{\gamma_k\}$ with a non-differentiable objective function (because of the perception metric of SRS), we adopt the simultaneous perturbation stochastic approximation (SPSA), which employs a gradient estimation method to optimize the large-scale unknown parameters, to solve (5). We set the uniform weight of each surrogate model [76]. To ensure the loss of each surrogate model is in the same range, we convert the cross-entropy loss into a probability via the softmax function. In this way, the loss of each model is in the range of $[0, 1]$.

## VI. Experimental Evaluations

In this section, we measure the impacts of our PT-AE attack in real-world settings. We first describe our setups and then present and discuss experimental results.

### A. Experimental Settings

**The settings of the PT-AE attack:** We select 3 CNN and 3 TDNN models to build $N = 6$ PT models with different parameters for ensembling in (5). Each PT model has the same one-sentence knowledge (8-second speech) of the target speaker, which is selected from the LibriSpeech [87] or VoxCeleb1 [83] datasets. We randomly choose 6-16 speakers from the VCTK dataset as other speakers to build each PT model. We choose $K = 50$ carriers from the 200 environmental sound carriers in [47] to form the candidate set for the attacker and can shift the pitch of a sound up/down by up to 25 semitones. The total energy threshold $\epsilon$ is set to be 0.08.

**Computational cost:** We observe that the ensemble loss in (5) typically converges after 500 steps of updating the carrier weights. However, we find that, like gradient descent, SPSA might not always reach the optimal solution and can get stuck in a local minimum. In addition, the presence of a large number of carrier weights can intensify this issue. To address it, we adopt the strategy from [27], and randomly initialize the weights of carriers $\gamma_k$ 50 times. We then select the carrier weights with the minimal ensemble loss to enhance the transferability of PT-AEs. The maximum computational cost during generating one PT-AE is 25,000 search steps.

**Target speaker recognition systems:** We aim to evaluate the attacks against two major types of speaker recognition systems: i) digital-line evaluations: we directly forward AEs to the open-source systems in the digital audio file format (16-bit PCM WAV) to evaluate the attack impact. ii) over-the-air evaluations: we perform over-the-air attack injections to the real-world smart devices.

**Evaluation metrics:** (i) Attack effectiveness: we use attack success rate (ASR) to evaluate the percentage of AEs that can be successfully recognized as the target speaker in a speaker recognition system. (ii) Perception quality: we evaluate the perception quality of an AE via the metric of SRS.

### B. Evaluations of Digital-line Attacks

**Digital-line setups:** We consider choosing 4 different target models from statistical-based, i.e., GMM-UBM and i-vector-PLDA [6], and DNN-based, i.e., DeepSpeaker [68] and ECAPA-TDNN [41] models. To increase the diversity of target models, we aim to choose 3 males and 3 females from LibriSpeech and VoxCeleb1. For each gender, we randomly select 1 or 2 speakers from LibriSpeech then randomly select the other(s) from VoxCeleb1. We choose around 15-second speech from each speaker to enroll with each speaker recognition model. The performance of each target speaker recognition model is shown in Appendix C.

**Results of digital-line attacks:** In digital-line evaluations, we measure the performance of each attack strategy by generating 240 AEs (40 AEs for each target speaker) against each target speaker recognition model. We separate the results by the intra-gender (i.e., the original speaker whose speech is used for AE generation is the same-gender as the target speaker) and inter-gender scenario (the original and target speakers are not the same-gender, indicating more distinct speech features). We also evaluate the attacks against three tasks: CSI, OSI, and SV.

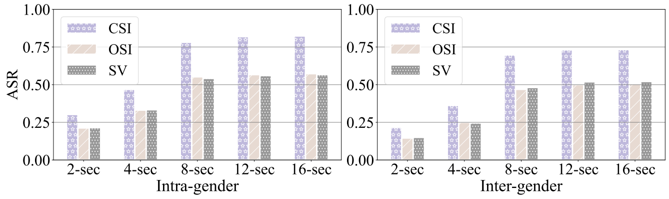

Fig. 8: Evaluation on different attack knowledge levels.

Table V shows the ASRs and SRSs of AEs generated by our PT-AE attack strategy, compared with other attack strategies, against CSI, OSI, and SV tasks. It is noted from Table V that in the intra-gender scenario, the PT-AE attack and QFA2SR (e.g., 60.2% for PT-AE attack and 40.0% for QFA2SR) can achieve higher averaged ASRs (over all three tasks) than other attacks (e.g., 11.3% for FakeBob, 19.2% for Occam, and 29.9% Smack). At the same time, the results of averaged SRS reveal that the perception quality of the PT-AE attack (e.g., 4.1 for PT-AE attack and 3.1 for Smack) is better than other attacks (e.g., 2.3 for QFA2SR, 2.1 for Occam, and 2.9 for FakeBob). In addition, it can be observed that in the inter-gender scenario, the ASRs and SRSs become generally worse. For example, the ASR of FakeBob changes from 11.3% to 6.9% from the intra-gender to inter-gender scenario. But we can see that our PT-AE attack is still effective in terms of both average ASR (e.g., 54.6% for PT-AE attack vs 29.7% for QFA2SR) and average SRS (e.g., 3.9 for PT-AE attack vs 3.2 for Smack). The results in Table V demonstrate that the PT-AE attack is the most effective in achieving both black-box attack success and perceptual quality.

### C. Impacts of Attack Knowledge Levels

**1) Impacts of speech length on attack effectiveness:** By default, we build each PT model in our attack using an 8-second speech sample from the target speaker. We are interested in how the attacker's knowledge affects the PT-AE effectiveness. We assume that the attacker knows the target speaker's speech from 2 to 16 seconds and constructs different PT models based on this varying knowledge to create PT-AEs.

**Results analysis:** Fig. 8 shows the ASRs of PT-AEs under different knowledge levels. We can see that more knowledge can increase the attacker's ASR. When the attack knowledge starts to increase from 2 to 8 seconds, the ASR increases substantially (e.g., 21.3% to 55.2% against OSI in the intra-gender scenario). When it continues to increase to 16 seconds, the ASR exhibits a slight increase. One potential explanation is that the ASR can be influenced by the differences in the architecture and training data between the surrogate and target models. Meanwhile, the one-shot VC method could also reach a performance bottleneck in converting parrot samples using even longer speech. In addition, increasing the speech length does not always indicate the increase of phoneme diversity, which can be also important in speech evaluation [81], [22].

