INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Robust Feature Decoupling in Voice Conversion by Using Locality-Based Instance Normalization

Yewei Gu[1,2], *Xianfeng Zhao*[1,2]✉, *Xiaowei Yi*[1,2]

[1]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100085, China
{guyewei,zhaoxianfeng,yixiaowei}@iie.ac.cn

Abstract

Extensive style transfer methods have shown that instance normalization (IN) is a simple yet effective way to remove style information from original inputs. However, few studies have focused on whether these channel-wise feature statistics, such as mean and standard deviation (std), are consistent locally and globally, which ultimately leads to insufficient feature decoupling. In this paper, we first propose locality-based instance normalization (LoIN) to impose statistical feature consistency constraints on latent feature maps. LoIN performs normalization using local feature statistics which are calculated on randomly selected frames rather than on the entire set of frames used in the training phase. Since the style representation is unique and stable, the feature statistics of the latent feature submaps will tend to be consistent as the training progresses. In particular, LoIN is lightweight, less computationally intensive, and transferable to any IN-driven VC method. Experimental results show the superiority of LoIN in disentanglement and transfer performance and show improvement in both speaker similarity and content consistency.

Index Terms: voice conversion, style transfer, instance normalization, feature disentanglement.

1. Introduction

Voice conversion is a technology that aims to modify the speaker style of a speech signal while preserving its content information. From an information-theoretic perspective, speech can be divided into two independent and complementary components: speaker-dependent information (SDI) and speaker-independent information (SII). To transfer arbitrary voice styles, the common framework is to build an encoder-decoder architecture, where the speaker encoder and content encoder decouple speech into SDI and SII respectively, and then the decoder predicts converted acoustic features conditioned on source SII and target SDI.

In terms of feature disentanglement, despite common information-constrained bottlenecks in encoder-decoder models [1, 2, 3, 4] and adversarial feedback in GAN-based models [5, 6, 7, 8], the instance normalization (IN) [9] module has also been proven to be a simple yet effective technique in the content encoder for removing style information, which is interpreted as a style normalization by normalizing feature maps using the channel-wise feature statistics as mean and standard deviation (std). These statistics are generally used to represent speaker styles in some typical methods. For example, AdaINVC [10] uses 11 IN layers in the content encoder, and AgaINVC [11] transmits the frame-wise mean and std as style information to the decoder based on the Unet [12] architecture. Successive works have also been continuously expanding IN, with the launch of adaptive instance normalization (AdaIN) [13] and weight adaptive instance normalization (WAdaIN) [14]. However, these methods are still limited to obtain more significant improvements due to the lack of consideration for both local and global consistency of style.

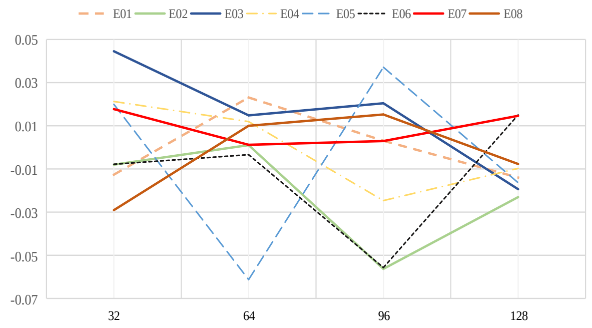


Figure 1: *The mean distributions across different feature submaps, with each line representing a specific utterance. The horizontal axis represents the number of frames included in the feature submap used for statistics. The figure shows that the channel-wise feature statistics used to represent speaker style are not consistent in submaps.*

To demonstrate the inconsistency in latent feature maps produced by AdaINVC, we conduct statistical analysis on the mean of latent feature submaps. Fig.1 depicts the average values of the channel-wise means across various feature submaps. The sharp changes in the mean distribution, such as E05, suggest that the style representations in different feature submaps are inconsistent, which hinders the stable decoupling of IN. To enhance the stability of feature decoupling, we delve deeper into the mechanism of IN, revealing that consistency constraints between local and global feature statistics in feature maps are crucial. In this paper, we explore how to enforce consistency constraints on the style representation of converted speeches, which has not been done before. We propose a new extension of IN, called locality-based instance normalization (LoIN), which does not rely on additional computational modules and only requires implementing a random frame-selection operation before IN. This modification computes channel-wise statistics on selected local frames rather than the original global statistics, enhancing feature decoupling by encouraging a more uniform distribution of styles in feature submaps. Our results show that this lightweight modification of IN significantly improves feature decoupling, leading to enhanced performance in voice conversion tasks.

