Advoice: Proactively Defend against AI-Synthesized Fake Voices via Adversarial Attacks

Pinji Chen
Wuhan University
Wuhan, Hubei, China
pinjichen@whu.edu.cn

Boheng Li
Wuhan University
Wuhan, Hubei, China
randy.bh.li@foxmail.com

Dongyu Yao
Wuhan University
Wuhan, Hubei, China
rain.dongyu.yao@foxmail.com

Mingming Zhang
Wuhan University
Wuhan, Hubei, China
mingmingzhang@whu.edu.cn

ABSTRACT

Voice synthesize systems using deep learning techniques have become overwhelmingly powerful and accessible in recent years. Given the speech of the target voice, such a system *e.g.* voice conversion (VC), text-to-speech (TTS), can generate synthetic fake speeches that speak in the voice of the target speaker, which can be natural and realistic, and thus are exhibiting real threats to society. Unfortunately, existing works that propose to defend against forgeries mainly focus on developing detectors for *passively* detecting whether a given speech is real or fake, which not only struggle to catch the rapid progress of forgery techniques in this cat-and-mouse game but also failed to block the spread of misinformation in advance.

In this paper, we propose *Advoice*, a novel approach that *proactively* defenses against deep learning-based synthesis techniques. Through this approach, users can protect their shared speeches from popular deep learning-based voice synthesis systems. Specifically, *Advoice* leverages the power of adversarial attacks, embeds audibly imperceptible adversarial perturbations to speech records, *etc.*, and disrupt the generation of fake voices to defend against forgeries. We conducted experiments on the *state-of-the-art* VC and TTS systems, results show that, while our protected voices are distinguishable from the original ones, training a voice synthesis model with protected voices can lead to a significant degradation in the quality of synthesized voices. This degradation is twofold. On the one hand, the quality of the synthesized voices is reduced with more audible artifacts such that the synthesized voices are more obviously fake or less convincing to human ears. On the other hand, the synthesized voices can easily be detected based on various metrics. Code is available at (*url omitted for double-blind review*).

PSTW '22, June 4–15, 2022, Wuhan, China

ACM ISBN 978-1-4503-1234-1/18/06…$00.00
https://doi.org/114514.1919810

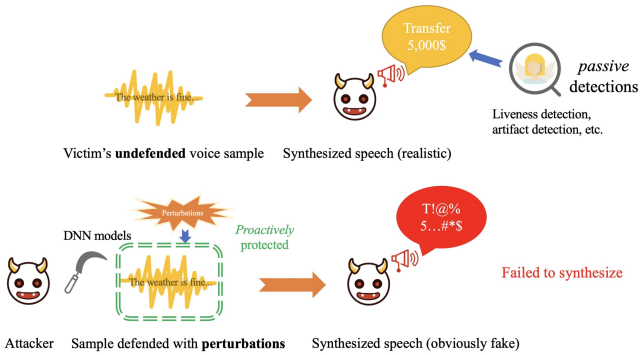

Figure 1: Comparison between our *proactive* defense approach and previous detection defense methods. Previous manners focus on detecting whether a voice sample is synthesized or not via techniques like liveness detection, *etc.* Our approach introduces perturbations to the original voice sample, making it hard for forgers to use the protected sound to synthesize speeches.

CCS CONCEPTS

• **Security and privacy** → **Human and societal aspects of security and privacy**; • **Information systems** → **Multimedia information systems**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

fake voice, voice conversion, text-to-speech, adversarial attack

ACM Reference Format:
Pinji Chen, Boheng Li, Dongyu Yao, and Mingming Zhang. 2022. *Advoice*: Proactively Defend against AI-Synthesized Fake Voices via Adversarial Attacks. In *Proceedings of the 1st Practical Science and Technology Writing (PSTW '22), June 4–15, 2022, Wuhan, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/114514.1919810

1 INTRODUCTION

Voice can deliver much more than we imagine and it is a fundamental part of our identity. The acquaintances can easily match the voice of the speaker, while the stranger can readily infer the gender,

the approximate age, the size and strength of the speaker. Nowadays, however, the human voice is no longer as unique as we would like to believe. Recent works in deep learning have led to a wide range of tools that produce synthetic speech spoken in the voice of a target speaker, such as text-to-speech (TTS) tools that transform the arbitrary text into spoken words [3, 7, 8, 10, 16, 22, 27], and voice conversion (VC) tools that reshapes existing voice samples into the same content spoken by the target [11, 17, 18, 21, 28].

