Voice Privacy Protection with Adversarial Examples Using Anchor Speaker Embedding

Shunya Ishikawa*†, Yuki Katsumata*†, and Toru Nakashika*
* The University of Electro-Communications, Japan
E-mail: {s.ishikawa, y.katsumata, nakashika}@uec.ac.jp

***Abstract*—Voice cloning technology has advanced rapidly due to its capacity for flexible speech customization, with applications spanning media content production and chatbots. However, this technology also poses significant risks to voice privacy, such as fraud via cloned voices and unauthorized access to speaker authentication systems. Recent research efforts have focused on preventing such imitation using adversarial examples. When applied to speech, such examples, or adversarial speech, can cause voice cloning models to synthesize speech signals that differ from the target voice, thereby protecting audio privacy. A previous study demonstrated the effectiveness of adversarial speech, which is optimized based on a simple loss function with only the target speaker's embedding. In this study, we propose a novel adversarial speech generation method, incorporating an anchor speaker embedding into a loss function to enhance the protective performance from objective and subjective perspectives. We also propose a method that uses Adam to optimize the noise added to the original speech to generate adversarial speech. Experimental results indicate that the proposed method is superior to the conventional method in objective and subjective metrics.**

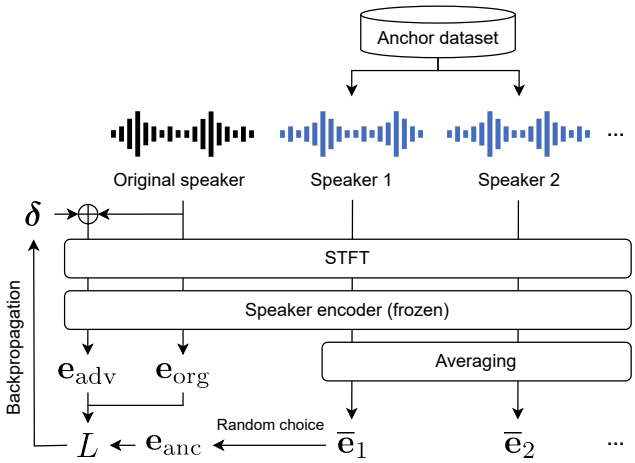

Fig. 1. The overview of our proposed method.

I. Introduction

Text-to-speech (TTS) and Automatic Speech Recognition (ASR) technologies have been rapidly improving their performance in recent years due to advances such as deep learning methods. In fact, synthetic speech using deep learning models [1], [2] greatly improves naturalness and intelligibility compared to conventional HMM-based methods [3] and is being used in a wide variety of systems, including smart speakers, voice assistants, and navigation systems.

However, this technological advancement in speech brings new security risks threatening voice privacy [4], [5]. Voice privacy, in this paper, refers to the state where a person's voice and audio matching its characteristics are not used by others without their permission. Deep-learning voice cloning technology [6]–[9] endangers voice privacy because it allows a TTS model to reproduce a person's speech with high accuracy based on a small number of speech samples. In addition, much content, including personal speech, is posted on social media platforms, and these speech data are more easily collectible than users realize. Therefore, the development of voice cloning technology increases the concern that malicious third parties will use imitated speech to commit criminal acts such as spoofing attacks, fraud, and unauthorized access to devices equipped with voice authentication functions. Against this backdrop, there is a growing need for technology to protect voice privacy.

To address these threats to voice privacy, the application of adversarial examples [10]–[12], which affect the performance of deep learning models, has received much attention. Adversarial examples are data produced by adding small amounts of noise to the original data; when input to a model, a resulting output differs from an intended output. This approach has been extensively studied, especially in the image recognition field, as a method for inducing misclassification of objects, and the efficacy of attacks using adversarial examples of image data has been widely reported [13]–[16]. In the speech field, the methods for adversarial example generation have also been investigated for speaker recognition [17]–[19] and ASR models [20]–[23], with progress made in degrading model performance [24]–[29].

