PSVRF: LEARNING TO RESTORE PITCH-SHIFTED VOICE WITHOUT REFERENCE

Yangfu Li   Xiaodan Lin*   Jiaxin Yang

School of Information Science and Engineering, Huaqiao University, Xiamen, China

ABSTRACT

Pitch scaling algorithms have a significant impact on the security of Automatic Speaker Verification (ASV) systems. Although numerous anti-spoofing algorithms have been proposed to identify the pitch-shifted voice and even restore it to the original version, they either have poor performance or require the original voice as a reference, limiting the prospects of applications. In this paper, we propose a no-reference approach termed PSVRF for high-quality restoration of pitch-shifted voice. Experiments on AISHELL-1 and AISHELL-3 demonstrate that PSVRF can restore the voice disguised by various pitch-scaling techniques, which obviously enhances the robustness of ASV systems to pitch-scaling attacks. Furthermore, the performance of PSVRF even surpasses that of the state-of-the-art reference-based approach.

Index Terms— voice restoration, anti-spoofing, Automatic Speaker Verification (ASV), pitch scaling

1. INTRODUCTION

Motivation: The recent emergence of Automatic Speaker Verification (ASV) systems in high-security-required fields like AoT, Voice Assistant, and multimedia forensic leads to an increasing concern for its security risks [1–3]. ASV systems utilize the distance between the extracted features of test speech and those of pre-collected reference speech to determine the speaker. However, the attackers could hide the real identity of a speaker through automatic voice disguise (AVD). In particular, a classic AVD technique termed pitch scaling [4] is extensively applied in various commercial software due to its excellent balance of effectiveness and efficiency, posing a great threat to the security of ASV.

Prior works and limitation: Early works [5–7] typically estimate the approximate range of pitch shifting rather than the precise degree of disguise, rendering them incapable of accurately restoring the pitch-shifted voice. Later, Pilia et al. propose a method achieving more accurate estimation results than previous work [8]. However, the model can only deal with time-domain pitch scaling. Recently, L. Zheng et al. propose a state-of-the-art method for detecting and restoring pitch-shifted voice [9]. This method utilizes an ASV system to achieve the estimation of disguising parameters and the restoration of disguised voice, which is capable of reliably working on various pitch scaling algorithms. However, this method still has two limitations: (1) due to the dependency on the ASV system, it cannot be adaptive to the situation without reference audio; and (2) it uses pitch scaling algorithms to achieve restoration, which doubles the artifacts introduced during disguising, decreasing the restoration quality.

Our approach: In this paper, we propose a Pitch-Shifted Voice Restoration Framework (PSVRF) for estimating disguising parameters in the absence of reference and restoring pitch-shifted voice in high quality. Specifically, PSVRF consists of three contributing components: (1) Estimator, which predicts the disguising parameter through the log Mel filter-bank (fbank) features of disguised voice without any reference; (2) Feature Reconstruction Network (FRN), which reconstructs the fbank features of original voice in high quality through fbank features of pitch-shifted voice and the predicted parameter; and (3) a neural vocoder, which converts the reconstructed features into waveforms, achieving end-to-end pitch-shifted voice restoration. The experiments conducted on AISHELL-1 and AISHELL-3 with various pitch scaling algorithms demonstrate that PSVRF overpasses the state-of-the-art reference-based restoration method in not only the accuracy of estimation but also the quality of restoration.

2. BACKGROUND

2.1. Pitch Scaling

Pitch scaling techniques can be mainly divided into two categories: frequency-domain (FD) disguise and time-domain (TD) disguise. FD disguise is usually operated by expanding or compressing the spectrum while keeping the content of the voice unchanged. TD disguise can be realized by adjusting the sampling rate, which changes the fundamental frequency of speech signal and hence the pitch. FD disguise and TD disguise can be formulated into a unified form as follows [9]:

$$p_s = 2^{\alpha/12} p_o,$$  (1)

where $p_o$ and $p_s$ represent the original pitch and scaled pitch. $\alpha$ is the semitone, i.e., the disguising parameter, which describes the degree of disguise.

