PSSRF: LEARNING TO RESTORE PITCH-SCALED SPEECH 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 due to their efficiency and effectiveness. Although numerous anti-spoofing algorithms have been proposed to identify the pitch-scaled speech and even restore it to the original version, they either have poor performance or require the original speech as a reference, limiting the prospects of applications. In this paper, we propose a no-reference approach termed PSSRF[1] for high-quality pitch-scaled speech restoration. Experiments on AISHELL1 and AISHELL3 demonstrate that PSSRF can precisely estimate the disguising degree of pitch-scaled speech and possesses great robustness to different pitch-scaling techniques. In addition, the restoration quality of PSSRF even surpasses that of the state-of-the-art reference-based approach.

***Index Terms*—** Speech recovery, Automatic Speaker Verification (ASV), Pitch scaling, speech signal processing

1. INTRODUCTION

Motivation: The recent emergence of Automatic Speaker Verification (ASV) in high-security-required fields like AloT, Voice Assistant, and multimedia forensic leads to an increasing concern for its security risks [1–3]. ASV uses the distance between the extracted features of test audio and those of pre-collected reference audio 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 used in various commercial software due to the excellent balance of disguising quality and implementation difficulty, which poses a great threat to the security of ASV.

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

Our approach: In this paper, we propose a Pitch-Scaled Speech Restoration Framework termed *PSSRF* for estimating disguising parameters in the absence of reference and restoring pitch-scaled speech in high quality. Specifically, PSSRF consists of three contributing components: (1) Estimaitor, which estimates the disguising parameter through the log Mel filterbank (fbank) features of scaled speech without any reference; (2) Feature Reconstruction Network (FRN), which reconstructs the fbank features of original speech in high quality through the estimated parameter and fbank features of pitch-scaled speech; and (3) a neural vocoder, which converts the reconstructed features into waveforms, achieving end-to-end pitch-shifted speech restoration. The experiments conducted on AISHELL-1 and AISHELL-3 with various pitch scaling algorithms demonstrate that PSSRF obtains state-of-the-art results in not only the accuracy of estimation but also the quality of restoration.

2. BACKGROUND

2.1. Pitch Scaling

Pitch scaling techniques can be mainlly 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 the speech signal and hence the pitch. FD disguise and TD disguise can be formulated into a unified form as follows [9]:

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

where p_o and p_s represent the original pitch and scaled pitch. α is the semitone, i.e., the diguising parameter, which describes the degree of disgusie.

[1] Code is avaliable on: `https://github.com/YChenL/PSSRF`.

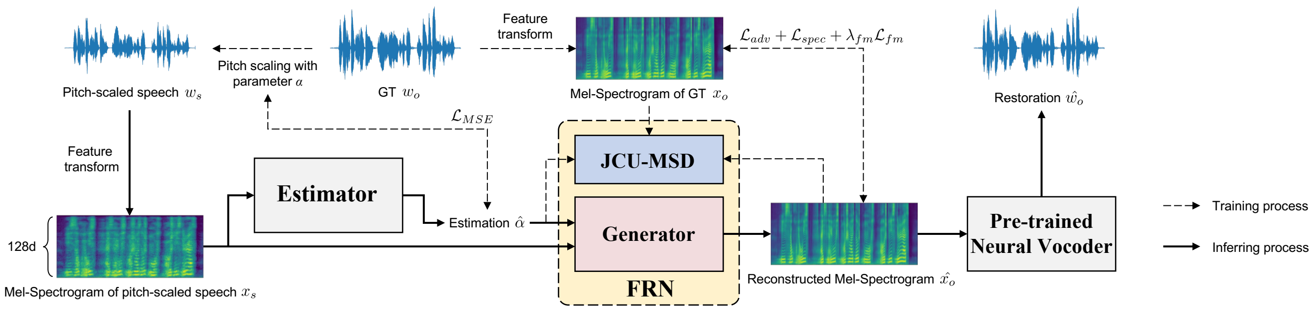


Fig. 1: Overview of PSSRF, which includes Estimator, Feature Reconstruction Network (FRN) composed of a generator and the associated Joint Conditional-Unconditional Multi-scale Discriminator (JCU-MSD), and a neural vocoder.