However, research using adversarial examples as a countermeasure against voice cloning has not been established yet. A conventional method [30] using an iterative fast gradient sign method (I-FGSM), which generates an adversarial example that increases the distance to the speaker embedding of the original speech, was limited to demonstrating its effectiveness using objective evaluation metrics. The example objectively increases the distance but is not guaranteed to do so subjectively. We found several cases where the speaker identity in its cloned speech was not sufficiently changed auditorily.

This work was supported by JSPS KAKENHI Grant Number 24H00715 and 25H01148.

†The first author (corresponding author) and the second author contributed equally to this work.

This suggests that the embedding space generated by a speaker encoder and human perception of speaker identity may have different properties, and objective and subjective evaluation metrics may not match. Therefore, there is a need for a method that not only excels in objective evaluation metrics but also allows subjective perception of changes in speaker identity.

In this study, we propose a new loss function and an improved method for optimizing the noise added to the original data when generating adversarial examples. The proposed loss function induces the speaker identity of an adversarial example to be closer to an anchor speaker other than its original speaker, unlike a conventional loss function that merely distances the identity of the adversarial example from its original speaker. Verifying the effectiveness of the proposed improvements by objective and subjective evaluation metrics, this study presents a new approach to voice privacy protection.

II. Voice Cloning and Adversarial Examples

A. Speaker Conditioning VITS

Variational inference text-to-speech (VITS) [31] is one of the advanced end-to-end TTS model. Conventional TTS methods typically employ a pipeline comprising several steps, such as text analysis, conversion to phoneme sequences, and speech waveform generation. In contrast, VITS integrates variational inference with adversarial training to synthesize high-quality speech directly from text using a single model. Moreover, VITS supports multiple speakers by using speaker IDs. Speaker conditioning VITS (SC-VITS[1]) extends this framework by incorporating a speaker encoder that inputs target speech and outputs a continuous latent speaker representation. Consequently, SC-VITS is capable of extracting speaker features from the target speech and reflecting these features in a generated speech sample, thereby enabling arbitrary voice conversion and high-quality voice cloning.

B. Iterative Fast Gradient Sign Method

A fast gradient sign method (FGSM) [10] is a method that uses the gradient information of the loss function of a classification model to generate adversarial examples by making minor modifications (i.e., adding noise) to input data. It is the most basic approach among adversarial example generation methods and is intended to generate adversarial examples as input data that would cause a model to misclassify (i.e., attack the model). The adversarial example $\mathbf{x}'$ is given by

$$\mathbf{x}' = \mathbf{x} - \epsilon \cdot \mathrm{sign}\left(\nabla_{\mathbf{x}} L\right), \tag{1}$$

where $\mathbf{x}$ is the original input data, ϵ is a scalar value controlling the amount of noise, $\mathrm{sign}(\cdot)$ is the sign function, and L is the negative loss function of the model. The sign function makes the noise uniform in magnitude and keeps the computation simple. FGSM is computationally inexpensive because it adds noise in a single step but has a low success rate for attacks on the model.

[1]https://github.com/hcy71o/SC-VITS

Iterative FGSM (I-FGSM) [11] is an extension of FGSM in which noise is added in multiple iterations to increase the effectiveness of adversarial examples generated. The adversarial example $\mathbf{x}'_i$ obtained at the i-th iteration is given by

$$\begin{aligned} \mathbf{x}'_0 &= \mathbf{x}, \\ \mathbf{x}'_{i+1} &= \mathrm{clip}_{\epsilon}\left(\mathbf{x}'_i - \alpha \cdot \mathrm{sign}\left(\nabla_{\mathbf{x}} L\right)\right), \end{aligned} \tag{2}$$

where $\mathrm{clip}_{\epsilon}(\mathbf{z})$ is a clipping function that calculates $\max(-\epsilon, \min(z, \epsilon))$ for each element $z \in \mathbf{z}$, and α is a scalar value controlling the amount of noise added per time $(0 < \alpha < \epsilon)$. I-FGSM has a higher attack success rate than FGSM but at the cost of increased computation.

