June 2024

# Advancing Adversarial Audio: Human-in-the-Loop Black-box Attacks

Rui Duan
*University of South Florida*

# Advancing Adversarial Audio: Human-in-the-Loop Black-box Attacks

by

Rui Duan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Electrical Engineering
College of Engineering
University of South Florida

Major Professor: Zhuo Lu, Ph.D.
Ismail Uysal, Ph.D.
Nasir Ghani, Ph.D.
Leah Ding, Ph.D.
Xinming Ou, Ph.D.

Date of Approval:
June 13, 2024

Keywords: Machine Learning, Software and Application Security, Music Copyright, Speaker Recognition, Internet of Things

## Dedication

This dissertation is dedicated to my esteemed supervisors, Dr. Zhuo Lu and Dr. Leah Ding, whose invaluable guidance and insightful feedback have been instrumental in shaping this work. To my brilliant and supportive fiancé, Hongxiang Yang, thank you for your unwavering encouragement and belief in me. Your steadfast support has been my anchor. I also extend my deepest gratitude to my loving parents, Jingjun Cai, and Jingfeng Duan, whose unconditional support and sacrifices have made this journey possible. Your constant love and encouragement have been my driving force. This achievement is as much yours as it is mine.

## Acknowledgments

First and foremost, I would like to express my highest respect and deepest appreciation to my advisor, Dr. Zhuo Lu. His selfless dedication and unwavering belief in my potential at the beginning of my PhD career were fundamental to my progress. Throughout the mid-stage of my PhD journey, his patience and understanding, especially during times when I was hindered by illness, provided me with the care and support I needed. During the most challenging periods of my PhD, his unparalleled academic support and encouragement were invaluable. Dr. Lu's contributions and assistance have significantly shaped my entire PhD journey.

I am also profoundly grateful to Dr. Leah Ding for her insightful guidance and continuous encouragement. Her expertise and advice have greatly contributed to the quality of my research. Her meticulous and conscientious attitude deeply influenced me, especially when I first started my research in machine learning.

My heartfelt thanks go to my colleagues and friends who have made this journey collaborative and enjoyable. Your intellectual companionship and moral support have been indispensable. Specifically, I have to thanks my research partner Zhe Qu, Shangqing Zhao, Zhengping Luo, and my colleagues Tao Hou, Keyu Chen, Mingchen Li, Hung Nguyen, Jiahao Xue, Wenwei Zhao, Xiaowen Liu, Kuan-Hsun Chou, Minh Nguyen, Yuwen Cui, Xiao Han, Junjie Xiong, Chi Zhang, Changjia Zhu, and Haiyun Liu.

Lastly, I extend my deepest gratitude to my family for their unconditional love and support. To my parents, Jingjun Cai and Jingfeng Duan, your unwavering belief in me and your sacrifices have been the bedrock of my achievements. To my brilliant fiancé, Hongxiang Yang, thank you for your patience, understanding, and steadfast encouragement throughout this journey.

# Table of Contents

## List of Tables

## List of Figures

# Abstract

Adversarial audio attacks pose significant security challenges to real-world audio applications. Attackers may manipulate speech to impersonate a speaker, gaining access to smart devices like Amazon Echo. In audio applications, there are two key areas: music and speech. In music, most attackers create a small noise-like perturbation on the original signal to evade copyright detection. However, this method degrades music's perceived quality for human listeners. In the speech, creating an adversarial example often requires many queries to the target model, a process too cumbersome for practical use in real-world scenarios, like interacting with smart devices numerous times.

In this dissertation, we first explore the integration of human factors into adversarial attack loops. Specifically, we conduct a human study to understand how participants perceive perturbations in music signals. Using regression analysis, we model the relationship between audio feature deviations and human-perceived deviations. Based on this human perception model, we propose, formulate, and evaluate a perception-aware attack framework for creating adversarial music.

Considering the black-box audio attack, we investigate adversarial attacks on real-world speaker recognition models using limited practical knowledge. We introduce the concept of the Parrot training model and utilize state-of-the-art voice conversion methods to generate parrot speech samples, enabling the construction of a surrogate model with knowledge of only a single sentence from the target speaker. We propose a two-stage PT-AE attack strategy that demonstrates greater effectiveness than existing strategies while minimizing the required attack knowledge.

# Chapter 1: Introduction

Recently, adversarial machine learning attacks have posed serious security threats against practical audio signal classification systems, including speech recognition, speaker recognition, and music copyright detection. Previous studies have mainly focused on ensuring the effectiveness of attacking an audio signal classifier via creating a small noise-like perturbation on the original signal. It is still unclear if an attacker is able to create audio signal perturbations that can be well perceived by human beings in addition to its attack effectiveness. This is particularly important for music signals as they are carefully crafted with human-enjoyable audio characteristics.

The adversarial attack wants to find a small perturbation that can be injected into the original input, causing the AI system to predict an incorrect label. However, most adversarial audio attacks have primarily concentrated on improving the effectiveness of attacking an audio signal classifier by introducing small, noise-like perturbations to the original audio signal. Our motivation is that can an attacker create an adversarial audio that is not only effective in spoofing the classification system but is also perceptible to human beings. This question becomes particularly significant in the context of music signals as they are carefully crafted with human-enjoyable audio characteristics. Meanwhile, it is interesting to see whether a piece of adversarial music with good quality can still bypass copyright detection systems (e.g., on YouTube).

## 1.1 Human-in-the-loop Adversarial Attack

### 1.1.1 Perception-aware Attack on Music Copyright Systems

Adversarial machine learning attacks, originated from the image domain [70, 94, 35, 148], have recently become a serious security issue in audio signal processing system designs leveraging machine learning, including speech recognition [34, 172, 41, 132, 140], speaker identification [38, 18], and music copyright detection [137].

Adversarial machine learning attacks attempt to create a small perturbation on the original audio signal such that a machine learning classifier can yield an incorrect output. For example, a small change in a speech command could make Amazon Echo [8] and Google assistant [9] recognize a different, yet malicious command [41, 176]. And manipulating copyrighted music might bypass the copyright detection in YouTube [137]. One key component in adversarial audio signals is the perturbation, which is designed to cause misclassification and at the same time be small enough to be hardly noticed. To quantify the perturbation, existing studies [38, 102] usually use a mathematical distance (e.g., the Euclidean distance [137], or more generally, the $L_p$ norm [35]) between the original and perturbed audio signals. As a result, the perturbed signal with the minimized distance to the original one could be considered as a good candidate under the constraint that it can successfully spoof the classifier.

However, the $L_p$ norm based methods only measure the magnitude distance between two signals; but the human perception is much more complex than computing the magnitude distance. There exists a gap between the mathematical distance and the eventual human perception. Although the two may be related in some way (e.g., zero distance meaning no signal perturbation), there is still no direct relation to indicate an increase or decrease of the distance in mathematics would be human-perceived as the same. For example, adding a perturbation that is the same as the original music signal is equivalent to increasing the volume of the music, which does not quite change the human perception of music quality.

Indeed, a few studies [35, 132] have pointed out similar issues and indicated that new methods are needed to measure the perceptual similarity between the original and perturbed signals; but there is limited work on systematically designing adversarial machine learning from the human perception perspective.

In this work, we create a new mechanism to craft adversarial audio signals. We focus on generating adversarial music signals to bypass a music copyright detector and hardly raise human attention. To this end, we formulate the relationship between signal perturbation and human perception with two key steps: 1) quantifying the change of human perception with respect to the change of a music signal; and 2) finding a new way to generate perturbations to minimize the change in human perception and fool a classifier.

To study how a change of a music signal affects human perception, we first conduct a human study where volunteers quantify their perceived deviations between the original and perturbed signals as ratings on a Likert scale [150]. We use regression analysis to build an approximate mathematical relation between the change of music and the human-perceived deviation rating obtained from the human study. Given a perturbed signal, we use the regressed model to predict the human rating on the perceived deviation. We call this output quantified deviation (qDev).

We then reformulate adversarial machine learning for music signals as a perception-aware attack problem of finding a perturbation that minimizes its qDev while misleading a target classifier. The reformulation, however, leads to a computationally intractable optimization with a non-convex and non-differentiable objective function. To solve this problem, we propose a method by reducing the search space for finding a feasible solution. We observe that a common process in music classification is to identify and extract audio fingerprints (e.g., high energy values on certain frequencies) from a signal's spectrogram [158, 33, 126]. Creating a perturbation may introduce additional frequencies and energy values, which will generate new fingerprints different from the original signal. Such difference can be used to fool the target classifier. Meanwhile, to make the perturbation less noticeable to humans, our proposed perception-aware attack is designed to create new frequencies and energy values as a perturbation to minimize the qDev metric. We show that the perception-aware attack can produce adversarial music more effectively in terms of attack success rate and human-perceived quality. We test our perception-aware attack on different genres of music against YouTube's copyright detection. Experimental results show that the perception-aware attack can produce effective adversarial music to bypass YouTube's detection while achieving a significantly higher perceptual quality compared to a recent $L_p$ norm based attack [137].

## 1.2 Practical Limited-Knowledge Adversarial Attack

### 1.2.1 Black-box Attacks on Speaker Recognition Models

Adversarial speech attacks against speech recognition [36, 172, 102, 149, 160, 41, 57, 176] and speaker recognition [57, 38, 176] 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 [36, 172, 102, 71] or black-box settings [160, 41, 57, 176, 38, 104, 18]. 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 [38, 176]. 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 [41, 57, 38] or the final output results [176, 171] of a classifier. The probing process usually requires a large number of interactions (e.g., over 1000 queries [171]), 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) or online commercial platforms (e.g., Microsoft Azure [6]). However, it can be even more cumbersome, if not possible, to probe physical devices because today's smart devices (e.g., Amazon Echo [8]) 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 [38, 171]). Two recent studies further limited the attacker's knowledge [176] to be only knowing the target speaker's one-sentence speech [176] and requiring probing to get the target model's hard-label (accept or reject) results (e.g., over 10,000 times) and recent work [39] only knowing one-sentence speech for each speaker enrolled in the target model.

In this work, 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 attacker's knowledge is the most restricted compared with prior work (in particular with the two recent attacks [176, 39]).

Centered around this one-sentence knowledge of the target speaker, our basic attack framework is to 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 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 [43, 108, 167, 40] 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 [110, 106]. 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.

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: noise carriers, which are the results of traditional methods [38, 176] during their search for the perturbation signals in the unrestricted $L_p$ space. Feature-twisted carriers that are perturbation signals generated by only varying the auditory features of the original signal itself [171, 58, 18, 39], environmental sound carriers that are produced by environmental sounds [53]. Based on the built PT model, we create and evaluate PT-AEs based on these three types of carriers.

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: transferability of PT-AEs to a black-box target model.

We adopt the match rate, which has been comprehensively studied in the image domain [110], 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. 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 [63, 164, 28, 21, 128, 47], 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 1.1 Summary of common attack strategies.

| Attack Strategy | Attack Scenario | Queries Needed | Knowledge Required | Human Perception |
|---|---|---|---|---|
| Carlini et al.[36] | White-box | ∼1000 | gradient info | ✗ |
| CommanderSong[172] | White-box | ∼100 | gradient info | ✗ |
| Psychoacoustic[132] | White-box | ∼5000 | gradient info | ✓ |
| AdvPulse[102] | White-box | ∼2000 | gradient info | ✗ |
| SpecPatch[71] | White-box | ∼1000 | gradient info | ✓ |
| Taori et al.[149] | Black-box | ∼300,000 | soft label | ✗ |
| SGEA[160] | Black-box | ∼300,000 | soft label | ✗ |
| Devil's Whisper[41] | Black-box | ∼1500 | soft label | ✗ |
| FakeBob[38] | Black-box | ∼5000 | soft label | ✗ |
| OCCAM[176] | Black-box | ∼10,000 | hard label | ✗ |
| TAINT[104] | Black-box | ∼1500 | hard label | ✓ |
| SMACK[171] | Black-box | ∼1000 | soft label | ✓ |
| QFA2SR [39] | Black-box | 0 | each speaker's sample | ✗ |
| PT-AE attack | Black-box | 0 | target speaker's sample | ✓ |

Queries: indicating the typical number of probes need to interact with the black-box target model. Soft level: the confidence score [41] or prediction score [149, 160, 41, 38, 171] from the target model. Hard label: accept or reject result [176, 104] from the target model. (iv) QFA2SR [39] 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 [106] 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 [171] and QFA2SR [39], 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 1.1 provides a comparison of the required knowledge between the proposed PT-AE attack and existing strategies.

## 1.3 Dissertation Overview

In Chapter 2, we provide a novel attack vector that opens a new avenue for generating adversarial examples (AEs) by integrating human factors into the attack design, and this approach is more like leveraging AI to better defeat both humans and AI systems. The results are as expected: Our AEs can improve music quality by over 80attack success rate as compared to existing representative works. Such an effective adversarial attack should receive more attention, as these music AEs can harm revenue (e.g., monetization from advertisements), and we are pushing for the development of more practical and robust AI systems to protect the rights of each copyright owner.

In Chapter 3, we explore the process of pushing the practicality of black-box attacks on the speaker recognition models. We will investigate how to use just a few seconds of speech knowledge to reproduce speech samples that sound like the target speaker and use these reproduced speeches can be used as the training dataset for our surrogate models. It is also intriguing to investigate how synthetic speech can enhance the transferability of AEs. Therefore, we built our surrogate model with synthetic data, and we also systematically evaluated the AEs of exiting speech attack methods and explored how to generate audio AEs with considerations for both transferability and human perception. Finally, the results show that our strategy against smart devices (e.g. Amazon Echo, Apple Homepod, and Google Home) achieves improvements of 95.9% in attack success and 44.9% in speech quality over existing works. This research work also exposes the practical black-box attack on real-world AI applications. In Chapter 4, we conclude the dissertation and discuss future work.

# Chapter 2: Creating Adversarial Music via Reverse-Engineering Human Perception

The perception-aware attack [1] investigates integrating the human factors in the adversarial attack loops to generate the adversarial examples with high audio quality.

## 2.1 Abstract

In this work, we formulate the adversarial attack against music signals as a new perception-aware attack framework, which integrates human study into adversarial attack design. Specifically, we conduct a human study to quantify the human perception with respect to a change of a music signal. We invite human participants to rate their perceived deviation based on pairs of original and perturbed music signals, and reverse-engineer the human perception process by regression analysis to predict the human-perceived deviation given a perturbed signal. The perception-aware attack is then formulated as an optimization problem that finds an optimal perturbation signal to minimize the prediction of perceived deviation from the regressed human perception model. We use the perception-aware framework to design a realistic adversarial music attack against YouTube's copyright detector[2]. Experiments show that the perception-aware attack produces adversarial music with significantly better perceptual quality than prior work.

[1] This chapter was published in ACM Conference on Computer and Communications Security (CCS) 2022. Permission is included in Appendix A

[2] Here we provide an anonymous YouTube link of the demos of our and prior attacks (https://www.youtube.com/watch?v=jK3ejLtx750)

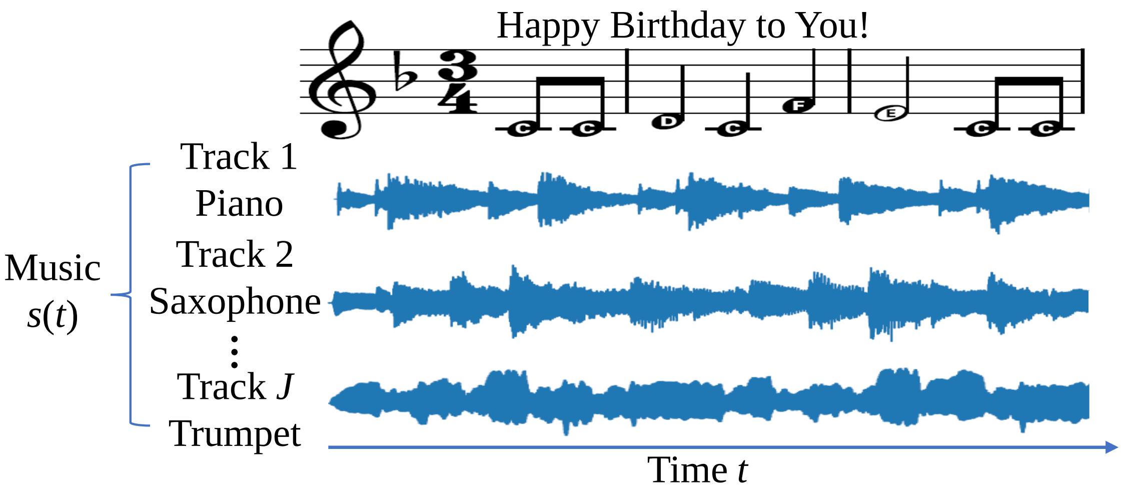


Figure 2.1 Music with multiple track signals.

## 2.2 Background of Music Signal and Adversarial Attacks

In this section, we briefly introduce the background and describe our motivation and design intuition.

### 2.2.1 Representation of Music Signal

In Fig. 2.1, a digital music signal $s(t)$ at sample time $t \in \{0, 1, 2, \cdots, T\}$ (where $T$ is the number of signal samples) can be represented as the sum of audio track signals [153], i.e., $s(t) = \sum_{j=1}^{J} s_j(t)$, where $J$ is the number of tracks, and the track signal $s_j(t)$ is a time-series of harmonic notes [115, 116, 87, 135]. A note, similar to a phoneme of speech [174, 172], is the smallest signal unit of a piece of music consisting of a fundamental frequency and a set of harmonics [65, 66, 157].

### 2.2.2 Adversarial Audio Attacks

Given a classifier with prediction function $f(\cdot)$ which takes the input audio signal $s(t)$ and outputs the correct label $f(s(t)) = y$, existing adversarial audio attacks [36, 169, 132] aim to add a small signal perturbation $\delta(t)$ to the original audio signal $s(t)$, and then supply the perturbed signal $\hat{s}(t) = s(t) + \delta(t)$ to the classifier that accordingly generates an incorrect label. The method of creating $\delta(t)$, which mainly inherits from the fundamental framework in the image domain [35, 148], can be formulated as

$$\begin{aligned} \text{minimize} \quad & \|\delta(t)\|_p \\ \text{subject to} \quad & f(\hat{s}(t)) \neq y, \end{aligned} \tag{2.1}$$

where $\|\delta(t)\|_p$ denotes the $L_p$ norm of the perturbation $\delta(t)$ [35, 70]. The objective of (2.1) is to minimize the change of the perturbed signal $\hat{s}(t)$ from the original $s(t)$. Since it is computationally difficult to solve (2.1), many variants of formulating the adversarial audio attacks have been proposed for distinct attack scenarios, such as speech recognition [36, 169, 132], speaker recognition [38, 176], and music copyright detection [137]. To still make $\hat{s}(t)$ look like $s(t)$, these formulations limit the $L_p$ norm of the perturbation $\delta(t)$ within a given threshold $\epsilon$, i.e., $\|\delta(t)\|_p \leq \epsilon$.

The $L_\infty$, $L_2$, and $L_0$ norms are commonly adopted in the literature to create adversarial attacks targeting various audio signal classifiers [36, 102, 176, 38, 93].

### 2.2.3 Motivation and Design Intuition

Although existing adversarial audio attacks mathematically limit the magnitude of the perturbation $\delta(t)$ via $\|\delta(t)\|_p \leq \epsilon$, it is still not clear whether such a constraint is the most effective to make the perturbation unnoticeable by human beings. For example, a few studies [35, 132] have noted the concern on whether the $L_p$ norm metric is appropriate to measure the signal similarity from the human perception perspective. In other words, there is no evidence to show that the deviation in human cognition can be represented by $\|\delta(t)\|_p$. As a result, we are motivated to investigate the problem. Our goals are twofold: 1) relating the change of a music signal to the deviation of human perception and 2) finding a new way to create the perturbation that is unnoticeable by human beings as much as possible. To achieve these goals, our design consists of three major components.

1) Reverse-engineering human perception of signal deviation, we treat human perception as a black box and design a human study to quantify human perceived deviations. Specifically, we invite volunteers to assign a rating of perceived deviation to measure the difference between the original and perturbed signals. Then, we reverse-engineer the black box via regression analysis to build a relationship between the signal deviation and the human-perceived deviation.

2) Reformulating the adversarial audio attack as the perception-aware attack, based on the relationship found in the human study, we establish the perception-aware attack framework with the objective to quantitatively minimize the perceived deviation while attacking audio classification.

3) Demonstrating a realistic attack against a music copyright detector, based on the new attack framework, we create adversarial music against YouTube's copyright detector. We demonstrate via experiments the effectiveness of the attack in terms of success rate and human-perceived deviation.

### 2.2.4 Threat Model

We consider an attacker that aims to find a perturbation $\delta(t)$ to a music signal $s(t)$ such that $\hat{s}(t) = s(t) + \delta(t)$ leads to an incorrect output of an audio signal classifier, which is similar to the goal of existing audio attacks [148, 36, 169, 102, 176, 137]. At the same time, the attacker is designed to be aware of how $\hat{s}(t)$ affects the human perception and minimizes its perceived deviation from $s(t)$. We assume that the attacker has no knowledge of the algorithm design or parameter choices in the classifier, but has access to the classification result of any input signal. We also assume that the attacker has no access to the classifier's training database. A representative commercial scenario is that an attacker wants to bypass YouTube's copyright detector [137] and use copyrighted music content in an unauthorized way to attract more online views for advertisement revenue gain.

## 2.3 Reverse-Engineering Human Perception of Music Signals

In this section, we present how to quantify the human perceived deviation of music signals. We first analyze the key features for the signal quality, then conduct the human study, and lastly present the study results and regression analysis.

### 2.3.1 Audio Features for Human Perception

Based on existing studies in audio engineering [130, 152, 114, 72, 95], there are four widely-used features: pitch, rhythm, timbre, and loudness. Pitch is the subjective perception of highness or lowness of a sound, and is referred to as the fundamental frequency $\omega_0$ of a note [76, 107]. Rhythm is described as the tempo of the musical sound [152], which depends on the length of each note and the time intervals between adjacent notes. Timbre is the mixture of the harmonics, which brings the ”color” to music [107, 163], and it is similar to the characteristics of the speech [51]. Loudness measures the intensity of an audio signal and can be seen as the energy level or the volume of the signal [152].

In the following, we briefly introduce the commonly-used methods to compute the feature deviations between two signals $s(t)$ and $\hat{s}(t)$ in the literature. For each feature, the procedure is the same and shown in Fig. 2.2: $s(t)$ and $\hat{s}(t)$ each will be separated into frames with a small time interval (e.g., 16ms [72]). The signal samples in each frame are used to generate a feature value (e.g., pitch value). The feature values from all frames constitute a time-series data vector. Then, an algorithm called Dynamic Timing Warping (DTW) [138, 139] is used to quantify the similarity between the time-series vector for $s(t)$ and the one for $\hat{s}(t)$, and generate a vector of frame-wise deviation values for the feature. The advantage of DTW over the Euclidean distance is that DTW can reduce the time distortion [133] via finding an optimal path between two time-series vectors. For instance, the red line in Fig. 2.2 indicates the DTW path between $s(t)$ and $\hat{s}(t)$.

