EFFECTIVE AND INCONSPICUOUS OVER-THE-AIR ADVERSARIAL EXAMPLES WITH ADAPTIVE FILTERING

Patrick O'Reilly[1], *Pranjal Awasthi*[2], *Aravindan Vijayaraghavan*[1], *Bryan Pardo*[1]

[1]Northwestern University [2]Google Research

ABSTRACT

While deep neural networks achieve state-of-the-art performance on many audio classification tasks, they are known to be vulnerable to adversarial examples - artificially-generated perturbations of natural instances that cause a network to make incorrect predictions. In this work we demonstrate a novel audio-domain adversarial attack that modifies benign audio using an interpretable and differentiable parametric transformation - adaptive filtering. Unlike existing state-of-the-art attacks, our proposed method does not require a complex optimization procedure or generative model, relying only on a simple variant of gradient descent to tune filter parameters. We demonstrate the effectiveness of our method by performing over-the-air attacks against a state-of-the-art speaker verification model and show that our attack is less conspicuous than an existing state-of-the-art attack while matching its effectiveness. Our results demonstrate the potential of transformations beyond direct waveform addition for concealing high-magnitude adversarial perturbations, allowing adversaries to attack more effectively in challenging real-world settings.

Index Terms— Adversarial examples, speaker verification

1. INTRODUCTION

Voice-based systems for authentication and control in products such as mobile devices (e.g. Google Assistant), household appliances (e.g. Amazon Alexa), vehicles (e.g. Tesla Voice Commands), and remote transaction interfaces (e.g. Chase Voice ID) have been widely adopted in recent years. Deep neural networks achieve state-of-the-art performance on many audio tasks, including speech command recognition [1, 2] and speaker recognition [3, 4], and as a result have been incorporated into numerous such systems [5, 6, 7]. While deep networks are powerful classifiers, they are known to be vulnerable to adversarial examples, artificially-generated perturbations of natural instances that cause a network to make incorrect predictions [8]. This vulnerability presents an opportunity for malicious actors to gain access to and influence the behavior of products like those mentioned above by surreptitiously passing adversarially-crafted audio to the underlying neural network systems.

The research community has developed a number of audio-domain adversarial attacks to evaluate the robustness of deep neural networks in tasks such as speech command recognition [9, 10], speaker recognition and verification [9, 11, 12, 13], and automatic speech recognition [14, 15, 16, 17]. Following Carlini & Wagner [14], many attacks adopt the standard image-domain approach of leveraging gradient information from the victim model or a similarly-constructed *surrogate* model in order to optimize a small additive perturbation of a benign instance. The perturbed instance is then fed to the victim model, producing a general misclassification (in the case of an *untargeted* attack) or a specific prediction chosen by the attacker (in the case of a *targeted* attack).

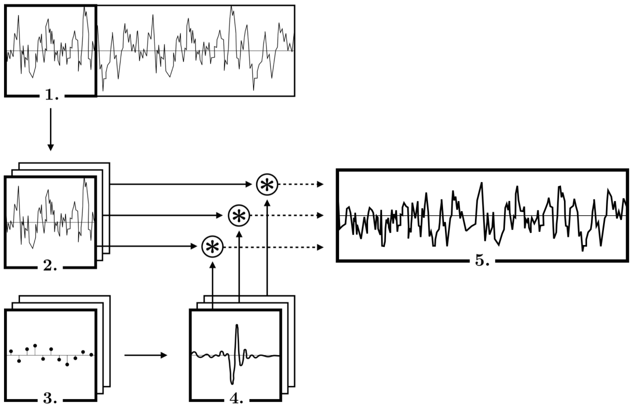

Fig. 1. We generate audio adversarial examples through adaptive filtering. **1-2.** Benign audio is divided into frames **3.** For each frame we parameterize an adversarial transfer function **4.** We convert each transfer function to an impulse response and convolve with the corresponding audio frame **5.** Individual filtered frames are combined via overlap-add to obtain the adversarial audio.