C. Non-Targeted Loss Function

A previous study applying adversarial examples to speech generation [30] focused on the loss of the original speaker identity by using a loss function consisting of a single term, as in the following equation:

$$L = \cos\left(\mathbf{e}_{\text{adv}}, \mathbf{e}_{\text{org}}\right) := \frac{\mathbf{e}_{\text{adv}}^{\mathsf{T}} \mathbf{e}_{\text{org}}}{\|\mathbf{e}_{\text{adv}}\|_2 \|\mathbf{e}_{\text{org}}\|_2}, \tag{3}$$

where $\mathbf{e}_{\text{org}}$ and $\mathbf{e}_{\text{adv}}$ are the speaker embeddings obtained by inputting the original input speech and an adversarial example for that speech (i.e., adversarial speech) into a speaker encoder, respectively. The speaker identity of the adversarial speech can be effectively distanced from the original speaker by employing this loss function in adversarial example generation methods such as I-FGSM.

D. Baseline

We assume that SC-VITS, an advanced model for high-quality speech generation, is used for voice cloning. We adopt the method as a baseline, generating adversarial speech using the loss function in (3) and I-FGSM [30]. This method aims to prevent the speaker encoder of SC-VITS from successfully extracting the speaker identity of input speech by derailing its inference.

III. Proposed Method

A. Noise Optimization Using Adam

The gradient of I-FGSM is prone to oscillation, and convergence may become unstable with an increasing number of iterations. Moreover, since α is a constant, it does not adapt to the loss range, which can result in insufficient updates in regions with sparse gradients. To address these issues, we propose a method for generating adversarial speech using Adam [32].

We define the relation between the original speech sample $\mathbf{x}$ and the adversarial speech sample $\mathbf{x}'$ by introducing noise $\boldsymbol{\delta}$, such that $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}$. In this context, our proposed method

iteratively optimizes $\boldsymbol{\delta}$ using Adam:

$$\mathbf{m}_i = \beta_1 \mathbf{m}_{i-1} + (1-\beta_1)\nabla_{\boldsymbol{\delta}} L, \tag{4}$$

$$\mathbf{v}_i = \beta_2 \mathbf{v}_{i-1} + (1-\beta_2)(\nabla_{\boldsymbol{\delta}} L)^2, \tag{5}$$

$$\widehat{\mathbf{m}}_i = \mathbf{m}_i/(1-\beta_1^i), \tag{6}$$

$$\widehat{\mathbf{v}}_i = \mathbf{v}_i/(1-\beta_2^i), \tag{7}$$

$$\boldsymbol{\delta}_{i+1} = \boldsymbol{\delta}_i - \alpha_{\text{adam}}\, \widehat{\mathbf{m}}_i/\sqrt{\widehat{\mathbf{v}}_i + \boldsymbol{\varepsilon}} \tag{8}$$

where L is a loss function, $\mathbf{m}_i$ and $\mathbf{v}_i$ denote the first moment (mean) and second moment (variance) of the gradient of the loss function at the i-th iteration, respectively, β_1 and β_2 are decay rates for the respective moments, α_{adam} is a step size, and $\boldsymbol{\varepsilon}$ is a small constant.

This dynamic calculation of $\boldsymbol{\delta}$ suppresses the gradient's oscillation, operates independently of the loss range, and handles sparse data effectively, resulting in faster and more stable convergence during adversarial speech generation compared to I-FGSM.

B. Targeted Loss Function Using Anchor Speaker Embedding

Equation (3) does not guide adversarial speech toward acquiring a new speaker identity. Consequently, the speech could be as far removed from human speech as noise, which is easily perceived. To address this issue, we propose a loss function that directs adversarial speech toward a specific speaker (anchor speaker) identity:

$$L = \cos\left(\mathbf{e}_{\text{adv}}, \mathbf{e}_{\text{org}}\right) - \lambda \cos\left(\mathbf{e}_{\text{adv}}, \mathbf{e}_{\text{anc}}\right). \tag{9}$$

Here, $\mathbf{e}_{\text{anc}}$ is the speaker embedding of an anchor speaker different from the original speaker, and λ is a parameter that controls the similarity to the anchor speaker. By distancing the speaker identity of the adversarial speech from the original speaker while bringing it closer to the anchor speaker, the proposed loss function not only disrupts the original speaker identity but also induces the adversarial speech to acquire a new speaker identity. In other words, by explicitly establishing the objective of approaching another specific speaker, our proposed method aims to change the speaker identity more intentionally and incorporates a subjective perspective.