10.21437/Interspeech.2023-1691

2. Method

2.1. Feature Disentanglement in IN

In this section, we investigate the decoupling mechanism of IN. Based on the pronunciation mechanism, an individual's unique, static vocal tract structure constitutes SDI, while the complex tonal movement determines the SII [15]. This suggests that for short-term stationary speech, the SDI is time-invariant static information, while the SII is time-varying [16]. We use $X \in \mathbb{R}^{D\times L}$ for latent feature maps and C for SII, where D stands for the channel depth and L for the frame number. A_{dl} denotes the feature value corresponding to the d-th channel and l-th frame. In view of IN, the time-variant SII conforms to a standard normal distribution, while the SDI acts as modulation factors μ_d and σ_d, which are channel-wise mean and std. For speech construction, it has

$$\begin{gathered} A_{dl} = C_{dl} * \sigma_d + \mu_d; 1 \le d \le D, 1 \le l \le L \\ \mu_d = \frac{1}{L}\sum_{i=1}^{L} A_{dl};\ \sigma_d = \sqrt{\frac{1}{L}\left(A_{dl} - \mu_d\right)^2 + \varepsilon} \end{gathered} \tag{1}$$

Eq.(1) reveals the working principle of AdaIN, which is commonly used in decoders to couple SDI and SII. When compared to other coupling functions such as addition and concatenation, it is a more effective way to explain the independence and complementarity of SDI and SII, which indicates that SDI should be consistent in each subsection. When rethinking the decoupling mechanism of IN in Eq.(2), we can deduce that the intrinsic correlation between μ_d, σ_d, and style representation is particularly important to obtain sufficient decoupling.

$$IN\left(A\right) = \frac{A - \mu_d\left(A\right)}{\sigma_d\left(A\right)};\ 1 \le d \le D \tag{2}$$

In fact, the impact of L on computing the μ_d and σ_d is often overlooked in IN. It can be observed that the styles in each subsection are consistent, implying that $\mu_d(A_{L1}) = \mu_d(A_{L2})$, where L_1 and L_2 correspond to the lengths of different feature submaps. However, in Fig.1, the local style representation of the converted speech lacks this consistency, indicating that existing methods lack constraints on local and global style consistency. Given that the length of the input speech varies during inference, ensuring local μ_d and σ_d consistency with the global values is essential for achieving robust VC.

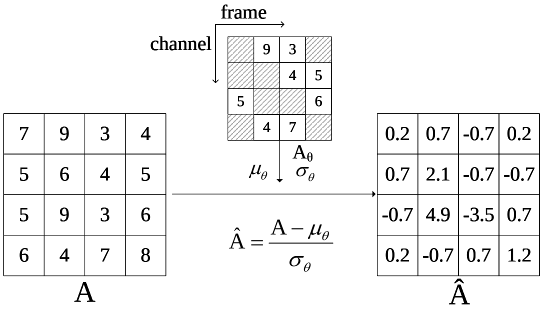
2.2. Robust Requirements for IN

In this section, we will investigate how IN affects the robustness of the model [17, 18]. Since the static information (SDI) is stably decoupled, the challenge lies in achieving sufficient decoupling of the time-varying information (SII) in the content encoder (E_c). It requires 1) the decoupling is content-independent; 2) the performance is not impacted by the variable lengths of the input speeches. To address the challenges, we divide the speech X into m segments, denoted as $X = X_1, X_2, ..., X_m$, where X_i corresponds to the time interval $X[t_{i-1} : t_i]$. C_i represents SII in X_i. The robust decoupling requires

$$\begin{aligned} E_c(X) &= E_c(\{X_1, X_2, \cdots, X_m\}) \\ &= \{E_c(X_1), E_c(X_2), \cdots, E_c(X_m)\} \end{aligned} \tag{3}$$

Considering the adaptability to variable-length segments, the maximum likelihood solution is

$$\begin{gathered} E_c(X[t_{i-1} : t_i]) = E_c(X_i) \\ C[t_{i-1} : t_i] = C_i \end{gathered} \tag{4}$$


frame
channel
μ_θ
σ_θ
$\mathrm{A_\theta}$
$\hat{\mathrm{A}} = \frac{\mathrm{A} - \mu_\theta}{\sigma_\theta}$
A
Â

7	9	3	4
5	6	4	5
5	9	3	6
6	4	7	8

▨	9	3	▨
▨	▨	4	5
5	▨	▨	6
▨	4	7	▨

0.2	0.7	-0.7	0.2
0.7	2.1	-0.7	-0.7
-0.7	4.9	-3.5	0.7
0.2	-0.7	0.7	1.2

Figure 2: *The workflow of LoIN. The statistics for each channel are calculated from randomly selected frames.*