Audio adversarial attacks typically optimize an additive perturbation of a benign input directly at the waveform representation. To ensure the attack remains inconspicuous to human listeners, the perturbation is often constrained in terms of a p-norm magnitude or through the use of an auxiliary loss function during optimization. Existing attacks can successfully balance the conspicuousness of perturbations against their effectiveness in an *over-the-line* setting, where the attack audio can be fed directly to the victim model over a purely digital channel [15, 11].

In many practical scenarios an attacker lacks such direct access and instead must craft adversarial examples in an *over-the-air* setting, where malicious audio is played through a speaker and received by a microphone before entering the victim model. The adversarial audio is therefore subject to various distortions introduced by the acoustic environment, such as background noise and reverberation, as well as to those introduced by the physical properties of the speaker and microphone. Today's waveform-additive attacks require large-magnitude perturbations to achieve high success rates in an over-the-air setting [15, 16]. Consequently, the perturbations introduced by these attacks often become clearly audible as noise when deployed in practical scenarios.

Recent works have sought to conceal over-the-air attacks through the use of short *universal* perturbations, adversarial sounds that can be played as a separate source simultaneously with live speech [9, 18, 12]. In the case of Li et al. [9], these perturbations can be disguised as environmental sounds (e.g. phone notifications). However, these attacks still optimize perturbations directly at the waveform, which can result in characteristic noise-like artifacts. Existing attacks at non-waveform representations either produce conspicuous, highly particular sounds [10] or are not demonstrated over-the-air [11].

In the image domain, researchers have sought to address an analogous perceptibility/effectiveness trade-off by moving away from bounded additive perturbations in pixel space. Instead, recent works use differentiable parametric transformations to modify the content of benign images in ways that, while perceptible, may be explained away as natural artifacts of the photographic process (e.g. motion-blur [19], shadows [20], and color adjustment [21, 22]). The resulting perturbations remain inconspicuous despite large p-norm magnitudes, allowing for more potent attacks. However, the transformations employed are specific to the image domain.

Inspired by these works, we propose a novel audio-domain attack (Section 2) that introduces perturbations through adaptive filtering rather than the standard approach of direct addition at the waveform. We then demonstrate the effectiveness of our proposed attack against a state-of-the-art speaker verification model in a challenging simulated over-the-air setting and corroborate our results in a real-world over-the-air setting. Finally, we describe the results of a listener study that establishes our proposed attack as less conspicuous than a waveform-additive baseline with a state-of-the-art frequency-masking loss [15, 23, 13, 24] while matching its effectiveness.[1]

2. ADAPTIVE FILTERING ATTACK

We propose to attack using adaptive filtering, a differentiable parametric method that dynamically shapes the frequency content of audio. By crafting adversarial features directly in the frequency domain, we avoid the noise-like artifacts associated with unstructured waveform perturbations.

2.1. Overview

We divide an audio input x into T fixed-length frames. For the t^{th} audio frame $x^{(t)}$, we parameterize a finite impulse response (FIR) filter by specifying the transfer function $H_t(f)$ over F frequency bands. We define H as the full adaptive filter over all T frames, i.e. $H = [H_1(f), ..., H_T(f)]$. The full adaptive filter H is thus specified by $T \times F$ total parameters, typically a fraction of the number required for an additive perturbation of the input at the waveform. To apply the filter, we map each transfer function to a time-domain impulse response via the inverse discrete Fourier transform and apply a Hann window, as in Engel et al. [25]. Each impulse response is then convolved with the corresponding input frame via multiplication in the Fourier domain, and the resulting filtered frames are synthesized via overlap-add with a Hann window to avoid boundary artifacts.