The case in which the anchor speaker's gender differs from or matches that of the original speaker may present advantages and disadvantages. When the gender is different, the adversarial speech is more likely to have a distinctly different voice quality from the original speaker, making it more subjectively perceived as a different person. However, a greater amount of noise may be required to achieve a change in speaker identity. Conversely, when the anchor speaker shares the same gender as the original speaker, the change in voice quality is less pronounced, although a smaller amount of noise may be needed to alter the speaker identity. The experiment will investigate how the choice of the anchor speaker's gender affects both subjective and objective evaluations.

C. Flow of Proposed Method

Fig. 1 depicts the flow of our proposed method for generating adversarial speech that attacks the speaker encoder of a pre-trained SC-VITS. $\mathbf{e}_{\text{anc}}$ is randomly selected from the average embeddings of each speaker in an anchor dataset, taking their genders into account. We employ this simpler random selection because our preliminary experiment revealed little difference in performance between the random selection and the selection method in which $\mathbf{e}_{\text{anc}}$ is the least similar to $\mathbf{e}_{\text{org}}$. The method adopts (9) for L in (4) and (5). $\boldsymbol{\delta}$ is obtained by backpropagating the L (i.e., computing $\nabla_{\boldsymbol{\delta}} L$) and applying the Adam method. Thus, by repeatedly inputting the adversarial speech into the encoder and updating the noise using the loss function, the finished adversarial speech with the optimized noise is obtained.

IV. Experiments

A. SC-VITS Training

SC-VITS was trained from scratch using the train-clean-100 and train-clean-360 subsets of the LibriTTS dataset [33] with the VITS loss function. It was trained for 237 epochs (equivalent to 10,000 iterations) with a batch size of 32 and a sampling frequency of 24 kHz. AdamW [34] was the optimizer with a learning rate of 0.0002 and the betas set to 0.8 and 0.99. For spectrogram computation, 1,024 frequency bins, a hop length of 256, and a window length of 1,024 were employed.

B. Inference: Noise Optimization

From the test-clean subset of the LibriTTS dataset, 10 samples were extracted per speaker to create a dataset of 20 male and 20 female speakers. α and α_{adam} were set to 0.0001, β_1 to 0.9, and β_2 to 0.999. To compare the quality of adversarial speech under equivalent conditions, we introduced ΔUT-MOS, defined as the UT-MOS [35] of the original speech minus that of the adversarial speech. Noise optimization was repeated until this value reached 0.1, 0.2, 0.3, 0.4, or 0.5, and the results were compared for each case.

C. Evaluation Metrics and Methods for Comparison

Two metrics were employed for objective evaluation. The similarity between the speaker embedding of the original speech and the adversarial speech was assessed using the cosine similarity defined in (3). In addition, following [30], we adopted the equal error rate (EER), an evaluation metric for automatic speaker verification (ASV). EER measures whether an ASV system, with the original speech registered, accepts or rejects the clone of the adversarial speech. Since the original speaker and the clone originate from the same speaker, a higher rejection rate of the clone results in an increased false rejection rate and EER. In other words, the more effective the adversarial speech is, the more likely the system will reject its clone, leading to a higher EER. For the calculation of EER, we employed a pretrained ECAPA-TDNN [19] available in SpeechBrain [36]. The compared methods included three approaches: the baseline (I-FGSM) described in Section II-D

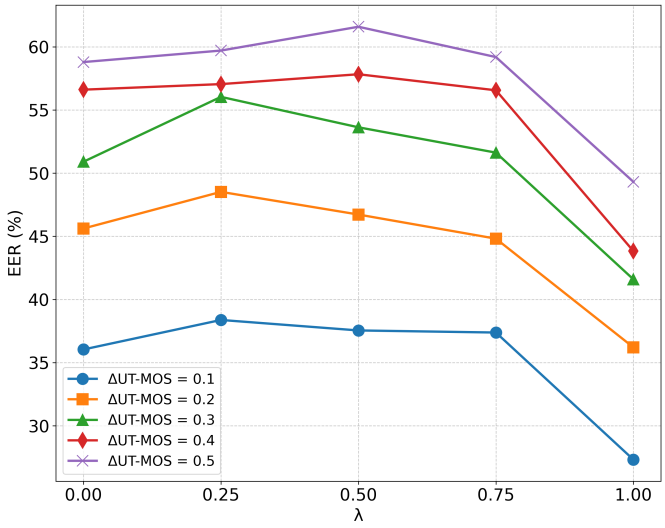

Fig. 2. Results of the preliminary experiment to determine an optimal λ.