The pitch value in each frame is the basic frequency $\omega_0$ obtained via pitch estimation, which is a maximum likelihood estimation problem [59] via finding $\omega_0$ from harmonics

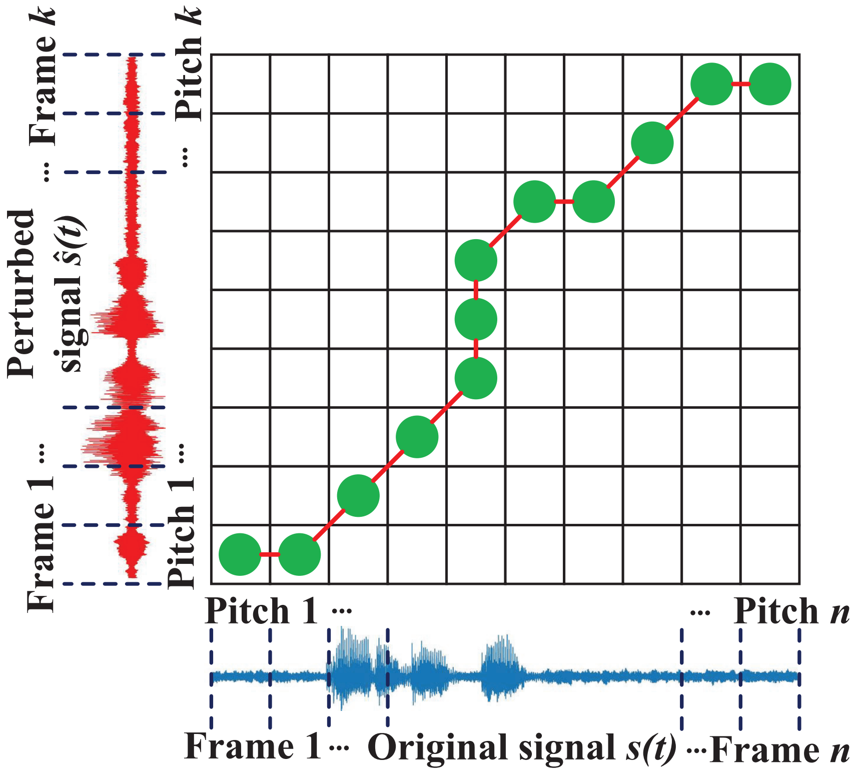


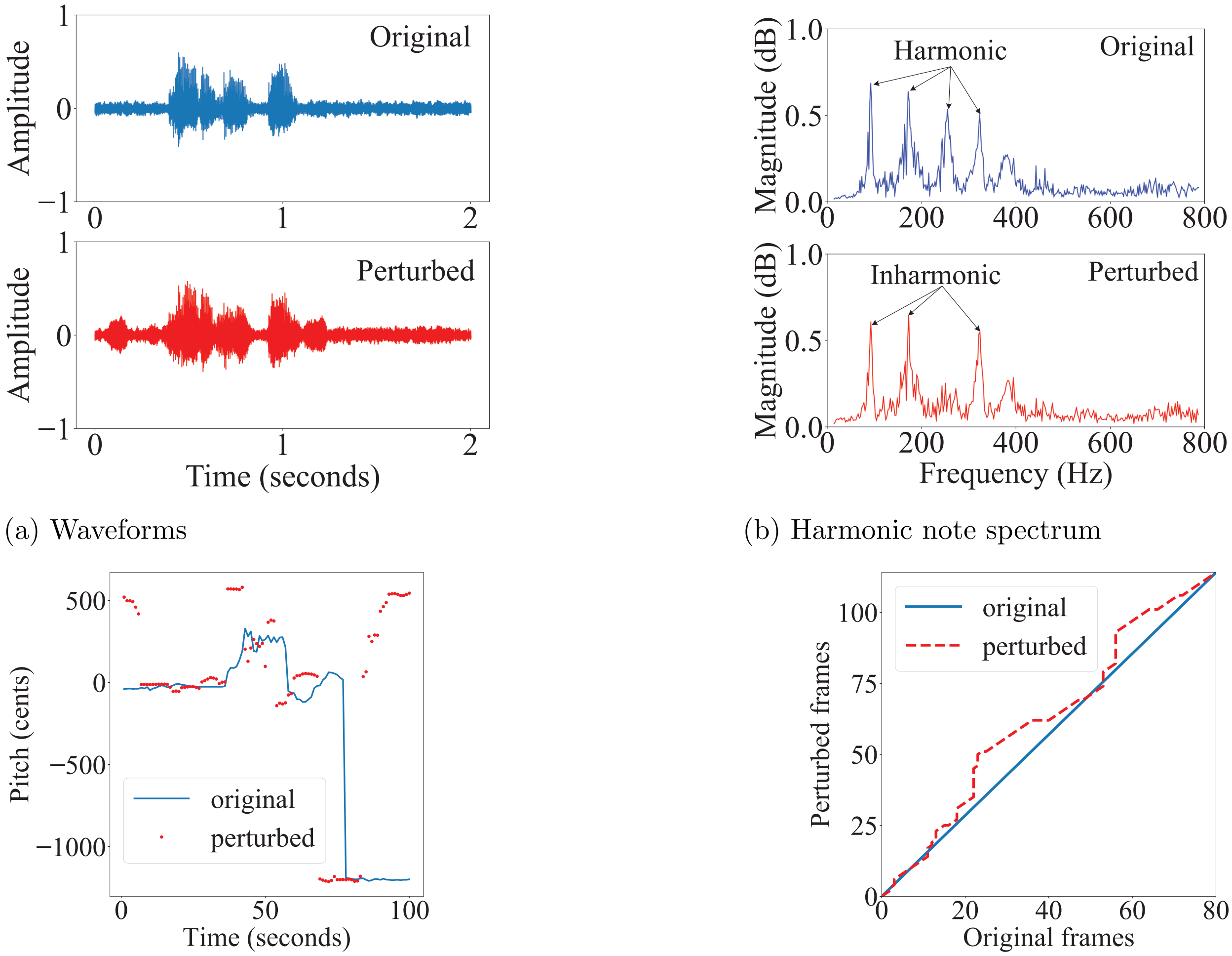
Figure 2.2 Computing deviation values via DTW.

$\sum_{m=1}^{M} m\omega_0$. The estimated pitch values from all frames form a time series for each signal and then DTW is used to generate the vector of frame-wise pitch deviation values between the two signals.

Rhythm computation is based on pitch estimation. A deviation value for rhythm between two frames is computed as the linear regression error in DTW during computing the deviation value for pitch [114]. All these values generated during DTW form the vector of frame-wise deviation values for rhythm.

The timbre value for each frame is computed as a Mel-Frequency Cepstrum Coefficient (MFCC) [50]. The vector of frame-wise deviation values for timbre is the result of the DTW between the MFCC vectors for $s(t)$ and $\hat{s}(t)$.

Loudness is closely related to the $L_p$ norm used in existing adversarial attack formulations (2.1). The loudness for each frame is usually calculated as the short-term log-energy [152],



(a) Waveforms

(b) Harmonic note spectrum

(c) Pitch deviation

(d) Rhythm deviation

Figure 2.3 Impacts of a noise-like perturbation on the music features.

which is the logarithm of the total energy of the frame. After two short-term log-energy vectors for $s(t)$ and $\hat{s}(t)$ are obtained, the DTW between them generates the vector of frame-wise deviation values for loudness.

The last step for each feature is to aggregate the computed vector of frame-wise deviation values into a single value to represent the overall feature deviation. According to existing studies [136, 72], the non-linear average calculation is commonly adopted for pitch and rhythm aggregations, and linear averaging is used for timbre and loudness. After the aggregations, the resultant four feature deviation values form a final feature deviation vector to describe the audio characteristic deviation from $s(t)$ to $\hat{s}(t)$.

### 2.3.2 Impacts of Audio Feature Deviations

To have a good sense of how pitch, rhythm, timbre, and loudness change in a perturbed music signal, we show the feature deviations caused by an adversarial example in [137] in Fig. 2.3.

As [137] adopted an $L_p$ norm based formulation to create adversarial audio and limited the $L_p$ norm of the perturbation, Fig. 2.3a shows that there is a minor waveform change in the time-domain between the original and perturbed music signal. This indicates that the perturbation only incurs a small energy or loudness change to the original signal.

Next, we look at the waveform change in the frequency-domain and compare the power spectrum in Fig. 2.3b. The observed change is more evident than the time domain in Fig. 2.3b: the third harmonic in the original harmonics is suppressed, which leads to inharmonicity in the signal and can negatively impact the timbre feature and accordingly the audio quality.

If we look at the pitch contours (i.e., the curves drawn by connecting all pitch values over time) for the original and perturbed signals in Fig. 2.3c, we observe the evident difference of the pitch features between the two signals. Similarly, Fig. 2.3d shows the optimal DTW path of the perturbed signal to the original one. Intuitively, a music signal with the minimal rhythm deviation should have a nearly straight line DTW path. Fig. 2.3d shows that the DTW path of the perturbed signal is tortuous compared with the original one.

Note that creating adversarial music inevitably causes some distortions of the original signal. Fig. 2.3 demonstrates that there may exist some way to better coordinate such distortions among all audio features to mimic the original signal's quality as much as possible since they are eventually perceived by humans. If we look at the basic adversarial audio attack formulation used in recent research [38, 102, 137], the $L_p$ norm of the additive noise is only relevant to the loudness feature without a clear relation to the other three features. It is evident that $L_p$ norm is much easier to compute than pitch, rhythm, and timbre via gradient descend. At the current stage, we do not focus on the computational aspect but on the human perception aspect and continue to understand how these features affect human perception.

### 2.3.3 Human Study Procedures and Setups

To understand how different features affect human perception. We conduct a human study with the procedure shown in Fig. 2.4: we first generate a dataset that consists of pairs of original and perturbed music signals. For each pair, we can compute (according to the procedure in Section 2.3.1) the deviation values for the four features, which form a feature deviation vector. Then, we invite every human participant to assign a deviation rating to each pair based on his/her perceived difference. Next, considering the feature deviation vectors as the inputs and the human ratings as the outputs, we use regression analysis to find the best model to describe the relation between the vectors and the ratings. In this way, we can reverse-engineer the human perception process to build an approximation model to quantitatively predict how much a perturbed signal is perceived by a human.

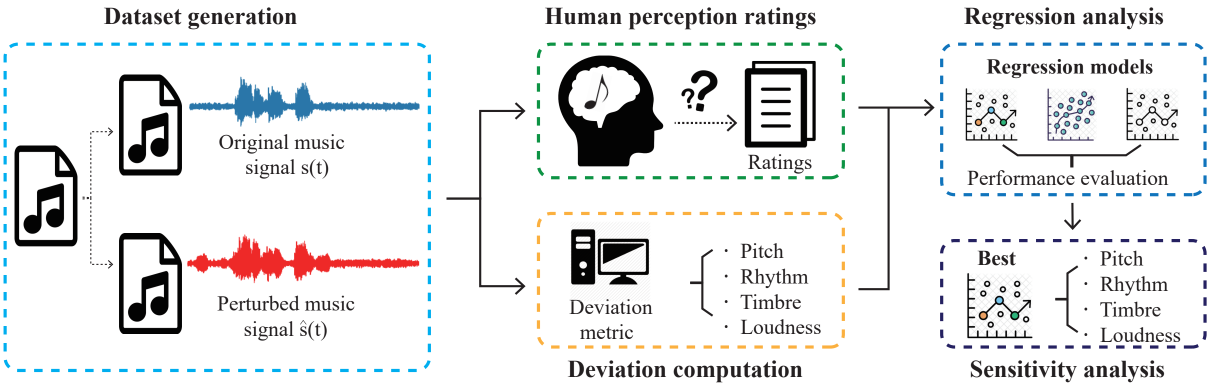


Figure 2.4 The human study procedure and steps.

Since there is no publicly available dataset that provides various versions of perturbed music signals, we propose to generate our own dataset with the following requirements:

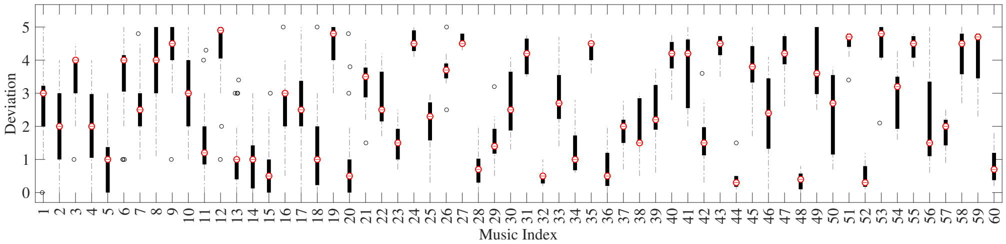


Figure 2.5 Distributions of human ratings of perceived deviation for all pairs of music clips.

sufficient diversity of music genres, sufficient perturbations from the pitch, rhythm, timbre, and loudness perspective, and slight or moderate perturbation to avoid making participants feel overly noisy.

We build a dataset of 60 pairs of original and perturbed music clips from the genres of Pop, Hip-hop, Rock, Jazz, Classical, R&B, Country, and Disco. To make participants concentrate on each small perturbation, we crop each music clip to a 5-second WAV format (16kHz, 16-bit PCM, Mono) to avoid audio compression. As there is no guideline or reference to standardize the dataset generation for our study, we aim to create perturbed signals with different feature deviations and varying intensities for human participants such that the data is diverse for regression analysis. Specifically, we use two main mechanisms to create perturbed music clips.

1) Additive noise: an intuitive method is to inject additive noise into the original music. The noise will affect all four features at the same time. To broadly affect the original music, we consider injecting the noise from three aspects: amplitude, frequency and time. To control the amplitude of the noise, we can choose the signal-to-noise (SNR) level from 0dB, 5dB, 10dB, and 15dB [154]. To inject frequency-sensitive noise, we use both white noise [155] (covering all frequencies with equal intensity) and colored noise (with the power concentrated at certain frequencies). To make noise time-varying, we set random duration and interval of the noise, but the total injection duration is less than the half of the original music length.

2) Additive notes: To ensure distinctive deviations among all music features, we also inject additive notes to the original music. To inject notes with the pitch manipulation, we randomly choose notes with the pitch value from 27.5Hz to 4186Hz [100] (88 notes space). For rhythm manipulation, we randomly select the additive notes with different lengths and ensure the intervals between adjacent notes are less than 50% of the original signal's length. To create timbre deviation, we select different instruments to play the additive notes as long as the notes are within the valid pitch ranges of those instruments.

We recruited 35 participants whose ages fell between 20 and 35. All the participants are volunteers without any compensation. Each participant was asked to listen to each pair of the original and perturbed music clips, and then assign a deviation rating on a Likert scale [150] according to his/her overall music perception: $0-1$ perfect perceptual quality with imperceptible noise, $1-2$ good perceptual quality with quiet noise, $2-3$ noticeable with slight noise, $3-4$ noticeable and noisy, and $4-5$ very noisy. More specifically, $1-2$ means volunteers can only notice some small perturbation after listening to a part of music clips many times, and $2-3$ indicates the deviation can be noticed by listeners but not noisy. During the experiments, all the volunteers were given the same earphone with the same initial volume setting. They can listen to a music clip as many times as they want.

Our study involved human participants that assigned ratings by listening to music. The full protocol was reviewed and exempted by our Institutional Review Board (IRB), which has determined that the study involves the minimal risk for human participants (i.e., the risk is no more than the one that they face during their daily lives). We follow the approved protocol to inform them of the full study procedure and protect their identities without publishing any personally identifiable information.

Given the computed feature deviations from the original and perturbed music clips as well as the human participant ratings of their perceived deviation, we aim to find the best regression model $M^* \in \mathcal{M}$ in the model set $\mathcal{M}$ to minimize the mean squared error (MSE) of regressed prediction, i.e.,

$$M^* = \underset{M \in \mathcal{M}}{\arg\min}\ \mathbb{E}\|r - M(d_p, d_r, d_t, d_l)\|_2^2, \tag{2.2}$$

where $r$ is the human participant rating, $d_p$, $d_r$, $d_t$, and $d_l$ are the deviation values (computed according to the procedure in Section 2.3.1) for pitch, rhythm, timbre, and loudness, respectively. In our study, we choose Linear Regression [152, 73], Support Vector Regression, Random Forest, Logistic Regression, and Bayesian Ridge to form the model set $\mathcal{M}$. With $M^*$ found in (2.2), we use it to quantitatively predict any human-perceived deviation given a pair of original and perturbed music signals.

### 2.3.4 Result Analysis and Discussion

Fig. 2.5 box-plots all the human ratings (ranging from 0 to 5) for individual pairs of music clips from our human study. We can find in Fig. 2.5 that human perception is indeed subjective: each pair of music clips has a range of deviation ratings by different participants; there are always rating outliers for a pair of music clips. Fig. 2.5 also shows that overall, the ratings and the 25%-75% boxes are roughly evenly distributed from 0 to 5, which offers sufficient data diversity for regression analysis.

We first use each of Linear Regression, Support Vector Regression (SVR), Random Forest, Logistic Regression, and Bayesian Ridge to model the relationship between feature deviation values and the average human rating, and find the best model with the minimum MSE. Table 2.1 shows the MSEs of different regression models during testing.

Table 2.1 MSEs of different regression models.

| **Model:** | Linear | SVR | Random Forest | Logistic | Bayesian |
|---|---|---|---|---|---|
| **MSE:** | 1.2351 | 0.8558 | 0.1541 | 1.6572 | 1.2628 |

Through regression analysis, we find that Random Forest performs the best among all the five regression models. As Table 2.1 shows, Random Forest leads to an MSE of 0.1541, which is substantially better than Support Vector Regression that achieves the second with an MSE of 0.8558, but an over 5 times increase from Random Forest. The other models result in even worse MSEs. As a result, we choose Random Forest as our regression model to predict the human-perceived deviation. Specifically, given a pair of original and perturbed signals, we name the prediction output of Random Forest as quantified deviation (qDev).

Then, we analyze to what extent qDev values and realistic human ratings move in tandem; that is, an increase or decrease of value for one will lead to the same for the other. This is important because when creating an adversarial attack against a classifier, we aim to reduce the qDev value of a perturbed signal (so its deviation rating by a human should also decrease) such that the perturbation is hardly noticed by a listener. We use Spearman's rank correlation coefficient [46, 141] to model the correlation in our study. Spearman's coefficient is a commonly used statistic measure to evaluate the relationship between two variables using a monotonic function, where value 1 or -1 indicates that the two always move in the same or opposite direction; value 0 means no correlation.

Table 2.2 Spearman's coefficient between the human rating and a deviation measure.

| **Deviation Measure:** | $L_2$ | $L_\infty$ | SNR | qDev |
|---|---|---|---|---|
| **Spearman's Coefficient:** | 0.3909 | 0.0893 | 0.0134 | 0.9608 |

Table 2.2 lists the Spearman's coefficients between the human rating and each of the following deviation measures: $L_2$ norm [137], $L_\infty$ norm [137, 38], SNR [172, 41], and qDev from Random Forest. It is seen from Table 2.2 that qDev has a very high correlation with the realistic human rating, indicating it can be quite useful for predicting a human-perceived deviation of a signal. In other words, minimizing qDev in a mathematical formulation to form an audio signal perturbation would be most likely suppress a human's attention to the signal deviation caused by the perturbation. Interestingly, we also observe that the

commonly used $L_p$ norms and SNR are in fact not well related to human perception (e.g. $L_2$ norm has the best correlation of 0.3909). Table 2.2 offers quantitative evidence to echo the concern raised in related studies [35, 132] that suggests new ways to measure the human perceptual similarity may be needed.

### 2.3.5 Sensitivity Analysis

To explore which feature is potentially more important than others in human perception, we conduct sensitivity analysis via the One-at-a-time (OAT) strategy [31, 22, 117]: we remove in turn pitch, rhythm, timbre, and loudness to form three-feature inputs for regression, and measure the MSE of the resultant regression. We find Random Forest is always the best in our OAT analysis to minimize the MSE with only three features reaming as the inputs.

Table 2.3 Sensitivity analysis for each feature.

| Excluding: | Pitch | Rhythm | Timbre | Loudness | None |
|---|---|---|---|---|---|
| MSE: | 0.1891 | 0.1581 | 0.1889 | 0.3539 | 0.1541 |

Table 2.4 Sensitivity analysis for each two features.

| Excluding: | R&T | T&L | R&L | P&T | P&L | P&R |
|---|---|---|---|---|---|---|
| MSE: | 0.3033 | 0.3855 | 0.2071 | 0.3482 | 0.2337 | 0.1748 |
| Correlation: | 0.9306 | 0.9189 | 0.9437 | 0.9210 | 0.9363 | 0.9575 |

Table 2.3 shows the MSE of Random Forest for each regression of excluding pitch, rhythm, timbre, and loudness in turn. From Table 2.3, loudness that represents the energy of the perturbation appears to be the most sensitive feature to human-perceived deviation. For example, removing loudness leads to a 129% MSE increase from 0.1541 to 0.3539. But it is clear that the other features individually contribute to the overall human perception, and removing one of them causes more MSE in the regression.

Overall, we find in the human study that Random Forest is the best regression model to yield the minimum MSE to predict the human rating as qDev. Simpler regression models, such as Linear Regression or SVR, do not perform as well as Random Forest. This may also confirm that human perception is indeed a complicated process. In addition, qDev is a much more appropriate metric than the conventional $L_p$ norm or SNR in terms of both MSE and Spearman's correlation with the human rating, and the features of pitch, rhythm, timbre, loudness all contribute to the overall perception.

## 2.4 Perception-Aware Attack Strategies

With the metric of qDev regressed via Random Forest from audio features, we reformulate the problem of creating adversarial music signals into a perception-aware attack framework. We then analyze how to narrow down the search space in the reformulation, and eventually find an efficient solution via dynamic clipping.

### 2.4.1 Problem Reformulation

Existing studies [36, 169, 102, 176] solve the original optimization problem in (2.1) via finding a sub-optimal yet efficient alternative solution. For our perception-aware reformulation, it is natural to think about reformulating existing alternative solutions by directly replacing its $L_p$ norm with the new metric of qDev. However, such a reformulation no longer offers the advantage of computational efficiency because the process of computing audio features in qDev is unfortunately non-linear, non-convex, and non-differentiable [59]. Accordingly, we formulate the perception-aware attack by replacing $L_p$ norm with qDev in the original form (2.1) as

$$\begin{aligned} \text{minimize} \quad & \text{qDev}(s(t), \hat{s}(t)), \\ \text{subject to} \quad & f(\hat{s}(t)) \neq y, \end{aligned} \tag{2.3}$$

where qDev$(s(t), \hat{s}(t))$ denotes the qDev between the perturbed signal $\hat{s}(t) = s(t) + \delta(t)$ and the original one $s(t)$. To ensure $\hat{s}(t)$ to be a valid waveform, we always constrain the normalized amplitude of each of its sample points to be in $[-1, 1]$ [38].