Given the strong ties between our voices and our identities, a tool that successfully mimics our voices can do severe damage to the computer system and human activities. The AI-synthesized fake voice can bypass the automatic speaker verification system which has already been deployed in the application log-in services like Wechat [4]. In addition, such tools can directly be used to deceive the end-user, for example, an adversary can utilize the synthetic voice of a corporate CEO to order a subordinate to issue an illegitimate money transfer[23]. Therefore, recognition and defense of synthetic voice are needed urgently.

Prior works provide some mitigation methods to prevent AI-synthesized fake voice attacks, we classify them into three mainstreams: 1) **Liveness detection**, which includes the measuring of human vocal tract movement[31] and breath detection via microphone [24]. However, the vocal tract movement requires precise static calibration during enrollment/testing, which poses restrictions on application scenarios and may mistakenly classify a legal user as an illegal one. And the breath detection requests users to speak extremely close to the microphone (<4 inches), which decreases the usability; 2) **Loudspeaker detection**, which detects the presence of magnetic fields produced by loudspeakers[5] or compares current audio environment with the ones of previously enrolled speakers[30], since the synthetic speech should be played by loudspeaker. Yet, it requires careful motion of smartphone during recording or precise static calibration during enrollment/testing; 3) **Artifact detection**, [1, 2, 26] which uses machine learning methods to train the recognition models of synthetic speech. Nevertheless, attackers can obviously create adversarial examples as machine learning has a natural weakness in adversarial examples.

Despite the variety of existing defense mechanisms, all the proposed defenses, in the everlasting cat-and-mouse game between attackers and defenders, have subsequently been broken. More importantly, every previous defense can merely detect the fake voice after the attack occurs, instead of preventing the spread of misinformation, let alone the generation of fake voices. Therefore, we instinctively think that *can we defend the AI-synthesized fake voice proactively*? We are inspired by the proactive defense method in the image domain. To protect an individual's photo from being changed the hair color by the facial manipulation system, previous works on the image domain add imperceptible perturbations to deepfake the input image into adversarial examples, and then the facial manipulation system fails to alter the person's hair color of the this adversarial examples[19]. To sum it up, adversarial examples can disrupt the output of the DNN-model. In an attempt to prevent unauthorized speech synthesis, we might as well introduce human imperceptible noise into the utterances of a speaker whose voice is to be defended.

In this work, we propose *Advoice*, a novel defense method in the audio domain that can proactively defend against AI-synthesized fake voices by adding imperceptible perturbations into original audio samples. We attack the common unit encoder in VC and TTS models.

Advoice is not a passive detection method, instead, it proactively defends against synthesis models via adversarial perturbations and therefore is not inherently subject to the limitations mentioned before. This intrinsic difference in the working principles helps *Advoice* effectively protect the speaker's voice.

In this paper, our key contributions are:

- For the first time, we propose *Advoice*, a novel method to defend against AI-based voice synthesis in a *proactive* manner.
- *Advoice* proactively defends against deep learning-based voice synthesis models via embedding audibly imperceptible perturbations into the voice samples to be defended. Such perturbations can disrupt the synthesis generation process on the *state-of-the-art* VC and TTS models.
- Experiments show that voice samples protected by *Advoice* are indistinguishable from the original ones while being effective in fooling *state-of-the-art* VC and TTS models to generate degraded synthesized fake voices which are obviously fake or less convincing to human ears, and such voices can easily be detected based on various metrics.

2 RELATED WORK

2.1 Voice Synthesis

Advances in deep learning have introduced a new wave of voice synthesis tools, capable of producing audio that sounds as if spoken by a target speaker. Voice Conversion (VC) and Text-to-speech (TTS) are two typical technologies for generating synthesized speeches.

Voice Conversion (VC) is a technology that modifies the speech of a source speaker and makes their speech sound like that of another target speaker without changing the linguistic information. Voice conversion models based on the variational autoencoder (VAE) [7, 12, 32] and cycle consistent generative adversarial network (GAN) [5, 23, 24, 31] can make use of non-parallel source-target speech corpora, which means it can take advantage of the victim's voice sample in the wild and generate synthesized speeches, and thus poses a threat to the real world. In our work, we defend against the *state-of-the-art* voice conversion model AutoVC[17], it leverages an autoencoder with a carefully designed bottleneck and can achieve distribution-matching style transfer by training only on a self-reconstruction loss. It is the first model to perform zero-shot voice conversion and achieves state-of-the-art results in many-to-many voice conversion with non-parallel data in 2019.