and the proposed method under two conditions, one in which the original speaker and an anchor speaker are of different genders (Anchor-diff) and one in which they are of the same gender (Anchor-same).

Two metrics were used for subjective evaluation, and eight subjects participated. Speech quality and naturalness were assessed using the mean opinion score (MOS) at ΔUT-MOS $=$ 0.1 on the original speech (Original), the adversarial speech generated by the two proposed methods (Anchor-diff, Anchor-same), and the baseline (I-FGSM). The similarity between the speaker identity of a presented speech sample and those of the original speech was evaluated using the same-different test on their corresponding cloned speech samples. Following the original speech, one of the four cloned speech samples was presented, to which subjects chose one of the four options: very similar (VS), similar (S), different (D), or very different (VD).

D. Hyperparameter Optimization

In (9), we set λ to 0, 0.25, 0.5, 0.75, and 1 and measured a mean EER across genders for each setting to determine the optimal λ. ΔUT-MOS was 0.1, 0.2, 0.3, 0.4, and 0.5. Fig. 2 shows the result of this preliminary experiment. It indicated that the maximum EER was achieved at $\lambda = 0.25$ when ΔUT-MOS ≤ 0.3 and at $\lambda = 0.5$ when ΔUT-MOS ≥ 0.4. Furthermore, for all values of ΔUT-MOS, EER was higher at $\lambda = 0.25, 0.5$ compared to the case when $\lambda = 0$, that is, when the conventional loss function in (3) was applied to the proposed noise optimization method using Adam. This result demonstrated the effectiveness of our method. The variation in the optimal λ concerning ΔUT-MOS is considered to be due to the early convergence of the second term in (9) when ΔUT-MOS is large for a given λ, which restricts the convergence of the first term and prevents sufficient separation of speaker identity. For practical purposes, since it is necessary to generate adversarial speech with minimal impact on speech quality, $\lambda = 0.25$ was adopted.

TABLE I
RESULTS OF THE OBJECTIVE EVALUATIONS. THE GENDERS LISTED IN THE FIRST COLUMN INDICATE THE ORIGINAL SPEAKER'S GENDER.

ΔUT-MOS	Method	Female		Male	
		Cos(↓)	EER(↑)	Cos(↓)	EER(↑)
0.1	I-FGSM	0.25	32.38	0.28	32.58
	Anchor-diff	0.13	**41.15**	**0.19**	35.60
	Anchor-same	**0.12**	33.99	**0.19**	**36.54**
0.2	I-FGSM	0.02	39.57	0.03	40.66
	Anchor-diff	−0.13	**50.74**	**−0.09**	46.29
	Anchor-same	**−0.14**	45.90	−0.07	**47.29**
0.3	I-FGSM	−0.14	42.34	−0.14	47.98
	Anchor-diff	−0.30	**57.78**	**−0.29**	54.29
	Anchor-same	**−0.32**	53.39	−0.27	**56.20**
0.4	I-FGSM	−0.24	50.98	−0.26	53.10
	Anchor-diff	−0.43	**58.41**	**−0.44**	55.68
	Anchor-same	**−0.44**	57.93	−0.41	**60.46**
0.5	I-FGSM	−0.31	54.04	−0.34	53.19
	Anchor-diff	−0.52	58.38	**−0.54**	61.02
	Anchor-same	**−0.54**	**63.22**	−0.52	**62.74**

TABLE II
RESULTS OF THE SUBJECTIVE EVALUATIONS. IN THE VS+S RESULT, THE VALUES IN PARENTHESES REPRESENT THE VS AND S VALUES FROM LEFT TO RIGHT. SIMILARLY, IN THE D+VD RESULTS, THE VALUES IN PARENTHESES REPRESENT THE D AND VD VALUES FROM LEFT TO RIGHT.