TABLE V: The evaluation of different attacks in digital line.

| | Intra-gender | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tasks | CSI | | | | | | | | OSI | | | | | | | | SV | | | | | | | | | |
| Models | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | | **Average** | |
| Metrics | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | **ASR** | **SRS** |
| FakeBob | 25.8% | 2.9 | 26.7% | 3.6 | 10.6% | 3.2 | 29.2% | 3.0 | 4.2% | 2.9 | 5.8% | 3.1 | 6.7% | 3.2 | 9.2% | 3.1 | 3.0% | 2.8 | 5.8% | 2.6 | 8.3% | 2.7 | 5.8% | 3.2 | **11.3%** | **2.9** |
| Occam | 45.8% | 2.1 | 41.7% | 2.1 | 46.7% | 2.2 | 47.5% | 2.4 | 5.0% | 1.6 | 5.8% | 1.9 | 4.2% | 2.1 | 2.5% | 2.4 | 5.8% | 2.0 | 5.8% | 1.9 | 5.0% | 2.2 | 4.2% | 2.1 | **19.2%** | **2.1** |
| Smack | 74.1% | 3.5 | 45.8% | 2.3 | 44.2% | 3.6 | 48.3% | 3.3 | 10.0% | 3.2 | 13.3% | 3.6 | 9.2% | 3.5 | 8.3% | 2.6 | 12.5% | 3.5 | 13.3% | 3.4 | 11.7% | 2.1 | 9.2% | 2.6 | **29.9%** | **3.1** |
| QFA2SR | 76.7% | 2.2 | 70.8% | 2.4 | 76.7% | 2.1 | 77.5% | 2.1 | 26.7% | 2.8 | 31.7% | 2.3 | 28.3% | 1.9 | 30.0% | 2.1 | 30.8% | 2.3 | 29.2% | 1.9 | 32.5% | 2.6 | 28.3% | 2.5 | **40.0%** | **2.3** |
| PT-AEs | 80.8% | 4.8 | 79.2% | 4.4 | 78.3% | 4.3 | 75.0% | 4.3 | 54.2% | 4.2 | 56.7% | 3.7 | 52.5% | 4.4 | 57.5% | 3.9 | 55.0% | 3.9 | 56.7% | 3.4 | 54.2% | 4.1 | 50.8% | 4.2 | **60.2%** | **4.1** |
| | Inter-gender | | | | | | | | | | | | | | | | | | | | | | | | | |
| Metrics | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | **ASR** | **SRS** |
| FakeBob | 17.5% | 2.9 | 18.3% | 3.6 | 13.3% | 3.0 | 12.5% | 2.3 | 2.5% | 2.9 | 1.7% | 2.7 | 4.2% | 2.6 | 2.5% | 2.4 | 2.5% | 2.1 | 1.7% | 2.8 | 3.3% | 2.7 | 2.5% | 2.9 | **6.9%** | **2.8** |
| Occam | 26.7% | 3.2 | 25.8% | 2.6 | 23.3% | 2.5 | 21.7% | 2.1 | 5.8% | 2.8 | 10.0% | 2.1 | 10.8% | 2.3 | 7.5% | 1.5 | 11.7% | 3.0 | 9.2% | 2.7 | 10.0% | 2.6 | 6.7% | 2.2 | **14.1%** | **2.6** |
| Smack | 21.7% | 3.4 | 26.7% | 3.4 | 19.2% | 3.0 | 17.5% | 3.6 | 12.5% | 3.2 | 14.2% | 3.1 | 13.3% | 2.8 | 15.8% | 2.7 | 11.7% | 3.3 | 15.8% | 3.1 | 14.2% | 2.9 | 15.0% | 2.7 | **16.9%** | **3.2** |
| QFA2SR | 46.7% | 1.9 | 35.8% | 2.2 | 43.3% | 2.6 | 35.8% | 2.4 | 21.7% | 1.5 | 24.2% | 1.6 | 25.8% | 2.6 | 27.5% | 2.8 | 26.7% | 2.1 | 23.3% | 2.3 | 26.7% | 2.4 | 27.5% | 2.2 | **29.7%** | **2.3** |
| PT-AEs | 71.7% | 4.3 | 70.8% | 4.3 | 70.0% | 4.6 | 66.7% | 5.1 | 45.8% | 3.8 | 48.3% | 3.6 | 46.7% | 3.5 | 49.1% | 3.7 | 46.7% | 3.9 | 48.3% | 3.6 | 49.1% | 3.8 | 48.3% | 4.1 | **54.6%** | **3.9** |

Existing studies [75], [104] highlighted that phonemes represent an important feature of the voiceprint to train the VC model. Thus, we aim to explore further how phoneme diversity (in addition to sentence length) can influence the ASR.

**2) Impacts of phoneme diversity:** Since there is no clear, uniform definition for phoneme diversity in previous VC studies [75], [104], we define it as the number of unique phonemes present in a given speech segment. It is worth noting that while some phonemes might appear multiple times in the segment, each is counted only once towards phoneme diversity. This approach is taken because, from an attacker's perspective, unique phonemes are more valuable than repeated ones. While unique phonemes contribute distinct voiceprint features to a VC model, repeated phonemes, can be easily replicated and offer less distinctiveness [75].

To evaluate the impact of phoneme diversity on ASR, we choose speech samples of target speakers that have different phoneme diversities but are of the same length (measured by seconds). From our observations in existing datasets (e.g., LibriSpeech), a shorter speech sample can exhibit a higher phoneme diversity than a longer speech sample. This allows us to select speech samples with significantly different levels of phoneme diversity under the same speech length constraint.

We establish low and high phoneme diversity groups in speech segments of the same length to better understand the impact of phoneme diversity on attack effectiveness. In particular, for each level of speech length (e.g., 8-second) in a dataset, we first rank the speech sample of each target speaker by phoneme diversity, then group the top half of all samples (with high values of phoneme diversity) as the high phoneme diversity group and the bottom half as the low diversity group. In this way, the low phoneme diversity group has fewer distinctive phonemes than the high group, offering enough difference regarding attack knowledge for comparison.

We construct our attack knowledge speech set using the speech samples of 3 male and 3 female speakers from LibriSpeech and VoxCeleb1, consistent with the digital-line setups detailed in Section VI-B. Our goal is to capture various phoneme diversities under different speech lengths. Table VI shows the average phoneme diversity and the total number of phonemes of speech samples in the low and high diversity groups under the same level of speech length (2 to 16 seconds). Table VI demonstrates that the phoneme diversity increases as the speech length increases. Moreover, we find that the phoneme diversity can vary evidently even when the number of total phonemes is similar. For the 8-second category, the low phoneme diversity group has an average diversity of 18.6, while the high diversity group has 24.2. Despite this difference, they have a similar total number of phonemes (80.4 vs 80.6).

TABLE VI: Phoneme diversities with different speech lengths.

| | 2-second | | 4-second | | 8-second | | 12-second | | 16-second | |
|---|---|---|---|---|---|---|---|---|---|---|
| Averaged | Diversity | Total | Diversity | Total | Diversity | Total | Diversity | Total | Diversity | Total |
| Low-diversity | 5.4 | 12.4 | 10.2 | 23.0 | 18.6 | 80.4 | 26.4 | 100.8 | 32.2 | 134.8 |
| High-diversity | 6.4 | 13.2 | 14.6 | 23.4 | 24.2 | 80.6 | 31.4 | 102.0 | 37.4 | 139.4 |

'Diversity' and 'Total' indicate the phoneme diversity and the number of total phonemes, respectively. 'Low-diversity' and 'High-diversity' indicate the groups with low and high phoneme diversities, respectively.