Code is available on: https://github.com/YChenL/PSSRF.
2.2. Self-supervised audio spectrogram transformer

Audio spectrogram transformer (AST) is the state-of-the-art purely attention-based model for audio tasks [10]. Recently, K. He et al. propose Masked AutoEncoder (MAE) for a large-scale self-supervised pre-train [11], which can obviously enhance the performance of the purely attention-based model in vision. Specifically, it masks numerous patches of the inputs and then utilizes the model to reconstruct the masked inputs. Later, Y. Gong et al. introduce the pre-train strategy proposed in MAE into AST and propose the Self-supervised audio spectrogram transformer (Ssast) [12], which makes two efficient improvements: (1) a frame-level masking strategy, which is more efficient than patch-level masking; (2) Joint Discriminative and Generative Masked Spectrogram Patch Modeling.

3. METHODOLOGY

The architecture and objective functions of PSVRF are shown in Fig. 1, which consists of three main components: Estimator, Feature Restoration Network (FRN), and a pre-trained Neural Vocoder, achieving the restoration of pitch-shifted voice end-to-end. We detail these components in 3.1 to 3.3.

3.1. Estimator

Estimator is similar to the AST, which is composed of a linear projection and a transformer encoder. Specifically, each fbank feature is partitioned into 16×16 patches, which are flattened into 1D 768-dimensional patch embeddings and fed into the linear projection. Then, the transformer encoder accepts the output of the linear projection plus the position embedding as the inputs. The transformer encoder has an embedding dimension of 768, 12 layers, and 12 heads, which are the same as those in the original AST [10]. During fine-tuning and inference, an average pooling followed by a fully connected layer is applied to yield the estimation of the disguising parameter. Notably, Estimator is pre-trained as Ssast in Voxceleb [13] and Voxceleb2 [14] with 400 epochs and fine-tuned in a supervised mode using MAE loss.

3.2. Feature Reconstruction Network

Generator: We specially design a model termed Feature Reconstruction Network (FRN) for high-quality restoration. To be specific, FRN is a type of Generator Adversarial Network (GAN) [15], which is composed of a generator $G$ and the associated discriminator $D_{G}$. As shown in Fig. 2 (a), $G$ is mainly composed of 20 residual blocks with a hidden dimension of 256, which is introduced in WaveNet [16]. Differently, we make the model non-causal and set the dilation rate to 1 since the inputs are spectrograms instead of waveforms.

Discriminator: Multi-scaled discriminators (MSD) [17] and Joint Conditional Unconditional discriminators (JCU) [18] have been proven as the most efficient models in audio synthesis tasks. Inspired of them, we propose a Joint Conditional Unconditional Multi-scale discriminator (JCU-MSD), which is shown in Fig. 1 (b).

Objective function: Adversarial loss $L_{adv}$, spectrogram reconstruction loss $L_{spec}$, and feature matching loss $L_{fm}$ are applied to constrain the generator. $L_{adv}$ is defined as follows:

$$L_{adv} = \mathbb{E}_{x \sim x_s}[D_{\phi}(x, \hat{\alpha})^2] + \mathbb{E}_{z \sim G}[(D_{\phi}(x, \hat{\alpha}) - 1)^2],$$

(2)

where $x_s$ is fbank feature of disguised speech, $\hat{\alpha}$ is estimated disguising parameters. $L_{spec}$ is measured by L2 distances between the real spectrogram and its reconstructed counterpart, which can be formulated as follows:

$$L_{spec} = \|x_o - G(x_s, \hat{\alpha})\|_2,$$

(3)

where $x_o$ is fbank feature of orginal speech. $L_{fm}$ is computed by summing L1 distances between every discriminator feature maps of real and generated samples, which is defined as follows:

$$L_{fm} = \sum_{i=0}^{N} \|D_{\phi}(x_o, \hat{\alpha}) - D_{\phi}(G(x_s, \hat{\alpha}), \hat{\alpha})\|_1,$$

(4)

where $N$ is the total number of hidden layers in the JCU-MSD. Finally, the total loss of PSVRF is defined as follows:

$$L_{total} = L_{adv} + L_{spec} + \lambda_{fm} L_{fm},$$

(5)

where $\lambda_{fm}$ is a scaled scalar equal to 0.5 in this work. $G$ aims to minimize $L_{total}$ while the objective of $D$ is opposite to $G$. 

Fig 1: Overview of PSVRF, which includes Estimator, Feature Reconstruction Network (FRN) composed of a generator and the associated Joint Conditional- Unconditional Multi-scale Discriminator (JCU-MSD), and a neural vocoder.
3.3. Neural Vocoder

We use the state-of-the-art neural vocoder termed DiffWave [19] to transform the reconstructed fbank features into the restorated waveform. This vocoder is pretrained in Voxceleb [13] and Voxceleb2 [14] with $10^6$ steps and directly applied in PSVRF. Differently, the dimension of the input spectrograms is set to 28 instead of 80 in this work.

