



Asynchronous Voice Anonymization Using Adversarial Perturbation On Speaker Embedding

Rui Wang¹, Liping Chen^{1*}, Kong Aik Lee², Zhen-Hua Ling¹

¹NERC-SLIP, University of Science and Technology of China (USTC), China

²Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong

wangrui256@mail.ustc.edu.cn, {lipchen, zhling}@ustc.edu.cn, kong-aik.lee@polyu.edu.hk

Abstract

Voice anonymization has been developed as a technique for preserving privacy by replacing the speaker's voice in a speech signal with that of a pseudo-speaker, thereby obscuring the original voice attributes from machine recognition and human perception. In this paper, we focus on altering the voice attributes against machine recognition while retaining human perception. We referred to this as the *asynchronous voice anonymization*. To this end, a speech generation framework incorporating a speaker disentanglement mechanism is employed to generate the anonymized speech. The speaker attributes are altered through adversarial perturbation applied on the speaker embedding, while human perception is preserved by controlling the intensity of perturbation. Experiments conducted on the LibriSpeech dataset showed that the speaker attributes were obscured with their human perception preserved for 60.71% of the processed utterances. Audio samples can be found in ¹.

Index Terms: voice privacy, human perception preservation, asynchronous anonymization, adversarial perturbation on speaker embedding

1. Introduction

With the rapid development of speech technology in recent years, the security of speech data is confronted with increasing risks. For example, the advances in speaker recognition techniques facilitate recognizing the speaker's identity in a speech utterance with high accuracy. When combined with speech recognition, the personal information conveyed in the speech utterances might be revealed and leaked. Also, the advancements in personalized speech generation techniques, including *voice conversion* (VC) [1–3] and *text-to-speech* (TTS) [4–6], enable the generation of artificial speech utterances in both high quality and speaker similarity. Such techniques can be maliciously used to generate fake speech imitating specific individuals, thereby impersonating their identities.

The aforementioned risks of speaker information raise the call for voice privacy protection techniques. In this regard, the anti-spoofing [7] methods were developed to detect artificially generated speech utterances. They operate after the speaker information is extracted and applied to generate the fabricated speech. By preventing the fabricated speech from being used for malicious purposes, the anti-spoofing methods protect the speaker information in a passive manner. On the other hand, the voice anonymization technology [8–11] protects the speaker information within a speech utterance by replacing the original

speaker with a pseudo-speaker by means of VC. It provides proactive protection to the information of the original speaker once it's generated. Such a speaker replacement approach alters not only the speaker information perceived by machine learning algorithms but also that perceived by human listeners.

However, application scenarios also call for the voice privacy protection under the condition that the human perception of the original speaker is preserved, e.g., the famous people who address the public, the content creators of YouTube short videos, etc. In this paper, we focus on anonymizing the voice when perceived by machine algorithms while preserving human perception. In contrast to the existing voice anonymization where both machine and human perceptions of speaker attributes are altered synchronously, the proposed task is asynchronous. To this end, a speech generation framework is adopted to incorporate an explicit mechanism for speaker information modeling. Within this framework, the speaker representation extracted from the original speech is modified and used to generate the protected speech. In our work, the VC function of the YourTTS [12] model is utilized given its capability in information disentanglement and speech generation. As a preliminary work, this paper focuses on protecting closed-set speakers. The *fast gradient step method* (FGSM) algorithm [13] is adopted to modify the speaker embedding. Experiments were conducted on the LibriSpeech dataset [14]. Listening test reveals that 60.71% of the processed utterances sounded alike to those utilizing original speaker attributes. In automatic speaker verification evaluations, higher equal error rates (EERs) of these utterances suggest their capability to obscure speaker attributes for machines while maintaining them to humans.

The rest of the paper is organized as follows. Section 2 describes the background of our work. The proposed protected speech generation method is illustrated in Section 3. Experimental results and analysis are presented in Section 4. Finally, Section 5 concludes the paper.