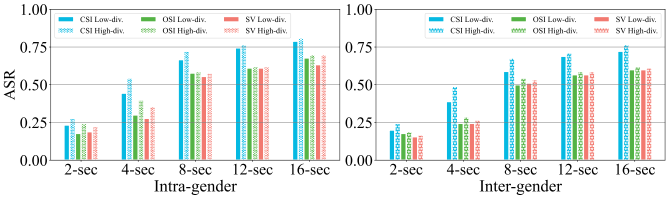

Fig. 9: Evaluation on phoneme diversity.

Then, under each level of speech length (2, 4, 8, 12, 16 seconds) for each target speaker (3 male and 3 female speakers), we use speech samples from the low and high phoneme diversity groups for parrot training and generate 90 PT-AEs from each group. This resulted in a total of 5,400 PT-AEs for the phoneme diversity evaluation.

**Results analysis:** Fig. 9 shows the ASRs of PT-AEs generated from low and high diversity groups against CSI, OSI, and SV tasks. It can be seen from the figure that the high-diversity group-based PT-AEs have a higher ASR than the low-diversity ones in both intra-gender and inter-gender scenarios. For example, the inter-gender ASRs are 47.70% (low-diversity) vs 55.56% (high-diversity). The largest difference in ASR is observed in the 4-second case in the CSI task for the intra-gender scenario, with a maximum difference of 10.0%. The results show that using speech samples with high phoneme diversity for parrot training can indeed improve the attack effectiveness of PT-AEs.

In addition, we calculate via Pearson's coefficients [54] the correlation of the ASR with each of the methods to measure the attack knowledge level, including measuring the speech length, counting the total number of phonemes, and using the phoneme diversity. We find that phoneme diversity achieves the highest Pearson's coefficient of 0.9692 in comparison with using speech length (0.9341) and counting the total number of phonemes (0.9574). As a result, the phoneme diversity for measuring the attack knowledge is the most related to the

TABLE VII: Experimental results on smart devices.

| Intra-gender | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Smart Devices | Methods | FakeBob | | Occam | | Smack | | QFA2SR | | PT-AEs | |
| | Tasks | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS |
| Amazon Echo | OSI | 0/12 | N/A | 1/12 | 1.89 | 2/12 | 4.45 | 3/12 | 2.60 | 7/12 | 4.33 |
| Amazon Echo | SV | 0/12 | N/A | 2/12 | 2.01 | 2/12 | 4.53 | 4/12 | 2.72 | 7/12 | 5.08 |
| Google Home | SV | 0/12 | N/A | 0/12 | N/A | 1/12 | 3.96 | 3/12 | 2.55 | 5/12 | 4.49 |
| Apple HomePod | SV | 2/12 | 2.15 | 3/12 | 3.16 | 3/12 | 5.09 | 5/12 | 3.12 | 9/12 | 5.16 |
| **Average** | **-** | **4.2%** | **2.15** | **12.5%** | **2.35** | **16.7%** | **4.51** | **31.3%** | **2.75** | **58.3%** | **4.77** |
| Inter-gender | | | | | | | | | | | |
| | Tasks | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS |
| Amazon Echo | OSI | 0/12 | N/A | 1/12 | 1.26 | 2/12 | 3.89 | 2/12 | 2.27 | 5/12 | 4.15 |
| Amazon Echo | SV | 0/12 | N/A | 1/12 | 1.35 | 1/12 | 4.12 | 3/12 | 2.03 | 6/12 | 4.27 |
| Google Home | SV | 0/12 | N/A | 0/12 | N/A | 1/12 | 3.11 | 2/12 | 1.92 | 4/12 | 4.53 |
| Apple HomePod | SV | 1/12 | 1.59 | 2/12 | 2.59 | 2/12 | 4.14 | 4/12 | 3.10 | 8/12 | 4.86 |
| **Average** | **-** | **2.1%** | **1.59** | **8.3%** | **1.73** | **12.5%** | **3.82** | **22.9%** | **2.33** | **47.9%** | **4.45** |

TABLE VIII: ASRs with removing each design component.

| | | Amazon-OSI | Amazon-SV | Google-SV | Apple-SV | **Average** |
|---|---|---|---|---|---|---|
| **No removing** | PT-AEs | 50.0% | 54.2% | 37.5% | 70.8% | **53.1%** |
| **1) No PT** | Non-PT AEs | 29.2% | 33.3% | 25.0% | 37.5% | **31.3%** |
| **2) No environmental sound** | Noise | 25.0% | 33.3% | 25.0% | 33.3% | **29.2%** |
| | Featute-twisted | 33.3% | 37.5% | 25.0% | 37.5% | **33.3%** |
| **3) No or insufficient ensemble learning** | Single PT-CNN | 29.2% | 33.3% | 20.8% | 41.7% | **31.3%** |
| | Single PT-TDNN | 29.2% | 37.5% | 20.8% | 41.7% | **32.3%** |
| | Multiple PT-CNN | 41.7% | 45.8% | 29.2% | 58.3% | **43.8%** |
| | Multiple PT-TDNN | 45.8% | 45.8% | 33.3% | 58.3% | **45.8%** |

attack effectiveness, while using the speech length or the total number of phonemes can still be considered adequate as they both have high Pearson's coefficients.

### *D. Evaluations of Over-the-air Attacks*

Next, we focus on attacking the smart devices in the over-the-air scenario. We consider three popular smart devices: Amazon Echo Plus [2], Google Home Mini[12], and Apple HomePod (Siri) [4]. For speaker enrollment, we use 3 male and 3 female speakers from Google's text-to-speech platform to generate the enrollment speech for each device. We only use an 8-second speech from each target speaker to build our PT models. We consider OSI and SV tasks on Amazon Echo, and the SV task on Apple HomePod and Google Home. Similarly, we evaluate the different attacks in both intra-gender and inter-gender scenarios. For each attack strategy, we generate and play 24 AEs using a JBL Clip3 speaker to each smart device with a distance of 0.5 meters.

**Results analysis:** Table VII compares different attack methods against the smart devices under various tasks. We can see that our PT-AE attack can achieve average ASRs of 58.3% (intra-gender) and 47.9% (inter-gender) and at the same time the average SRSs of 4.77 (intra-gender) and 4.45 (inter-gender). By contrast, QFA2SR has the second-best ASRs of 31.3% (intra-gender) and 22.92% (inter-gender); however, it has a substantially lower perception quality compared with the PT-AE attack and Smack, e.g., 2.75 (QFA2SR) vs 4.51 (Smack) vs 4.77 (PT-AE attack) in the intra-gender scenario. We also find that FakeBob and Occam appear to be ineffective with over-the-air injection as zero ASR is observed against Amazon Echo and Google Home. Overall, the over-the-air results demonstrate that the PT-AEs generated by the PT-AE attack can achieve a high ASR with good perceptual quality. Additionally, we also evaluated the robustness of PT-AEs over distance, the results can be found in Table X in Appendix D.