The optimal solution suggests that the local C_i in the submaps should be equal to the corresponding intervals in C. In terms of IN, this can be inferred from Eq.(2) as follows.

$$IN(A) = \frac{A - \mu}{\sigma} = \left\{\frac{A_i - \mu}{\sigma}\right\}, i \in [1, m] \tag{5}$$

$$\{IN(A_i)\} = \left\{\frac{A_i - \mu_i}{\sigma_i}\right\}, i \in [1, m] \tag{6}$$

To ensure that the global normalization $IN(A)$ is equivalent to the set of local normalizations $IN(A_i)$, it is necessary for $\mu_i = \mu$ and $\sigma_i = \sigma$. This implies that consistency constraints, which maintain the consistency of both local and global statistical characteristics, are crucial for enhancing the robustness of IN decoupling.

2.3. Locality-Based Instance Normalization

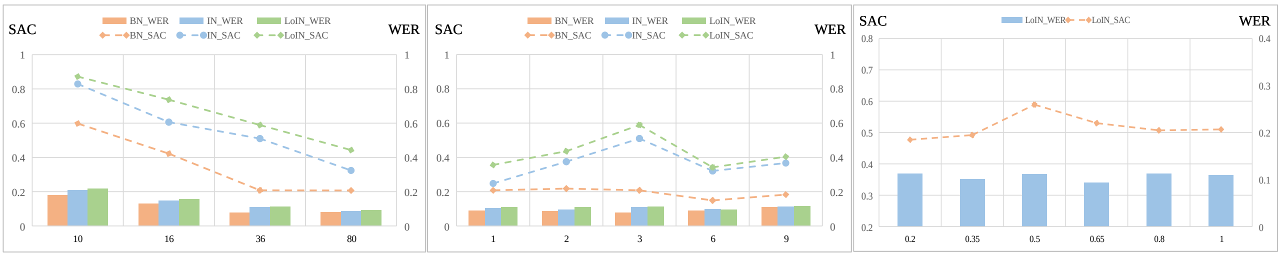
To improve local consistency of style information, we propose a simple extension to IN called LoIN. As shown in Fig.2, we randomly select frames at a fixed ratio θ for each channel (vertical axis) in the feature map $A \in \mathbb{R}^{D\times L}$. The selected frames (unmasked) form the feature submap $\hat{A} \in \mathbb{R}^{D\times L_\theta}$ ($L_\theta = L * \theta$). For example, half of the frames are randomly selected for each channel in Fig.2 when θ=0.5. Then normalization is performed on A by using the local channel-wise mean μ_θ and standard deviation σ_θ computed on $\hat{A}$. The process is recorded as Eq.(7), where $\mathcal{R}$ and $\mathcal{F}$ stand for random selection and channel-wise feature statistical methods, respectively.

$$\begin{gathered} A_\theta = \mathcal{R}(A, \theta); \\ \mu_\theta, \sigma_\theta = \mathcal{F}(A_\theta) \\ \hat{A} = \frac{A - \mu_\theta}{\sigma_\theta} \end{gathered} \tag{7}$$

LoIN is a simple yet efficient alternative to IN, which even reduces computational cost. Due to the relative stability of individual styles, the random local feature style distribution gradually converges to a stable point during the reconstruction task training. Additionally, the random frame selection strategy can also be seen as a data augmentation method that effectively combats overfitting.

3. EXPERIMENTS

To confirm the effectiveness of LoIN, we conduct two types of experiments: EA and EB. In EA, the performance of LoIN is evaluated on a self-designed model called EAM. To minimize the impact of other complex decoupling modules, EAM utilizes



(a)Evaluation on bottleneck

(b) Evaluation on cascaded number

(c) Evaluation on frame-selection ratio

Figure 3: *Performance comparison of EAM using LoIN and IN. The histogram represents WER(right), and the line chart represents SAC(left).*