2.2. Self-supervied audio spectrogram transformer

Audio sepctrogram 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-supervied 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 PSSRF are shown in Fig. 1, which consists of three main components: Estimator, Feature Restoration Network (FRN), and a pre-trained Nerual Vocoder, achieving the restoration of pitch-scaled speech end-to-end. We detail these components in **3.1** to **3.3**.

3.1. Estimator

Estimator is similar as 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 disgusing 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 this task. 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_ϕ. As shown in Fig. 1 (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. The forward propagation of G is definded as follows:

$$\hat{x_o} = G(x_s, \hat{\alpha}). \tag{2}$$

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

Objective function: we apply another two loss functions besides the adversarial loss, i.e., spectrogarm reconstruction loss $\mathcal{L}_{spec}$, and feature matching loss $\mathcal{L}_{fm}$. The $\mathcal{L}_{spec}$ is measured by L2 distances between the real spectrogram and its reconstructed counterpart, which can be formulated as follows:

$$\mathcal{L}_{spec} = ||x_o - G(x_s, \hat{\alpha})||_2. \tag{3}$$

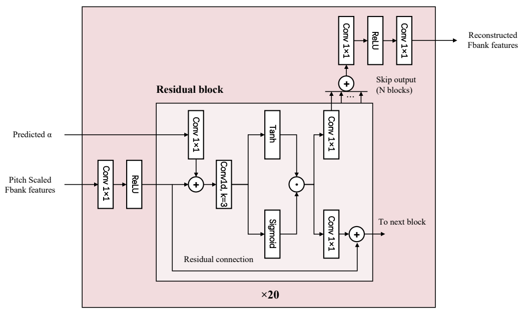
The $\mathcal{L}_{fm}$ is computed by summing L1 distances between every discriminator feature maps of real and generated samples, which is defined as follows:

$$\mathcal{L}_{fm} = \sum_{i=0}^{N} ||D_\phi^i(x_o, \hat{\alpha}) - D_\phi^i(G(x_s, \hat{\alpha}), \hat{\alpha})||_1, \tag{4}$$

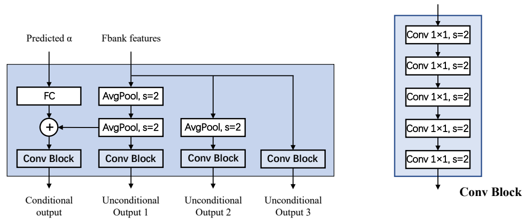
where N is the total number of hidden layers in the JCU-MSD. Finally, the total loss of generator G is defined as follows:

$$\mathcal{L}_G = \mathcal{L}_{adv} + \mathcal{L}_{spec} + \lambda_{fm}\mathcal{L}_{fm}, \tag{5}$$

where λ_{fm} is a scaled scalar equal to 0.5 in this work.



(a) Generator



(b) JCU-MSD

Fig. 2: Detailed structure of the generator and JCU-MSD, where Conv1 $\times$ 1 represents an one-dimensional (1D) convolutional operation with kernel size 1.

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 PSSRF. Differently, the dimension of the input spectorgrams is set to 128 instead of 80 in this work.

4. EXPERIMENTS

4.1. Implement details

Table 1: Datasets composition

	Dataset	# Speakers	# Utterances	α Range
T1	AISHELL-3 train	138 female / 36 male	63262	$\mathcal{U}(-8, 8)$
T2	AISHELL-1 train AISHELL-3 train	356 female / 158 male	123471	$\mathcal{U}(-8, 8)$
A_1 Unseen	AISHELL-1 test AISHELL-1 dev	35 female / 25 male	651 $\times$ 33	$[-8, 8, 0.5]$
A_3 Seen	AISHELL-3 test	134 female / 36 male	270 $\times$ 33	$[-8, 8, 0.5]$
A_3 Unseen	AISHELL-3 test	38 female / 6 male	480 $\times$ 33	$[-8, 8, 0.5]$