### *E. Contribution of Each Component to ASR*

As the PT-AE generation involves three major design components, including parrot training, choosing carriers, and ensemble learning, to enhance the overall transferability, we propose to evaluate the contribution of each individual component to the ASR. Our methodology is similar to the One-at-a-time (OAT) strategy in [44]. Specifically, we remove and replace each design component with an alternative, baseline approach (as a baseline attack), while maintaining the other settings the same in generating PT-AEs, and then compare the resultant ASR with the ASR of no-removing PT-AEs (i.e., the PT-AEs generated without removing/replacing any design component). Through this method, we can determine how each component contributes to the overall attack effectiveness.

We use the same over-the-air attack setup as described in Section VI-D. For each baseline attack, we craft 96 AEs for both intra and inter-gender scenarios. These AEs are played on each smart device by the same speaker at the same distance. We present the experimental setup and results regarding evaluating the contribution of each design component as follows.

**1) Parrot training:** Rather than training the surrogate models with parrot speech, we directly use the target speaker's one-sentence (8-second) speech for enrollment with the surrogate models. These surrogate models, which we refer to as non-parrot-training (non-PT) models, are trained on the datasets that exclude the target speakers' speech samples.

**Results:** As shown in Table VIII (the "No PT" row), we observe a significant ASR difference between non-PT-based AEs and no-removing PT-AEs. For example, in the Amazon-SV task, PT-AEs achieve an ASR of 54.2%, which is 20.9% higher than the 33.3% ASR of non-PT AEs. Overall, the average ASR for PT-AEs is 21.8% higher than that of non-PT AEs. This substantial performance gap is primarily filled by adopting parrot training.

**2) Environmental sound carrier:** To understand the contribution of the feature-twisted environment sound carrier, we use two baseline attacks related to noise and feature-twisted carriers. i) Noise carriers, we employ the PGD attack to generate the AEs based on the PT models through ensemble learning, setting $\epsilon = 0.05$ to control the $L_\infty$ norm. ii) Feature-twisted carriers, as discussed in Section V-A, we shift the pitch of the original speech up or down by up to 25 semitones to create a pitch-twisted set. We use this set to solve the problem in (5) via finding the optimal weights for the twisted-pitch carriers, with a total energy threshold of $\epsilon = 0.08$.

**Results:** Table VIII (the "no environmental sound" rows) indicates that environmental-sound-based PT-AEs hold a distinct advantage over other carriers in terms of attack effectiveness. We note that when we exclude the feature-twisted environmental sound carriers and rely solely on either the noise or feature-twisted carriers, the average ASR drops by 23.9% (vs. noise carrier) and 19.8% (vs. feature-twisted carrier). These findings show that utilizing feature-twisted environmental sounds can significantly enhance the attack effectiveness.

**3) Ensemble learning:** We note that our ensemble-based model in (5) combines multiple CNN and TDNN models.

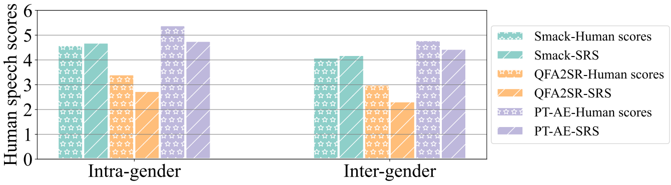

Fig. 10: Human evaluation on the AEs.

To evaluate the contribution of ensemble learning, we design two sets of experiments. First, we replace the ensemble-based model in (5) with just a single PT-CNN or PT-TDNN model to compare the ASRs. Second, we replace (5) with an ensemble-based model, which only consists of multiple (in particular 6 in experiments) surrogate models under the same CNN or TDNN architecture (i.e., no ensembling across different architectures).

**Results:** We can observe in Table VIII (the "no or insufficient ensemble learning" rows) that the single PT-CNN and PT-TDNN models only have average ASRs of 31.3% and 32.3%, respectively. If we do adopt ensemble learning but combine surrogate models under the same architecture, the average ASRs can be improved to 43.8% and 45.8% under multiple PT-CNN and PT-TDNN models, respectively. By contrast, no-removing PT-AEs achieve the highest average ASR of 53.1%.

In summary, the three key design components for PT-AEs, i.e., parrot training, feature-twisted environmental sounds, and ensemble learning, improve the average ASR by 21.8%, 21.9%, and 21.3%, respectively, when compared with their individual baseline replacements. As a result, they are all important towards the black-box attack and have approximately equal contribution to the overall ASR.

### *F. Human Study of AEs Generated in Experiments*

We have used the metric of SRS based on regression prediction built upon the human study in Section IV-B to assess that the PT-AEs have better perceptual quality than AEs generated by other attack methods in experimental evaluations. We now conduct a new round of human study to see whether PT-AEs generated in the experiments are indeed rated better than other AEs by human participants. Specifically, we have recruited additional 45 student volunteers (22 females and 23 males), with ages ranging from 18 to 35. They are all first-time participants and have no knowledge of the previous human study in Section IV-B. Following the same procedure, we ask each volunteer to rate each pair of original and PT-AE samples.

Fig. 10 shows the average human speech scores of Smack, QFA2SR, and our attack. We can see that PT-AEs generated by our attack are rated higher than Smack and QFA2SR. In the intra-gender scenario, the average human score of our attack is 5.39, which is higher than Smack (4.61) and QFA2SR (3.62). The score for each method drops slightly in the inter-gender scenario. The results align with the SRS findings in Table VII. We also find SRS scores are close to human scores. In the inter-gender scenario, SRS predicts our PT-AEs perceptual quality as 4.45, close to the human average of 4.8. The results of Fig. 10 further validates that the PT-AEs have better perceptual quality than AEs generated by other methods.

### *G. Discussions*

**Ethical concerns and responsible disclosure:** Our smart device experiments did not involve any person's private information. All the experiments were set up in our local lab. We have reported our findings to manufacturers (Amazon, Apple, and Google). All manufacturers thanked our research and disclosure efforts aimed at safeguarding their services. Google responded promptly to our investigations, confirming that there is a voice mismatch issue and closed the case as they stated that the attack requires the addition of a malicious node. We are still in communication with Amazon and Apple.

We also discuss potential defense strategies against PT-AEs. Due to the page limit, we have presented the defense discussion in Appendix E.