We can optimize the parameters of H adversarially in the way we would a traditional additive perturbation. Let $f : \mathbb{R}^d \mapsto \mathbb{R}^n$ be a neural network model mapping d-dimensional audio input to an n-dimensional representation (e.g. embeddings), and let $H(x)$ denote the application of the adaptive filter to the full input audio x (all T frames), as discussed above. Given a benign input $x \in \mathbb{R}^d$ and a target $y \in \mathbb{R}^n$, we optimize $H \in \mathbb{R}^{T\times F}$ by performing gradient descent on an objective of the form:

$$\underset{H\in\mathbb{R}^{T\times F}}{\text{argmin}}\ \mathcal{L}_{adv}(f, x, H, y) + \mathcal{L}_{aux}(x, H) \tag{1}$$

Here, the adversarial loss function $\mathcal{L}_{adv}$ measures the success of the attack in achieving the outcome $f(H(x)) = y$ and the auxiliary loss function $\mathcal{L}_{aux}$ encourages the filtered audio $H(x)$ to resemble the benign audio x. To constrain the magnitude of the filter parameters, we perform *projected gradient descent* on the objective [26]. For projection bound ϵ, at each optimization iteration the parameters of H are projected onto a 2-norm ball of radius ϵ centered at 0. A trade-off between adversarial success and imperceptibility is achieved using the *selective* variant of projected gradient descent proposed by Bryniarski et al. [27]. All losses $\mathcal{L}$ are formulated such that $\mathcal{L} \leq 0$ implies the associated constraint has been satisfied (e.g. the adversary achieves the target output, the filter introduces no perceptible perturbation). Update steps are then taken along the gradient of the adversarial loss when $\mathcal{L}_{adv} > 0$ and along the auxiliary loss otherwise. This allows us to avoid manually tuning a weighting of the terms in the objective function.

Finally, to produce adversarial examples that are robust in the over-the-air setting, we use the expectation-over-transformation approach of Athalye et al. [28] and optimize over a set of simulated distortions (Section 2.3) designed to mimic those present in real-world acoustic environments.

2.2. Adversarial and Auxiliary Losses

To perform attacks on a speaker-verification model which produces n-dimensional embeddings $f(x) \in \mathbb{R}^n$ and classifies inputs according to a thresholded cosine distance in the embedding space, we use the adversarial loss proposed by Zhang et al. [12]. Let $S_{\cos}$ denote the cosine similarity between two embedding vectors:

$$S_{\cos}(u, v) = \frac{u \cdot v}{\|u\|_2\,\|v\|_2} \tag{2}$$

Then we define our adversarial loss as

$$\mathcal{L}_{adv}(f, x, H, y) = (1 - S_{\cos}(f(H(x)), y) - \tau + \kappa)^+ \tag{3}$$

where $(\cdot)^+ = \max(\cdot, 0)$, y is an embedding of the target speaker, τ is the cosine-distance verification threshold of the model, and κ is the margin by which we wish to cross the threshold. For our auxiliary loss, we use the cosine distance computed frame-wise on the mel-frequency cepstral coefficient (MFCC) representation of the original and filtered audio, as proposed by Shamsabadi et al. [11]:

$$\mathcal{L}_{aux}(x, H) = \sum_{t=1}^{T} 1 - S_{\cos}(H_t(x^{(t)})_{\text{mfcc}}, x^{(t)}_{\text{mfcc}}) \tag{4}$$

2.3. Over-the-Air Simulation

Drawing from previous over-the-air attacks [15, 16, 9], we model the over-the-air setting by applying the following distortions: **time-domain offsets** sampled uniformly from -150ms to +150ms; **Gaussian noise** generated with signal-to-noise ratio (SNR) values from +30.0 to +40.0 dB; **bandpass filtering** applied with a high-pass cutoff of 400Hz and a low-pass cutoff of 6000Hz; **environmental noise samples** drawn from the Room Impulse Response and Noise Database [29] and scaled to SNR values from -5.0 to $+10.0$ dB; and **reverberation** applied via recorded room impulse responses [29].