Finding the optimal solution to (2.3) becomes even more difficult than the original one in (2.1) because computing qDev involves a much more complicated process than the $L_p$ norm. Our strategy is to analyze what properties the perturbation signal $\delta(t)$ should have towards finding a solution to (2.3).

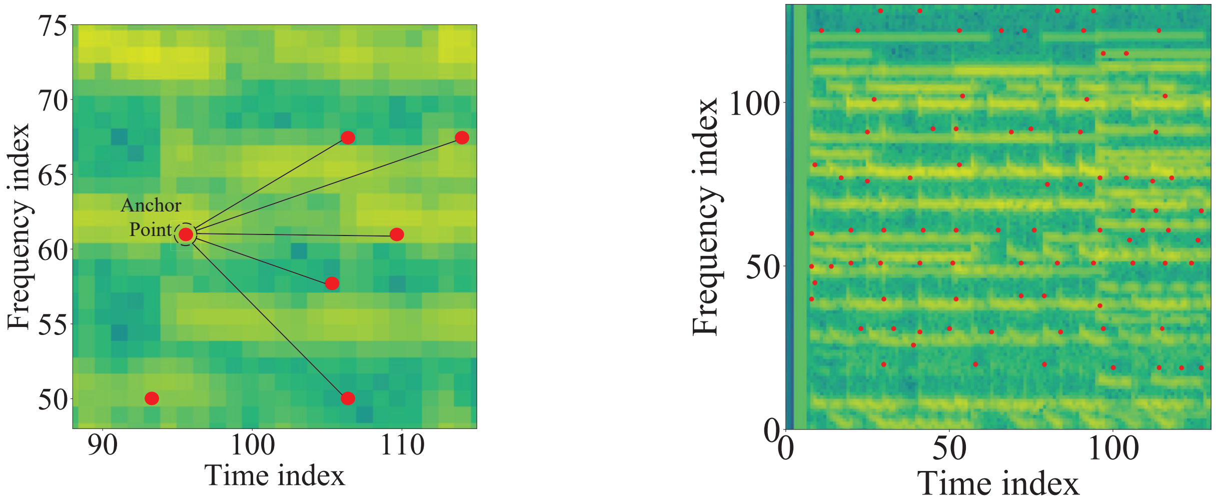
### 2.4.2 Perturbation Signal Property Analysis

Since the solution to (2.3) is computationally intractable, we have to narrow down the search space for the perturbation signal $\delta(t)$ by analyzing what properties it should have.

The reformulation (2.3) means two obvious goals that the perturbation signal $\delta(t)$ should achieve: 1) misclassification (i.e., the attack should fool the classifier) and 2) minimized qDev (i.e., it also produces good perceptual quality a human can perceive). At first glance, the two goals seem to contradict with each other (as the best perceptual quality of music indicates no change of its signal and thus no attack success). We need to explore one step further to understand what audio features $\delta(t)$ needs as a result of each of the two goals, then consider all needed features jointly to reconcile any conflict to construct a search space of $\delta(t)$ that is sufficiently narrowed down towards a feasible solution.

Here we discuss some properties for attacking audio fingerprinting. First, we consider what feature properties $\delta(t)$ should have towards launching a successful attack. A key technique for audio signal classification is audio fingerprinting [32]. The technique and its variants have been widely adopted in audio signal watermarking [26, 44], integrity verification [68], music information retrieval [158, 37, 126], broadcast monitoring [20, 121, 75] and copyright detection [137].

The essential idea in audio fingerprinting is to consider certain high-energy areas of an audio signal in the spectrogram as its fingerprints. As an example shown in Fig. 2.6(a)



(a) Fingerprinting generation.

(b) Distribution of peaks.

Figure 2.6 Fingerprinting generation via finding all peaks in a signal's spectrogram.

[158]: an energy peak (anchor point) is paired with other peaks within a certain target area in a signal's spectrogram, then the fingerprints are computed based on the frequency information of the peaks and the time intervals between them. Fig. 2.6(b) shows there are many peaks in a signal's spectrogram that lead to a large number of fingerprints for audio signal classification and identification.

As we can observe from Fig. 2.6, peaks in the spectrogram are a key feature for audio signal classification. These peaks are usually the results of a mixture of high-energy points of audio signal harmonics [74, 68, 158]. From the attacker's perspective, creating new positions of harmonics in the spectrogram should be a direct way to manipulate the fingerprints, which can lead to the misclassification of the signal. In the audio features, timbre is the most relevant to the harmonics of the signal [107, 135]. Given an energy threshold (that represents the loudness) for perturbation $\delta(t)$, a good way to create the attack is to affect the feature of timbre for the signal.

Next, we consider what feature properties $\delta(t)$ should have for good music perceptual quality. From the sensitivity analysis in the human study in Section 2.3.5, all features, pitch, rhythm, timbre, affect the human perception of signal deviation or the metric of qDev. The change of any of them may result in an increase of qDev and accordingly a noticeable change by human perception. Hence, a reasonable strategy for creating $\delta(t)$ given an energy threshold is to incur the change of only one feature while making the other feature deviations small such that the qDev aggregated over all features remains small.

To summarize, it would be good to 1) change the feature of timbre for a potentially successful attack, and 2) manipulate only one feature while keeping the others unchanged as much as possible to maintain the perceptual quality. To reconcile the two requirements: we propose to change timbre much more than the other features.

Now the question becomes how to create $\delta(t)$ with a quite different timbre feature while maintaining almost the same pitch and rhythm features. The traditional perturbation design in (2.1) usually generates a noise-like perturbation and is not able to create this required signal because it causes all distortions of pitch, rhythms, and timbre (as shown in Fig. 2.3). As a music signal consists of well-crafted, human-enjoyable musical notes, we propose to create $\delta(t)$ by reproducing the same music notes via new instruments. The timbre feature is always associated with the harmonics, and we can find these natural harmonics from the instruments. In this way, the timbre of $\delta(t)$ can be changed substantially due to different harmonic characteristics of new instruments; but pitch and rhythm may deviate less if we find appropriate instruments to play the same notes. To demonstrate the feasibility of our design, we compare the feature deviations of a perturbed music signal mixed by randomly-generated noise and instrument-generated music notes.

Table 2.5 Noise vs notes played by a different instrument.

| | Pitch | Rhythm | Timbre | Loudness | qDev |
|---|---|---|---|---|---|
| **Instrument:** | 0 | 0.85 | 25320 | 2873 | 2.23 |
| **Noise:** | 0.9049 | 7.239 | 19521 | 1988 | 3.86 |

As shown in Table 2.5, the additive instrument produces a higher loudness value than noise (indicating a more energy level); at the same time, it generates more timbre deviations

(25320 vs 19521) but less pitch and rhythm deviations than the noise. Depending on the difference between $s(t)$ to $\hat{s}(t)$, the non-linearly aggregated pitch and rhythm deviations have values commonly in the range from 0 to 50, and the linearly aggregated timbre and loudness deviations usually range from zero to tens of thousands. There exists an obvious deviation gap between instrument-generated notes and randomly-generated noise of different features. We also use qDev to quantify the deviations, and the instrument-generated notes have a clearly lower qDev value than random noise (2.23 vs 3.86). This makes it a much more desirable signal component for $\delta(t)$ in terms of both human-perceived quality (low qDev) and attack effectiveness (more timbre variation).

Consequently, we can effectively narrow down the search space by considering $\delta(t)$ as a linear combination of signals consisting of the same music notes played by different instruments for the original music signal. Then, generating the perturbed signal $\hat{s}(t)=s(t)+\delta(t)$ is like finding "subtle" instrumental track signals then optimally remixing them (based on qDev) into the original music.

It is worth mentioning that a music signal can consist of both instrumental and vocal tracks. It is possible to add a new vocal track (i.e., the same vocal notes sung by a different voice to change the feature of timbre) into the perturbation $\delta(t)$. As it is easier to generate instrumental signals by computer music synthesis, we only use instrumental tracks to form $\delta(t)$ in this paper.

### 2.4.3 Perception-Aware Attack Formulation

With the shrunk search space, we write $\delta(t) = \sum_{k=1}^{K} \theta_k \delta_k(t)$, where $K$ denotes the number of different instrumental tracks, $\delta_k(t)$ is the $k$-th instrumental track signal, and $\theta_k$ is the non-negative weight for $\delta_k(t)$. Next, we reformulate (2.3) into a perception-aware attack of finding the best linear weights $\theta_k$ in $\delta(t)$ to minimize the qDev:

$$\begin{aligned} \underset{\{\theta_k\}_{k\in[1,K]}}{\text{minimize}} \quad & \mathrm{qDev}\Big(s(t), s(t) + \textstyle\sum_{k=1}^{K}\theta_k\delta_k(t)\Big) && (2.4) \\ \text{subject to} \quad & f\Big(s(t) + \textstyle\sum_{k=1}^{K}\theta_k\delta_k(t)\Big) \neq y, && \\ & \textstyle\sum_{k=1}^{K}\theta_k \leq \epsilon, && (2.5) \\ & \mathcal{P}_{s(t)} \subseteq \mathcal{P}_{\delta_k(t)}\, \forall\, k \in \{k | k \in [1,K], \theta_k \neq 0\}, && (2.6) \end{aligned}$$

where (2.5) ensures the energy level of the perturbation signal $\delta(t)$ is less than a threshold $\epsilon$, $\mathcal{P}_{s(t)}$ and $\mathcal{P}_{\delta_k(t)}$ in (2.6) represent the sets of pitch values in the original signal $s(t)$ and the $k$-th track signal $\delta_k(t)$, respectively; (2.6) ensures that $\delta_k(t)$ covers the pitch range of $s(t)$ so the pitch feature of $\delta_k(t)$ does not deviate much from $s(t)$.

The optimization (2.4) is a problem of finding the optimal linear weights. Although still non-differentiable, (2.4) opens a door for a grid search based heuristic solution. Specifically, we can let each linear weight $\theta_k$ be a multiple of a small step $\Delta$ (that is a fraction of the threshold $\epsilon$ in (2.5)), then enumerate all combinations of possible values for $\{\theta_k\}_{k\in[1,K]}$ to find a solution to (2.4). For example, setting $\Delta = 0.1\epsilon$ and $K = 10$ produces 92,378 combinations in total. Iterating through them, though not very efficient, is quite feasible for an attacker's computing capability today.

### 2.4.4 Dynamic Clipping

The optimization in (2.4) finds out a perturbation signal $\delta(t)$ based on the entire duration of the original signal $s(t)$. However, a piece of music can consist of multiple segments with audio characteristics varying within a wide range of instruments and vocals, creating distinct timbre features. For better perceptual quality and attack effectiveness, it is necessary to segment $s(t)$ into $N$ clips according to evident timbre changes and create the perturbation for each clip using the clip-wise optimization based on (2.4). We call this procedure dynamic clipping.

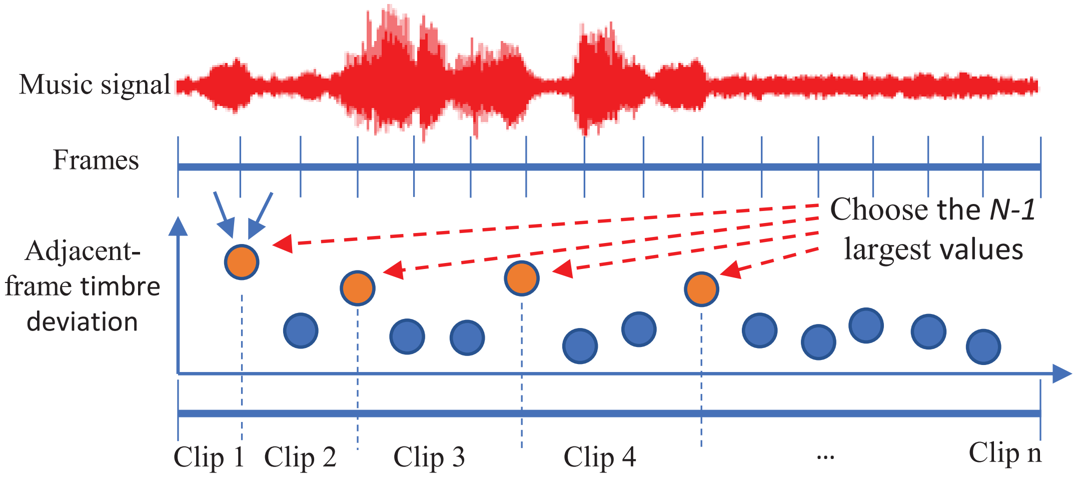

Music signal
Frames
Adjacent-frame timbre deviation
Choose the N-1 largest values
Clip 1
Clip 2
Clip 3
Clip 4
...
Clip n

Figure 2.7 Overview of dynamic clipping.

Fig. 2.7 shows the overall process of dynamic clipping: in order to dynamically segment $s(t)$ into $N$ clips, we first separate $s(t)$ into small frames and compute the timbre deviation between each pair of adjacent frames (using the timbre deviation calculation discussed in Section 2.3.1). Then, we identify $N-1$ pairs which have the $N-1$ largest adjacent-frame deviation values, as they contain the most evident $N-1$ changes of timbre over the duration of the music. We use the timing boundary between two frames in a pair as a timing position to segment $s(t)$. In this way, $s(t)$ is segmented into $N$ clips, each of which will be used to find a corresponding perturbation based on (2.4).

## 2.5 Realistic Black-box Attack against Copyright Detector

In this section, we create a realistic attack based on the perception-aware attack framework in Section 2.4. We choose the YouTube copyright detector as our target as YouTube has exhibited some robustness against noise and perturbations [137]. Because there is no knowledge of YouTube's design, we create our own detector based on open-source information for an adversarial transfer attack. We first present how to generate additional instrumental tracks for the perturbation signal given a music signal, then describe the design of our detector as a surrogate model for YouTube's detector.

### 2.5.1 Perturbation Signal Generation

Perturbation signals generated by (2.4) require the detailed music notes of the original music. For a popular piece of music, its Musical Instrument Digital Interface (MIDI) file is usually available in online databases (e.g., FreeMidi.org and Nonstop2k[3]). The MIDI file contains all instrumental tracks with music notes. We use Music21[4] to play a downloaded MIDI file with different instruments to form a perturbation for (2.4). To achieve the diversity of the timbre feature for (2.4), we consider an instrument set of instruments across the four families *stringed* (Guitar, Electric Guitar, Violin, Viola, Cello, Bass, Electric Bass), *woodwind* (Clarinet, Flute, Saxophone, Oboe, Bassoon), *brass* (Trumpet, Baritone, Tuba, Horn, Trombone), *keyboard* (Piano, Electric Piano). We empirically select at most two instruments from each family based on a music genre to reduce the computational complexity and the pitch range requirement for perturbation generation in (2.6).

### 2.5.2 Surrogate Detector

A copyright detector takes audio fingerprinting features as the input. We select the fingerprints and their extraction method introduced in [158]. We extract fingerprints by considering the time, frequency, and amplitude data of the audio. Specifically, we use Fast Fourier Transform (FFT) to generate a spectrogram of an audio signal and extract the spectral peaks of acoustic harmonics, which are shown invariant and reproducible from signal degradation [33] and robust to noise and distortion [158]. We then apply the fast combinatorial hashing method [158] to form these fingerprints to hashes for the similarity comparison later.

[3] FreeMidi.org:https://freemidi.org/, Nonstop2k:https://www.nonstop2k.com/

[4] Music21 is a Python-based toolkit for computer-aided musicology. In this work, we use it to produce different instrumental tracks playing the same musical notes

### 2.5.3 Detection Design

The detection is built to compute the similarity of the fingerprints of an input signal to the detector's database. If the similarity score is higher than a similarity threshold, the detector will raise an alarm. To ensure our surrogate detector has a degree of transferability to YouTube's detector, we must adopt a threshold that is similar to YouTube's. We note that our objective is not to precisely rebuild YouTube's model, but to choose an appropriate threshold (even in a rough way) such that we can use the surrogate detector to predict the output label during minimizing qDev in (2.4). Because music consists of diversities of audio features, we choose one threshold for each of 8 music genres: Pop, Hip-hop, Rock, Classical, Jazz, R&B, Country, and Disco.

Fig. 2.8 shows the process we use to approximately calibrate the surrogate detector's threshold towards YouTube's. This process is similar to the one proposed in [38] that estimates the threshold of a black-box model. In particular, to obtain the threshold for a music genre, we choose a song from the genre, crop it into clips, choose the most representative clip that contains the highest number of fingerprints among all the clips. Then, we randomly add instrumental track signals with different energy levels to this clip, generating a number of clips with perturbations of varying energy levels. We send these clips to YouTube to see the copyright detection results, and set the detection threshold for the surrogate detector such that it yields the same results as YouTube does.

## 2.6 Experiments and Results

In this section, we present the experiments and results. We first describe the experimental settings, then discuss the audio perceptual quality and attack effectiveness of generated adversarial music.

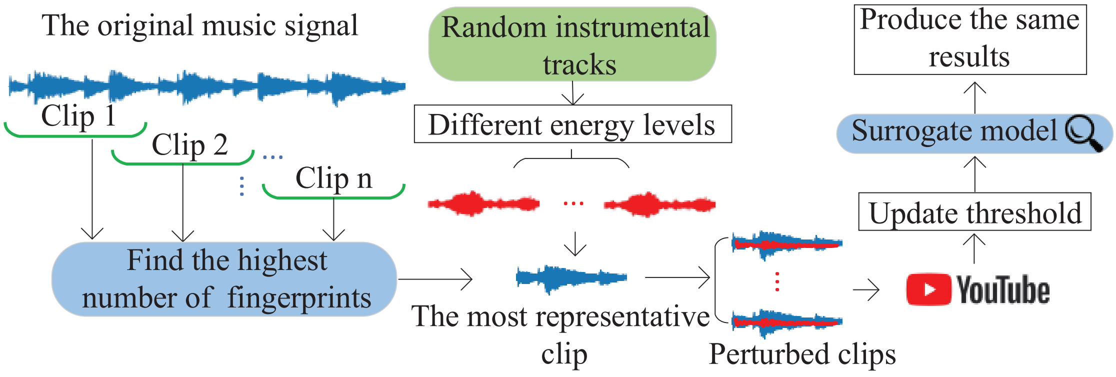


Figure 2.8 Process of obtaining the threshold from YouTube.

### 2.6.1 Experiments Setup

To cover a wide range of music data, we selected 32 top hits songs of the last 20 years from 8 genres: Pop, Hip-hop, Rock, Classical, Jazz, R&B, Country, and Disco. We created 40 clips of 5–10 seconds and 160 clips of 30 seconds for evaluations. We have verified that all the clips were copyright-detected by YouTube.

The default settings in (2.4) for the perception-aware attack include the search step $\Delta = 0.1\epsilon$, the number of instruments for perturbation generation $K = 7$, and the number of clips in dynamic clipping $N = 6$.

We compare the perception-aware attack with a recent attack method (the ICML20 method) against YouTube in [137], which tried to limit the $L_p$-norm of the perturbation and force the perturbation look like a natural signal. We directly adopted the source code provided by the authors of [137] in our experiments. We also implemented a random noise attack method that adds random noise to music as a baseline case.

It is worth noting that [137] found YouTube exhibited some degree of robustness against noise-like adversarial perturbations. During our research, we also find that YouTube has been continuously improving its copyright detector. For example, both the adversarial sample originally provided in [137] and our early examples no longer succeed against YouTube. We suspect that YouTube has a de-noising or noise-resilient mechanism and keeps improving it for robust copyright detection.

Here we provide an anonymous YouTube link that demonstrates adversarial music clips created by the perception-aware attack in comparison with the ICML20 attack and the random noise attack: https://www.youtube.com/watch?v=jK3ejLtx750.

### 2.6.2 Perceptual Quality of Adversarial Music

We first evaluate the perceptual quality of adversarial music created by the perception-aware, ICML20, and random noise attacks. In the experiments, given original music, we created perturbed music clips of 5–10 seconds under each attack by increasing the energy threshold of the perturbation such that the perturbed clip exactly bypassed YouTube's detector. For each perturbed clip, we used the Random Forest regressed qDev in Section 2.3.4 to predict its deviation from the original clip. We also involved human participants and let each of them assign an actual deviation rating based on their perceived difference between the original and perturbed music clips. The same earphone and the same rating guideline were used.

Fig. 2.9 illustrates the average human ratings and qDev values of the perception-aware, ICML20, and random noise attacks for each music genre. It is evident from the figure that the perception-aware attack always achieves much smaller deviation ratings and qDev values than the other two attacks. For example, for classical music, the perception-aware attack obtains a rating of 0.71 (indicating nearly perfect perceptual quality according to the rating guideline) while the ICML20 and noise attacks get 3.15 and 3.40, respectively (indicating noticeable and noisy). It is also observed that rock music seems harder to perturb for the perception-aware attack and has a rating of 2.90 (noticeable with slight noise). Overall, Fig. 2.9 shows that the perception-aware attack achieves substantially better perceptual quality than the ICML20 and random noise attacks.

By comparing the qDev value with the human rating in every music genre in Fig. 2.9, we can see that qDev is a good prediction to the human rating as the qDev does not deviate much from the average human rating for each genre. For example, the Hip-hop music created by

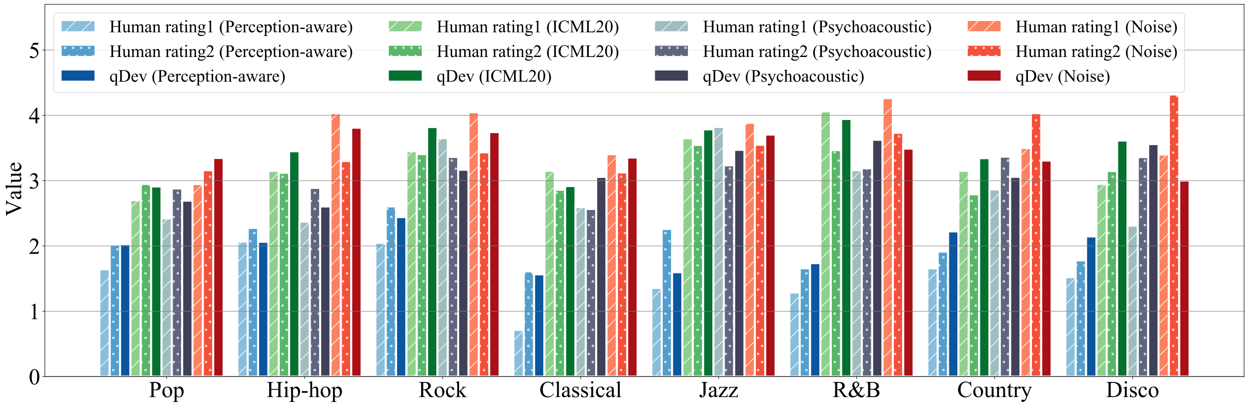


Figure 2.9 Human ratings and qDev values: Perception-aware, ICML20, psychoacoustic and random noise attacks.

the perception-aware attack has the qDev of 2.06 compared with the average human rating of 2.07. Table 2.6 shows the MSE for qDev prediction in each music genre. The MSE of qDev averaged over all music genres is 0.3294 in our experiments, which is higher than the training MSE of 0.1541 in regression analysis in Section 2.3.4 but is better than other regression models.