2. Background

In this section, we will first briefly go through the speaker embedding modeling. Then, the inference flow of the VC function within YourTTS model is described. Finally, the FGSM algorithm for adversarial attack is presented.

2.1. Speaker embedding

Speaker embedding refers to the representation of a variable-length utterances as a fixed-length vector [15]. Mathematically, denoting an utterance as $\mathcal{O} = \{\mathbf{o}_1, \dots, \mathbf{o}_T\}$ with T denote the number of frames, a speaker embedding is estimated as:

$$\mathbf{x} = \mathcal{E}(\mathbf{o}_1, \dots, \mathbf{o}_T), \quad (1)$$

*: corresponding author. This work was supported in part by the National Natural Science Foundation of China under Grant U23B2053.

¹<https://voiceprivacy.github.io/asynchronous-voice-anonymization>

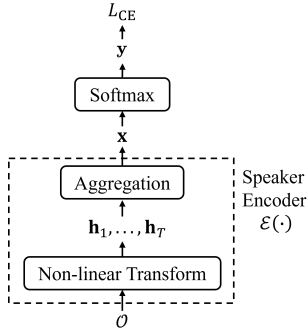


Figure 1: The architecture of the discriminative speaker attributes modeling. The input speech utterance is represented as \mathcal{O} . Its speaker embedding vector \mathbf{x} is extracted with the speaker encoder $\mathcal{E}(\bullet)$.

where $\mathcal{E}(\bullet)$ represents the speaker embedding extractor.

In the neural network framework, $\mathcal{E}(\bullet)$ is always specified to be the speaker encoder, trained in a discriminative manner. Fig. 1 describes the neural network architecture for speaker embedding modeling. As illustrated in Fig. 1, $\mathcal{E}(\bullet)$ firstly performs non-linear transforms on \mathcal{O} and outputs $\{\mathbf{h}_1, \dots, \mathbf{h}_T\}$ as the representations of the speaker attributes within each frame. Then it aggregates the frame-level embedding vectors into an utterance-level embedding vector \mathbf{x} . Following that, a softmax layer is applied on \mathbf{x} to predict the speaker class of \mathcal{O} as follows:

$$\mathbf{y} = \text{softmax}(\mathbf{x}), \quad (2)$$

with \mathbf{y} denoting the output of the softmax layer.

The cross-entropy (CE) between \mathbf{y} and the label \mathbf{t} is computed as the loss function as follows:

$$L_{CE} = - \sum_{c=1}^C t_c \log y_c, \quad (3)$$

with C denoting the number of speaker classes. As such, the model is trained by minimizing L_{CE} . The speaker encoder $\mathcal{E}(\bullet)$ can be obtained as a result. In inference, given a speech utterance, the speaker encoder is applied to estimate the speaker \mathbf{x} , referred to as the speaker embedding.

2.2. YourTTS

YourTTS is a model used for personalized speech generation in both TTS and VC functions [12]. The VC function is adopted in our work, which converts the speaker in the source speech to that of the reference speech. The inference flow of the VC function can be decomposed into information disentanglement and waveform construction as illustrated in Fig. 2.

In detail, the model takes a pair of source and reference speech utterances as input, denoted as \mathcal{O}_s and \mathcal{O}_r , respectively. In the disentanglement stage, the content information in \mathcal{O}_s is disentangled via the content disentanglement module. Specifically, this is fulfilled by the posterior encoder [16] and the reverse pass of the flow-based decoder f^{-1} . The content information is represented as \mathbf{z}_p in Fig. 2. In this process, the speaker embedding vector \mathbf{x}_s is extracted from \mathcal{O}_s first and then used to remove the speaker information from the content representation. Meanwhile, the speaker information within \mathcal{O}_r is extracted via the speaker encoder, represented with the speaker embedding vector \mathbf{x}_r . In our work, the F0 values of the source