4. EXPERIMENTS

4.1. Implement details

| Dataset  | Speakers | Utterances | $\alpha$ Range |
|----------|----------|------------|----------------|
| T1       | AISHELL-3 train | 138 female / 36 male | 63262, $\ell(-8, 8)$ |
| T2       | AISHELL-3 train | 356 female / 158 male | 123471, $\ell(-8, 8)$ |
| A1 Unseen| AISHELL-1 dev  | 35 female / 25 male  | 561 × 33, $[-8, 8, 0.5]$ |
| A3 Seen  | AISHELL-3 test | 134 female / 36 male | 270 × 33, $[-8, 8, 0.5]$ |
| A3 Unseen| AISHELL-3 test | 38 female / 8 male   | 480 × 33, $[-8, 8, 0.5]$ |

**Evaluation:** We evaluate the performance of the Estimator in PSVRF trained with T1 / T2 and that of the baseline [9] in A1 Unseen, A3 Seen, and A3 Unseen with the Mean Absolute Error (MAE) between the predicted $\alpha$ and the ground truth, which is recorded in TABLE 2. In addition, we investigate the relationship between the estimation deviation i.e., $|\hat{\alpha} - \alpha|$, of PSVRF$_{T2}$ and the disguising parameters $\alpha$ applied in A1 Unseen, which is shown in Fig. 2.

**Results:** From TABLE 2, we can find that although only the Phase Vocoder implemented by librosa is used to yield the training data, PSVRF can still precisely estimate the disguising parameters from different pitch scaling algorithms. In addition, a larger scale dataset can further boost the performance and generalization ability of PSVRF. It is noteworthy that PSVRF does not require any reference while it still achieves competitive results compared to the reference-based method. Fig. 3 reveals that the estimation of negative $\alpha$ is

![Diagram](image-url)
more accurate than that of the positive, which is consistent with the conclusion in [6, 7] that lowering pitch is easier to detect than raising pitch. Besides, the tiny $\alpha$ is prone to be estimated as zero, resulting in a linear deviation in the neighbourhood of zero.

**Discussion:** There are two explanations of how Estimator works: (1) the estimator learns a mapping from the artifacts introduced by pitch-scaling algorithms to $\alpha$; (2) Estimator learns a manifold which is composed of the original voice and its pitch-shifted counterpart, and mappings testing samples to the learned manifold for generalization. The results in Table 2 reveal that Estimator can be generalized to various disguising algorithms. However, the artifacts introduced by different algorithms are usually different. In addition, more speakers’ information can boost the performance of PSVRF. Therefore, we believe explanation (2) may be more correct, which will be further studied in our future work.

### 4.3. Evaluation of restoration quality

**Evaluation:** We apply a typical ASV model termed ECAPA-TDNN [26] as an effective tool to evaluate the restoration quality of PSVRF and the baseline. Specifically, we qualitatively evaluate the improvement provided by different methods for the ASV model when faced with pitch-shifted samples from $A_1$ Unseen, which is shown in Fig. 4. In addition, we provide a visual comparison of the restoration obtained by different methods to further explain the advantage of PSVRF, which is shown in Fig. 5.

**Results:** Fig. 4 reveals that both the baseline and PSVRF can clearly enhance the performance of ASV when faced with pitch-shifted voice, while PSVRF provides higher restoration quality, which is reflected in the lower ERR of ASV. The main reason is that pitch-shifting algorithms will introduce artifacts during the disguising phase, and the baseline utilizes the pitch scaling algorithm to achieve the restoration, doubling the unpleasant artifacts and degrading the quality of restored speech. Differently, FRN in PSVRF is specifically designed to fit a mapping from fbank features of disguised speech to fbank features of original speech, which restores the pitch and removes the artifacts. This issue is further indicated in Fig. 5, where these two methods can both reconstruct the fundamental frequency, i.e., the pitch, of the original speech exactly, but PSVRF can reconstruct more clear formant and high-frequency information, resulting in the more realistic timbre.

**Discussion:** Notably, the performance of PSVRF will obviously decline under tiny $\alpha$, which is similar to or even worse than that of the baseline. The main reason is the estimation deviation of tiny $\alpha$, which is mentioned in 4.2.