Method	MOS(↑)	Same-different	
		VS+S(↓)	D+VD(↑)
Original	4.03	86.3 (45.0+41.3)	13.8 (11.3+ 2.5)
I-FGSM	3.56	**48.8** (12.5+36.3)	**51.3** (37.5+13.8)
Anchor-diff	**3.66**	**48.8** (15.0+33.8)	**51.3** (27.5+23.8)
Anchor-same	3.49	51.3 (18.8+32.5)	48.8 (31.3+17.5)

E. Results

Table I presents the results of the objective evaluations. The proposed method had lower cosine similarities, which strongly suggests that it more effectively distances the adversarial speech from the original speaker in the speaker embedding space. The proposed method had higher EERs, indicating that the clone of the adversarial speech generated by the proposed method exhibits greater changes in speaker identity compared with one produced by I-FGSM. For cosine similarity, Anchor-same tended to perform better when the original speaker was female, whereas Anchor-diff tended to perform better when the original speaker was male. In contrast, the opposite trend was observed for EER. These results suggest that selecting a female anchor speaker results in better cosine similarity while choosing a male anchor speaker results in better EER.

Table II presents the results of the subjective evaluations. For adversarial speech, the MOS for Anchor-diff was the highest, suggesting that the adversarial speech of the proposed method is superior in practical use. In other words, it suggests that guiding the adversarial speech to acquire a new speaker identity, as in the proposed method, prevents the noise added to the speech from getting out of control and thus prevents the

speech from being easily perceived as adversarial. The higher MOS of Anchor-diff compared with Anchor-same indicates less noise in Anchor-diff, which is contrary to our hypothesis.

Regarding the same-different test, achieving a higher VD is more important than a lower VS to enhance the security of cloned speech. Prioritizing a lower VS implies generating cloned speech that is merely distant from the original speaker, as in the case of I-FGSM. However, individuals abusing voice cloning should attempt to restore and utilize the degraded cloned speech. If speech restoration techniques such as distortion reduction are applied to the cloned speech, it could once again recover the original speaker, compromising its security. In contrast, prioritizing a higher VD, as in the proposed method, implies generating cloned speech closer to another speaker. Even if speech restoration techniques are applied to the cloned speech, it remains nearer to a different speaker than the original one and is thus more secure. Therefore, achieving a higher VD is more important than a lower VS from the perspective of cloned speech security.

The proposed method efficiently increased VD for the increase in VS. Compared with I-FGSM, Anchor-diff increased VD by 72% while limiting the increase in VS by 20%. This result suggests that the proposed method is suited for enhancing the security of the cloned speech. The proposed method performed comparably to I-FGSM in the results for VS+S and D+VD in the same-different test. However, it demonstrated superiority in cloned speech security, and the proposed method outperformed I-FGSM in MOS and objective evaluation metrics. In summary, the experimental results supported the superiority of the proposed method over I-FGSM.

V. Conclusion

In this study, we proposed a method for generating adversarial speech that employs a novel loss function utilizing an anchor speaker's embedding and noise optimized using Adam. We assumed that SC-VITS, an advanced open-source TTS model, would be used for voice cloning, and we evaluated the quality of adversarial speech generated by the conventional I-FGSM method and that produced by the proposed method using objective and subjective metrics. The experimental results demonstrated that the proposed method outperformed the conventional method on many objective and subjective evaluation metrics. Future work may include the evaluation and performance enhancement of voice privacy protection for speaker encoders not used in the adversarial speech generation, as well as improvements in the computational efficiency of the noise optimization method.

References

[1] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild," in *Proc. ACL*, 2024, pp. 12 442–12 462.

[2] Z. Ju *et al.*, "NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models," in *Proc. ICML*, 2024, pp. 22 605–22 623.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical Parametric Speech Synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[4] M. Alkaeed, A. Qayyum, and J. Qadir, "Privacy Preservation in Artificial Intelligence and Extended Reality (AI-XR) Metaverses: A Survey," *J. Netw. Comput. Appl.*, vol. 231, p. 103 989, 2024.