## VII. Related Work

**White-box attacks:** Adversarial audio attacks [28], [114], [72], [101], [105], [32], [43], [118], [43], [29], [118] can be categorized into white-box and black-box attacks depending on their attack knowledge level. White-box attacks [28], [95] assumed the knowledge of the target model and leveraged the gradient information of the target model to generate highly effective AEs. Some recent studies aimed at improving the practicality of white-box attacks [72], [52] via adding the perturbation to the original speech signal without synchronization, albeit still assuming nearly full knowledge of the target model.

**Query-based black-box attacks:** Existing black-box attacks [29], [118], [101], [105], [74], [113] assumed no access to the internal knowledge of target models, and most black-box attacks attempted to know the target model via a querying (or probing) strategy. The query-based attacks [29], [43], [118], [113], [74] needed to interact with the target model to get the internal prediction scores [29], [105], [32], [113] or hard label results [118], [74]. A large number of queries were necessary for the black-box attack to be effective. For example, Occam [118] needed over 10,000 queries to achieve a high ASR. This makes the attack strategy cumbersome to launch, especially in over-the-air scenarios. The PT-AE attack does not require any probing to the target model.

**Transfer-based black-box attacks:** The transfer-based attacks [17], [44], [30] commonly assumed no interaction or limited probing [32] to the target model. For example, Kenansville [17] manipulated the phoneme of the speech to achieve an untargeted attack. QFA2SR [30] focused on building the surrogate models with specific ensemble strategies to enhance the transferability of AEs by assuming knowing several speech samples of all the enrolled speakers of the target model. Compared with QFA2SR, we further minimize the knowledge and only assume a short speech sample of the target speaker for the attacker. Even with the most limited attack knowledge, we propose a new PT-AE strategy that creates more effective AEs against the target model.

**Audio attacks considering the perception quality:** Some recent studies [95], [52], [74] leveraged the psychoacoustic feature to optimize the carriers and improve the perception of AEs. Meanwhile, [44], [113] manipulated the features of an audio signal to create AEs with good perceptual quality. In addition, there are audio attack strategies [116], [26], [16],

[114] focusing on improving the stealthiness of the AEs. For example, dolphin attack [116] used ultrasounds to generate imperceptible AEs. The human study in this work defines the metric of SRS to quantify the speech quality using a similar regression procedure motivated by the qDev model in [44] that was created to measure the music quality. We then design a new TPR framework built upon the SRS metric to jointly evaluate both the transferability and perception of PT-AEs.

## VIII. Conclusion

In this work, we investigated using the minimum knowledge of a target speaker's speech to attack a black-box target speaker recognition model. We extensively evaluated the feasibility of using state-of-the-art VC methods to generate parrot speech samples to build a PT-surrogate model and the generation methods of PT-AEs. It is shown that PT-AEs can effectively transfer to a black-box target model and the proposed PT-AE attack has achieved higher ASRs and better perceptual quality than existing methods against both digital-line speaker recognition models and commercial smart devices in over-the-air scenarios.

## References

[1] Alexa Voice ID. https://www.amazon.com/gp/help/customer/display.html?nodeId=GYCXKY2AB2QWZT2X/. 2022-12-13.

[2] Amazon Activities. https://www.digitaltrends.com/news/alexa-check-my-balance-amazon-echo-can-now-bank-for-you//. 2023-04-18.

[3] Amazon Alexa. https://developer.amazon.com/en-US/alexa. 2022-01-07.

[4] Apple Siri. https://support.apple.com/en-us/HT204389/. 2022-12-13.

[5] Fidelity-MyVoice. https://www.fidelity.com/security/fidelity-myvoice/overview/. 2022-12-13.

[6] Kaldi. https://github.com/kaldi-asr/kaldi/. 2022-12-13.

[7] Microsoft Azure. https://azure.microsoft.com/en-ca/products/cognitive-services/speech-to-text//. 2023-02-07.

[8] Tencent VPR. https://cloud.tencent.com/product/vpr/. Accessed: 2022-12-13.

[9] AGAIN-VC. https://github.com/KimythAnly/AGAIN-VC/, 2023. Accessed: 2023-01-07.

[10] AutoVC. https://github.com/auspicious3000/autovc/, 2023. Accessed: 2023-01-07.

[11] FreeVC. https://github.com/OlaWod/FreeVC/, 2023. Accessed: 2023-01-07.

[12] Google Home. https://home.google.com/welcome/, 2023. 2023-5-05.

[13] PPG-VC. https://github.com/liusongxiang/ppg-vc/, 2023. Accessed: 2023-01-07.

[14] Semitone. https://en.wikipedia.org/wiki/Semitone/, 2023. Accessed: 2023-04-20.

[15] VQMIVC. https://github.com/Wendison/VQMIVC/, 2023. Accessed: 2023-01-07.

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

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

[18] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1), 2021.

[19] Supraja Anand, Lisa M Kopf, Rahul Shrivastav, and David A Eddins. Objective indices of perceived vocal strain. *Journal of Voice*, 33(6):838–845, 2019.

[20] Yogesh Balaji, Tom Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*, 2019.

[21] Mikołaj Bińkowski, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C Cobo, and Karen Simonyan. High fidelity speech synthesis with adversarial networks. *arXiv preprint arXiv:1909.11646*, 2019.

[22] Shelley B Brundage and N. Ratner. Measurement of stuttering frequency in children's speech. *Journal of Fluency Disorders*, 14:351–358, 1989.

[23] Kate Bunton, Raymond D Kent, Joseph R Duffy, John C Rosenbek, and Jane F Kent. Listener agreement for auditory-perceptual ratings of dysarthria. 2007.

[24] Lei Cai, Hongyang Gao, and Shuiwang Ji. Multi-stage variational auto-encoders for coarse-to-fine image generation. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 630–638. SIAM, 2019.

[25] Qi-Zhi Cai, Min Du, Chang Liu, and Dawn Song. Curriculum adversarial training. *In Proc. of IJCAI*, 2018.

[26] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *Proc. of USENIX Security*, 2016.

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

[28] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *Proc. of SPW*, 2018.

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

[30] Guangke Chen, Yedi Zhang, Zhe Zhao, and Fu Song. Qfa2sr: Query-free adversarial transfer attacks to speaker recognition systems. *arXiv preprint arXiv:2305.14097*, 2023.

[31] Yen-Hao Chen, Da-Yi Wu, Tsung-Han Wu, and Hung-yi Lee. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5954–5958. IEEE, 2021.

[32] Yuxuan Chen, Xuejing Yuan, Jiangshan Zhang, Yue Zhao, Shengzhi Zhang, Kai Chen, and XiaoFeng Wang. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices. In *Proc. of USENIX Security*, 2020.

[33] Zhehuai Chen, Andrew Rosenberg, Yu Zhang, Gary Wang, Bhuvana Ramabhadran, and Pedro J Moreno. Improving speech recognition using gan-based speech synthesis and contrastive unspoken text selection. In *Interspeech*, pages 556–560, 2020.