[1] Audio examples are available at https://interactiveaudiolab.github.io/project/audio-adversarial-examples.html

To enable the computation of gradients, we implement bandpass filtering and reverberation through a differentiable FIR convolution operation. At each optimization iteration, the above distortions are randomly initialized and applied in sequence to all benign and adversarial audio. As is common, we omit all distortions when computing the auxiliary loss [9, 15].

3. EXPERIMENTS

Using our proposed method, we perform attacks against a state-of-the-art speaker verification model. We evaluate the effectiveness of the generated attacks in simulated and real over-the-air environments and conduct a listener study to judge the conspicuousness of our attacks.

3.1. Speaker Verification Model

In the speaker recognition task, a model must predict the identity of the speaker of a given utterance. This may involve multiclass classification against a set of known speakers, or confirmation of a speaker's claimed identity through comparisons against known (or "enrolled") utterances from the speaker using a distance measure computed on utterance embeddings. Recent works have proposed attacks against both task variants [11, 9, 12, 13]. We focus on the latter, referred to as speaker verification, and use the pre-trained ResNetSE34V2 model [30] provided in the VoxCeleb-Trainer repository [4]. The model was trained on the VoxCeleb2 dataset [31], comprising 1.1 million utterances from 6112 speakers sampled at 16kHz. We perform attacks on the "test-clean" partition of the LibriSpeech dataset [32], comprising 2620 utterances from 40 speakers (20 male, 20 female). To standardize comparisons, we trim or pad all utterances to four seconds. As in Zhang et al. [12], we set the model verification threshold according to the equal error rate (EER) computed on a set of 14240 pairwise comparisons drawn from the "test-clean" partition. The model achieves an EER of 1.36% when evaluated in an over-the-line setting and 2.98% when evaluated in our simulated over-the-air setting.

3.2. Attacks

Proposed: For our proposed adaptive filtering attack, we set the frame size to 1024 samples and the number of bands to 128, resulting in a total of 8064 parameters per input across 63 frames. We use the adversarial cosine loss discussed in Section 2.2 with margin $\kappa = 0.5$ to encourage perturbations which produce high-confidence predictions, and set $\epsilon = 40$ for projected gradient descent with learning rate 0.005 on the filter parameters to discourage extreme filter configurations. We compute our auxiliary loss using the 128-coefficient MFCC representation.

Baseline: To compare against a strong waveform-additive baseline attack with a perceptually-inspired auxiliary loss, we implement the two-stage frequency-masking attack of Wang et al. [13]. Originally proposed by Qin et al. [15] for automatic speech recognition, this attack first optimizes a waveform perturbation through L_∞ projected gradient descent with an adaptive projection bound in an attempt to find a minimal-magnitude noisy perturbation. A second stage of optimization is then performed in which the L_∞ projection is removed and an auxiliary frequency-masking loss is introduced to encourage perturbations which are imperceptible to humans. The loss is constructed using frame-by-frame approximations of the frequency-masking thresholds of the human ear induced by the benign input, allowing the adversarial perturbation to "hide" in spectro-temporal regions of diminished auditory sensitivity. Given the computational expense involved, these thresholds must be computed and cached before the attack begins. An adaptive weighting schedule is used to balance the adversarial and auxiliary losses, increasing the weight α of the perceptual loss as the attack achieves its objective and decreasing the weight when the attack fails. It has been shown that this attack can be successfully applied to speaker recognition by replacing the adversarial loss term with an appropriate alternative [13], and that the attack can achieve success in real and simulated over-the-air settings through expectation-over-transformation hardening [24, 23]. We therefore set the attack's adversarial loss to the cosine loss mentioned above while leaving the attack's adaptive weighting schedule unchanged, following Wang et al. [13]. We set the initial value of the second-stage tradeoff parameter to $\alpha = 0.005$ and second-stage learning rate to 0.0001, as we find this combination minimizes perceptible artifacts without compromising effectiveness in our simulated over-the-air setting (see Section 3.3).