[5] H. Xu *et al.*, *Sok: Comprehensive Security Overview, Challenges, and Future Directions of Voice-Controlled Systems*, 2024. arXiv: 2405.17100.

[6] Z. Qin, W. Zhao, X. Yu, and X. Sun, *OpenVoice: Versatile Instant Voice Cloning*, 2024. arXiv: 2312 . 01479.

[7] S. Chen *et al.*, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," *IEEE TASLP*, vol. 33, pp. 705–718, 2025.

[8] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone," in *Proc. ICML*, vol. 162, 2022, pp. 2709–2720.

[9] S. Kim, H. Kim, and S. Yoon, *Guided-TTS 2: A Diffusion Model for High-Quality Adaptive Text-to-Speech with Untranscribed Data*, 2022. arXiv: 2205.15370.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *Proc. ICLR*, 2015.

[11] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Examples in the Physical World," in *Proc. ICLR Workshop*, 2017.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial Machine Learning at Scale," in *Proc. ICLR*, 2017.

[13] C. Szegedy *et al.*, "Intriguing Properties of Neural Networks," in *Proc. ICLR*, 2014.

[14] C. Xie *et al.*, "Improving Transferability of Adversarial Examples with Input Diversity," in *Proc. CVPR*, 2019, pp. 2730–2739.

[15] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks," in *Proc. CVPR*, 2019, pp. 4312–4321.

[16] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into Transferable Adversarial Examples and Black-Box Attacks," in *Proc. ICLR*, 2017.

[17] N. R. Koluguri, T. Park, and B. Ginsburg, "TitaNet: Neural Model for Speaker Representation with 1D Depth-Wise Separable Convolutions and Global Context," in *Proc. ICASSP*, 2022, pp. 8102–8106.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Prop-

agation and Aggregation in TDNN Based Speaker Verification," in *Proc. InterSpeech*, 2020, pp. 3830–3834.
[20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006, pp. 369–376.
[21] A. Graves, "Sequence Transduction with Recurrent Neural Networks," in *Proc. ICML Workshop on Representation Learning*, 2012.
[22] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1240–1253, 2017.
[23] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A Comparison of Techniques for Language Model Integration in Encoder-Decoder Speech Recognition," in *Proc. IEEE SLT*, 2018.
[24] Y. Gong and C. Poellabauer, "Crafting Adversarial Examples for Speech Paralinguistics Applications," in *Proc. DYNAMICS Workshop*, 2018.
[25] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling Deep Structured Prediction Models," in *Proc. NIPS*, 2017, pp. 6980–6990.
[26] X. Yuan *et al.*, "CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition," in *Proc. USENIX Security*, 2018, pp. 49–64.
[27] Y. Qin, N. Carlini, I. Goodfellow, G. Cottrell, and C. Raffel, "Imperceptible, Robust, and Targeted Adversarial Examples for Automatic Speech Recognition," in *Proc. ICML*, 2019, pp. 5231–5240.
[28] N. Carlini and D. Wagner, "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text," in *Proc. IEEE SPW*, 2018, pp. 1–7.
[29] X. Zhang *et al.*, "Adversarial Example Attacks Against ASR Systems: An Overview," in *Proc. IEEE DSC*, 2022, pp. 470–477.
[30] S. Chen *et al.*, "Adversarial Speech for Voice Privacy Protection from Personalized Speech Generation," in *Proc. ICASSP*, 2024, pp. 11 411–11 415.
[31] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in *Proc. ICML*, 2021, pp. 5530–5540.
[32] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. ICLR*, 2015.
[33] H. Zen *et al.*, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. InterSpeech*, 2019, pp. 1526–1530.
[34] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *Proc. ICLR*, 2019.
[35] T. Saeki *et al.*, "UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022," in *Proc. InterSpeech*, 2022, pp. 4521–4525.
[36] M. Ravanelli *et al.*, *SpeechBrain: A General-Purpose Speech Toolkit*, 2021. arXiv: 2106.04624.