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

[35] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[36] Frederic L Darley, Arnold E Aronson, and Joe R Brown. Differential diagnostic patterns of dysarthria. *Journal of speech and hearing research*, 12(2):246–269, 1969.

[37] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240, 2006.

[38] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[39] Jiangyi Deng, Yanjiao Chen, and Wenyuan Xu. Fencesitter: Black-box, content-agnostic, and synchronization-free enrollment-phase attacks on speaker recognition systems. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 755–767, 2022.

[40] Jiangyi Deng, Yanjiao Chen, Yinan Zhong, Qianhao Miao, Xueluan Gong, and Wenyuan Xu. Catch you and i can: Revealing source voiceprint against voice conversion. *arXiv preprint arXiv:2302.12434*, 2023.

[41] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*, 2020.

[42] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[43] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, pages 357–369, 2020.

[44] Rui Duan, Zhe Qu, Shangqing Zhao, Leah Ding, Yao Liu, and Zhuo Lu. Perception-aware attack: Creating adversarial music via reverse-engineering human perception. In *Proc. of ACM CCS*, pages 905–919, 2022.

[45] César Ferri, Peter Flach, and José Hernández-Orallo. Learning decision trees using the area under the roc curve. In *Icml*, volume 2, pages 139–146, 2002.

[46] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *European Conference on Computer Vision*. Springer, 2020.

[47] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[48] Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.

[49] Tom Goldstein, Christoph Studer, and Richard Baraniuk. A field guide to forward-backward splitting with a fasta implementation. *arXiv preprint arXiv:1411.3406*, 2014.

[50] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[51] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[52] Hanqing Guo, Yuanda Wang, Nikolay Ivanov, Li Xiao, and Qiben Yan. Specpatch: Human-in-the-loop adversarial audio spectrogram patch attack on speech recognition. 2022.

[53] William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational auto-encoders. *arXiv preprint arXiv:2102.12037*, 2021.

[54] Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011.

[55] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Yuxuan Wang, Yuan Cao, Ye Jia, Zhifeng Chen, Jonathan Shen, et al. Hierarchical generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1810.07217*, 2018.

[56] Wenbin Huang, Wenjuan Tang, Hongbo Jiang, Jun Luo, and Yaoxue Zhang. Stop deceiving! an effective defense scheme against voice impersonation attacks on smart devices. *IEEE Internet of Things Journal*, 9(7):5304–5314, 2021.

[57] Sergey Ioffe. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pages 531–542. Springer, 2006.

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

[59] Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Ijima, Kaoru Hiramatsu, and Kunio Kashino. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4910–4914. IEEE, 2017.

[60] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6820–6824. IEEE, 2019.

[61] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279*, 2019.

[62] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[63] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *16th annual conference of the international speech communication association*, 2015.

[64] Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5220–5224. IEEE, 2017.

[65] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33, 2020.

[66] Alexandru Korotcov, Valery Tkachenko, Daniel P Russo, and Sean Ekins. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Molecular pharmaceutics*, 14(12):4462–4475, 2017.

[67] Kong Aik Lee, Qiongqiong Wang, and Takafumi Koshinaka. The coral+ algorithm for unsupervised domain adaptation of plda. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[68] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

[69] Jason Li, Ravi Gadde, Boris Ginsburg, and Vitaly Lavrukhin. Training neural speech recognition systems with synthetic speech augmentation. *arXiv preprint arXiv:1811.00707*, 2018.

[70] Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion.

[71] Yuanchun Li, Ziqi Zhang, Bingyan Liu, Ziyue Yang, and Yunxin Liu. Modeldiff: testing-based dnn similarity comparison for model reuse detection. In *Proceedings of the 30th ACM SIGSOFT International Symposium on Software Testing and Analysis*, pages 139–151, 2021.

[72] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations. In *Proc. of ACM CCS*, pages 1121–1134, 2020.

[73] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*, 2019.

[74] Han Liu, Zhiyuan Yu, Mingming Zha, XiaoFeng Wang, William Yeoh, Yevgeniy Vorobeychik, and Ning Zhang. When evil calls: Targeted adversarial voice over ip network. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 2009–2023, 2022.

[75] Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng. Any-to-many voice conversion with location-relative sequence-to-sequence modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1717–1728, 2021.

[76] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[77] Hui Lu, Zhiyong Wu, Dongyang Dai, Runnan Li, Shiyin Kang, Jia Jia, and Helen Meng. One-shot voice conversion with global speaker embeddings. In *Interspeech*, pages 669–673, 2019.

[78] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *In Proc. of ICML Work Shop*, 2017.

[79] Yuhao Mao, Chong Fu, Saizhuo Wang, Shouling Ji, Xuhong Zhang, Zhenguang Liu, Jun Zhou, Alex X Liu, Raheem Beyah, and Ting Wang. Transfer attacks revisited: A large-scale empirical study in real computer vision settings. *arXiv preprint arXiv:2204.04063*, 2022.

[80] Agnieszka Mikołajczyk and Michał Grochowski. Data augmentation for improving deep learning in image classification problem. In *2018 international interdisciplinary PhD workshop (IIPhDW)*, pages 117–122. IEEE, 2018.

[81] M. Mines, Barbara F. Hanson, and J. Shoup. Frequency of occurrence of phonemes in conversational english. *Language and Speech*, 21:221 – 241, 1978.

[82] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *arXiv preprint arXiv:2104.09494*, 2021.

[83] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

[84] Preeti Nagrath, Rachna Jain, Agam Madan, Rohan Arora, Piyush Kataria, and Jude Hemanth. Ssdmnv2: A real time dnn-based face mask detection system using single shot multibox detector and mobilenetv2. *Sustainable cities and society*, 66:102692, 2021.

[85] Mahesh Kumar Nandwana, Luciana Ferrer, Mitchell McLaren, Diego Castan, and Aaron Lawson. Analysis of critical metadata factors for the calibration of speaker recognition systems. In *INTERSPEECH*, pages 4325–4329, 2019.

[86] Phani Sankar Nidadavolu, Vicente Iglesias, Jesús Villalba, and Najim Dehak. Investigation on neural bandwidth extension of telephone speech for improved speaker recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6111–6115. IEEE, 2019.

[87] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[88] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[89] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

[90] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[91] Sona Patel, Rahul Shrivastav, and David A Eddins. Perceptual distances of breathy voice quality: A comparison of psychophysical methods. *Journal of Voice*, 24(2):168–177, 2010.

[92] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[93] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[94] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.

[95] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *Proc. of ICML*, pages 5231–5240. PMLR, 2019.

[96] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.

[97] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *In Proc. of NIPS*, 2019.

[98] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[99] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, volume 2017, pages 999–1003, 2017.

[100] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[101] Rohan Taori, Amog Kamsetty, Brenton Chu, and Nikita Vemuri. Targeted adversarial examples for black box audio systems. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 15–20. IEEE, 2019.