3.3. Effectiveness Evaluation

We perform 400 targeted attacks using both the proposed and baseline method, drawing 10 examples for each of the 40 speakers and randomly assigning a target class (speaker identity) for each. We apply a different simulated over-the-air distortion set at each optimization iteration as described in Section 2.3; the baseline and proposed attacks see an identical distortion set at each iteration of optimization. In all cases, we use the Adam optimizer [33] and optimize for 8000 iterations to ensure exposure to a sufficiently large number of simulated distortions. In the two-stage baseline attack, this is divided between 2000 iterations for the first stage and 6000 for the second.

For each of the 400 final generated attacks, we evaluate its effectiveness in achieving the target class over a further 2000 randomly sampled simulated distortion sets. We randomly select 40 generated pairs of baseline and proposed attacks, balanced across classes, with which we perform a limited real-world over-the-air evaluation and listener study (see Section 3.4).

We conduct our real-world over-the-air evaluation in a room with dimensions $3.3\text{m} \times 3.6\text{m} \times 2.3\text{m}$. Ambient noise is measured at 41 dB_{SPL}. We use a small commercial Bluetooth speaker (Bose SoundLink) to play the attack and a laptop with a generic RealTek dual microphone array to record the attack audio passed to the victim. We consider two configurations. In the **near** configuration, the speaker is placed 1m from the microphone, and attack audio is measured at an average of 60 dB_{SPL} at the victim device. In the **far** configuration, the speaker is placed 3.5m from the microphone, across the room, and attack audio is measured at an average of 55 dB_{SPL} at the victim device. In all cases we normalize attack audio to ensure a fair comparison across examples. Average attack success rates for the simulated and real over-the-air settings are reported in Table 1 ("Over-the-Air Effectiveness").

3.4. Perceptual Study

If two adversarial attacks achieve roughly equal effectiveness in fooling a system, the attacker should choose the one less likely to be detected by a third party. In this work, we assume the third party to be a human end-user hearing the attack or a non-expert company employee listening to input audio for their system as a quality control. We claim that our proposed adaptive filtering method (see Section 2) is capable of generating targeted attacks which match or exceed the effectiveness of existing state-of-the-art attacks based on additive noise while remaining less perceptible. To show this, we compare

Attack	**Over-the-Air Effectiveness**			**Perceptual Study**	**Perturbation Magnitude**	
	Simulated	Near	Far	Forced Choice	L_∞	L_2
Baseline	94.1%	92.5%	90.0%	34.1%	0.08	1.97
Proposed	96.9%	100.0%	90.0%	65.9%	0.23	6.59

Table 1. We compute the average **over-the-air effectiveness** rate of the baseline and proposed attacks in a simulated setting and in two real-world settings, as described in Section 3.3. For comparison, in our experiments the corresponding unperturbed audio is misclassified as the target at a rate of 0.006% in the simulated setting and not at all in the real-world settings. The conspicuousness of the selected attacks is evaluated through a human **perceptual study**, and we report the proportion of trials in which each attack is rated as less conspicuous in a two-way forced choice. Finally, we report the average L_∞ and L_2 **perturbation magnitude** of each attack measured at the waveform.

attacks generated using our proposed approach against attacks generated using a complex perceptually-regularized baseline attack (see Section 3.2).

We perform a web-based pairwise ABX (forced choice) study, which previous work [34] has shown to be an effective way of comparing audio quality. For each of the 40 generated attacks selected, we create a comparison triplet consisting of the baseline attack audio (A), the proposed filtering attack audio (B), and the unperturbed reference audio (X). We then perform a study in which we ask listeners to compare A and B to X and select the recording that sounds most like X. The proportion of listeners that prefer A to B provides an indication of the relative perceptibility of the baseline and proposed attacks.