We also evaluate the impact of dynamic clipping in Section 2.4.4 on the overall perceptual quality of the perturbed music. We compare its performance with a static clipping design in which a clip is uniformly segmented into 6 smaller clips with equal length for perturbation generation.

Table 2.9 shows the qDev values of the two designs for different music genres. We can observe that dynamic clipping achieves uniformly better perceptual quality in all genres.

Previous experiments were conducted in a formal lab setting to quantify the perceived deviation via actual human ratings and qDev estimates. When a person listens to music during the daily life, there is no reference for him/her to perceive a deviation. The person may or may not notice an issue if the music is perturbed.

We conducted another experiment to measure how a human participants perceive perturbed music without reference. In particular, we selected 16 30-second music clips, and asked two questions to each participant for each clip: If familiar with the music: Assign a

Table 2.6 The MSE of qDev averaged among genres

| | Pop | Hip-hop | Rock | Classical | - |
|---|---|---|---|---|---|
| **MSE-G1:** | 0.7256 | 0.1607 | 0.2347 | 0.4803 | - |
| **MSE-G2:** | 1.0268 | 0.3158 | 0.3818 | 0.2303 | - |
| **MSE-G1*:** | 1.8894 | 2.0828 | 3.1533 | 0.7655 | - |
| **MSE-G2*:** | 2.4649 | 1.8025 | 2.8255 | 1.1740 | - |
| | Jazz | R&B | Country | Disco | Overall |
| **MSE-G1:** | 0.1531 | 0.4263 | 0.4442 | 0.7352 | 0.4107 |
| **MSE-G2:** | 0.3326 | 1.2768 | 0.2991 | 1.0670 | 0.5848 |
| **MSE-G1*:** | 2.3529 | 3.0245 | 0.4880 | 0.7493 | 2.0054 |
| **MSE-G2*:** | 2.6149 | 2.9379 | 2.7360 | 1.5489 | 2.2944 |

Note that, ”MSE-G1” means the MSE result of the previous group participants who took the user study in 2.3 before the evaluation. To evaluate the generalizability, we invite a new group volunteer which is denoted as ”MSE-G2”. ”MSE-G1*” and ’MSE-G2*” denotes the MSE result of the different qDev model which is trained without additive noise dataset in 2.3.3.

deviation rating based on your memory using the same rating guideline. Otherwise: Do you feel abnormal about the music? Please answer 1) Yes, 2) No, or 3) Not Sure.

Table 2.8 shows the average human ratings without reference and average qDev values for different music genres. We can find that the rating distribution among music genres is quite similar to Fig. 2.9. For example, the Classical music can still achieve nearly perfect perceptual quality of 0.86, and Rock and Hip-Hop are the worst genres to perturb and make human participants feel noticeable with slight noise deviations. Interestingly, we find that the human rating for R&B music is 0.5 (nearly perfect perceptual quality) without reference, which is an improvement from the experiments with reference. The potential reason is that the additive instrumental track signals sound natural and embedded to the original music. It becomes hard for humans to recognize these different timbre features without any reference.

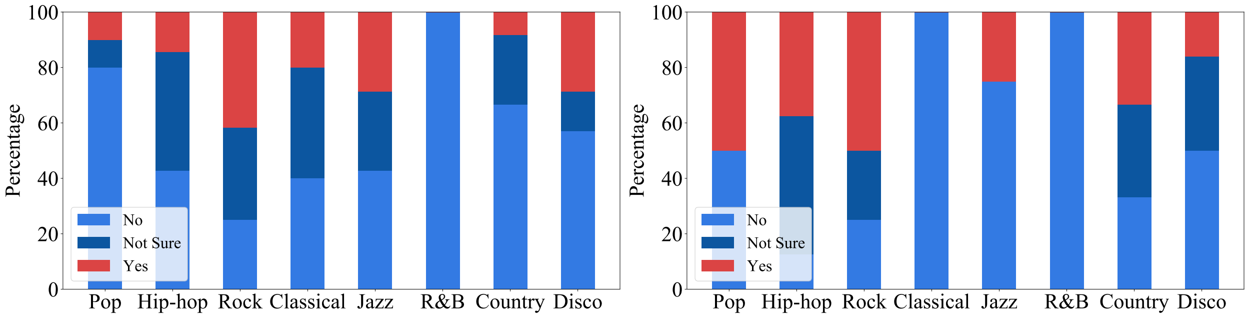
Fig. 2.10 depicts a more interesting result of the percentages of different answers by participants unfamiliar with the given music. We can see that most participants do not notice

Table 2.7 qDev values in dynamic vs static clipping.

| | Pop | Hip-hop | Rock | Classical |
|---|---|---|---|---|
| **Dynamic:** | 1.8953 | 2.9250 | 2.6051 | 1.4956 |
| **Static:** | 2.2522 | 3.1854 | 3.1955 | 1.7558 |
| | Jazz | R&B | Country | Disco |
| **Dynamic:** | 1.8653 | 1.3897 | 1.6925 | 2.1933 |
| **Static:** | 2.9192 | 2.0925 | 2.0230 | 2.2588 |

Table 2.8 Human ratings without reference and qDev.

| | Pop | Hip-hop | Rock | Classical |
|---|---|---|---|---|
| **Human rating G1:** | 1.4500 | 2.4428 | 2.4867 | 0.8583 |
| **Human rating G2:** | 1.9993 | 2.6408 | 2.7988 | 1.7367 |
| **qDev:** | 1.7850 | 2.7133 | 2.5653 | 1.6255 |
| | Jazz | R&B | Country | Disco |
| **Human rating G1:** | 1.7500 | 0.5000 | 1.4458 | 1.4821 |
| **Human rating G2:** | 1.8909 | 1.4333 | 2.6606 | 2.0053 |
| **qDev:** | 2.5679 | 1.4905 | 1.6925 | 2.1178 |



(a) Group1

(b) Group2

Figure 2.10 Percentages of answers by participants of different groups unfamiliar with the given music.

any abnormality in perturbed soft music (e.g., R&B, Pop, and Classical). For example, no audience finds any issue in any R&B music; and 30% or more answers for Rock, Disco and Jazz say that music clips are abnormal. Fig. 2.10 shows that the majority of participants do not notice the subtle perturbation generated by the perception-aware attack. Considering the fact that participants may form a cognitive bias in the study (i.e., they might feel “obliged” or “mentally-focused” to identify an abnormality), we think that a casual listener without reference might be more unlikely to notice the perturbation of adversarial music created by the perception-aware attack.

### 2.6.3 Attack Effectiveness vs qDev

Next, we measure the attack success rates of the perception-aware, ICML20, and random noise attacks against YouTube. As discussed in Section 2.5.3, the fingerprinting similarity thresholds in our surrogate detector were set roughly according to YouTube’s detection results using a few music samples. But an adversarial music clip bypassing the surrogate detector does not necessarily mean that it will also evade YouTube’s detection. In this experiment, we used the perception-aware, ICML20, and random noise attacks to each create 240 adversarial clips of 30 seconds (that 100% bypassed the surrogate detector), and then uploaded them to a private YouTube channel to test YouTube’s copyright detection.

It is clear that we can always get a 100% attack success rate by generating a sufficiently large perturbation and adding it to the original music, which can, unfortunately, produce extremely noisy sound. Hence, it is also necessary to pair the attack success rate with music perceptual quality.

To this end, we focus on comparing the average qDev values of adversarial music clips created by perception-aware, ICML20, and random noise attacks under the same attack success rates against YouTube. Fig. 2.11 shows the comparison results. As shown in Fig. 2.11, higher attack success rates come with lower music perceptual quality in general. The qDev values of the perception-aware attack are always better than ICML20 and random noise attacks for the same attack success rate. In particular, its qDev increases from 1.64 (good quality with quiet noise) to 2.53 (noticeable with slight noise) when the attack success rate goes from 20% to 80%; in contrast, the ICML20 attack has the qDev value increasing from 2.70 (noticeable with slight noise) to nearly 4 (very noisy). The random noise attack has the

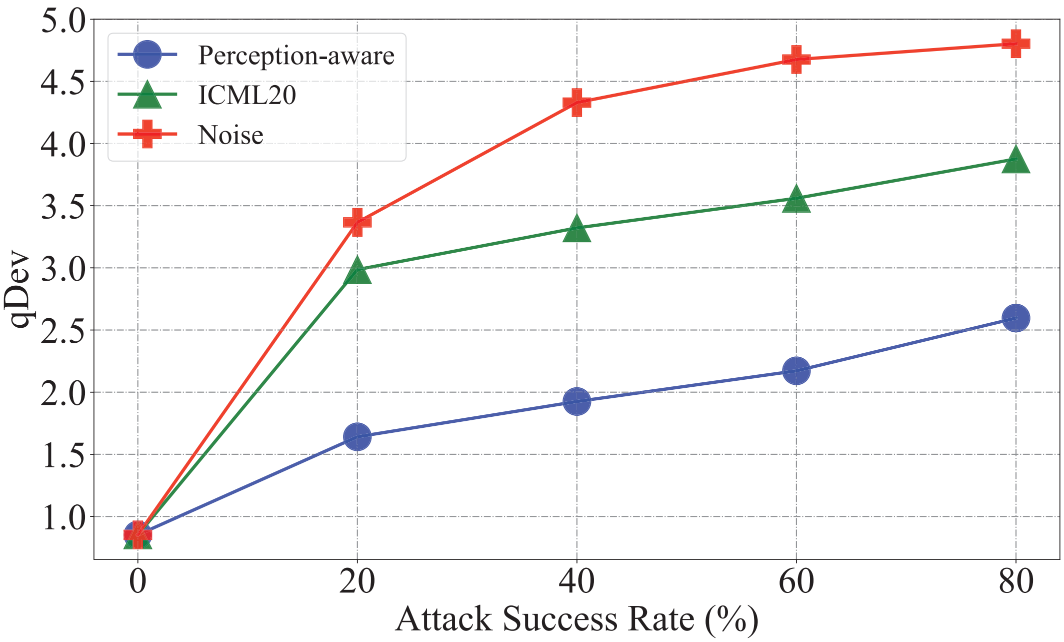


Figure 2.11 Comparisons of attack success rates and qDev.

highest qDev value almost reaching 5 when the attack success rate is 80%. Overall, Fig. 2.11 offers very intuitive comparisons and demonstrates that the perception-aware attack is able to create more effective attacks against YouTube with better music quality.

In our experiments for the perception-aware attack, the number of instruments used to generate the perturbation was set to be $K = 7$. It means that (2.4) always tries to find 7 weights assigned to 7 instrumental tracks. We can reduce the computational complexity by restricting the number of instrumental tracks. The less the number, the less the computational complexity (2.4) incurs. We conducted experiments to evaluate the impact of this number. Specifically, we still used 7 instruments but only choose 1, 3, or 5 out of 7 to form the instrumental track(s) as the perturbations to create the adversarial music clips. Under approximately the same attack success rates against YouTube, we show the average qDev values of 160 adversarial music clips for each various instrument selection method in Table 2.9.

We find in Table 2.9 that the qDev value gradually decreases from 2.8901 to 2.5902 when we choose 1 to 7 out of 7 instruments to create the perturbations. This is expected as the

Table 2.9 Attack success rates and qDev values for different numbers of instruments.

| **Number of instruments:** | 1 | 3 | 5 | 7 |
|---|---|---|---|---|
| **Success rate:** | 78.13% | 80.00% | 79.38% | 80.63% |
| **qDev:** | 2.8901 | 2.7256 | 2.6713 | 2.5902 |

objective of (2.4) is to minimize qDev and more instrument selections lead to a lower qDev value. One interesting observation is that choosing less instruments does not quite affect the attack success rate against YouTube. However, using only one instrument creates a quite loud music signal played by the instrument that is more identifiable to humans. Adding more instruments and distributing weights among them help suppress one single loud perturbation signal and makes the overall perturbation less identifiable.

### 2.6.4 Discussions

Though the perception-aware attack produce better-quality perturbations, we can still notice deviations (some are minor and others more noticeable) from the perturbed music. One may further improve the attack as discussed below.

The metric of qDev based on current data regression of human ratings is not sufficiently sensitive to a small value difference. For example, a qDev value decrease from 4 to 1 should indicate an evident music perceptual quality improvement; however, a decrease from 2.1 to 2.0 may well fall into the error range of subjective judgements and is not fully correlated with music quality improvement. This may indicate that within this subtle qDev range, there might exist other improvements to make the perturbation sound more natural and attached to the original music. For example, some instruments (e.g., trumpet during our observations) can produce audio characteristics more identifiable to humans than some others, making its track evidently comparable to the foreground tracks (e.g., the main vocal track) in the original music. It may be necessary for (2.4) to select such an instrument to beat the classification via creating more timbre variations and minimize the qDev. There may exist other benchmarks in this case to further differentiate the selection of instruments as a small qDev difference may no longer help the selection.

Dynamic clipping segments a music signal into multiple clips and finds the optimal additional instruments for each clip. When the instrument sets for adjacent clips are chosen in a distinct way, human participants may be sharp enough to notice an instrumental transition. Smoothing this transition may result in a better experience; but the smoothing still needs to take suppressing audio fingerprints into consideration.

Our human-in-the-loop methodology can be extended to the speech domain. There are technical differences between fingerprinting music and recognizing speech. We expect this leads to non-trivial efforts to rebuild the qDev based on human perception of speech difference, perform sensitivity analysis for speech features, and then shrink the search space by considering qDev-friendly signals (e.g., from a set of synthetic speech phonemes) to minimize the non-differentiable qDev, which can form another full research study adopting a similar methodology.

We build the qDev model aims to evaluate the relative quality of the perturbed music rather than an absolute musical expertise ratings. Although there exists some bias in the human subjects, qDev can still be an effective predictor to evaluate a better or worse quality of the perturbed music with its own criteria, which is satisfied to achieve the goal of minimizing the music perception deviation in our attack.

The perception-aware attack does not cause an immediate operational impact, such as denial of service; but it can pose as an evident abusive risk for YouTube's copyright detector whose main purpose is to combat monetizing copyrighted content without proper authorization. Currently, we keep all the adversarial music clips private in YouTube for the research purpose only. We plan to disclose our results to YouTube according to the disclosure window of 45–90 days ahead of publication. During the initial disclosure, we plan to provide a detailed description of the attack strategy along with YouTube links to the created music clips. We will work with YouTube during any follow-up communication.

## 2.7 Discussions on Defense Strategies

Audio pre-processing is a potential method to reduce the effectiveness of adversarial examples, as the small perturbation could be mitigated during the audio squeezing [172, 41, 38, 176] and audio compression [102, 48]. These defense methods are unlikely effective against the perception-aware attack as squeezing/compression does not quite change the spectrogram feature (e.g., the high energy harmonics will not be revised during the processing). On the other hand, these defense methods may not be desirable in some scenarios. For example, YouTube does not downgrade the music quality via squeezing and compression.

The advantage of audio fingerprinting is its computational efficiency [74, 68, 158]. Exiting research [158, 60, 147] focused mostly on extracting spectrogram features in a robust way for fingerprinting based detection. Although these fingerprints can be made robust to noise and pitch-shifting [60], the perception-aware attack creates additional harmonics and spectrogram features that can be extracted as fingerprints and fool the detection. We can potentially improve audio fingerprinting against the perception-aware attack by adding the pitch and rhythm features as other types of fingerprints. This, however, will incur substantially more costs because estimating pitch and rhythm incurs complicated maximum-likelihood estimation [59] than spectrogram based fingerprinting. There is a need to achieve a balanced tradeoff between detection accuracy and computational complexity.

Another possible way to defend against the perception-aware attack is to leverage existing defense strategies from the machine learning community. In particular, adversarial training [70, 109, 23, 30, 142, 151, 166] and certified defense [24, 165, 113, 86] are popular among the methods to provide more robustness against adversarial attacks. Adversarial training primarily focuses on making the model robust to the adversaries via solving a min-max optimization problem that finds the model parameters to minimize the cost results from strong adversary examples. Given a bounded $L_p$ ball, the re-trained model becomes more robust against the adversarial attacks. However, the perception-aware attacker uses qDev instead of $L_p$ norm to craft adversarial examples. This creates a model mismatch [143] and can make the re-trained model ill-suited. A potential way to solve the issue is to use qDev to guide the adversarial training. However, computing qDev is a non-differentiable process. Initial efforts can be focused on finding a differentiable function to approximate qDev to efficiently finish the adversarial training. Certified defense is to find an upper bound of the adversarial loss which guarantees the robustness to any attack in the same threat model. Existing work [24] can provide a provable defense to the neural networks via convex layerwise adversarial training. To use certified defense against the perception-aware attack, we need to find a differential upper bound to characterize the adversarial loss based on the qDev modeling, which, similar to using adversarial training, involves non-trivial research efforts.

## 2.8 Related Work

Most adversarial attacks [36, 93, 38, 102, 176] control the energy of the perturbation within a bounded $L_p$ ball such that a created adversarial audio example resembles the original signal in its waveform format. In this paper, we show that limiting the waveform change is not fully related to human-perceived change. Instead of using the $L_p$ norm, we propose to use qDev based on the comprehensive human study to create adversarial signals with better quality. A few recent studies on speech recognition attacks [140, 132, 100] have also discussed adopting psycho-acoustic hiding methods to embed low energy perturbations near the frequency of a louder signal. These efforts are orthogonal to ours as they mainly focused on where to hide the perturbations and we aim at making the perturbations themselves better perceived. There are also a few recent studies [174, 34, 17, 172, 41] focusing on creating inaudible signals as attacks. These studies generally use various strategies to effectively hide the presence of the attack. The perception-aware attack adopts a different strategy that creates perturbation signals to minimize the human-perceived deviation. The ICML20 method [137] focused on creating a neural network based black-box attack against copyright detectors. It proposed a mathematical attempt that enforces the perturbation to be similar to a signal of certain frequencies to make it more natural based on $L_p$ norm. But how it indeed affects human perception was not studied. By contrast, the perception-aware attack integrates the proposed qDev into its formulation, creates effective adversarial music while suppressing human attention, and achieves better perceptual quality than the ICML20 method in the experimental results.

Human perception studies [34, 172, 38, 41, 176] have been adopted to evaluate the stealthiness of adversarial audio examples as the SNR metric may not be appropriate to well reflect the human perception [41, 176]. Exiting work [34, 172, 38, 41, 176] designed human perception studies from different perspectives and evaluated the attack performance based on the results of human study. For instance, [162] conducted a comprehensive human study to evaluate the synthetic speech quality to reveal the impact of deep-learning based speech synthesis to human. These studies focused on analyzing the results of the human evaluation, rather than integrating human factors into the designs. There are few studies [152, 72] focusing on defining human-involved metrics for singing scoring systems. The systems were designed to generate an absolute score to indicate the singing performance given the recording of a human's singing via linear weighting [152] or non-linear neural network [72] on audio features. By contrast, our strategy focuses on modeling the human-perceived deviation between original and perturbed music signals, compares different regression models, and analyzes how each audio feature affects the overall human perception of music deviation.

## 2.9 Summary

In this work, we conducted a human study to reverse-engineer the human perception of music deviation via regression analysis. Based on the analysis, we proposed the perception-aware attack framework to create adversarial music that can mislead a music classifier while preserving the perceptual quality. Experimental results have shown that the perception-aware attack is effective and achieves better music perceptual quality compared to prior work. Our work demonstrates that perceptual quality of adversarial attacks can be significantly improved by integrating human factors into the adversarial audio attack design process.

# Chapter 3: Parrot-Trained Attacks against Speaker Recognition Models

The Parrot-Trained attack [5] explores using the minimal attack knowledge against the real-world speaker recognition models.

## 3.1 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

[5] This chapter was published in Network and Distributed System Security (NDSS) Symposium 2024. Permission is included in Appendix A

scenario and 47.9%–58.3% against smart devices, including Apple HomePod (Siri), Amazon Echo, and Google Home, in the over-the-air scenario.

## 3.2 The Background of Speaker Recognition and Design 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.

### 3.2.1 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 [4] 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 multiple-speaker-based speaker identification (SI) or single-speaker-based speaker verification (SV). Specifically, SI can be divided into close-set identification (CSI) and open-set identification (OSI) [53, 38].

### 3.2.2 The Mechanisms of Speaker Recognition

Speaker recognition models [7, 5, 122, 96] are typically categorized into statistical models, such as Gaussian-Mixture-Model (GMM) based Universal Background Model (UBM) [134] and i-vector probabilistic linear discriminant analysis (PLDA) [52, 120], and deep neural network (DNN) models [97, 55]. There are three phases in speaker recognition.

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 [52] and X-vectors [146]). Then, these features can be trained by a classifier (e.g., PLDA [81]) to recognize different speakers.

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 [3] and Amazon Echo [1]) or text-independent speech samples to the speaker recognition system. Depending on the number of enrolled speakers, speaker recognition tasks [38, 176, 171] can be multiple-speaker-based speaker identification (SI) or single-speaker-based speaker verification (SV).

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) [53, 38]. 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.

### 3.2.3 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 $y_i \in \mathcal{Y}$.

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.,

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

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 $\theta$.

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

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

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.

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

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

### 3.2.4 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

$$f(x + \delta) = y_t, \quad D(x, x + \delta) \leq \epsilon, \tag{3.1}$$

the $y_t \neq y$ 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 [38, 176] or a measure of auditory feature difference (e.g., qDev [58] and NISQA [171]); and $\epsilon$ limits the change from $x$ to $x + \delta$.