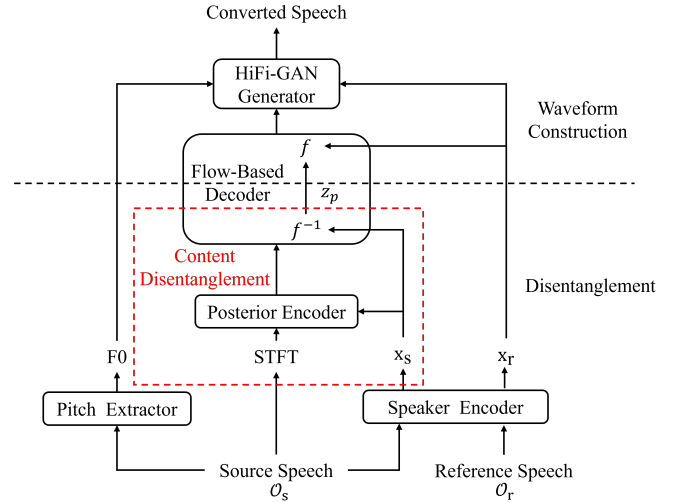


Figure 2: Inference flow of the voice conversion function of YourTTS. In comparison with the official version as described in [12], the F0 extracted from the source speech is used. The black dotted line separates the modules into information disentanglement and waveform construction. The modules in the rectangular box of red dotted line are used for disentangling the content information from the source speech.

speech \mathcal{O}_s are extracted with a pitch extractor, representing its prosody information. Finally, the prosody and content representation of \mathcal{O}_s combined with the speaker representation of \mathcal{O}_r to construct the waveform. This is achieved via the forward pass of the flow-based decoder followed by the HiFi-GAN [17] generator. Finally, the speech is generated by converting the speaker attributes within the source speech \mathcal{O}_s into that of the reference speech \mathcal{O}_r . For details, readers are suggested to read [12].

2.3. FGSM

The fast gradient step method (FGSM) was proposed in [13] to generate adversarial samples to attack neural network models. Given an input sample \mathbf{s} and the network model $\mathbf{y} = f_\theta(\mathbf{s})$ where $f_\theta(\bullet)$ is the network function with the subscript θ denoting the model parameter set. Denote the loss function as $L(\mathbf{y}, \mathbf{y}^{\text{true}})$, where \mathbf{y}^{true} is the true label associated with \mathbf{s} . The adversarial sample is obtained by taking a small step in the direction that maximizes the loss function with respect to \mathbf{s} . Mathematically, $\bar{\mathbf{s}}$ is obtained as:

$$\bar{\mathbf{s}} = \mathbf{s} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{s}} L(f_\theta(\mathbf{s}), \mathbf{y}^{\text{true}}; \theta)). \quad (4)$$

In (4), ϵ is a variable controlling the intensity of the attack, $\text{sign}(\bullet)$ is the operation that takes the sign of the value, $\nabla_{\mathbf{s}}$ is the derivative with respect to the original sample \mathbf{s} .

3. Protected speech generation

This section describes the proposed method for speaker protected speech generation, including the overall model architecture and the adversarial perturbation on the speaker embedding.

3.1. Overall architecture

Fig. 3 depicts the diagram of the generation of the protected speech. In detail, given the VC flow of a YourTTS model, the

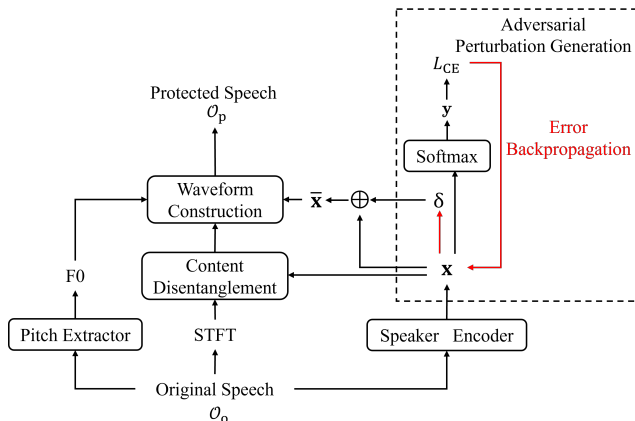


Figure 3: Protected speech generation based on the VC function of YourTTS. The content disentanglement module inherits that from Fig. 2. The rectangular box of the dotted line is the adversarial attack to generate the perturbed speaker embedding \bar{x} . The error backpropagation process to obtain the perturbation δ is denoted by the red arrow line.