Text-to-speech (TTS) devotes to synthesizing natural-sounding human-like speech from text. The traditional approaches to TTS can be divided into two categories: a waveform concatenation approach and a statistical parametric approach. Recently, along with developments in deep learning, neural-network-based end-to-end approaches have been proposed to achieve better performance, such as DeepVoice/Clarinet [7], and Char2Wav[8]. In particular, WaveNet [22] developed by DeepMind in 2016, and Tacotron [27] created by Google in 2017 are two milestones in speech synthesis. Our proposed model mainly focuses on SV2TTS[10], it consists of three independently trained networks, a speaker encoder network, a sequence-to-sequence synthesis network based on Tacontron 2

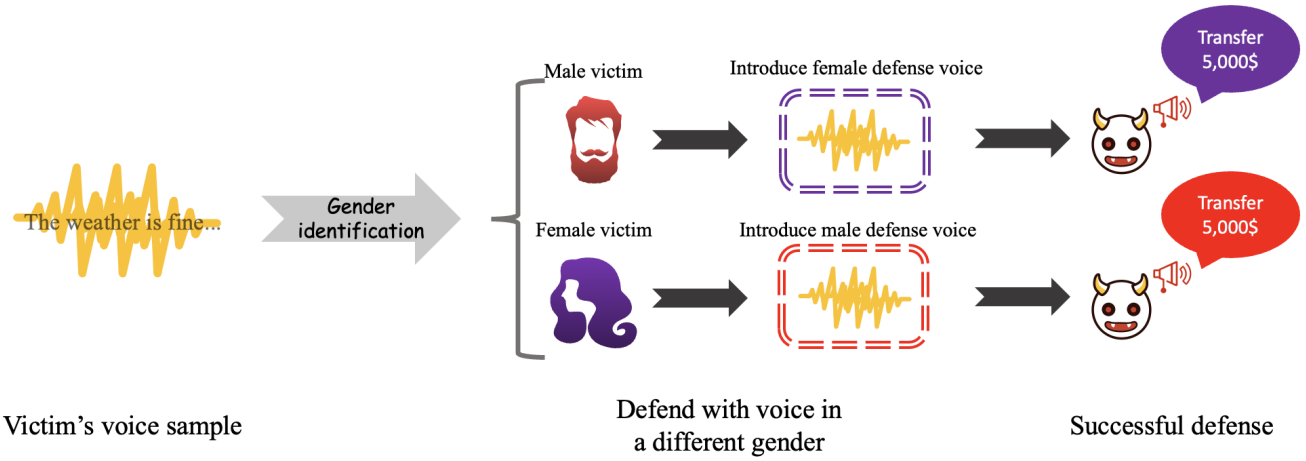

Figure 2: Overview of our *Advoice* method. First, we detect the gender of the voice we want to protect, then we choose a voice of the opposite sex to be the defense voice(we will introduce this concept later) and introduce an imperceptible perturbation into the voice sample. Afterward, if an attacker tries to take advantage of this protected voice sample to generate synthetic fake speeches via VC or TTS techniques, they will get an obviously fake sound (because of opposite gender or degraded quality).

and an auto-regressive WaveNet-based vocoder network. Combining the *state-of-the-art* TTS techniques, this model can synthesize natural speech from unseen speakers with high quality.

2.2 Fake Voice Detection

Given high threats to voice synthesis techniques, many *passive* detection attempts for distinguishing real or fake were made in recent years. Prior works such as liveness detection[1, 31] and loudspeaker detection[5] focus on the subtle differences between voices made by humans and loudspeakers and thus can defeat loudspeaker-made fake voices. However, these techniques failed in cyberspace, especially when fake voices only spread in the digital world *e.g.*on social media. In recent works like artifact detection, researchers keenly noticed the unnatural artifact in forgeries, and thus they train DNN models to detect artifacts. For instance, Wang *et al.*[26] leverages the power of layer-wise neuron activation patterns with a conjecture to distinguish forgeries, and achieves high accuracy in the digital world. However, these artifacts can be removed by reconstruction methods or get lost in even simple compressions, let alone carefully-designed evasion methods like adversarial noise attacks. More importantly, these methods all intend to detect fake voices rather than impeding the generation process or spreading process and thus are not able to prevent the spread of misinformation. In conclusion, current *passive* detection methods are not ready for meeting real-world threats.