A common white-box attack formulation [36, 102] to solve (3.1) can be written as

$$\arg\min_{\delta \in \Omega} \mathcal{J}(x + \delta, y_t) + c\, D(x, x + \delta), \tag{3.2}$$

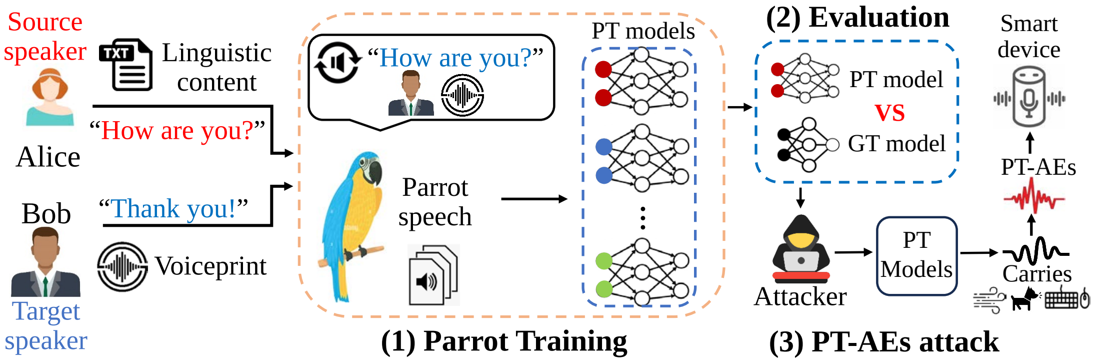


Figure 3.1 The procedure of parrot training based black-box attack.

the $\mathcal{J}(\cdot,\cdot)$ 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 (3.2) 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 [176, 104] can be formulated as

$$\arg\min_{\delta\in\Omega} \mathcal{L}(x+\delta)=\begin{cases} D(x,x+\delta) & \text{if } f(x+\delta)=y_t, \\ +\infty & \text{otherwise.} \end{cases} \tag{3.3}$$

Since (3.3) 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 [176]) are required, which makes real-world attacks less practical against commercial speaker recognition models that accept speech signals over the air.

### 3.2.5 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 [176] that the attacker possesses a very short audio sample of the target speaker (note that [176] 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 (3.2) and try to build a local function $\mathscr{J}^*$ similar to the loss prediction function $\mathscr{J}$ in (3.2), then replace $\mathscr{J}$ with $\mathscr{J}^*$ to create an audio AE. This may look like a traditional transfer attack strategy [41]. But the key difference is that the traditional transfer attack still needs to keep probing the classifier (e.g., 1500 queries [41]) to build the local model $\mathscr{J}^*$; in contrast, the attacker here only has a very short sample of the target speaker to construct $\mathscr{J}^*$ without probing.

As a result, the first challenge we need to solve is how to build $\mathscr{J}^*$ 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 [43, 108, 167, 40], 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 $\mathscr{J}^*$ 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: Can a PT model approximate a GT model? Secondly, are PT-AEs built upon a PT model as transferable as GT-AEs against a black-box GT model? At last, how to optimize the generation of PT-AEs towards an effective black-box attack? Fig. 3.1 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 3.3; (2) we study different types of PT-AE generations from a PT model regarding their transferability and perception quality in Section 3.4; and (3) we formulate an optimized black-box attack based on PT-AEs in Section 3.5. Then, we perform comprehensive evaluations to understand the impact of the proposed attack on commercial audio systems in Section 3.6.

### 3.2.6 Threat Model

In this work, 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 [176], 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 [171, 38, 176] 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).

## 3.3 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.

### 3.3.1 One-shot Voice Conversion

Generating data with certain properties is commonly used in the image domain, including transforming the existing data via data augmentation [129, 144, 111, 144], generating similar training data via Generative Adversarial Networks (GAN) [69, 45, 19], and generating new variations of the existing data by Variational Autoencoders (VAE) [88, 64, 77, 29]. These approaches can also be found in the audio domain, such as speech augmentation [89, 127, 90, 98], GAN-based speech synthesis [42, 91, 83, 25], and VAE-based speech synthesis [88, 79, 175]. Specifically, VC [131, 105, 99, 159, 40] 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 [162, 54] have revealed that it can be difficult for humans to distinguish whether the speech generated by a VC method is real or fake.

Recent VC has been developed by only using one-shot speech [43, 108, 167, 40] (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 work 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. 3.1, 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.

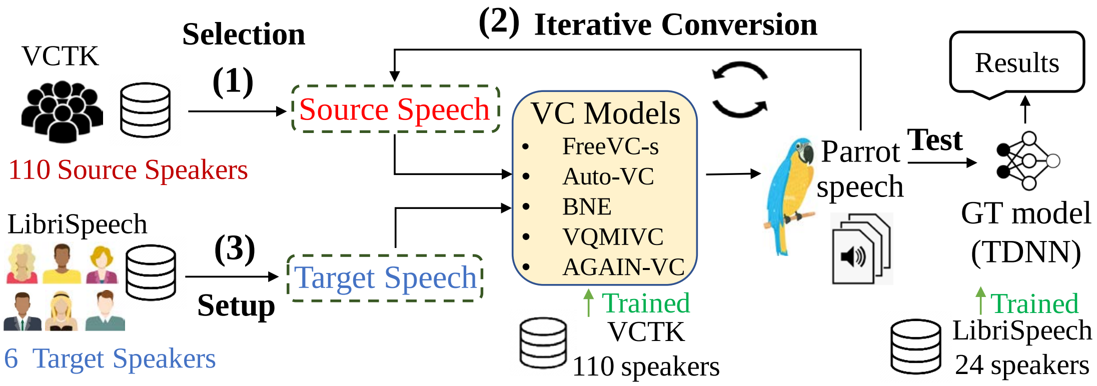


Figure 3.2 Parrot speech generation: setups and evaluations.

### 3.3.2 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 [84, 85, 54].

Existing VC studies [84, 85] 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 [99, 159, 105], 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.

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 [54], 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 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. 3.2.

There are a wide range of one-shot VC methods recently available for parrot speech generation. We consider and compare the performance of AutoVC [11], BNE [14], VQMIVC [16], FreeVC-s [12], and AGAIN-VC [10]. As shown in Fig. 3.2, we use the VCTK dataset [156] 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 [123], 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%.

We use the False Positive Rate (FPR) [80, 38] 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)

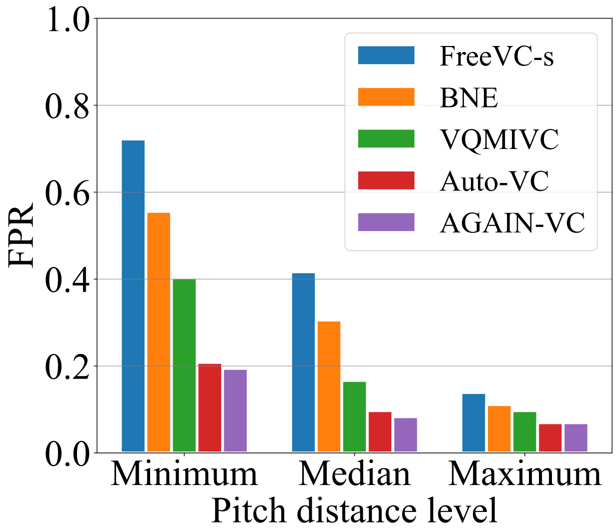


Figure 3.3 FPRs under different initial source speakers.

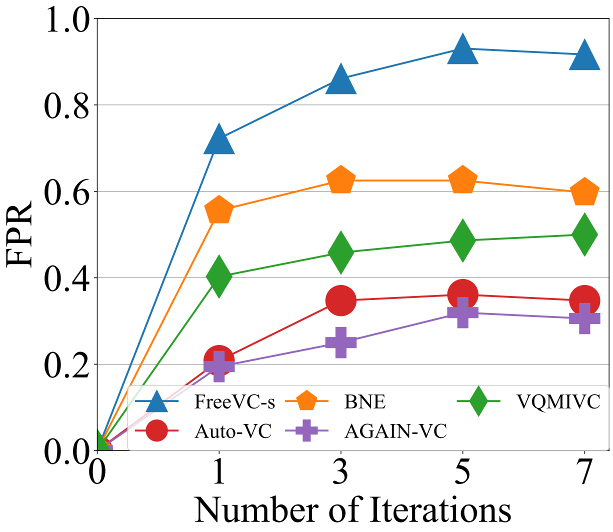


Figure 3.4 FPRs under different numbers of iterations.

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.

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. 3.3 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

Table 3.1 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 |

in Fig. 3.2. 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).

Next, we evaluate the impact of iterative conversions on the FPR. Fig. 3.4 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. 3.4, 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: 1) 2-second length level (around 5 words), 2) 4-second level (10 words), 3) 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 3.1 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 3.1 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.

### 3.3.3 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.

### 3.3.4 Comparison of PT and GT 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.

We aim to compare the 12 PT models with the 2 GT models when recognizing the attacker's target speaker. Existing studies [101, 92] 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 [49], Precision [61], and F1-Score [119], 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 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.

Fig. 3.5 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

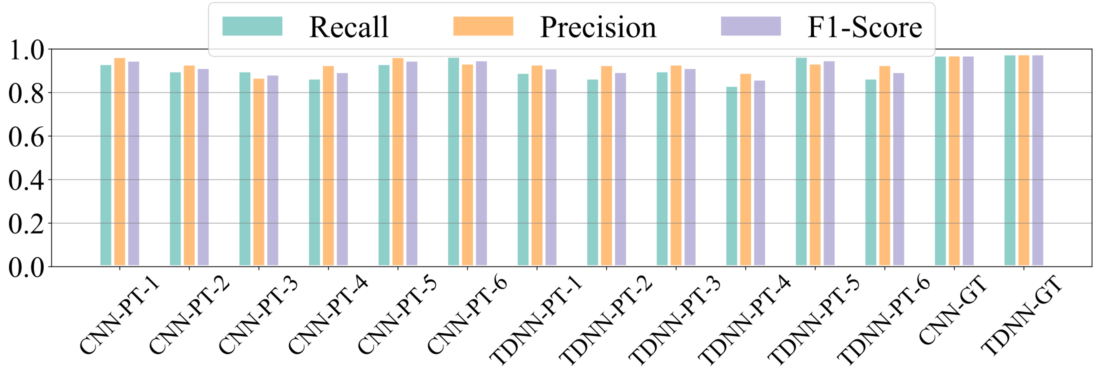


Figure 3.5 Comparison of PT and GT models.

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.

## 3.4 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.

### 3.4.1 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.

Traditional methods [38, 104] 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 [132, 71, 104] that can be used to alleviate the noisy effect.

Directly manipulating the auditory feature of a speech signal could make a classifier sensitive but stealthy to the human ears. Existing works [18, 171] have found that modifying the phonemes or changing the prosody of the speech can also spoof the audio classifier while preserving the perception quality.

The enrollment phase attack [53] employed environmental sounds (e.g., traffic) to create the perturbation signal to poison a speaker recognition model.

### 3.4.2 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 [58, 171] have pointed out that traditional metrics, such as signal-to-noise ratio (SNR) [41] and the $L_p$ norm [172, 38, 176], 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 [58] and NISQA for feature-twisted AEs [171]). In addition, we also notice that the harmonics-to-noise ratio (HNR) [173] 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.

We create the human study dataset with noise carriers [36, 132, 71, 38, 176, 104], feature-twisted carriers [171], and environmental carriers [53]. We choose 30 original speech signals (with length from 5 to 15 seconds) from the existing speech dataset [**?**]. 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.

We first consider building a dataset with noise carriers. The dataset 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.

Then, we introduce the feature-twisted carriers in AE generation. For feature-twisted speech signals, we shift the tone (i.e., the pitch) [171] to generate pitch-twisted carriers. Specifically, we shift up/down by 25 semitones[6] 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.

At last, we also consider using environmental sound carriers to generate the AEs. Environmental sound carriers are selected from the large-scale human-labeled environmental sound datasets [63] 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.

[6] 1 semitone $= 12\log_2(f'/f)$, where $f$ and $f'$ are the original and perturbed speech frequencies, respectively [15].

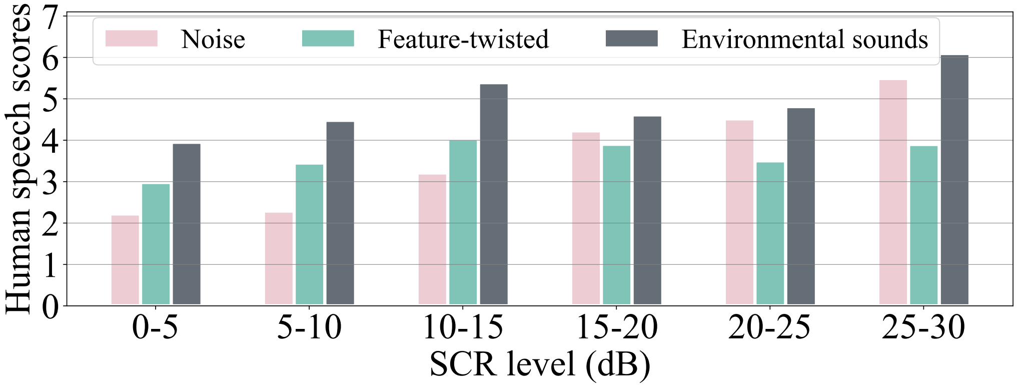


Figure 3.6 Human scores for carrier-perturbed speech signals.

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 [63, 164, 28, 21, 128, 47], 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).

Fig. 3.6 compares the average human scores at varying SCR levels for different 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. 3.6 also shows that overall, environmental sound carriers yield the better human scores than the feature-twisted carriers and noise carriers.

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 [172, 38, 176], SCR (equivalent to SNR [41]), HNR [173], audio-feature-regression-based qDev [58], and DNN-based NISQA [171, ?]. Note that the qDev model [58] was originally trained using music instead of speech. We follow the procedure in [58] 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 [78], to measure the correlation between the metric and the human score. Table 3.2 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 [?], which may not be effective for feature-twisted speech as the diversity of training data is important to the prediction performance [58]. Based on Table 3.2, we use the metric of SRS to measure the perpetual quality of an audio AE.

### 3.4.3 Measuring Transferability of PT-AEs

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

#### *3.4.3.1 Building Target and Surrogate Models*

The first step in evaluating the transferability is to build 1) target models, which refer to the models to be attacked by the attacker using PT-AEs, and 2) 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 [106].

We consider building a diversity of target models with 4 DNN-based speaker recognition models including 2 CNN [82] and 2 TDNN models [145, 146]. 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

Table 3.2 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 |

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.

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 $\{\text{PT-CNN-C-}i\}_{i\in[1,6]}$, $\{\text{PT-CNN-D-}i\}_{i\in[1,6]}$, $\{\text{PT-TDNN-C-}i\}_{i\in[1,6]}$, and $\{\text{PT-TDNN-D-}i\}_{i\in[1,6]}$.

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

Table 3.3 Match rates between surrogate and target models (Noise).

| AE Carrier Type: | Noise | | | | |
|---|---|---|---|---|---|
| Target Model: | CNN-A | CNN-B | TDNN-A | TDNN-B | Average |
| GT-CNN-C | 0.2167 | 0.1500 | 0.1167 | 0.1417 | 0.1563 |
| PT-CNN-C | 0.1917 | 0.1417 | 0.0917 | 0.1250 | 0.1375 |
| GT-CNN-D | 0.0917 | 0.2167 | 0.0833 | 0.1917 | 0.1458 |
| PT-CNN-D | 0.0417 | 0.1667 | 0.0583 | 0.1583 | 0.1063 |
| GT-TDNN-C | 0.1000 | 0.1500 | 0.1750 | 0.1583 | 0.1458 |
| PT-TDNN-C | 0.0917 | 0.1417 | 0.1667 | 0.1333 | 0.1333 |
| GT-TDNN-D | 0.1333 | 0.1000 | 0.2083 | 0.2083 | 0.1625 |
| PT-TDNN-D | 0.1250 | 0.0833 | 0.1750 | 0.1667 | 0.1375 |

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.

### *3.4.3.2 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.

First, we solve the white-box problem (3.2) via projected gradient descent (PGD) [67], and we choose $L_\infty$ norm as the distance metric, which shows a good performance in Table 3.2. We set $\epsilon = 0.05$ to control the $L_\infty$ norm.

Secondly, we twist the pitch and rhythm of the original speech [171, 58] using the perception metric SRS as the distance measurement. As the random-forest-based SRS is non-differentiable, we use grid search to solve 3.2. 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.

Table 3.4 Match rates between surrogate and target models (Feature-twisted).

| AE Carrier Type: | Feature-twisted | | | | |
|---|---|---|---|---|---|
| Target Model: | CNN-A | CNN-B | TDNN-A | TDNN-B | Average |
| GT-CNN-C | 0.2333 | 0.2083 | 0.1583 | 0.1750 | 0.1937 |
| PT-CNN-C | 0.2083 | 0.1750 | 0.1083 | 0.1583 | 0.1625 |
| GT-CNN-D | 0.1667 | 0.1917 | 0.1500 | 0.1833 | 0.1729 |
| PT-CNN-D | 0.1417 | 0.1500 | 0.1417 | 0.1583 | 0.1479 |
| GT-TDNN-C | 0.1500 | 0.1833 | 0.2583 | 0.1417 | 0.1833 |
| PT-TDNN-C | 0.1167 | 0.1750 | 0.2500 | 0.1333 | 0.1688 |
| GT-TDNN-D | 0.1583 | 0.2750 | 0.2833 | 0.2917 | 0.2520 |
| PT-TDNN-D | 0.1417 | 0.2500 | 0.2500 | 0.2583 | 0.2225 |

Lastly, we choose 30 environmental sounds from [63] which includes natural sounds, sounds of things, human sounds, animal sounds, and music. Based on the SRS to represent the distance $D$ in (3.2), we solve (3.2) via finding the best linear weights [58] 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).

### *3.4.3.3 Evaluation Metric for Transferability*

The transferability has been extensively studied in the image domain [124, 106, 125, 110]. One important evaluation metric in the transfer attacks [110, 106] 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

Table 3.5 Match rates between surrogate and target models (Environmental sound).

| AE Carrier Type: | Environmental sound | | | | |
|---|---|---|---|---|---|
| Target Model: | CNN-A | CNN-B | TDNN-A | TDNN-B | Average |
| GT-CNN-C | 0.3500 | 0.3250 | 0.2417 | 0.2250 | 0.2854 |
| PT-CNN-C | 0.3083 | 0.2583 | 0.2000 | 0.1750 | 0.2353 |
| GT-CNN-D | 0.1833 | 0.3250 | 0.2417 | 0.2917 | 0.2604 |
| PT-CNN-D | 0.1583 | 0.2167 | 0.2750 | 0.2583 | 0.2271 |
| GT-TDNN-C | 0.3500 | 0.1833 | 0.3583 | 0.3417 | 0.3083 |
| PT-TDNN-C | 0.3167 | 0.1750 | 0.2833 | 0.3083 | 0.2708 |
| GT-TDNN-D | 0.1417 | 0.3083 | 0.3917 | 0.4083 | 0.3125 |
| PT-TDNN-D | 0.1250 | 0.2667 | 0.3417 | 0.3333 | 0.2667 |

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.

#### *3.4.3.4 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 $\{\text{PT-CNN-C-}i\}_{i\in[1,6]}$ models against a target model.

Table 3.3, 3.4 and 3.5 show 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.

Table 3.3, 3.4 and 3.5 also compare 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.

### 3.4.4 Defining Transferability-Perception Ratio for Evaluation

Now, given an AE carrier type $C \in \{\text{noise, feature-twisted, environmental sounds}\}$, 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

$$\mathrm{TPR}(C) = m(C)/(8 - \mathrm{SRS}(C)), \tag{3.4}$$

where $8 - \mathrm{SRS}(C)$ 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. 3.7 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

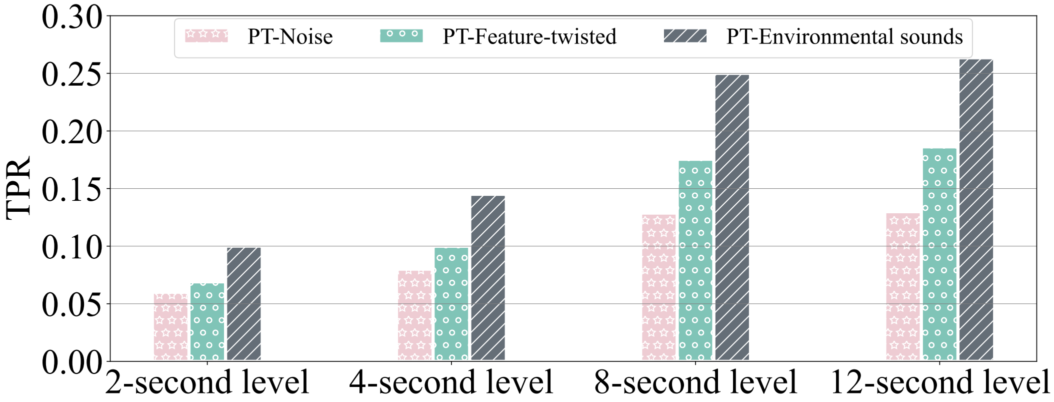


Figure 3.7 TPRs of carriers with different attack knowledge levels.

observed in Fig. 3.7 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. 3.6, Table 3.3,3.4 and 3.5. 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.

## 3.5 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.

### 3.5.1 Combining Carriers for Optimized PT-AEs

The findings in Fig 3.7 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 [53] 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: Feature-twisted environmental sounds, and manipulating the pitch [171] or the rhythm [58] is a straightforward way to twist the features of environmental sounds. We follow the same feature-twisting procedure in Section 3.4.3.2 to twist the pitch and rhythm features of environmental sounds to generate PT-AEs. Noise-based environmental sounds. We first add environmental sounds to the original speech and then use the noise attack procedure in Section 3.4.3.2 to generate PT-AEs.

Fig. 3.8 shows the TPRs of various PT-AEs generated based on adding noise, twisting the rhythm, and 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 [63]. Moreover, Fig. 3.8 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. 3.8 shows that adding noise is the least effective way to improve the TPR. Based on the results in Fig. 3.8, we consider generating PT-AEs against a black-box target model via twisting the pitch feature of environmental sounds.