[102] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *In Proc. of ICLR*, 2018.

[103] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. 2016.

[104] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng. Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion. *arXiv preprint arXiv:2106.10132*, 2021.

[105] Qian Wang, Baolin Zheng, Qi Li, Chao Shen, and Zhongjie Ba. Towards query-efficient adversarial attacks against automatic speech recognition systems. *IEEE Transactions on Information Forensics and Security*, 16:896–908, 2020.

[106] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.

[107] Emily Wenger, Max Bronckers, Christian Cianfarani, Jenna Cryan, Angela Sha, Haitao Zheng, and Ben Y Zhao. " hello, it's me": Deep learning-based speech synthesis attacks in the real world. In *Proc. of ACM CCS*, pages 235–251, 2021.

[108] M. Wester and R. Karhila. Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5372–5375, 2011.

[109] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

[110] Da-Yi Wu and Hung-yi Lee. One-shot voice conversion by vector quantization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7734–7738. IEEE, 2020.

[111] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.

[112] Zhuolin Yang, Bo Li, Pin-Yu Chen, and Dawn Song. Characterizing audio adversarial examples using temporal dependency. *arXiv preprint arXiv:1809.10875*, 2018.

[113] Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao. Smack: Semantically meaningful adversarial audio attack. 2023.

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

[115] Eiji Yumoto, Wilbur J Gould, and Thomas Baer. Harmonics-to-noise

ratio as an index of the degree of hoarseness. *The journal of the Acoustical Society of America*, 71(6):1544–1550, 1982.

[116] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proc. of ACM CCS*, pages 103–117, 2017.

[117] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE, 2019.

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

# APPENDIX

## A. Speaker Recognition Models

*1) Speaker Recognition Mechanisms:* Speaker recognition models[8], [6], [86], [67] are typically categorized into statistical models, such as Gaussian-Mixture-Model (GMM) based Universal Background Model (UBM) [96] and i-vector probabilistic linear discriminant analysis (PLDA) [38], [85], and deep neural network (DNN) models [68], [41]. There are three phases in speaker recognition.

1) In the training phase, one key component is to extract the acoustic features of speakers, which are commonly represented by the encoded low-dimensional speech features, (e.g., i-vectors [38] and X-vectors [100]). Then, these features can be trained by a classifier (e.g., PLDA [57]) to recognize different speakers.
2) During the enrollment phase, to make the classifier learn a speaker's voice pattern, the speaker usually needs to deliver several text-dependent (e.g., Siri [4] and Amazon Echo [1]) or text-independent speech samples to the speaker recognition system. Depending on the number of enrolled speakers, speaker recognition tasks [29], [118], [113] can be (i) multiple-speaker-based speaker identification (SI) or (ii) single-speaker-based speaker verification (SV).
3) In the recognition phase, the speaker recognition model will predict the speaker's label or output a rejection result based on the similarity threshold. Specifically, SI can be divided into close-set identification (CSI) and open-set identification (OSI) [39], [29]. The former predicts the speaker's label with the highest similarity score, and the latter only outputs a prediction when the similarity score is above the similarity threshold or gives a rejection decision otherwise. SV only focuses on identifying one specific speaker. If the similarity exceeds a predetermined similarity threshold, SV returns an accepted decision. Otherwise, it will return a rejection decision.

*2) Speaker Recognition Formulations:* Let $y_i$ denote the $i$-th speaker enrolled in group set $\mathcal{Y}$, where $\mathcal{Y} = \{y_1, y_2, \cdots, y_i\}$. Let $S(x, y_i)$ represent the similarity score function which takes the test speech signal $x$ as the input and outputs the similarity score based on the enrolled speaker $y_i \in \mathcal{Y}$.

- **CSI:** The CSI task assumes the test speech $x$ always belongs to a speaker in $\mathcal{Y}$, and there is no outsider speaking. The classification function of CSI $f_{\text{CSI}}(x)$ will output the speaker's label with the highest similarity score, i.e.,

$$f_{\text{CSI}}(x) = \arg\max_{y_i \in \mathcal{Y}} S(x, y_i).$$

- **OSI:** Different from the CSI task, OSI is able to judge whether the test speech $x$ belongs to $\mathcal{Y}$ or not. And its classification function $f_{\text{OSI}}(x)$ only outputs a speaker's label when the highest similarity score exceeds the threshold $\theta$.

$$f_{\text{OSI}}(x) = \begin{cases} \arg\max\limits_{y_i \in \mathcal{Y}} S(x, y_i), & \text{if } \max\limits_{y_i \in \mathcal{Y}} S(x, y_i) \geq \theta_{\text{OSI}}, \\ \text{Reject}, & \text{otherwise}, \end{cases}$$

where $\theta_{\text{OSI}}$ is the similarity threshold to reject in OSI.

- **SV:** The enrollment set of SV is only one speaker $y_1$ but not multiple speakers, and it also requires the similarity score greater than the threshold.

$$f_{\text{SV}}(x) = \begin{cases} \text{Accept}, & \text{if } S(x, y_1) \geq \theta_{\text{SV}}, \\ \text{Reject}, & \text{otherwise}, \end{cases}$$

where $\theta_{\text{SV}}$ is the threshold to accept or reject in SV.

## B. Comparison of PT and GT Models

**Constructing PT models:** There are multiple ways to set up and compare PT and GT models. We set up the models based on our black-box attack scenario, in which the attacker knows that the target speaker is trained in a speaker recognition model but does not know other speakers in the model. We first build a GT model using multiple speakers' speech samples, including the target speaker's. To build a PT model for the attacker, we start from the only information that the attacker is assumed to know (i.e., a short speech sample of the target speaker), and use it to generate different parrot speech samples. Then, we use these parrot samples, along with speech samples from a small set of speakers (different from the ones used in the GT model) in an open-source dataset, to build a PT model.

We use CNN and TDNN to build two GT models, called CNN-GT and TDNN-GT, respectively. Each GT model is trained with 6 speakers (labeled from 1 to 6) from LibriSpeech (90 speech samples for training and 30 samples for testing for each speaker). We build 6 CNN-based PT models, called CNN-PT-$i$, and 6 TDNN-based PT models, called TDNN-PT-$i$, where $i$ ranges from 1 to 6 and indicates that the attacker's targets speaker $i$ in the GT model and uses only one of his/her speech samples to generate parrot samples, which are used together with samples from other 3 to 8 speakers randomly selected from VCTK (none is in the GT models), to train a PT model.