Different from previous methods, in our work we leverage the power of adversarial attacks and embed imperceptible adversarial perturbations that can disrupt the generation process. This *proactive* manner focuses on defeating the source of misinformation, and thus is feasible for protecting voices in real-world cyberspace.

2.3 Adversarial Attack

The very first adversarial attack was discovered by Goodfellow *et al.*[6]. In their work, they found that imperceptible adversarial perturbations can easily fool the advanced DNN models to misclassify the adversarial examples with wrong labels. Further, algorithms like FGSM[6] and PGD[14] are proposed to generate imperceptible adversarial perturbations. Adversarial attacks can not only fool the DNN classifiers but also can interfere with the learning process of generative models like GANs and VAEs.

In the latest research, the potential of adversarial examples in protecting privacy and preventing the abuse of artificial intelligence has been gradually tapped. Huang *et al.*[9] leverages adversarial attacks to design unlearnable examples, Ruiz *et al.*[20] applies such perturbations to protect facial images from GAN-based Deepfake systems.

To the best of our knowledge, our proposed *Advoice* is the first method that uses adversarial attacks to defend against voice synthesis models.

3 METHODOLOGY

To make our speech sample protected by *Advoice* efficient in defeating both AutoVC voice conversion attack and SV2TTS text-to-speech attack, we focus on their common ground and make use of its weakness. We designed an attack on their adopted encoder-decoder structure. Given a speech sample x, an encoder network extracts the speaker's voice features from x to a latent vector $E(x)$, which presents the speaker's voice characteristics. Then the decoder network does a conversion from the latent space vector to a synthesized speech. Specifically, our key insight is to attack the encoder part of the model to make the extracted features change into another person's voice features $E(y)$. With this attack, the decoder

network would generate an fake or degradted speech sample y from $E(y)$. Here, we call this voice sample y defense voice. The main challenge of our proposed *Advoice* is to construct a well designed adversarial example $x + \delta$, which can be similar to the original x, and devote it to making a situation such that the encoder encodes $x+\delta$ to a latent space vector $E(x+\delta)$ similar to $E(y)$ and as far away as it can be from $E(x)$, which is the victim's voice characteristics. We formulate this as an optimization problem:

$$\min_{\delta} L(E(x+\delta), E(y)) - \lambda L(E(x+\delta), E(x))$$

Here, $L(\cdot,\cdot)$ is the distance metric between two vectors, we use L_2 norm in our approach. λ is the strength-control parameter. Specifically, to make the protected voice sample similar to the original ones, we need to limit the perturbation size to a constraint ϵ. Thus, the final equation is:

$$\min_{\delta} L(E(x+\delta), E(y)) - \lambda L(E(x+\delta), E(x)), where \parallel \delta \parallel_2 < \epsilon$$

Finally, $x + \delta$ is the adversarial example, in other words, the defended voice we want to create.

Yet we have successfully got the adversarial examples to make the features of synthetic voice changed, however, we can see that the synthetic voice will sound like the defense voice. So, if the defense voice is nearly the same as the target's voice, our defense method will become less effective. The solution is easy, we should always choose the most dissimilar voice samples to protect the potential victims. It's a good idea to select a person's voice of different gender as the defense voice because the male's voice is extremely distinct from the female one's. As a result, we add a judgment of gender before our defense is launched. The defense is shown in Figure2

4 EXPERIMENTAL EVALUATION

In this section, we first outline the experiment setup. Then we evaluate adversarial attack effectiveness and imperceptibility performance of our protected voice sample via both human perception study and machine perception study.

4.1 Experimental Setup

4.1.1 Threat model. We test our protection method on both VC and TTS models. In particular, we chose the *state-of-the-art* AutoVC[17] and SV2TTS[10] models to evaluate our performance. Here, we adopted a gray-box setting: we used the official pre-trained models, which were trained separately on their data sets.

4.1.2 Perception study setup. For the human perception study, we invited 150 volunteers to listen to the sound samples, and judge 1) whether the protected voice sample is similar to the original one, 2) whether the AI-synthesized speech taking advantage of our unprotected voice sample is trustworthy, and 3) whether the AI-synthesized speech taking advantage of our protected voice sample is trustworthy.