3.4.1. Participants

We recruited participants through Amazon Mechanical Turk, a micro-task labor market where workers take on small tasks (e.g. labeling images, proofreading sentences). Participants were limited to workers from North America with 97% task success rates and a minimum of 1000 successfully-completed tasks. All participants were 18 or older. Participants were not screened for gender or age over 18, although a preliminary questionnaire did ask those with hearing loss to self identify. Participants were given information on the study's nature and purpose and gave consent at the time of accepting the task. The entire task, including the pre-survey, took an average of 14 minutes to complete, and participants were rewarded $2.30 for participation. Only participant answers to questions, duration of the study and Amazon worker ID were collected. At the completion of data collection and payment, worker ID was replaced with a fully anonymous ID.

3.4.2. Study Task

The set of 40 triplet comparisons was divided between two 24-question survey forms. Each form consisted of 20 comparison questions and an additional 4 control questions where the closer example was made intentionally obvious, with the aim of estimating confidence in a participant's answers. The answers to these control questions were not used in our reported results, and our analysis only includes responses passing all controls. For each of the 24 ABX questions, the participant was given a two-way forced-choice with the following instructions:

> Please listen to the recordings Reference, A, and B in
> their entirety. Choose which of A and B sounds most
> like the Reference.

The ordering of A and B was counterbalanced across questions. Participants were free to play any recording repeatedly and could alter their answer on any question until the final submission of the form. We recruited 30 workers per form, resulting in 30 judgements per triplet and 1200 judgements overall.

4. RESULTS

We present the results of our effectiveness evaluation and perceptual study in Table 1. **Our proposed adaptive filtering attack matches the effectiveness of the baseline attack in all settings. Listeners rate, by a two-to-one margin, the proposed attack as less conspicuous than the baseline**. This is noteworthy given the relative simplicity of our novel proposed attack. Recall that our proposed method utilizes a basic adaptive filtering implementation, requires only a single stage of optimization, and optimizes against a simple spectral loss. By comparison, the baseline attack requires a two-stage optimization procedure and a computationally expensive frequency-masking loss [13, 15]. Moreover, our attack is rated as less conspicuous despite producing perturbations to the audio which are much larger on average than those introduced by the baseline attack, as measured in the L_∞ or L_2 norm. These results suggest that the choice of attack representation (e.g. waveform vs. filter) can significantly affect attack success and conspicuousness, and more specifically that high-magnitude adversarial perturbations can be concealed within simple frequency-domain manipulations.

5. CONCLUSION

We propose a novel adaptive filtering attack for generating audio adversarial examples. Our attack achieves high success rates in simulated and real over the air settings against a state-of-the-art speaker verification model, and does so without introducing conspicuous noise-like artifacts. This work opens the door to explore new classes of attack transformations beyond direct waveform addition. We expect this exploration will lead to a deeper understanding of ways to secure audio classifiers against perturbations, both adversarial and natural. For future work, we look to build upon our results by evaluating the effectiveness of filter-based attacks against more robust speaker verification pipelines that mimic the voice authentication systems used in remote transaction interfaces [12, 35], as well as other speech systems (e.g. automatic speech recognition).

6. REFERENCES

[1] Jimmy Lin and Raphael Tang, "Deep residual learning for small-footprint keyword spotting," in *ICASSP*, 2018.

[2] Tara Sainath and Carolina Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Interspeech*, 2015.

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

[4] Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han, "In defence of metric learning for speaker recognition," in *Interspeech*, 2020.

[5] E. Marchi, S. Shum, K. Hwang, S. Kajarekar, S. Sigtia, H. Richards, R. Haynes, Y. Kim, and J. Bridle, "Generalised discriminative transform via curriculum learning for speaker recognition," in *ICASSP*, 2018.

[6] Qingming Tang, Ming Sun, Chieh-Chi Kao, Viktor Rozgic, and Chao Wang, "Hierarchical residual-pyramidal model for large context based media presence detection," in *ICASSP*, 2019.