Dataset and processing: We utilize two multi-speaker mandarin datasets, AISHEEL-1 [20] and AISHELL-3 [21], to synthesize the training and testing sets, which are shown in TABLE 1. To be specific, original training sets of AISHELL-1 and AISHELL-3 are used to construct two different scale datasets for training PSSRF, i.e., T1 and T2. Then, original testing and validation sets of AISHEEL-1 and AISHELL-3 are divided into two categories based on the visibility of contained speakers, which are used to build three testing sets. Each testing set is shuffled and divided into 33 subsets, and each subset is disguised using not only FD algorithms such as Phase Vocoder and FD-PSOLA but also TD algorithms such as TD-PSOLA and WSOLA. The disguising parameter α increases from -8 to 8 with a step of 0.5, yielding the A_1 Unseen, A_3 Seen, and A_3 Unseen. In addition, three types of commercial software with excellent disguising effects, i.e., Audacity, Audition, and iZotope, are used to simulate more challenging and practical application scenarios.

Training setup: The pitch-scaling algorithm applied in training phase is Phase Vocoder implemented by `librosa` [22], where the disguising parameter α follows a uniform distribution $\mathcal{U}(-8, 8)$. In Estimator the time dimension of the input is fixed to 500. In feature transform, the sample rate is 16kHz, fft points are 1024, the window length is 1024, and the hop length is 256. The Adam optimizer [23] with a fixed learning rate of 1×10^{-4} is applied to train PSSRF. The batch size is 32, and epochs are equal to 400. The experiments are implemented with 2$\times$ NVIDIA A100 40GB.

4.2. Evaluation of estimation accuracy

Table 2: MAE of estimated α, red is the smallest, blue is the second. (PSSRF$_{\mathrm{T1}}$ / PSSRF$_{\mathrm{T2}}$ / L. Zheng [9])

Implementation	A_1 Unseen	A_3 Seen	A_3 Unseen	Algorithm
librosa [22]	0.691 / 0.188 / 0.675	0.324 / 0.319 / 1.020	0.385 / 0.356 / 0.890	Phase Vocoder
MATLAB*	0.698 / 0.213 / 0.590	0.448 / 0.417 / 0.979	0.515 / 0.460 / 0.848	Phase Vocoder
RTISI [24]	0.771 / 0.661 / 0.684	0.961 / 0.929 / 0.939	0.982 / 0.946 / 0.952	FD-PSOLA
PRAAT	0.717 / 0.616 / 0.619	0.932 / 0.912 / 0.919	0.968 / 0.933 / 0.930	TD-PSOLA
SoundTouch [25]	0.753 / 0.543 / 0.607	1.249 / 1.161/ 0.880	1.251 / 1.162 / 0.929	WSOLA
Audacity 3.1.3	0.784 / 0.754 / 0.631	1.317 / 1.173 / 1.091	1.393 / 1.198 / 0.932	UNKNOWN
Audition 2022	0.767 / 0.402 / 0.586	1.112 / 1.076 / 0.927	1.272 / 1.202 / 0.956	UNKNOWN
iZotope RX9	0.828 / 0.476 / 0.575	1.445 / 1.311 / 0.983	1.644 / 1.465 / 0.974	UNKNOWN

* shiftPitch function.

Evaluation: We evaluate the performance of the Estimatior in PSSRF trained with T1 / T2 and that of the baseline [9] in A_1 Unseen, A_3 Seen, and A_3 Unseen with the Mean Absolute Error (MAE) between the predicted α and the ground truth, which is recoreded in TABLE 2. In addition, we investigate the relationship between the estimation deviation i.e.,$\hat{\alpha} - \alpha$, of PSSRF$_{\mathrm{T2}}$ and the disgusing parameters α applied in A_1 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, PSSRF 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 PSSRF. It is noteworthy that PSSRF 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 α is