For the machine perception study, we tested voice sample's similarities on IFLYTEK Voiceprint Recognition System[15], and evaluate our attack effectiveness and imperceptibility performance via this similarity coefficient. It should be emphasized that this system

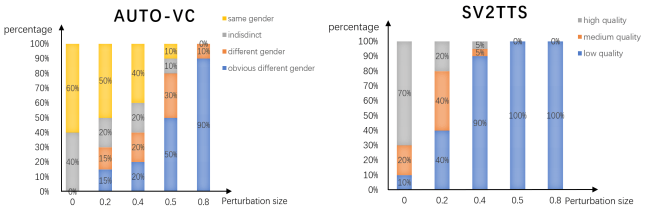

Figure 3: effectiveness of our defense on AUTO-VC and SV2TTS model

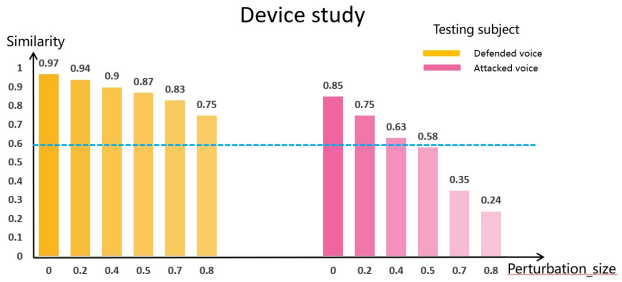

Figure 4: Tests were done on IFLYTEK Voiceprint Recognition System. If similarity >= 0.6, the platform recognizes two voices as spoken by one person.

believes that the voice samples with a similarity of more than 60% are from the same person, and vice versa.

4.2 Effectiveness Study

We test our protected voice sample's effectiveness in disrupting the VC and TTS models. Specifically, we randomly choose 5 clean female voice samples from CSTR VCTK Corpus[29] test set, which are the original voice samples that need protection. We set the perturbation size ϵ vary from 0.2 to 0.8, and we randomly choose a male voice sample from the test set as the defense voice, which is in different gender from the original. We set the distance between the speaker and the tester(microphone in machine perception study) to 1.0m, which is a circumstance close to the real world. Then, in the human study, we ask our testers to distinguish the gender of the AutoVC-converted speech samples(choosing between male and female) and the sound quality of the SV2TTS-synthesized speech samples(choosing between low quality, medium quality and high quality).

As in Figure 3, the human experimental results show that: in VC, when the perturbation reaches 0.4, the defense has started to take effect and 60% volunteers are unable to determine the gender of the voice. When the perturbation reaches 0.5, the defense has been fully successful, where all volunteers think that the voice synthesized by the attacker is a male voice instead of the female voice expected by the attacker. And in TTS, the defense has also begun to take effect when the perturbation reaches 0.5. The volunteers all think that the voice they heard is degraded to low quality, and can easily tell that this synthesized sample is untrustworthy and fake. This indicates that the synthesized voices already showed a large distortion and our defense was successful.

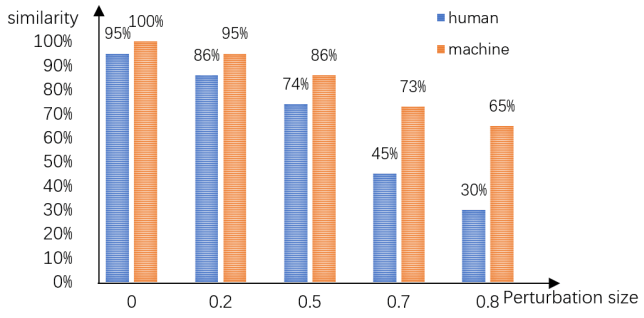

Figure 5: Test on imperceptibility of added perturbation. If the similarity is under 50%, we consider the perturbation is unacceptable

And finally, as mentioned earlier, we test our effectiveness on machines via similarity coefficient.

As shown in Figure 4, when the perturbation size is greater than or equal to 0.5, the similarity between the original voice sample and the synthetic voice sample is less than 0.6, and the similarity is 0.85 without perturbation. The introduced perturbation greatly reduces the similarity between the synthetic speech and the original speech, in other words, the attacker cannot pretend to be the same voice as the victim.