**Evaluation metrics:** We aim to compare the 12 PT models with the 2 GT models when recognizing the attacker's target speaker. Existing studies [71], [66] have investigated how to compare different machine learning models via the classification outputs. We follow the common strategy and validate whether PT models have the performance similar to GT models via common classification metrics, including Recall [37], Precision [45], and F1-Score [84], where Recall measures the percentage of correctly predicted target speech samples out of the total actual target samples, Precision measures the proportion of the speech which is predicted as the target label

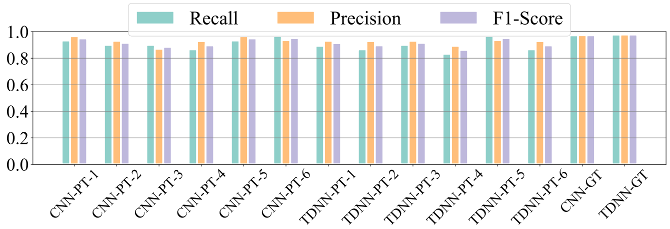

Fig. 11: Comparison of PT and GT models.

TABLE IX: Performance of speaker recognition systems.

| Task | CSI | OSI | | | SV | |
|---|---|---|---|---|---|---|
| | Accuracy | FAR | FRR | OSIER | FAR | FRR |
| DeepSpeaker | 98.89% | 11.42% | 1.11% | 0.83% | 6.96% | 0.41% |
| ECAPA-TDNN | 99.58% | 9.74% | 0.42% | 0.03% | 4.87% | 0.42% |
| GMM-UBM | 99.44% | 10.72% | 5.15% | 2.65% | 10.02% | 5.01% |
| i-vector-PLDA | 99.72% | 7.93% | 2.36% | 0.27% | 12.25% | 0.97% |

indeed belongs to the target speaker, and F1-Score provides a balanced measure of a model's performance which is the harmonic mean of the Recall and Precision. To test each PT model (targeting speaker $i$) and measure the output metrics compared with GT models, we use 30 ground-truth speech samples of speaker $i$ from LibriSpeech and 30 samples of every other speaker from VCTK in the PT model.

**Results analysis and discussion:** Fig. 11 shows the classification performance of PT and GT models. It is observed from the figure that CNN-GT/TDNN-GT achieves the highest Recall, Precision, and F1-Score, which range from 0.97 to 0.98. We can also see that most PT models have slightly lower yet similar classification performance as the GT models. For example, CNN-PT-1 has similar performance to TDNN-GT (Recall: 0.93 vs 0.98; Precision: 0.96 vs 0.98; F1-Score 0.95 vs 0.98). The results indicate that a PT model, just built upon one speech sample of the target speaker, can still recognize most speech samples from the target speaker, and also reliably reject to label other speakers as the target speaker at the same time. The worst-performing model TDNN-PT-4 achieves a Recall of 0.82 and a Precision of 0.86, which is still acceptable to recognize the target speaker. Overall, we note that the PT models can achieve similar classification performance compared with the GT models. Based on the findings, we are motivated to use a PT model to approximate a GT model in generating AEs, and aim to further explore whether PT-AEs are effective to transfer to a black-box GT model.

### *C. Performance of Digital-line Speaker Recognition Models*

Table IX shows the performance of the target speaker recognition models, where accuracy indicates the percentage of speech samples that are correctly labeled by a model in the CSI task; False Acceptance Rate (FAR) is the percentage of speech samples that belong to unenrolled speakers but are accepted as enrolled speakers; False Rejection Rate (FRR) is the percentage of samples that belong to an enrolled speaker but are rejected; Open-set Identification Error Rate (OSIER) is the equal error rate of OSI-False-Acceptance and OSI-False-Rejection.

TABLE X: Evaluation of different distances.

| Attack Scenarios | Smart Devices | Distance | 0.25 (m) | 0.5 (m) | 1.0 (m) | 2.0 (m) | 4.0 (m) |
|---|---|---|---|---|---|---|---|
| Intra-gender | Amazon Echo | OSI | 58.3% | 58.3% | 41.7% | 25.0% | 16.7% |
| | Amazon Echo | SV | 58.3% | 58.3% | 50.0% | 33.3% | 16.7% |
| | Google Home | SV | 50.0% | 41.7% | 41.7% | 25.0% | 16.7% |
| | Apple HomePod | SV | 75.0% | 75.0% | 75.0% | 58.3% | 33.3% |
| | **Average** | - | **60.4%** | **58.3%** | **52.1%** | **35.4%** | **20.8%** |
| Inter-gender | Amazon Echo | OSI | 41.7% | 41.7% | 25.0% | 16.7% | 8.3% |
| | Amazon Echo | SV | 50.0% | 50.0% | 33.3% | 25.0% | 16.7% |
| | Google Home | SV | 33.3% | 33.3% | 25.0% | 16.7% | 8.3% |
| | Apple HomePod | SV | 66.7% | 66.7% | 66.7% | 50.0% | 25.0% |
| | **Average** | - | **47.9%** | **47.9%** | **37.5%** | **27.1%** | **14.5%** |

### *D. Robustness of PT-AEs over Distance*

We aim to further evaluate the robustness of the PT-AE attack in the over-the-air scenario with different distances from the attacker to the target. We set different levels of distance between the attacker (i.e., the JBL Clip3 speaker) and a smart device from 0.25 to 4 meters. The results in Table X show that the ASR of the PT-AE attack changes over the distance. In particular, we can see that there is no significant degradation of ASR when the distance goes from 0.25 to 0.5 meters as the ASR slightly decreases from 60.4% to 58.3% in the inter-gender scenario. There is an evident degradation in ASR when the distance increases from 2.0 to 4.0 meters (e.g., 27.1% to 14.5% in the inter-gender scenario). This is due to the energy degradation of PT-AEs when they propagate over the air to the target device. Overall, PT-AEs are quite effective within 2.0 meters given the perturbation energy threshold of $\epsilon = 0.08$ set for all experiments.

### *E. Discussion on Defense*

**Potential defense designs:** To combat PT-AEs, there are two major defense directions available: (i) audio signal processing and (ii) adversarial training. Audio signal processing has been proposed to defend against AEs via down-sampling [74], [118], quantization [112], and low-pass filtering [72] to preserve the major frequency components of the original signal while filtering out other components to make AEs ineffective. These signal processing methods may be effective when dealing with the noise carrier [118], [72], [52], but are not readily used to filter out PT-AEs based on environment sounds, many of which have similar frequency ranges as human speech. Adversarial training [51], [78], [20], [25], [97], [102], [109] is one of the most popular methods to combat AEs. The key idea behind adversarial training is to repeatedly re-train a target model using the worst-case AEs to make the model more robust. One essential factor in adversarial training is the algorithm used to generate these AEs for training. For example, recent work [118] employed the PGD attack to generate AEs for adversarial training, and the model becomes robust to the noise-carrier-based AEs. The PT-AEs used in this paper adopt feature-twisted environmental sounds as the carrier. Thus, one potential way for defense is to generate enough AEs that cover a diversity of carriers and varying auditory features for training. Significant designs and evaluations are needed to find optimal algorithms to generate and train AEs to fortify a target model.