original speech utterance \mathcal{O}_o is used as both the source and reference utterances. The original speaker embedding vector \mathbf{x} is firstly extracted from \mathcal{O}_o . The content information is disentangled from \mathcal{O}_o via the content disentanglement module with the help from \mathbf{x} , and denoted with \mathbf{z}_p . Meanwhile, the F0 values are extracted, representing the prosody information in \mathcal{O}_o . To generate the protected speech \mathcal{O}_p , the speaker embedding vector \mathbf{x} is modified to be \bar{x} . In our work, the modification is accomplished by the adversarial perturbation generation module as illustrated in the rectangular box of the dotted line. To be specific, the perturbation is generated through the adversarial attack and is described in the following.

3.2. Adversarial perturbation on speaker embedding

Given the speaker model within which the speaker encoder is trained, it is perturbed using the FGSM algorithm. First of all, assume that the speaker of \mathcal{O}_o is among the speaker classes that the speaker model can recognize, its speaker can be predicted with the softmax function on \mathbf{x} as follows:

$$\mathbf{y} = \text{softmax}(\mathbf{x}), \quad (5)$$

where \mathbf{y} is the prediction output. The cross-entropy between \mathbf{y} and the label is computed to be L_{CE} according to (3). Following that, as shown in (4), the adversarial perturbation δ is obtained on \mathbf{x} via error backpropagation. The adversarial speaker embedding vector \bar{x} is obtained by adding δ to \mathbf{x} as follows:

$$\bar{x} = \mathbf{x} + \epsilon \cdot \delta, \quad (6)$$

with ϵ denotes the intensity. The protected speech waveform \mathcal{O}_p is constructed by combining the F0 values, the speech content representation \mathbf{z}_p and the adversarial speaker embedding \bar{x} .

4. Experiments

4.1. Dataset & configurations

Our experiments were conducted on the LibriSpeech corpus. Within which, the dev-clean and test-clean datasets were used for evaluation, including 5,323 utterances from 40 female and

40 male speakers. In our experiments, YourTTS model was trained on the LibriTTS [18] train-clean-100 dataset. The speaker encoder utilized the pre-trained H/ASP [19] model available in the open-source repository². The open-source H/ASP speaker encoder was trained on the VoxCeleb2 [20]. However, the FGSM algorithm in our protected speech generation requires that the original speaker must be one of the training speakers for the speaker encoder. To this end, a softmax layer is retrained on the speaker encoder with utterances from the test speakers. For each test speaker, 70% utterances were selected to train a softmax layer. In this process, only the parameters of the softmax layer were trained on the speaker encoder, while the remaining parameters of the speaker encoder were kept frozen. The rest 1,522 utterances of the speakers were used in our evaluations. In the FGSM algorithm implemented on the speaker embedding during protected speech generation, the intensity ϵ was set to 0.02.

In the evaluations, three kinds of speech utterances were evaluated, including: 1) the original speech; 2) the regenerated speech that the YourTTS model generated without modifying the speaker embedding vector; 3) the protected speech generated by adding perturbation to speaker embedding. Initially, a subjective listening test was applied to assess the ability of protected utterances to maintain human perception of speaker attributes. Given the precondition for the protected utterance to preserve the human perception of speaker attributes, utterances maintaining the human perception of the speaker attributes were chosen for subsequent evaluations. Next, a *speaker mean opinion score* (SMOS) test was conducted further to evaluate the preservation of speaker attributes in human perception. Following this, ASV evaluations were applied to measure speaker identity concealment from machine algorithms. Finally, ASR evaluations were conducted to measure the preservation of content information in the protected speech.