[7] Minhua Wu, Sankaran Panchapagesan, Ming Sun, Jiacheng Gu, Ryan Thomas, Shiv Naga Prasad Vitaladevuni, Bjorn Hoffmeister, and Arindam Mandal, "Monophone-based background modeling for two-stage wake word detection," in *ICASSP*, 2018.

[8] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.

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

[10] "Adversarial music: Real world audio adversary against wake-word detection system," in *NeurIPS*, 2019.

[11] Ali Shahin Shamsabadi, Francisco Sepúlveda Teixeira, Alberto Abad, Bhiksha Raj, Andrea Cavallaro, and Isabel Trancoso, "Foolhd: Fooling speaker identification by highly imperceptible adversarial disturbances," in *ICASSP*, 2021.

[12] Weiyi Zhang, Shuning Zhao, Le Liu3, Jianmin Li, Xingliang Cheng, Thomas Fang Zheng, and Xiaolin Hu, "Attack on practical speaker verification system using universal adversarial perturbations," in *ICASSP*, 2021.

[13] Qing Wang, Pengcheng Guo, and Lei Xie, "Inaudible adversarial perturbations for targeted attack in speaker recognition," in *Interspeech*, 2020.

[14] Nicholas Carlini and David Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *IEEE Security and Privacy Workshops*, 2018.

[15] Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel, "Imperceptible, robust, and targeted adversarial examples for automatic speech recognition," in *ICML*, 2019.

[16] Tao Chen, Longfei Shangguan, Zhenjiang Li, and Kyle Jamieson, "Metamorph: Injecting inaudible commands into over-the-air voice controlled systems," in *Proceedings of NDSS*, 2020.

[17] Paarth Neekhara, Shehzeen Hussain, Prakhar Pandey, Shlomo Dubnov, and Julian McAuley, "Universal adversarial perturbations for speech recognition systems," in *Interspeech*, 2019.

[18] Yi Xie, Cong Shi, Zhuohang Li, Jian Liu, Yingying Chen, and Bo Yuan, "Real-time, universal, robust adversarial attacks against speaker recognition systems," in *ICASSP*, 2020.

[19] Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Jian Wang, Bing Yu, Wei Feng, and Yang Liu, "Watch out! motion is blurring the vision of your deep neural networks," in *NeurIPS*, 2020.

[20] Amin Ghiasi, Ali Shafahi, and Tom Goldstein, "Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates," in *ICLR*, 2020.

[21] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and D. A. Forsyth, "Unrestricted adversarial examples via semantic manipulation," in *ICLR*, 2020.

[22] Hossein Hosseini and Radha Poovendran, "Semantic adversarial examples," in *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018.

[23] Joseph Szurley and J. Zico Kolter, "Perceptual based adversarial audio attacks," *arXiv preprint arXiv:1906.06355*, 2019.

[24] Tom Dörr, Karla Markert, Nicolas M. Müller, and Konstantin Böttinger, "Towards resistant audio adversarial examples," in *SPAI '20: Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, 2020.

[25] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts, "Ddsp: Differentiable digital signal processing," in *ICLR*, 2020.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2017.

[27] Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, and Nicholas Carlini, "Evading adversarial example detection defenses with orthogonal projected gradient descent," *arXiv preprint arXiv:2106.15023*, 2021.

[28] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, "Synthesizing robust adversarial examples," in *ICML*, 2018.

[29] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 5220–5224.

[30] Hee Soo Heo, Bong-Jin Lee, Jaesung Huh, and Joon Son Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

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

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[34] Mark Cartwright, Bryan Pardo, and Gautham J Mysore, "Crowdsourced pairwise-comparison for source separation evaluation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 606–610.

[35] Andre Kassis and Urs Hengartner, "Practical attacks on voice spoofing countermeasures," *arXiv preprint arXiv:2107.14642*, 2021.