4.3 Imperceptibility Study

To test the imperceptibility of the added perturbation, we deliver a perception study both on human beings and machinery. We randomly selected 5 clean voice samples from CSTR VCTK Corpus[29] test set. We defend these samples using our proposed *Advoice*, with different perturbation sizes ϵ from 0.2 to 0.8, defense voices are chosen randomly from the test set(but not the same as the original), respectively. Then, we conduct human perception experiments. We ask our volunteers to hear 1) the original voice sample(undefended), and 2) defended voice sample(with different perturbation size) at a distance of 1.0m, which is close to the real listening distance. Tested volunteers are required to give the percentage of the similarity between the protected sound and the original sound, choosing from 0% to 100%, while 100% refers to no difference.

As shown in the orange cylinder of figure5, the imperceptibility of the perturbation drops to an unacceptable level (below 50%) only at the perturbation size is larger than 0.7. At the 1.0m listening distance for human beings, the imperceptibility always remains if the added perturbation is smaller than this threshold. In this way, the imperceptibility of the defense is guaranteed. More importantly, due to our defense can be obviously effective when the perturbation size is larger than 0.5, imperceptibility and effectiveness can be insured simultaneously when ϵ varies from 0.5 to 0.7.

Similarly, we test our effectiveness on machines via similarity coefficient. As also shown in the figure5, the blue cylinder shows that under any perturbation, the similarity between the original audio and the audio with the added perturbation is above 60%, which can be considered as the same person's voice. The imperceptibility of our defense method can be considered good, which is also consistent with the results of the human study.

5 CONCLUSION AND DISCUSSION

5.1 Limitations

As shown in the result, adding imperceptible perturbations to the audio samples does thwart VC attack and TTS attack to some extent. But it has two main problems:

- It corrupts the defended audio samples. If we want to keep the defense performance, we should increase the perturbation size which makes the source files have audible distortion. In this case, the source speaker is unwilling to use this technique because his/her speech will sound unclear. The pivotal reason for degrading the source voice is that we add perturbations on all time domain and frequency domain of the audio samples, which makes people easy to realize the changes though we use the perturbation size to constrain the alternations.
- It needs a recorded audio sample and then process it into an anti-synthetic source. Therefore, it is unable to launch a defense in some in the flesh scenarios, for example, an attacker calls you and records your voice. As a result, you can't add perturbations to your utterance. Scenario restriction occurs because the generation model can just generate the exclusive perturbations based on one previously recorded audio sample, however, can't generate a perturbation that is universal and asynchronous.

5.2 Future Work

Above all, we have two ways to improve our defense effect.

For problem 1, it's instinctive to ask can we just add perturbations to several slices of time and have almost the same performance? Prior work[13] has verified the feasibility of adding a short segment of adversarial perturbation to fool the speaker verification model. But we consider that it's not enough to make the perturbations really inaudible. So, we propose that we should also do the energy analysis of the source audio samples and adaptively embed the perturbations into it based on the auditory masking effect[25]. Additionally, we take into account the human's weak impact on high-frequency signals(e.g.17 - 20KHz). If the perturbations are high frequency, we can not only increase the perturbation size which leads to robustness but also make it inaudible to human beings.

For problem 2, to launch our defense method in much more scenarios, we should generate a universal and asynchronous perturbation. Then, when we want to protect our voice, we can just play the prepared perturbations or embed it into the audio files.

6 ACKNOWLEDGEMENTS

Boheng Li's research work is partly supported by Ziheng Huang and A.P. Run Wang.

7 CONTRIBUTION STATEMENT

Idea & Implementation: Pinji Chen (70%) & Boheng Li (30%)

Experiment Evaluation: Mingming Zhang

Manuscript Writing: Pinji Chen (Introduction & Methodology & Conclusion and Discussion), Boheng Li (Abstract & Related Work & Methodology & Experimental Evaluation), Mingming Zhang (Experimental Evaluation), and Dongyu Yao (Introduction)

Major Review & LaTeX typesetting: Boheng Li(All chapters, the post-major revision version is almost the same as the current version. The original manuscript is also available in the attachment.)