### 3.5.2 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.

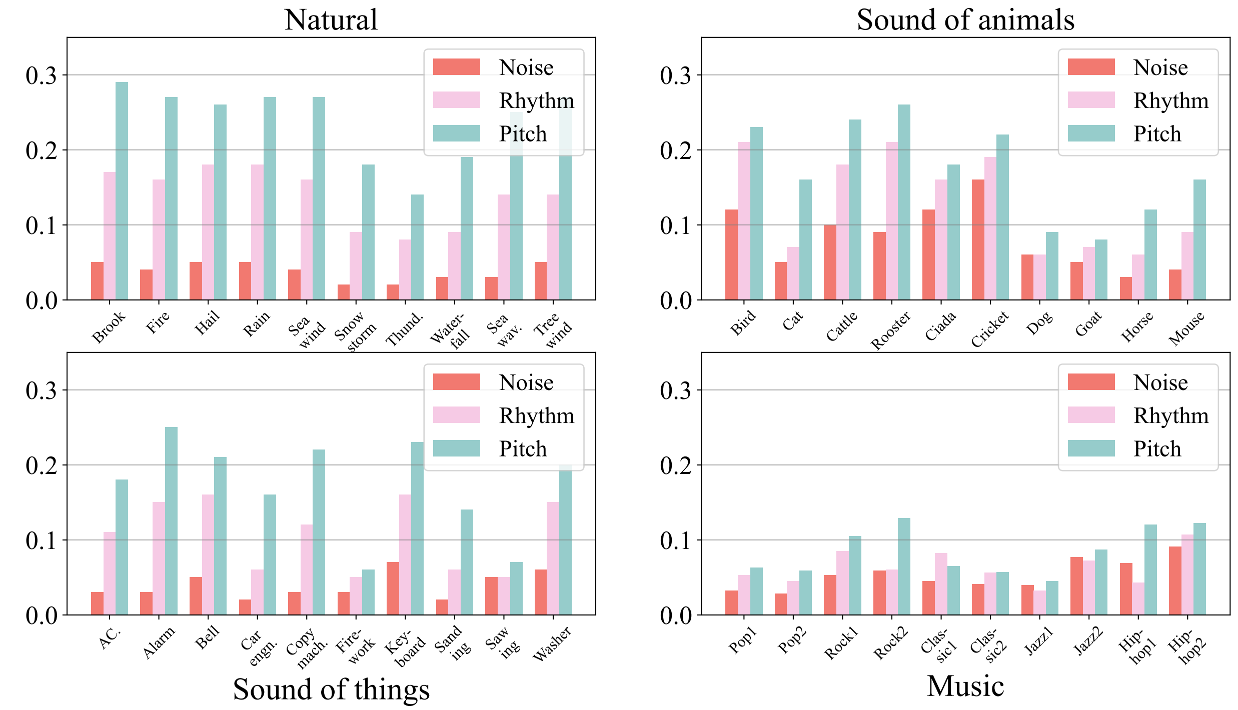


Figure 3.8 TPR of different optimized carriers.

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. 3.8). 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 [56, 62, 103, 106, 161, 168], 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:

$$\begin{aligned}\text{Objective: } & \arg\min_{\gamma_k} \Sigma_{n=1}^{N} w_n \mathcal{J}_n \left(x + \Sigma_{k=1}^{K}\gamma_k\delta_k, y_t\right) + \\ & \qquad\qquad c\,\mathrm{SRS}\big(x, x + \Sigma_{k=1}^{K}\gamma_k\delta_k\big) \qquad (3.5)\\ \text{Subject to: } & \Sigma_{k=1}^{K}\gamma_k \le \epsilon \qquad (3.6)\end{aligned}$$

where $x$ is the original speech to be perturbed to generate the attack speech; $y_t$ is the target speaker's label; (3.6) 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 (3.5) 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 (3.5). We set the uniform weight of each surrogate model [106]. 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 $[0, 1]$.

## 3.6 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.

### 3.6.1 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 (3.5). Each PT model has the same one-sentence knowledge (8-second speech) of the target speaker, which is selected from the LibriSpeech [123] or VoxCeleb1 [118] 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 [63] 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.

We observe that the ensemble loss in (3.5) 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 [35], 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.

We aim to evaluate the attacks against two major types of speaker recognition systems: 1) 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. 2) over-the-air evaluations: we perform over-the-air attack injections to the real-world smart devices.

Here is the details of evaluation metrics. 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. we evaluate the perception quality of an AE via the metric of SRS.

### 3.6.2 Evaluations of Digital-line Attacks

We consider choosing 4 different target models from statistical-based, i.e., GMM-UBM and i-vector-PLDA [5], and DNN-based, i.e., DeepSpeaker [97] and ECAPA-TDNN [55] 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.

Table 3.6 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% |

### 3.6.3 Performance of Digital-line Speaker Recognition Models

Table 3.6 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.

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.

Table 3.7 and Table 3.10 show 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 3.7, Table 3.8, and Table 3.9 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

Table 3.7 The evaluation of different attacks in digital line (Intra-gender CSI).

| Intra-gender | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tasks | CSI | | | | | | | |
| Models | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | |
| Metrics | 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 |
| Occam | 45.8% | 2.1 | 41.7% | 2.1 | 46.7% | 2.2 | 47.5% | 2.4 |
| Smack | 74.1% | 3.5 | 45.8% | 2.3 | 44.2% | 3.6 | 48.3% | 3.3 |
| QFA2SR | 76.7% | 2.2 | 70.8% | 2.4 | 76.7% | 2.1 | 77.5% | 2.1 |
| PT-AEs | 80.8% | 4.8 | 79.2% | 4.4 | 78.3% | 4.3 | 75.0% | 4.3 |

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 3.10, Table 3.11, and Table 3.12 demonstrate that the PT-AE attack is the most effective in achieving both black-box attack success and perceptual quality.

### 3.6.4 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.

Table 3.8 The evaluation of different attacks in digital line (Intra-gender OSI).

| Intra-gender | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tasks | OSI | | | | | | | |
| Models | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | |
| Metrics | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS |
| FakeBob | 4.2% | 2.9 | 5.8% | 3.1 | 6.7% | 3.2 | 9.2% | 3.1 |
| Occam | 5.0% | 1.6 | 5.8% | 1.9 | 4.2% | 2.1 | 2.5% | 2.4 |
| Smack | 10.0% | 3.2 | 13.3% | 3.6 | 9.2% | 3.5 | 8.3% | 2.6 |
| QFA2SR | 26.7% | 2.8 | 31.7% | 2.3 | 28.3% | 1.9 | 30.0% | 2.1 |
| PT-AEs | 54.2% | 4.2 | 56.7% | 3.7 | 52.5% | 4.4 | 57.5% | 3.9 |

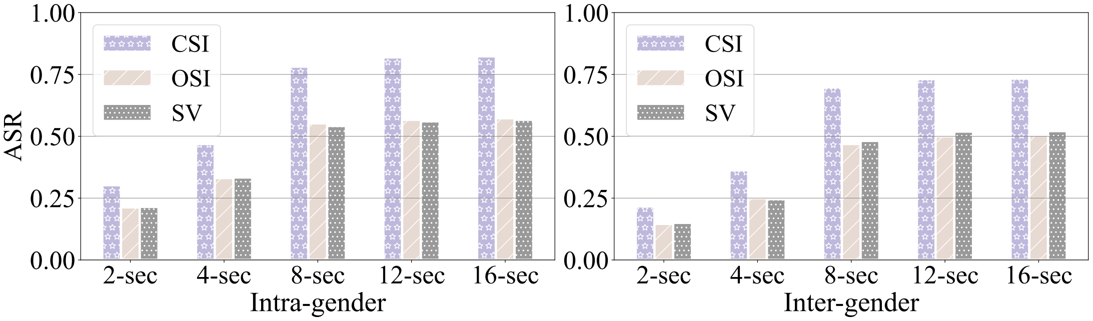


Figure 3.9 Evaluation on different attack knowledge levels.

Fig. 3.9 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 [112, 27].

Table 3.9 The evaluation of different attacks in digital line (Intra-gender SV).

| Intra-gender | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tasks | SV | | | | | | | |
| Models | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | |
| Metrics | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS |
| FakeBob | 3.0% | 2.8 | 5.8% | 2.6 | 8.3% | 2.7 | 5.8% | 3.2 |
| Occam | 5.8% | 2.0 | 5.8% | 1.9 | 5.0% | 2.2 | 4.2% | 2.1 |
| Smack | 12.5% | 3.5 | 13.3% | 3.4 | 11.7% | 2.1 | 9.2% | 2.6 |
| QFA2SR | 30.8% | 2.3 | 29.2% | 1.9 | 32.5% | 2.6 | 28.3% | 2.5 |
| PT-AEs | 55.0% | 3.9 | 56.7% | 3.4 | 54.2% | 4.1 | 50.8% | 4.2 |

Existing studies [105, 159] 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 [105, 159], 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 [105].

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.

Table 3.10 The evaluation of different attacks in digital line (Inter-gender CSI).

| Inter-gender | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tasks | CSI | | | | | | | |
| Models | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | |
| Metrics | 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 |
| Occam | 26.7% | 3.2 | 25.8% | 2.6 | 23.3% | 2.5 | 21.7% | 2.1 |
| Smack | 21.7% | 3.4 | 26.7% | 3.4 | 19.2% | 3.0 | 17.5% | 3.6 |
| QFA2SR | 46.7% | 1.9 | 35.8% | 2.2 | 43.3% | 2.6 | 35.8% | 2.4 |
| PT-AEs | 71.7% | 4.3 | 70.8% | 4.3 | 70.0% | 4.6 | 66.7% | 5.1 |

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 3.6.2. Our goal is to capture various phoneme diversities under different speech lengths. Table 3.13 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 3.13 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

Table 3.11 The evaluation of different attacks in digital line (Inter-gender OSI).

| Inter-gender | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tasks | OSI | | | | | | | |
| Models | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | |
| Metrics | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS |
| FakeBob | 2.5% | 2.9 | 1.7% | 2.7 | 4.2% | 2.6 | 2.5% | 2.4 |
| Occam | 5.8% | 2.8 | 10.0% | 2.1 | 10.8% | 2.3 | 7.5% | 1.5 |
| Smack | 12.5% | 3.2 | 14.2% | 3.1 | 13.3% | 2.8 | 15.8% | 2.7 |
| QFA2SR | 21.7% | 1.5 | 24.2% | 1.6 | 25.8% | 2.6 | 27.5% | 2.8 |
| PT-AEs | 45.8% | 3.8 | 48.3% | 3.6 | 46.7% | 3.5 | 49.1% | 3.7 |

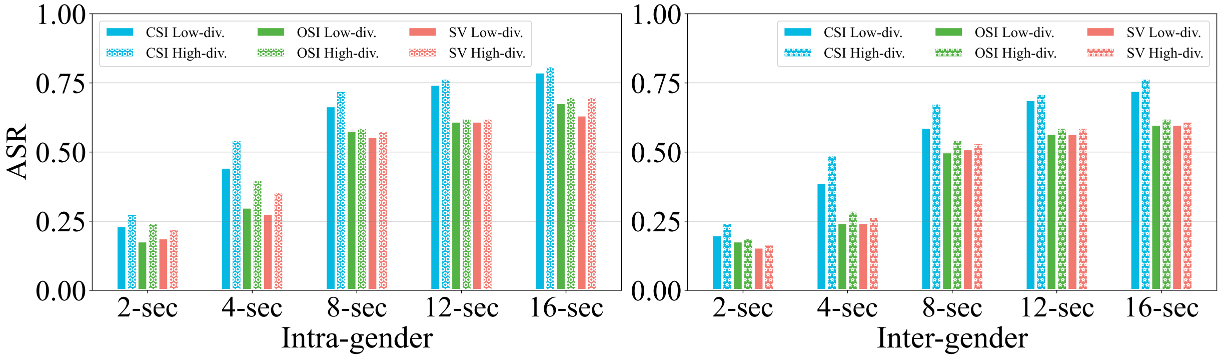


Figure 3.10 Evaluation on phoneme diversity.

group has 24.2. Despite this difference, they have a similar total number of phonemes (80.4 vs 80.6).

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.

Fig. 3.10 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

Table 3.12 The evaluation of different attacks in digital line (Inter-gender SV).

| Inter-gender | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Tasks | SV | | | | | | | |
| Models | Deep Speaker | | ECAPA-TDNN | | GMM-UBM | | i-vector-PLDA | |
| Metrics | ASR | SRS | ASR | SRS | ASR | SRS | ASR | SRS |
| FakeBob | 2.5% | 2.1 | 1.7% | 2.8 | 3.3% | 2.7 | 2.5% | 2.9 |
| Occam | 9.2% | 2.7 | 10.0% | 2.6 | 10.0% | 2.6 | 6.7% | 2.2 |
| Smack | 11.7% | 3.3 | 15.8% | 3.1 | 14.2% | 2.9 | 15.0% | 2.7 |
| QFA2SR | 26.7% | 2.1 | 23.3% | 2.3 | 26.7% | 2.4 | 27.5% | 2.2 |
| PT-AEs | 46.7% | 3.9 | 48.3% | 3.6 | 49.1% | 3.8 | 48.3% | 4.1 |

Table 3.13 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.

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 [78] 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

Table 3.14 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 |

the 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.

### 3.6.5 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[13], and Apple HomePod (Siri) [3]. 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.

Table 3.14 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.

### 3.6.6 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 3.15 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.

### 3.6.7 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 pro-

Table 3.15 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% |

pose to evaluate the contribution of each individual component to the ASR. Our methodology is similar to the One-at-a-time (OAT) strategy in [58]. 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 3.6.5. 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.

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.

As shown in Table 3.16 (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

Table 3.16 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% |

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.

To understand the contribution of the feature-twisted environment sound carrier, we use two baseline attacks related to noise and feature-twisted carriers. 1) 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. 2) Feature-twisted carriers, as discussed in Section 3.5.1, 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 (3.5) via finding the optimal weights for the twisted-pitch carriers, with a total energy threshold of $\epsilon = 0.08$.

Table 3.16 (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.

We note that our ensemble-based model in (3.5) combines multiple CNN and TDNN models. To evaluate the contribution of ensemble learning, we design two sets of experiments. First, we replace the ensemble-based model in (3.5) with just a single PT-CNN or PT-TDNN

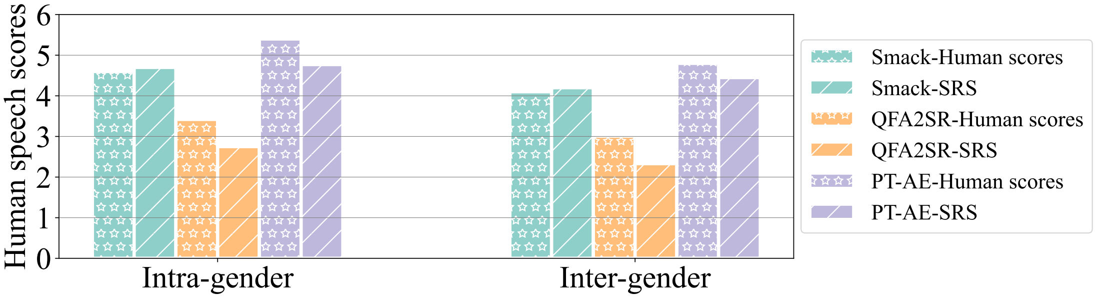


Figure 3.11 Human evaluation on the AEs.

model to compare the ASRs. Second, we replace (3.5) 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).

We can observe in Table 3.16 (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.

### 3.6.8 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 3.4.2 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 3.4.2. Following the same procedure, we ask each volunteer to rate each pair of original and PT-AE samples.

Fig. 3.11 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 3.14. 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. 3.11 further validates that the PT-AEs have better perceptual quality than AEs generated by other methods.

### 3.6.9 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.

## 3.7 Related Work

Adversarial audio attacks [36, 172, 102, 149, 160, 41, 57, 176, 57, 38, 176] can be categorized into white-box and black-box attacks depending on their attack knowledge level.

White-box attacks [36, 132] 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 [102, 71] via adding the perturbation to the original speech signal without synchronization, albeit still assuming nearly full knowledge of the target model.

Existing black-box attacks [38, 176, 149, 160, 104, 171] 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 [38, 57, 176, 171, 104] needed to interact with the target model to get the internal prediction scores [38, 160, 41, 171] or hard label results [176, 104]. A large number of queries were necessary for the black-box attack to be effective. For example, Occam [176] 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.

The transfer-based attacks [18, 58, 39] commonly assumed no interaction or limited probing [41] to the target model. For example, Kenansville [18] manipulated the phoneme of the speech to achieve an untargeted attack. QFA2SR [39] 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.

Some recent studies [132, 71, 104] leveraged the psychoacoustic feature to optimize the carriers and improve the perception of AEs. Meanwhile, [58, 171] manipulated the features of an audio signal to create AEs with good perceptual quality. In addition, there are audio attack strategies [174, 34, 17, 172] focusing on improving the stealthiness of the AEs. For example, dolphin attack [174] 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 [58] 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.

## 3.8 Discussion on Defense

To combat PT-AEs, there are two major defense directions available: audio signal processing and adversarial training. Audio signal processing has been proposed to defend against AEs via down-sampling [104, 176], quantization [170], and low-pass filtering [102] 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 [176, 102, 71], 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 [70, 109, 23, 30, 142, 151, 166] 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 [176] 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 work 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.

## 3.9 Summary

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.

# Chapter 4: Conclusion and Future Work

In this dissertation, we first investigate how to integrate human factors into the adversarial attack loops. Specifically, we conduct a human study to understand how human participants perceive the music signal perturbation. We use regression analysis to model the relationship between the audio feature deviation and the human-perceived deviation for music signals. Then, Based on the regressed human perception model, we propose, formulate, and evaluate the perception-aware attack framework to create adversarial music. The perception-aware attack is able to perturb music signals with better perceptual quality and achieve higher attack success rates than conventional $L_p$ norm based attacks against YouTube's copyright detector. To the best of our knowledge, our study presents the first systematic work that integrates human factors into the internals of adversarial audio attacks. We believe the results will encourage further human-in-the-loop research.

Furthermore, we keep exploring using limited practical knowledge to launch adversarial attacks on the real-world speaker recognition models. Specifically, 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. We propose a new TPR framework to jointly evaluate the transferability and perceptual quality for PT-AE generations with different types of carriers. 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.

For future work, it is worth exploring AI models rooted in human-centric principles, ethics, and user-centricity, collaborating with interdisciplinary teams to infuse transparency, interpretability, and ethical values. My research spans various data domains like audio, image, and NLP, with a focus on user-centered design, expert feedback, and human-computer interaction research. The goal is to create AI models that prioritize human well-being, foster trust, and uphold ethical standards in their deployment across diverse data modalities.

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

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

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

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

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

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

[9] Google Assistant. https://assistant.google.com/, 2022. Accessed: 2022-01-07.

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

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

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

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

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

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

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

[17] 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 Proc. of NDSS*, 2019.

[18] 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.

[19] 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.

[20] Eric Allamanche. Audioid: Towards content-based identification of audio material. In *Proc. of AES*, 2001.

[21] 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.

[22] Robert Bailis, Majid Ezzati, and Daniel M Kammen. Mortality and greenhouse gas impacts of biomass and petroleum energy futures in africa. *In Proc. of Science*, 2005.

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

[24] Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *Proc. of ICLR*, 2019.

[25] 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.

[26] Laurence Boney, Ahmed H Tewfik, and Khaled N Hamdy. Digital watermarks for audio signals. In *Proc. of ICMCS*, 1996.

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

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

[29] 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.

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

[31] J Elliott Campbell, Gregory R Carmichael, T Chai, M Mena-Carrasco, Y Tang, DR Blake, NJ Blake, Stephanie A Vay, G James Collatz, I Baker, et al. Photosynthetic control of atmospheric carbonyl sulfide during the growing season. *In Proc. of Science*, 2008.

[32] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. A review of audio fingerprinting. *Journal of VLSI signal processing systems for signal, image and video technology*, 41(3):271–284, 2005.

[33] Pedro Cano, Eloi Batlle, Harald Mayer, and Helmut Neuschmied. Robust sound modeling for song detection in broadcast audio. *In Proc. AES 112th Int. Conv*, 2002.

[34] 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.

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

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

[37] Michael A Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-based music information retrieval: Current directions and future challenges. *In Proc. of IEEE*, 2008.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Ingemar J Cox, Matthew L Miller, Jeffrey Adam Bloom, and Chris Honsinger. *Digital watermarking*, volume 53. Springer, 2002.

[45] 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.

[46] Wayne W Daniel. The spearman rank correlation coefficient. *In Proc. of Biostatistics: A Foundation for Analysis in the Health Sciences*, 1987.

[47] 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.

[48] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Li Chen, Michael E Kounavis, and Duen Horng Chau. Adagio: Interactive experimentation with adversarial attack and defense for audio. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 677–681. Springer, 2018.

[49] 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.

[50] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *In Proc. of IEEE TASSP*, 1980.

[51] Franz De Leon and Kirk Martinez. Enhancing timbre model using mfcc and its time derivatives for music similarity estimation. In *Proc. of EUSIPCO*, 2012.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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 *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 905–919, 2022.

[59] Zhiyao Duan, Bryan Pardo, and Changshui Zhang. Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(8):2121–2133, 2010.

[60] Sébastien Fenet, Gaël Richard, Yves Grenier, et al. A scalable audio fingerprint method with robustness to pitch-shifting. In *Proc. of ISMIR*, pages 121–126, 2011.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] Simon Godsill and Manuel Davy. Bayesian harmonic models for musical pitch estimation and analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–1769. IEEE, 2002.

[66] SIMON J Godsill and M Davy. Bayesian harmonic models for musical signal analysis. *In Proc. of Bayesian Statistics*, 7:105–124, 2003.

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

[68] Emilia Gomez, Pedro Cano, L Gomes, Eloi Batlle, and Madeleine Bonnet. Mixed watermarking-fingerprinting approach for integrity verification of audio recordings. In *Proc. of ITelCon*, 2002.

[69] 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.

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

[71] 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.

[72] Chitralekha Gupta, Haizhou Li, and Ye Wang. Perceptual evaluation of singing quality. In *Proc. of APSIPA ASC*, pages 577–586, 2017.

[73] Chitralekha Gupta, Haizhou Li, and Ye Wang. A technical framework for automatic perceptual evaluation of singing quality. *In Proc. of APSIPA Transactions on Signal and Information Processing*, 7, 2018.

[74] Jaap Haitsma and Ton Kalker. A highly robust audio fingerprinting system. In *Proc. of Ismir*, volume 2002, pages 107–115, 2002.

[75] Jaap Haitsma, Ton Kalker, and Job Oostveen. Robust audio hashing for content identification. In *Proc. of CBMIW*, 2001.

[76] William M Hartmann. *Signals, sound, and sensation.* In Proc. of Springer Science & Business Media, 2004.

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

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.

[87] Corey Kereliuk, Bertrand Scherrer, Vincent Verfaille, Philippe Depalle, and Marcelo M Wanderley. Indirect acquisition of fingerings of harmonic notes on the flute. In *Proc. of ICMC*, 2007.

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

[89] 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.

[90] 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.

[91] 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.

[92] 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.

[93] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. Fooling end-to-end speaker verification with adversarial examples. In *Proc. of ICASSP*, pages 1962–1966. IEEE, 2018.

[94] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016.

[95] Lily NC Law and Marcel Zentner. Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PloS one*, 7(12):e52508, 2012.

[96] 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.

[97] 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.

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

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

[100] Juncheng B Li, Shuhui Qu, Xinjian Li, Zico Kolter, and Florian Metze. Real world audio adversary against wake-word detection systems. *In Proc. of NIPS*.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

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

[107] Dominik B Loeffler. *Instrument timbres and pitch estimation in polyphonic music.* PhD thesis, Georgia Institute of Technology, 2006.