### 5. CONCLUSION

We propose a no-reference method termed PSVRF to estimate the disguising parameters of pitch scaling and restore pitch-shifted voice into original revisions, which has great significance for the security of ASV. The experiments reveal that even compared with the reference-based baseline, PSVRF still obtains competitive results in both the estimation accuracy and the restoration quality. Furthermore, as a no-reference method, PSVRF can directly make existing ASV applications more resistant to pitch scaling without additional modifications. Future work would be investigating the improvement of PSVRF when faced with tiny disguising parameters.
6. REFERENCES

[1] A. Gomez-Alanis, J. A. Gonzalez-Lopez, S. P. Dubagunta, A. M. Peinado, and M. M. Doss, “On joint optimization of automatic speaker verification and anti-spoofing in the embedding space,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 1579–1593, 2020.

[2] M. Aljasem, A. Irtaza, H. Malik, N. Saba, A. Javed, K. M. Malik, and M. Meharmohammadi, “Secure automatic speaker verification (sasv) system through small features and asymmetric bagging,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 3524–3537, 2021.

[3] H. Tak, M. Todisco, X. Wang, J.-w. Jung, J. Yamagishi, and N. Evans, “Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation,” arXiv preprint arXiv:2202.12233, 2022.

[4] J. Laroche, “Time and pitch scale modification of audio signals,” in Applications of digital signal processing to audio and acoustics, pp. 279–309, Springer, 2002.

[5] H. Wu, Y. Wang, and J. Huang, “Blind detection of electronic disguised voice,” in 2013 IEEE ICASSP, pp. 3013–3017, 2013.

[6] H. Liang, X. Lin, Q. Zhang, and X. Kang, “Recognition of spoofed voice using convolutional neural networks,” in 2017 IEEE GlobalSIP, pp. 293–297, 2017.

[7] L. Wang, H. Liang, X. Lin, and X. Kang, “Revealing the processing history of pitch-shifted voice using cnns,” in 2018 IEEE WIFS, pp. 1–7, 2018.

[8] M. Pilia, S. Mandelli, P. Bestagini, and S. Tubaro, “Time scaling detection and estimation in audio recordings,” in 2021 IEEE WIFS, pp. 1–6, 2021.

[9] L. Zheng, J. Li, M. Sun, X. Zhang, and T. F. Zheng, “When automatic voice disguise meets automatic speaker verification,” IEEE Trans. Inf. Forensics Security, vol. 16, pp. 824–837, 2021.

[10] Y. Gong, Y.-A. Chung, and J. Glass, “Ast: Audio spectrogram transformer,” arXiv preprint arXiv:2104.01778, 2021.

[11] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in Proc. CVPR, pp. 16000–16009, 2022.

[12] Y. Gong, C.-L. Lai, Y.-A. Chung, and J. Glass, “Ssaast: Self-supervised audio spectrogram transformer,” in Proc. AAAI, vol. 36, pp. 10699–10709, 2022.

[13] A. Nagrani, J. S. Chung, and A. Zisserman, “Voxceleb: a large-scale speaker identification dataset,” arXiv preprint arXiv:1706.08612, 2017.

[14] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” arXiv preprint arXiv:1806.05622, 2018.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

[16] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[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. 17022–17033, 2020.

[18] S. Liu, D. Su, and D. Yu, “Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans,” arXiv preprint arXiv:2201.11972, 2022.

[19] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” arXiv preprint arXiv:2009.09761, 2020.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, “Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline,” in 2017 O-COCOSDA, pp. 1–5, IEEE, 2017.

[21] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, “Aishell-3: A multi-speaker mandarin tts corpus and the baselines,” arXiv preprint arXiv:2010.11567, 2020.

[22] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “librosa: Audio and music signal analysis in python,” in Proc. SciPy, vol. 8, pp. 18–25, CiteSeer, 2015.

[23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[24] X. Zhu, G. T. Beauregard, and L. L. Wyse, “Real-time signal estimation from modified short-time fourier transform magnitude spectra,” IEEE Trans. Audio, Speech, and Language Processing, vol. 15, no. 5, pp. 1645–1653, 2007.

[25] “Soundtouch audio processing library.” [Online]. Available: http://www.surina.net/soundtouch/. Accessed: Jul. 2022.

[26] B. Desplanques, J. Thienpondt, and K. Demuynck, “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification,” arXiv preprint arXiv:2005.07143, 2020.