4.2. Human perception evaluation

First of all, a subjective listening test was conducted on the protected utterances to assess their ability to maintain the human perception of the speaker attributes. As the protected speech was generated using the YourTTS model, potentially introducing quality reduction compared to the original ones, regenerated utterances were used as the reference for a fair speaker comparison. In this test, 1,522 protected utterances were generated from the rest 1,522 utterances, three listeners were asked to decide whether the protected utterances sounded alike to the corresponding regenerated ones. The listeners provided binary responses (yes or no) for each pair. The utterance pairs that obtained yes from all three listeners were decided to have the same speaker attributes.

Through listening test, 60.71% of the protected speech utterances were labeled to match the speakers within the regenerated utterances. This shows that protected speech is capable of preserving the human perception of the speaker attributes. These utterances were selected and used in our remaining evaluations, called the labeled dataset, including 455 utterances from 36 female and 469 utterances from 37 male.

4.3. SMOS test

The *speaker-similarity mean opinion score* (SMOS) test was conducted by 15 paid native speakers. They were asked to score the speaker similarity between utterance pairs. SMOS scores

²https://github.com/clovaai/voxceleb_trainer

ranged from 1 to 5 with a 0.5 step. The higher scores indicated higher speaker similarity. The SMOS tests were carried out in three scenarios, where the utterance pair were: *original-original(ori-ori)*, *original-regenerated(ori-regen)* and *regenerated-protected(regen-prote)*. The SMOS scores are presented in Table 1.

Table 1: *The SMOS scores with 95% confidence intervals among the original(ori), regenerated(regen), and protected(prote) utterances generated with out proposed method.*

	ori-ori	ori-regen	regen-prote
SMOS	4.17 ± 0.14	4.08 ± 0.10	4.16 ± 0.10

As shown in Table 1, by comparing the score of ori-ori and ori-regen, it can be seen that the speech generation of our YourTTS model did not significantly affect the speaker attributes within the original speech utterances in terms of human hearing. Furthermore, the protected utterances were compared to the regenerated utterances, instead of the original ones, to mitigate the speaker similarity degradation caused by the YourTTS model, ensuring a fair evaluation. The SMOS of regen-prote was close to ori-ori, which serves as the benchmark for SMOS. This indicates the protected utterances sounded alike to humans with the regenerated utterances and the protected speech was able to preserve the speaker attributes for humans.

4.4. ASV evaluations

To evaluate the capability of the protected speech in protecting the speaker privacy, the ASV evaluations were carried out with the performances being measured by the *equal error rate* (EER). The ASV tests were conducted on the x-vector [21] model, i-vector [22] models, and the speaker encoder of the YourTTS model. According to whether the evaluation model was seen during speech generation, the first two models were black-box while the third one is white-box. The x-vector and i-vector models were trained following the Kaldi recipe³ [23] using the original speech and the regenerated speech of train-clean-360 dataset. The scores were computed with a *probabilistic linear discriminant analysis* (PLDA) [24] backend. The ASV evaluations were conducted in a gender-dependent manner on the 60.71% test utterances within which the speaker attributes sounded similar between the regenerated and the protected utterances. In the evaluation trial configuration, 25% of the test utterances of each speaker were used for enrollment with the other 75% used for test. The target and nontarget trials were composed in a ratio of 1:35 for each speaker. Overall, the evaluation on females involved 3,003 target and 107,158 nontarget trials, while the evaluation on males included 2,559 target and 91,463 nontarget trials. The EERs are given in Table 2.

Based on the findings in Table 2, it can be observed that higher EERs were obtained when there was a mismatch in the utterance types between the model training data and the evaluation data, i.e., the original and the regenerated utterances. This was due to the channel mismatch issue between the two types of utterances. As observed in Table 2, the protected utterances achieved higher EERs on all evaluation conditions compared to the regenerated utterances. This suggests that within our test dataset, the machine’s perception of the speaker attributes in 60.71% of utterances was obscured, while human perception remained unchanged.