Figure Drawing: Dongyu Yao (Schematic Diagrams, Figure 1 and 4), Mingming Zhang (Figures of the experiment part), and Boheng Li (Revision)

Grammar Checks: Dongyu Yao

Bibtex References: Mingming Zhang(40%), Dongyu Yao(40%), and Boheng Li(20%)

Slides of the Speech: Dongyu Yao(60%) and Boheng Li(40%)

Speaker: Boheng Li and Pinji Chen

REFERENCES

[1] Muhammad Ejaz Ahmed, Il-Youp Kwak, Jun Ho Huh, Iljoo Kim, Taekkyung Oh, and Hyoungshick Kim. 2020. Void: A fast and light voice liveness detection system. In *29th USENIX Security Symposium (USENIX Security 20)*. 2685–2702.

[2] Ehab A AlBadawy, Siwei Lyu, and Hany Farid. 2019. Detecting AI-Synthesized Speech Using Bispectral Analysis.. In *CVPR Workshops*. 104–109.

[3] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. 2018. Neural voice cloning with a few samples. *Advances in Neural Information Processing Systems* 31 (2018).

[4] WeChat's Official Blog. 2022. Announcing WeChat VoicePrint. https://blog.wechat.com/2015/05/21/voiceprint-the-new-wechat-password.

[5] Si Chen, Kui Ren, Sixu Piao, Cong Wang, Qian Wang, Jian Weng, Lu Su, and Aziz Mohaisen. 2017. You can hear but you cannot steal: Defending against voice impersonation attacks on smartphones. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 183–195.

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

[7] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. 2019. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5901–5905.

[8] Qiong Hu, Erik Marchi, David Winarsky, Yannis Stylianou, Devang Naik, and Sachin Kajarekar. 2019. Neural text-to-speech adaptation from low quality public recordings. In *Speech Synthesis Workshop*, Vol. 10.

[9] Hanxun Huang, Xingjun Ma, Sarah Monazam Erfani, James Bailey, and Yisen Wang. 2020. Unlearnable Examples: Making Personal Data Unexploitable. In *International Conference on Learning Representations*.

[10] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems* 31 (2018).

[11] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. 2018. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 266–273.

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

[13] Zhuohang Li, Yi Wu, Jian Liu, Yingying Chen, and Bo Yuan. 2020. AdvPulse: Universal, Synchronization-free, and Targeted Audio Adversarial Attacks via Subsecond Perturbations. In *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, Jay Ligatti, Xinming Ou, Jonathan Katz, and Giovanni Vigna (Eds.). ACM, 1121–1134. https://doi.org/10.1145/3372297.3423348

[14] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*.

[15] IFLYTEK open platform. 2022. Voice recognition. https://console.xfyun.cn/services/ivp.

[16] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2017. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654* (2017).

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

[18] Yurii Rebryk and Stanislav Beliaev. 2020. Convoice: Real-time zero-shot voice style transfer with convolutional network. *arXiv preprint arXiv:2005.07815* (2020).

[19] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. 2020. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European Conference on Computer Vision*. Springer, 236–251.

[20] Nataniel Ruiz, Sarah Adel Bargal, and Stan Sclaroff. 2020. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European Conference on Computer Vision*. Springer, 236–251.

[21] Joan Serrà, Santiago Pascual, and Carlos Segura Perales. 2019. Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion. *Advances in Neural Information Processing Systems* 32 (2019).

[22] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 4779–4783.

[23] Catherine Stupp. 2019. Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *The Wall Street Journal* 30, 08 (2019).

[24] Qian Wang, Xiu Lin, Man Zhou, Yanjiao Chen, Cong Wang, Qi Li, and Xiangyang Luo. 2019. Voicepop: A pop noise based anti-spoofing system for voice authentication on smartphones. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2062–2070.

[25] Qian Wang, Kui Ren, Man Zhou, Tao Lei, Dimitrios Koutsonikolas, and Lu Su. 2016. Messages behind the sound: real-time hidden acoustic signal capture with smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking, MobiCom 2016, New York City, NY, USA, October 3-7, 2016*, Yingying Chen, Marco Gruteser, Y. Charlie Hu, and Karthik Sundaresan (Eds.). ACM, 29–41. https://doi.org/10.1145/2973750.2973765

[26] Run Wang, Felix Juefei-Xu, Yihao Huang, Qing Guo, Xiaofei Xie, Lei Ma, and Yang Liu. 2020. Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1207–1216.

[27] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135* (2017).

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

[29] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).

[30] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. 2019. The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 1215–1229.

[31] Linghan Zhang, Sheng Tan, and Jie Yang. 2017. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 57–71.

[32] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. 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)*. IEEE, 6945–6949.