[108] 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.

[109] 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.

[110] 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.

[111] 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.

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

[113] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *Proc. of ICML*, pages 3578–3586. PMLR, 2018.

[114] Emilio Molina, Isabel Barbancho, Emilia Gómez, Ana Maria Barbancho, and Lorenzo J Tardón. Fundamental frequency alignment vs. note-based melodic similarity for singing voice assessment. In *Proc. of International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, pages 744–748. IEEE, 2013.

[115] James Anderson Moorer. Signal processing aspects of computer music: A survey. *In Proc. of the IEEE*, 65(8):1108–1137, 1977.

[116] Meinard Muller, Daniel PW Ellis, Anssi Klapuri, and Gaël Richard. Signal processing for music analysis. *IEEE Journal of selected topics in signal processing*, 5(6):1088–1110, 2011.

[117] James M Murphy, David MH Sexton, David N Barnett, Gareth S Jones, Mark J Webb, Matthew Collins, and David A Stainforth. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *In Proc. of Nature*, 2004.

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

[119] 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.

[120] 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.

[121] Helmut Neuschmied, Harald Mayer, and Eloi Batlle. Content-based identification of audio titles on the internet. In *Proc. of WEDELMUSIC*, 2001.

[122] 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.

[123] 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.

[124] 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.

[125] 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.

[126] Bryan Pardo. Finding structure in audio for music information retrieval. *IEEE Signal Processing Magazine*, 23(3):126–132, 2006.

[127] 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.

[128] 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.

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

[130] Hervé Platel, Cathy Price, Jean-Claude Baron, Richard Wise, Jany Lambert, Richard S Frackowiak, Bernard Lechevalier, and Francis Eustache. The structural components of music perception. a functional anatomical study. *Brain: a journal of neurology*, 120(2):229–243, 1997.

[131] 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.

[132] 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.

[133] Chotirat Ann Ratanamahatana and Eamonn Keogh. Making time-series classification more accurate using learned constraints. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 11–22. SIAM, 2004.

[134] 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.

[135] Jean-Claude Risset and David L Wessel. Exploration of timbre by analysis and synthesis. In *The psychology of music*, pages 113–169. Elsevier, 1999.

[136] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, volume 2, pages 749–752. IEEE, 2001.

[137] Parsa Saadatpanah, Ali Shafahi, and Tom Goldstein. Adversarial attacks on copyright detection systems. In *Proc. of ICML*, pages 8307–8315. PMLR, 2020.

[138] Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

[139] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *In Proc. of Intelligent Data Analysis*, 11(5):561–580, 2007.

[140] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. *In Proc. of NDSS*, 2019.

[141] Philip Sedgwick. Spearman's rank correlation coefficient. *In Proc, of Bmj*, 349, 2014.

[142] 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.

[143] Yash Sharma and Pin-Yu Chen. Attacking the madry defense model with $l$_1-based adversarial examples. *In Proc. of ICLR Work Shop*, 2018.

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

[145] 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.

[146] 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.

[147] Reinhard Sonnleitner and Gerhard Widmer. Robust quad-based audio fingerprinting. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(3):409–421, 2015.

[148] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[149] 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.

[150] Thilo Thiede, William C Treurniet, Roland Bitto, Christian Schmidmer, Thomas Sporer, John G Beerends, and Catherine Colomes. Peaq-the itu standard for objective measurement of perceived audio quality. *Journal of the Audio Engineering Society*, 48(1/2):3–29, 2000.

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

[152] Wei-Ho Tsai and Hsin-Chieh Lee. Automatic evaluation of karaoke singing based on pitch, volume, and rhythm features. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1233–1243, 2011.

[153] Stefan Uhlich, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi, and Yuki Mitsufuji. Improving music source separation based on deep neural networks through data augmentation and network blending. In *Proc. of ICASSP*, pages 261–265. IEEE, 2017.

[154] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pages 146–152, 2016.

[155] Saeed V Vaseghi. *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

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

[157] Paul J Walmsley, Simon J Godsill, and Peter JW Rayner. Multidimensional optimisation of harmonic signals. In *9th European Signal Processing Conference (EUSIPCO 1998)*, pages 1–4. IEEE, 1998.

[158] Avery Wang et al. An industrial strength audio search algorithm. In *Proc. of Ismir*, volume 2003, pages 7–13. Washington, DC, 2003.

[159] 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.

[160] 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.

[161] 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.

[162] 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.

[163] David L Wessel. Timbre space as a musical control structure. *Computer music journal*, pages 45–52, 1979.

[164] 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.

[165] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proc. of ICML*, pages 5286–5295. PMLR, 2018.

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

[167] 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.

[168] 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.

[169] Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. *Proc. of IJCAI*, 2018.

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

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

[172] 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.

[173] 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.

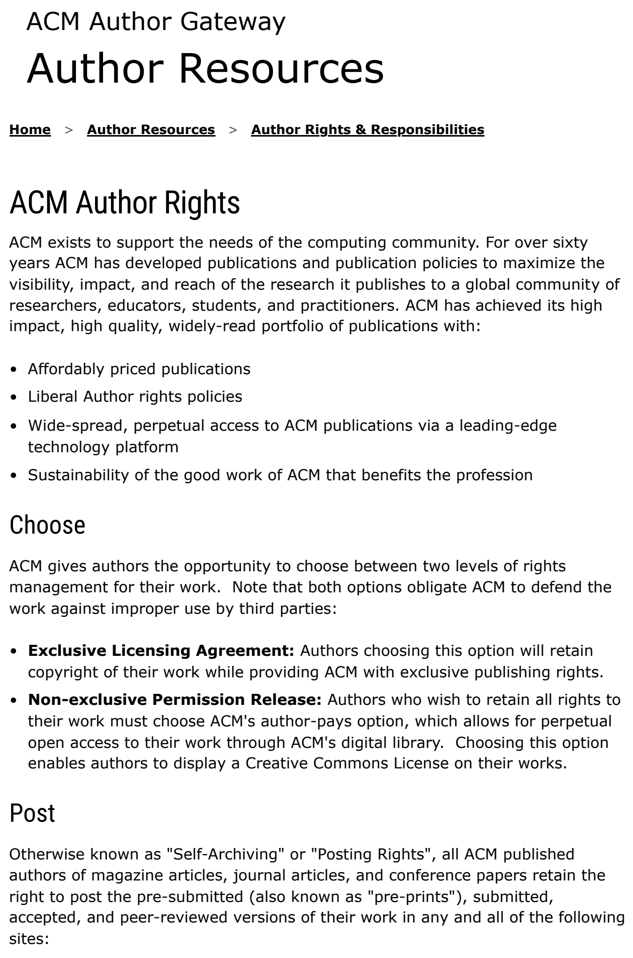
[174] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proc. of ACM CCS*, pages 103–117, 2017.

[175] 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.

[176] 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: Copyright Permissions

The permission below is for the reproduction of material in Chapter 2.



ACM Author Gateway

## Author Resources

Home > Author Resources > Author Rights & Responsibilities

### ACM Author Rights

ACM exists to support the needs of the computing community. For over sixty years ACM has developed publications and publication policies to maximize the visibility, impact, and reach of the research it publishes to a global community of researchers, educators, students, and practitioners. ACM has achieved its high impact, high quality, widely-read portfolio of publications with:

- Affordably priced publications
- Liberal Author rights policies
- Wide-spread, perpetual access to ACM publications via a leading-edge technology platform
- Sustainability of the good work of ACM that benefits the profession

#### Choose

ACM gives authors the opportunity to choose between two levels of rights management for their work. Note that both options obligate ACM to defend the work against improper use by third parties:

- **Exclusive Licensing Agreement:** Authors choosing this option will retain copyright of their work while providing ACM with exclusive publishing rights.
- **Non-exclusive Permission Release:** Authors who wish to retain all rights to their work must choose ACM's author-pays option, which allows for perpetual open access to their work through ACM's digital library. Choosing this option enables authors to display a Creative Commons License on their works.

#### Post

Otherwise known as "Self-Archiving" or "Posting Rights", all ACM published authors of magazine articles, journal articles, and conference papers retain the right to post the pre-submitted (also known as "pre-prints"), submitted, accepted, and peer-reviewed versions of their work in any and all of the following sites:

- Author's Homepage
- Author's Institutional Repository
- Any Repository legally mandated by the agency or funder funding the research on which the work is based
- Any Non-Commercial Repository or Aggregation that does not duplicate ACM tables of contents. Non-Commercial Repositories are defined as Repositories owned by non-profit organizations that do not charge a fee to access deposited articles and that do not sell advertising or otherwise profit from serving scholarly articles.

For the avoidance of doubt, an example of a site ACM authors may post all versions of their work to, with the exception of the final published "Version of Record", is ArXiv. ACM does request authors, who post to ArXiv or other permitted sites, to also post the published version's Digital Object Identifier (DOI) alongside the pre-published version on these sites, so that easy access may be facilitated to the published "Version of Record" upon publication in the ACM Digital Library.

Examples of sites ACM authors may not post their work to are ResearchGate, Academia.edu, Mendeley, or Sci-Hub, as these sites are all either commercial or in some instances utilize predatory practices that violate copyright, which negatively impacts both ACM and ACM authors.

After an ACM journal submission has been accepted and has entered the production process, ACM makes the Author’s Accepted Manuscript (AAM) available for preview under the ACM “Just Accepted” program until the “Version of Record” is available and assigned to its proper issue. The AAM carries the article’s permanent DOI and can be cited immediately.

## Distribute

Authors can post an Author-Izer link enabling free downloads of the Definitive Version of the work permanently maintained in the ACM Digital Library.

- On the Author's own Home Page or
- In the Author's Institutional Repository.

## Reuse

Authors can reuse any portion of their own work in a new work of their own (and no fee is expected) as long as a citation and DOI pointer to the Version of Record in the ACM Digital Library are included.

- Contributing complete papers to any edited collection of reprints for which the author is notthe editor, requires permission and usually a republication fee.
- Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).
- Commercially produced course-packs that are sold to students require permission and possibly a fee.

## Create

ACM's copyright and publishing license include the right to make Derivative Works or new versions. For example, translations are "Derivative Works." By copyright or license, ACM may have its publications translated. However, ACM Authors continue to hold perpetual rights to revise their own works without seeking permission from ACM.

Minor Revisions and Updates to works already published in the ACM Digital Library are welcomed with the approval of the appropriate Editor-in-Chief or Program Chair.

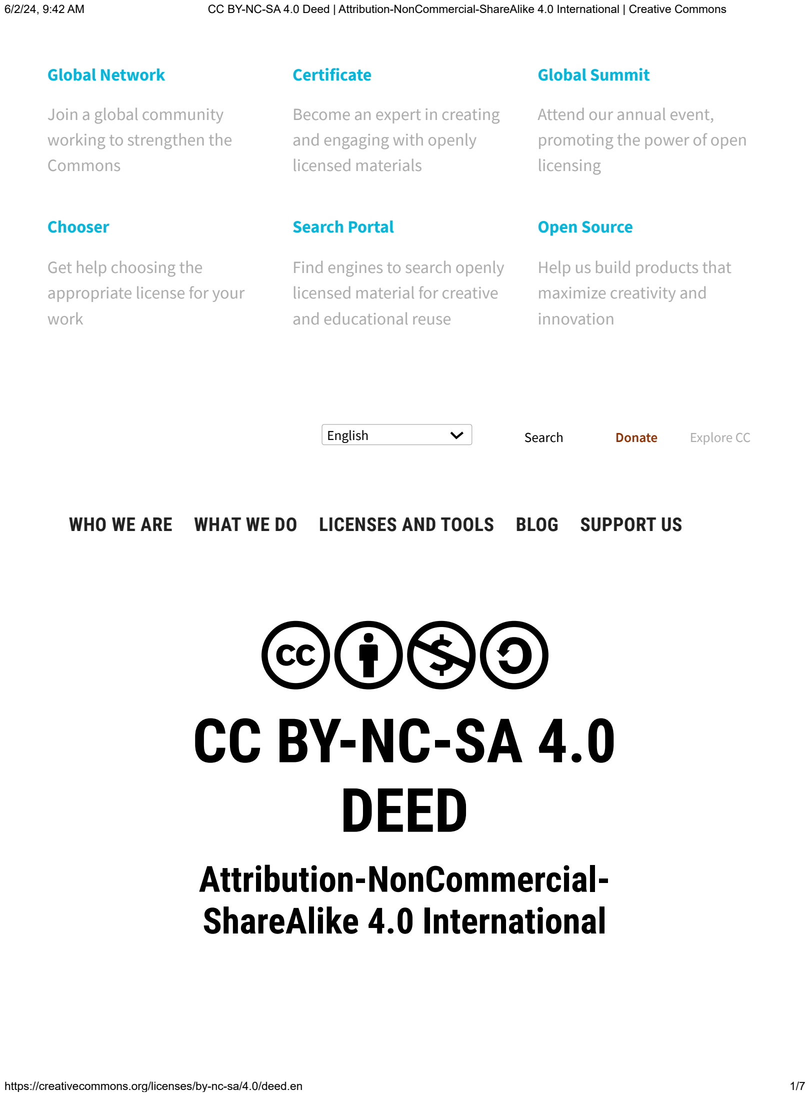
- If the revision is minor, i.e., less than 25% of new substantive material, then the work should still have ACM's publishing notice, DOI pointer to the Definitive Version, and be labeled a "Minor Revision of"
- If the revision is major, i.e., 25% or more of new substantive material, then ACM considers this a new work in which the author retains full copyright ownership (despite ACM's copyright or license in the original published article) and the author need only cite the work from which this new one is derived.

## Retain

Authors retain all perpetual rights laid out in the ACM Author Rights and Publishing Policy, including, but not limited to:

- Sole ownership and control of third-party permissions to use for artistic images intended for exploitation in other contexts
- All patent and moral rights
- Ownership and control of third-party permissions to use of software published by ACM

The permission below is for the reproduction of material in Chapter 3.



**Global Network**

Join a global community working to strengthen the Commons

**Certificate**

Become an expert in creating and engaging with openly licensed materials

**Global Summit**

Attend our annual event, promoting the power of open licensing

**Chooser**

Get help choosing the appropriate license for your work

**Search Portal**

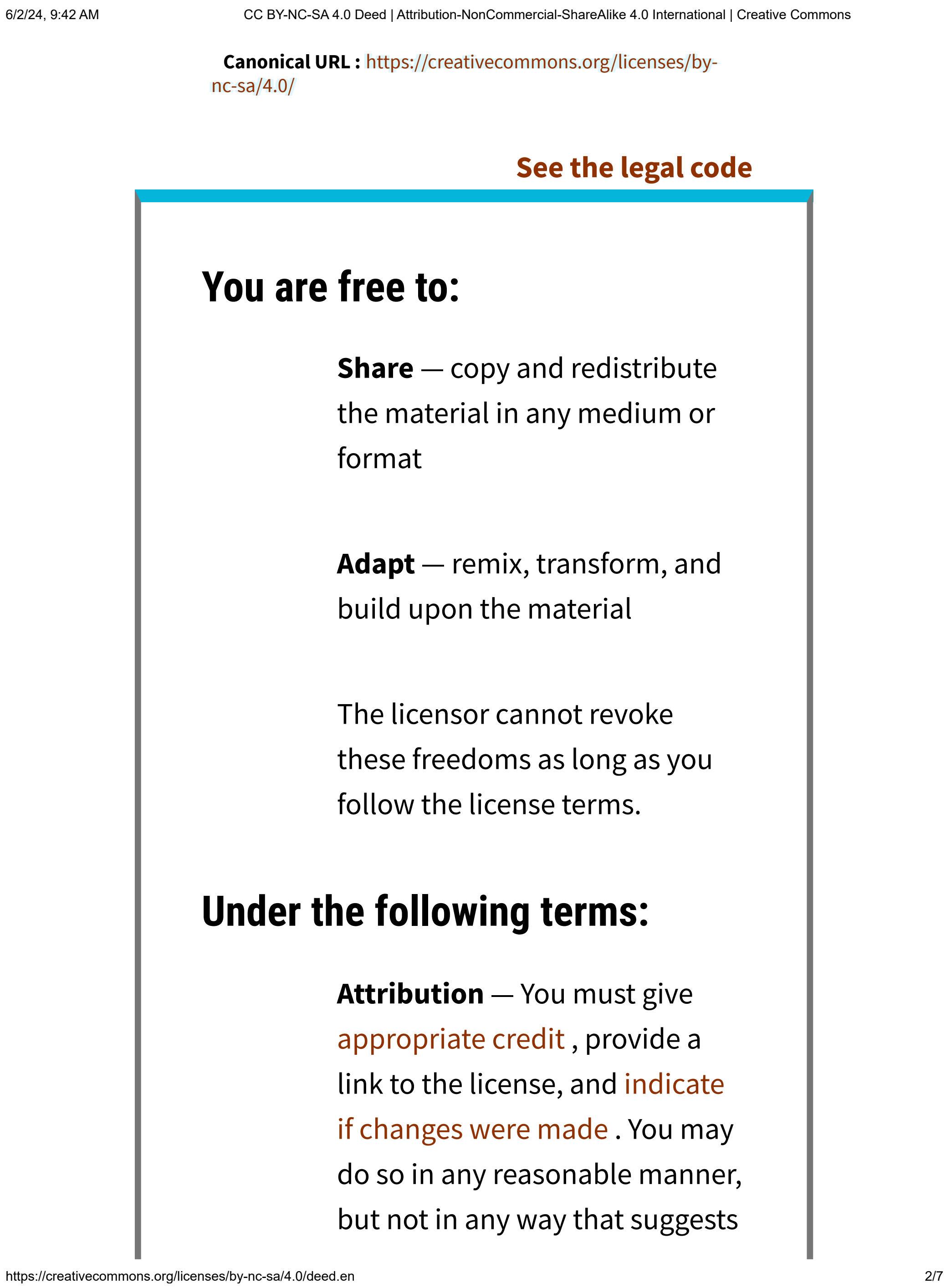
Find engines to search openly licensed material for creative and educational reuse

**Open Source**

Help us build products that maximize creativity and innovation

# CC BY-NC-SA 4.0 DEED

## Attribution-NonCommercial-ShareAlike 4.0 International



**Canonical URL :** https://creativecommons.org/licenses/by-nc-sa/4.0/

See the legal code

## You are free to:

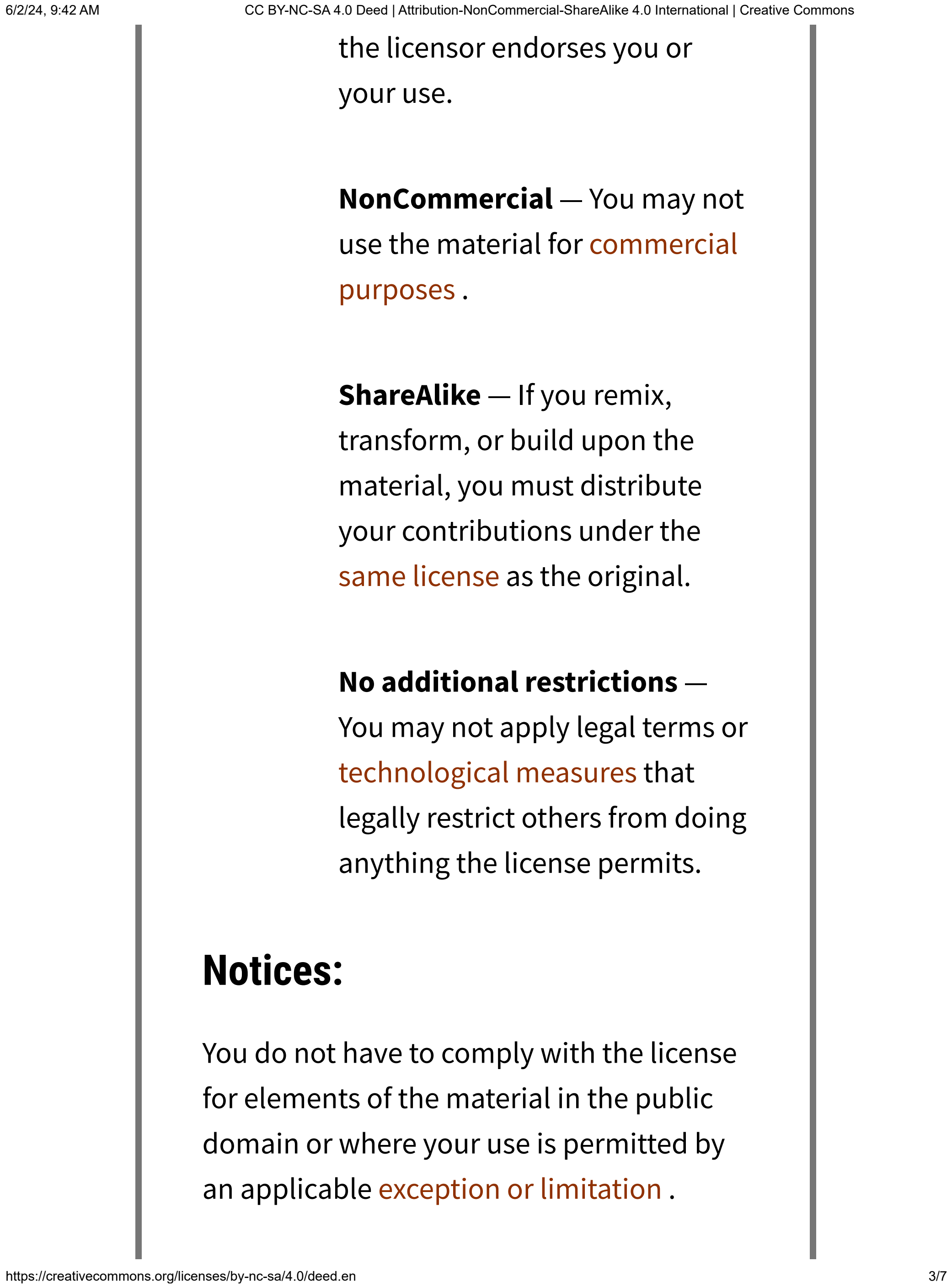
**Share** — copy and redistribute the material in any medium or format

**Adapt** — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

## Under the following terms:

**Attribution** — You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests



the licensor endorses you or your use.