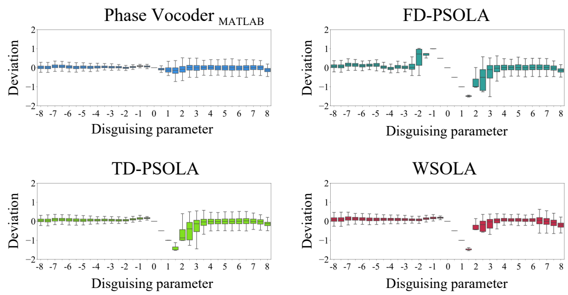


Fig. 3: Box-plot of the estimation deviation under various disguising parameters. The idea results should be always zeros.

more accurate than that of the positive, which is consistent with the conclusion in [6, 7] that lowing pitch is easier to detect than raising pitch. Besides, the tiny α is prone to be estimatied as zero, resulting a linear deviation in the neighbourhood of zero.

Discussion: There are two explanations of how Estimator works: (1) the esitimator learns a mapping from the artifacts introduced by pitch-scaling algorithms to α; (2) Estimator learns a manifold which is composed of the original speech and its pitch-scaled 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 PSSRF. 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: ASV is an effective tool to evaluate the quality of the restored pitch-scaled speech. We apply a typical model termed ECAPA-TDNN [26] to compare the restoration quality of PSSRF and that of the baseline. Specifically, we qualitatively evaluate the improvement provided by different methods for the ASV model when faced with pitch-scaled 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 PSSRF, which is shown in Fig. 5.

Results: Fig. 4 reveals that both the baseline and PSSRF can clearly enhance the performance of ASV when faced with pitch-scaled speeches, while PSSRF 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 PSSRF is specifically designed to fit a mapping from noised fbank features to noise-

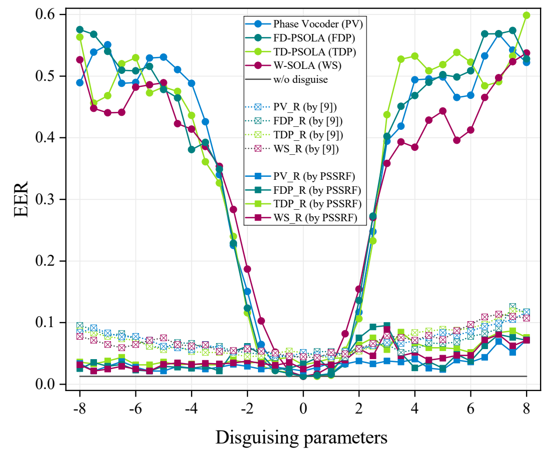
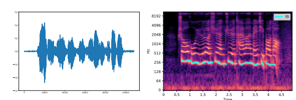
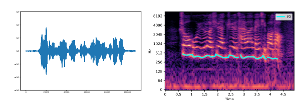


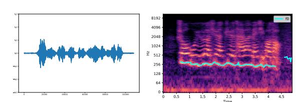
Fig. 4: ERR of the ASV model when faced with the pitch-scaled / restorated samples from different subsets of A_1 Unseen.



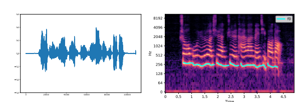
(a) Original speech



(b) Pitch-scaled speech (α=7.648)



(c) Restoration of L. Zheng [9]



(d) Restoration of PSSRF

Fig. 5: Waveforms and spectrograms of an example utterance in A_1 Unseen.

free fbank features, which is combined with a neural vocoder for high-quality restoration. This issue is further indicated in Fig. 5, where these two methods can both reconstruct the fundamental frequency of the original speech exactly, but PSSRF can reconstruct more clear formant and high-frequency information, resulting in a higher quality of the restoration.

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

5. CONCLUSION

We propose a no-reference method termed *PSSRF* to estimate the disguising parameters of pitch scaling and restore pitch-scaled speeches into original revsions, which has great significance for the security of ASV. The experiments reveal that even compared with the reference-based baseline, PSSRF still obtains competitive results in both the estimation accuracy and the restoration quality. Furthermore, as a no-reference method, PSSRF can directly make existing ASV applications more resistant to pitch scaling without additional modifications. Future work would be investigating the improvement of PSSRF 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 antispoofing 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 smaltp 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.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: 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, "Realtime 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.