³<https://github.com/kaldi-asr/kaldi/egs/voxceleb>

Table 2: *The EERs(%) calculated on the original(ori), regenerated(regen), and protected(prote) utterances generated with our proposed method. Evaluation results on the speaker encoder within the YourTTS model, x-vector, and i-vector are included.*

Model	training data	gender	ori	regen	prote
speaker encoder	ori	male	0.81	4.30	9.20
		female	0.96	4.20	9.53
x-vector	ori	male	1.93	3.97	8.22
		female	1.63	4.30	8.09
	regen	male	3.12	3.34	7.15
		female	3.19	3.04	6.90
i-vector	ori	male	1.87	4.97	8.51
		female	1.03	4.08	7.64
	regen	male	3.40	3.29	6.92
		female	1.63	2.22	5.27

4.5. ASR evaluations

Finally, in our experiments, the ASR evaluations were conducted on the protected utterances. In this evaluation, the ASR service whisper⁴ [25] provided by OpenAI was called for speech recognition. The performances were measured with *word error rate* (WER). Our evaluations were conducted on the libri-dev and libri-test datasets as officially published within the LibriSpeech corpus. The original, regenerated, and protected utterances were evaluated. The WERs are presented in Table 3.

Table 3: *Automatic speech recognition results WERs(%) on original(ori), regenerated(regen) and protected(prote) utterances generated with out proposed method.*

	ori	regen	prote
WER	4.26	7.47	7.66

As shown in Table 3, Compared with the original utterances, a higher WER was obtained on the regenerated speech. This might be caused by that our regenerated utterances were unseen by the ASR model of whisper. Meanwhile, the protected utterances obtained a close WER with the regenerated ones, indicating that the protection on the speaker attributes didn’t hurt the content information within the speech utterances.

5. Conclusions

This paper focuses on the *asynchronous voice anonymization* technique. In our work, a speech generation framework that incorporates the speaker disentanglement mechanism is adopted. The proposed method involves extracting the speaker embedding from the original speech and modifying it with adversarial perturbation to obscure speaker attributes from machine recognition, while human perception is preserved by controlling the magnitude of the perturbation. Experiments conducted on the LibriSpeech dataset showed that the machine recognition of speaker attributes was obscured with a 60.71% success rate while human perception was preserved. Our future research will focus on improving the modification of speaker embedding to conduct asynchronous voice anonymization on all utterances.

⁴<https://platform.openai.com/docs/api-reference/introduction>

6. References

- [1] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.
- [2] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1–6.
- [3] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only auto-encoder loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.
- [4] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, p. 4050, 2019.
- [5] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.
- [6] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.
- [7] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *speech communication*, vol. 66, pp. 130–153, 2015.
- [8] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, and et al, "Introducing the voiceprivacy initiative," in *Proc. Interspeech*, 2020, pp. 1693–1697.
- [9] S. E. McAdams, *Spectral Fusion, Spectral Parsing and the Formation of Auditory Images*. Ph.D. dissertation, Stanford university, 1984.
- [10] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 155–160.
- [11] B. M. L. Srivastava, M. Maouche, M. Sahidullah, E. Vincent, A. Bellet, M. Tommasi, N. Tomashenko, X. Wang, and J. Yamagishi, "Privacy and utility of x-vector based speaker anonymization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2383–2395, 2022.
- [12] 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 *International Conference on Machine Learning*, 2022, pp. 2709–2720.
- [13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2014.
- [14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.
- [15] K. A. Lee, Q. Wang, and T. Koshinaka, "Xi-vector embedding for speaker recognition," *IEEE Signal Processing Letters*, vol. 28, pp. 1385–1389, 2021.
- [16] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, 2021, pp. 5530–5540.
- [17] J. Kong, J. Kim, and J. Bae, "Hifi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.
- [18] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019, pp. 1526–1530.
- [19] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.
- [20] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, 2018, pp. 1086–1090.
- [21] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Interspeech*, 2017, pp. 999–1003.
- [22] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *Proc. IEEE workshop on automatic speech recognition and understanding (ASRU)*, 2011.
- [24] S. J. Prince, *Computer Vision: Models, Learning, and Inference*. Cambridge University Press, 2012.
- [25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, 2023, pp. 28 492–28 518.