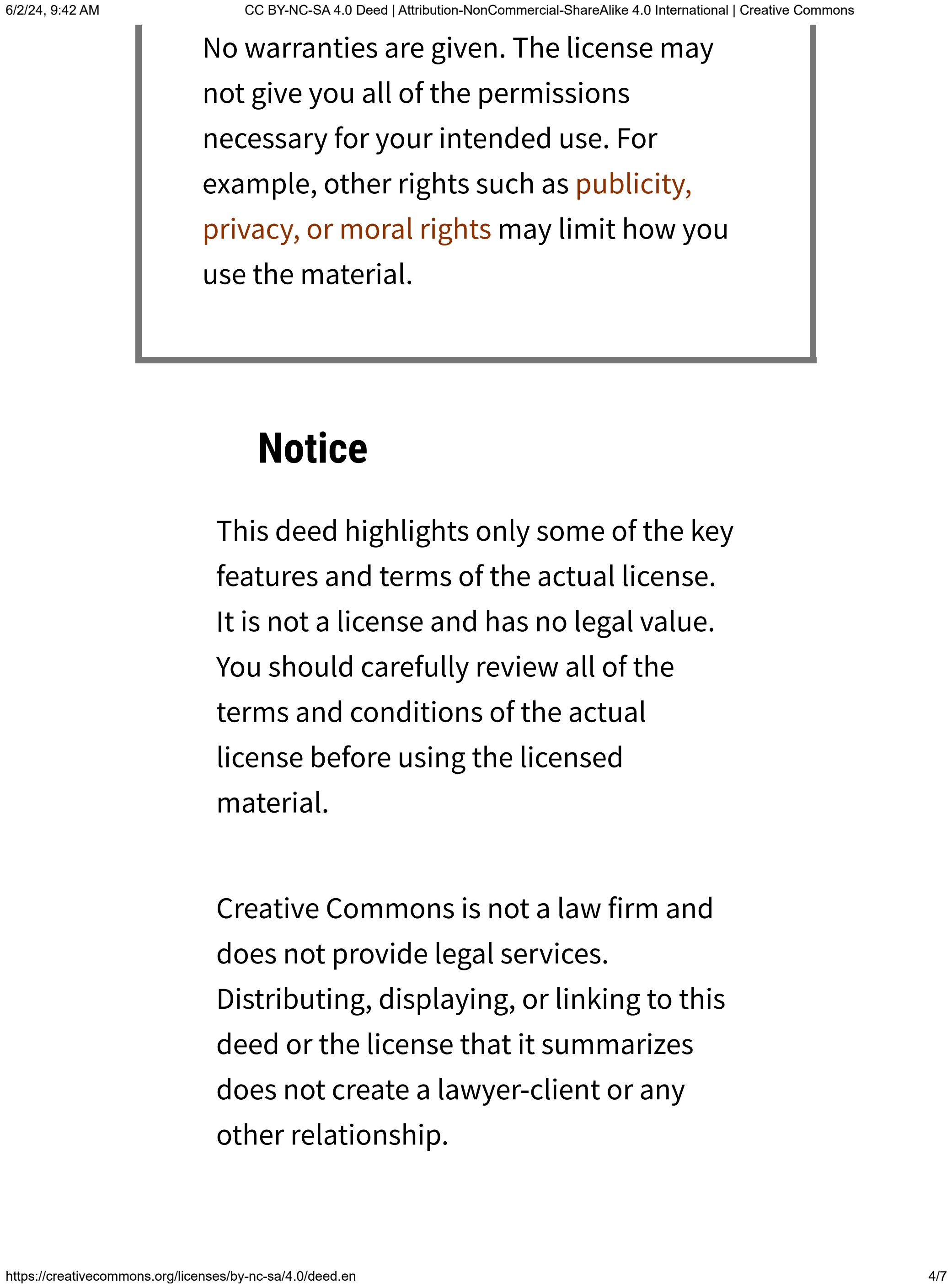
**NonCommercial** — You may not use the material for commercial purposes .

**ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

**No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

## Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation .

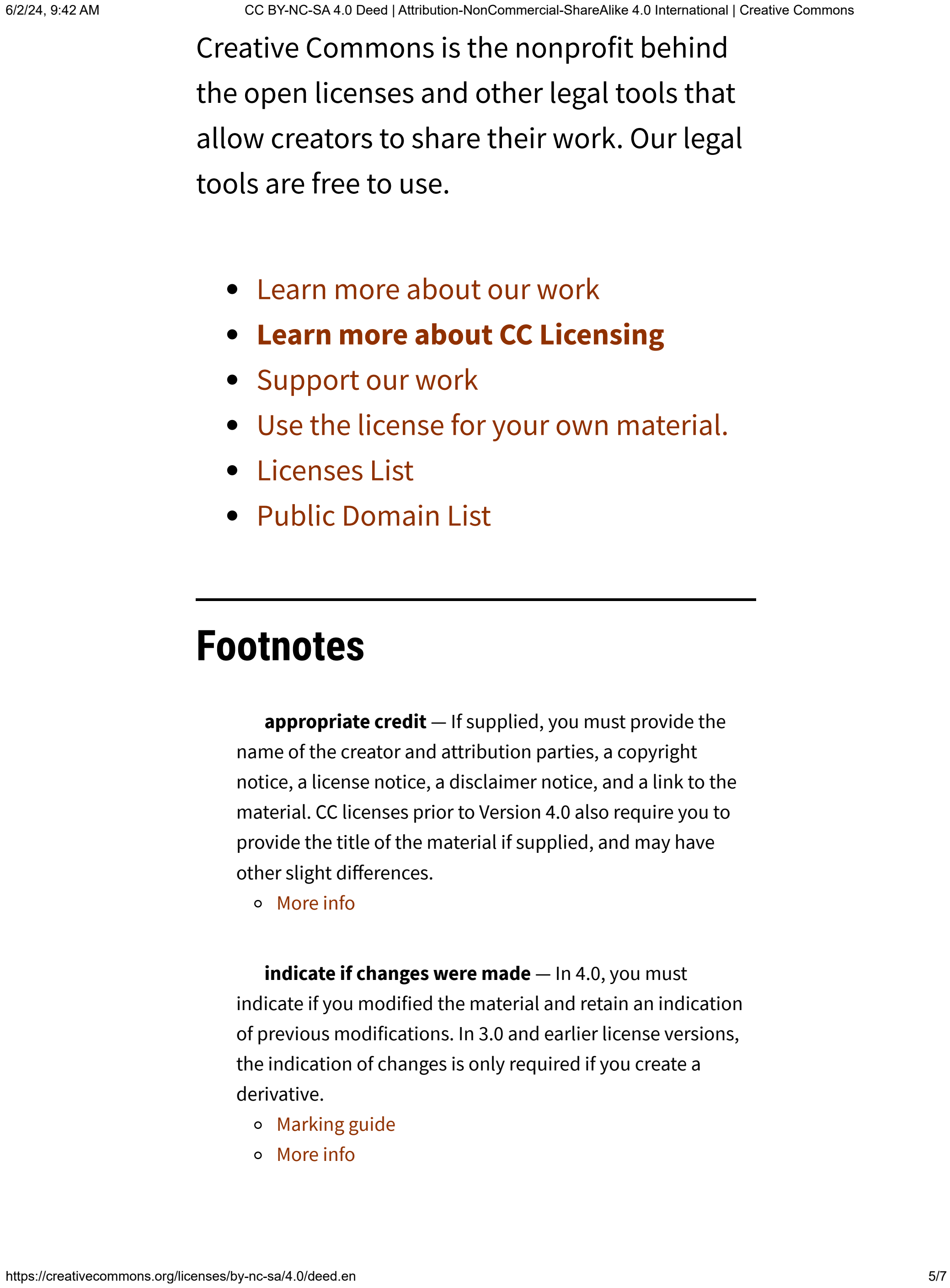


No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

## Notice

This deed highlights only some of the key features and terms of the actual license. It is not a license and has no legal value. You should carefully review all of the terms and conditions of the actual license before using the licensed material.

Creative Commons is not a law firm and does not provide legal services. Distributing, displaying, or linking to this deed or the license that it summarizes does not create a lawyer-client or any other relationship.



Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

- Learn more about our work
- **Learn more about CC Licensing**
- Support our work
- Use the license for your own material.
- Licenses List
- Public Domain List

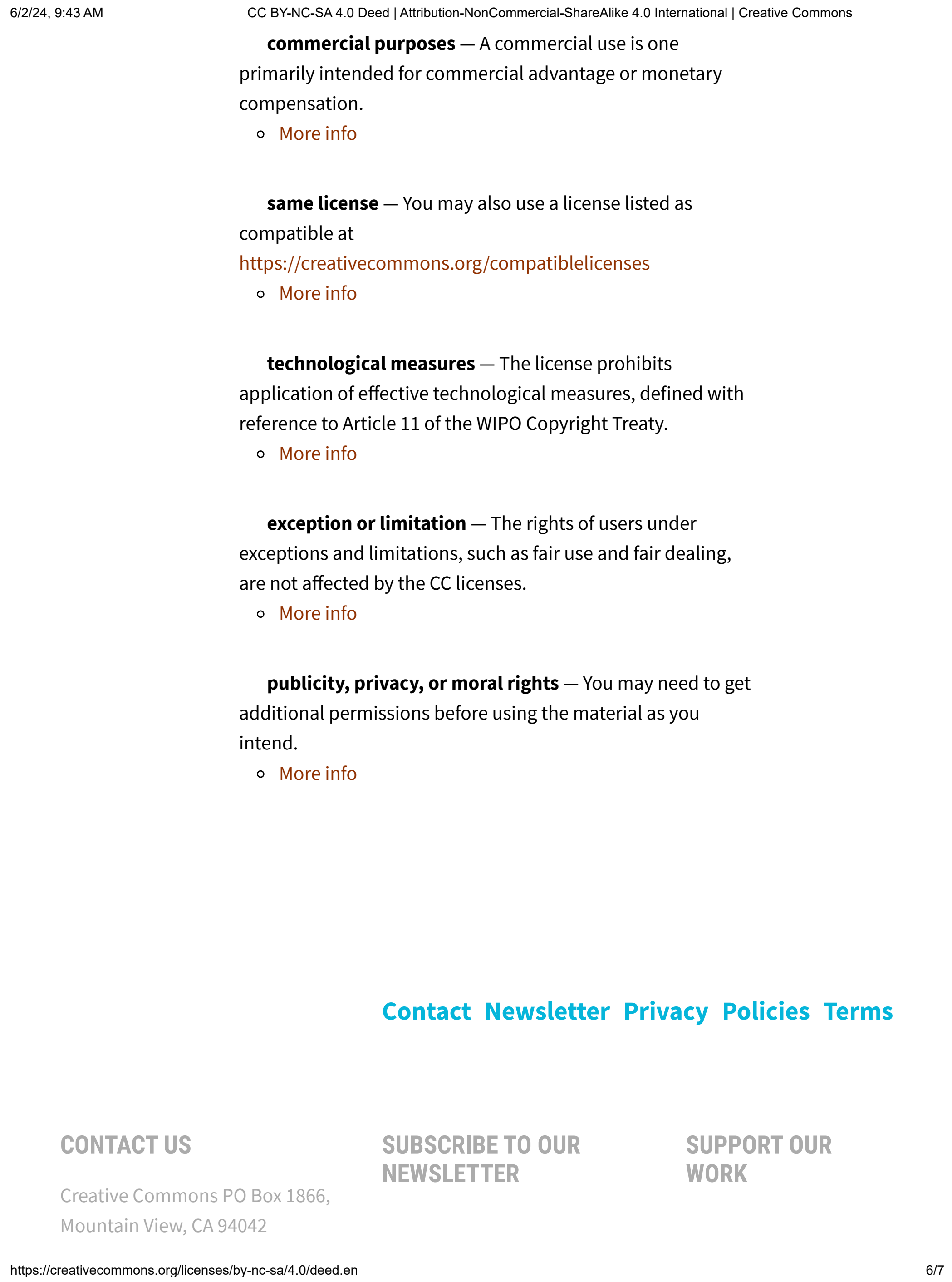
## Footnotes

**appropriate credit** — If supplied, you must provide the name of the creator and attribution parties, a copyright notice, a license notice, a disclaimer notice, and a link to the material. CC licenses prior to Version 4.0 also require you to provide the title of the material if supplied, and may have other slight differences.

- More info

**indicate if changes were made** — In 4.0, you must indicate if you modified the material and retain an indication of previous modifications. In 3.0 and earlier license versions, the indication of changes is only required if you create a derivative.

- Marking guide
- More info



**commercial purposes** — A commercial use is one primarily intended for commercial advantage or monetary compensation.

- More info

**same license** — You may also use a license listed as compatible at https://creativecommons.org/compatiblelicenses

- More info

**technological measures** — The license prohibits application of effective technological measures, defined with reference to Article 11 of the WIPO Copyright Treaty.

- More info

**exception or limitation** — The rights of users under exceptions and limitations, such as fair use and fair dealing, are not affected by the CC licenses.

- More info

**publicity, privacy, or moral rights** — You may need to get additional permissions before using the material as you intend.

- More info

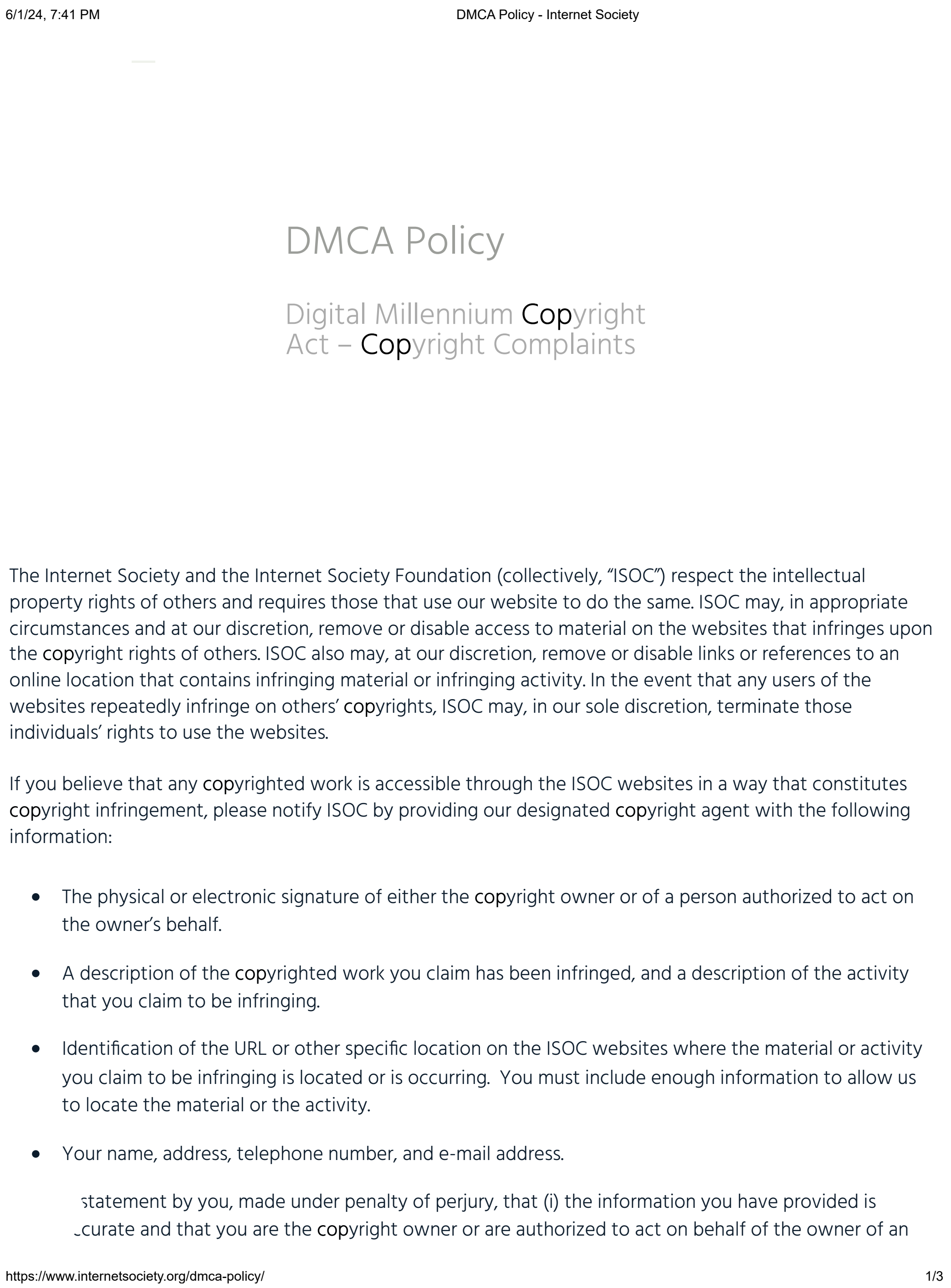
Contact Newsletter Privacy Policies Terms

CONTACT US

Creative Commons PO Box 1866,
Mountain View, CA 94042

SUBSCRIBE TO OUR NEWSLETTER

SUPPORT OUR WORK



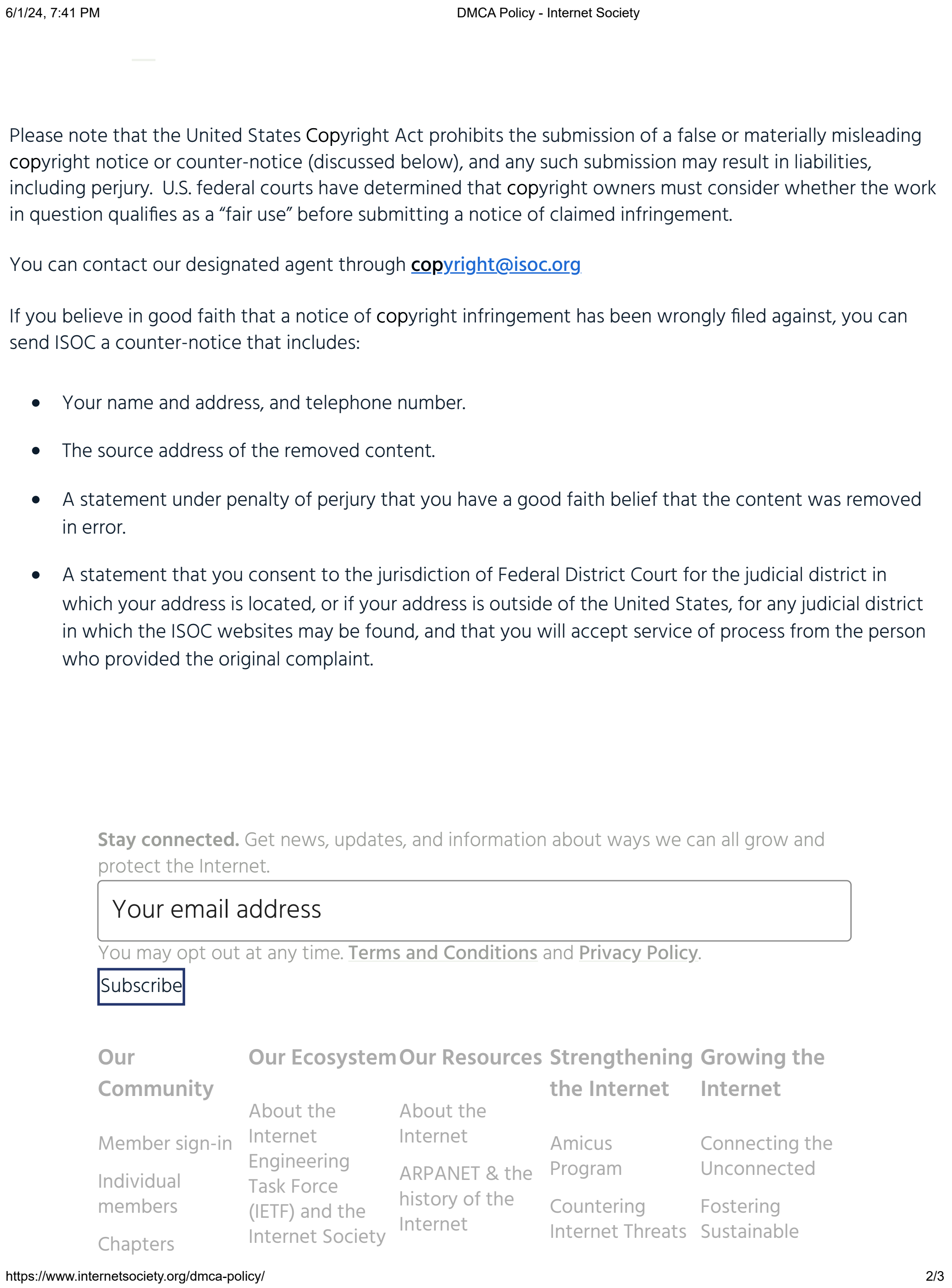
# DMCA Policy

## Digital Millennium Copyright Act – Copyright Complaints

The Internet Society and the Internet Society Foundation (collectively, "ISOC") respect the intellectual property rights of others and requires those that use our website to do the same. ISOC may, in appropriate circumstances and at our discretion, remove or disable access to material on the websites that infringes upon the copyright rights of others. ISOC also may, at our discretion, remove or disable links or references to an online location that contains infringing material or infringing activity. In the event that any users of the websites repeatedly infringe on others' copyrights, ISOC may, in our sole discretion, terminate those individuals' rights to use the websites.

If you believe that any copyrighted work is accessible through the ISOC websites in a way that constitutes copyright infringement, please notify ISOC by providing our designated copyright agent with the following information:

- The physical or electronic signature of either the copyright owner or of a person authorized to act on the owner's behalf.
- A description of the copyrighted work you claim has been infringed, and a description of the activity that you claim to be infringing.
- Identification of the URL or other specific location on the ISOC websites where the material or activity you claim to be infringing is located or is occurring. You must include enough information to allow us to locate the material or the activity.
- Your name, address, telephone number, and e-mail address.
- statement by you, made under penalty of perjury, that (i) the information you have provided is ccurate and that you are the copyright owner or are authorized to act on behalf of the owner of an



Please note that the United States Copyright Act prohibits the submission of a false or materially misleading copyright notice or counter-notice (discussed below), and any such submission may result in liabilities, including perjury.  U.S. federal courts have determined that copyright owners must consider whether the work in question qualifies as a "fair use" before submitting a notice of claimed infringement.

You can contact our designated agent through copyright@isoc.org

If you believe in good faith that a notice of copyright infringement has been wrongly filed against, you can send ISOC a counter-notice that includes:

- Your name and address, and telephone number.
- The source address of the removed content.
- A statement under penalty of perjury that you have a good faith belief that the content was removed in error.
- A statement that you consent to the jurisdiction of Federal District Court for the judicial district in which your address is located, or if your address is outside of the United States, for any judicial district in which the ISOC websites may be found, and that you will accept service of process from the person who provided the original complaint.

**Stay connected.** Get news, updates, and information about ways we can all grow and protect the Internet.

Your email address

You may opt out at any time. **Terms and Conditions** and **Privacy Policy**.

Subscribe

**Our Community**

Member sign-in

Individual members

Chapters

**Our Ecosystem**

About the Internet Engineering Task Force (IETF) and the Internet Society

**Our Resources**

About the Internet

ARPANET & the history of the Internet

**Strengthening the Internet**

Amicus Program

Countering Internet Threats

**Growing the Internet**

Connecting the Unconnected

Fostering Sustainable