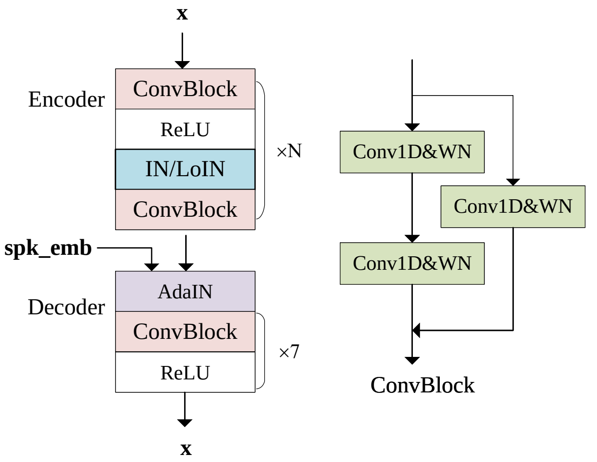


Figure 4: *The architecture of EAM. spk_emb is produced by pretrained speaker verification system Dvector. N is cascaded number. WN stands for weight normalization.*

a simple architecture consisting of only a few fundamental units shown in Fig.4. In EB, we replaced existing IN-related methods with LoIN to evaluate its impact on VC performance.

3.1. Experiment Conditions

Implementation Details For EA, silent segments in each utterance are trimmed, and utterances are randomly divided into segments of $L = 128$. The 80-bin mel-spectrograms are used as inputs, which are extracted like HiFiGAN [19], and then normalized to 0˜1. The model is trained on VCTK [20] with 20 speakers. The batch size is set to 4. AdamW ($\beta_1 = 0.8$, $\beta_2 = 0.99$, weight decay $\lambda = 0.00015$) and CosineAnnealingLR (learning rate decay $lr_d = 0.995$, initial learning rate $lr = 0.0001$) are used for optimization. Only L1 loss is required in EA. For EB, all models in EB are trained on the same dataset and strictly followed the default configuration provided in the open source code.

3.2. Metrics

Objective Metrics To evaluate the style similarity, Speaker Similarity Accuracy (SAC) is computed by the speaker verification system Dvector[1] [21] to determine the ratio at which the target and converted speech belong to the same speaker. To evaluate the content consistency, Word Error Rate (WER) between the source and converted speech is measured by an automatic speech recognition system wav2vec2.0[2] [22]. We use VCTK, VCC2020 [23], and LibriSpeech [24] datasets for zero-shot test in EB. Each dataset provides 1000 source-target pairs (only 280 for VCC2020). Only VCTK is used for EA.

Subjective Metrics Mean Opinion Score (MOS) [25] is used to assess naturalness and similarity. The converted samples are evaluated by 12 raters who are asked to assign a score of 1˜5. Statistical results are reported along with 95% confidence intervals (CI) to ensure the accuracy of the findings. For each dataset, the converted speeches are divided into 4 groups: F2F, F2M, M2F, and M2M (F for female and M for male). Our audio samples are available on the demo page[3].

4. Results and Discussion

4.1. Evaluation on EA

The experiment aims to investigate the impact of LoIN on model performance across three distinct conditions. The EAM series are denoted as EAM_bot_N_θ, where *bot* refers to the bottleneck dimension; N corresponds to the cascaded number in Fig.4; θ means frame-selection ratio.

Evaluation on Bottleneck. Bottlenecks are crucial structures for VC [27]. Experiments are conducted on EAM_*bot*_3_0.5, where *bot* is set to 10, 16, 36, and 80. As Fig.3(a) shows, IN and LoIN are conducive to style decoupling compared with Batch Normalization (BN) [28]. In particular, when $bot \geq 16$, LoIN outperforms IN by about 10% improvements in SACs, which directly benefits from the more stable local consistency. However, when $bot = 10$, the superiority of LoIN has decreased. This can be interpreted as narrow bottlenecks tending to maintain the consistency between submaps and the whole one while the gains from LoIN are limited. This further reflects the importance of local representation consistency to robustness from the side. Moreover, WERs remain consistently low, indicating that compression does not damage content information, and most likely, source style information is also retained.

Evaluation on Cascaded Number. Cascaded INs are commonly used to provide normalization instead of BN. The experiments are conducted on EAM_36_N_0.5 using BN, IN, and LoIN, where N corresponds to the cascaded number in Fig.4. As Fig.3(b) shows, as the N increases, the SACs of IN and LoIN steadily improve in the early stages before decreasing substantially. Compared to IN, LoIN continuously maintains ap-

[1] https://github.com/yistLin/dvector

[2] https://huggingface.co/docs/transformers/model_doc/wav2vec2

[3] https://brightgu.github.io/LoINVC/

Table 1: *Comparison of results by applying IN and LoIN in different methods. SIM and NAT respectively represent the subjective MOS value for evaluating the similarity and naturalness of speech.*

Methods	VCTK [20]				VCC2020 [23]				LibriSpeech [24]			
	SAC	WER	SIM	NAT	SAC	WER	SIM	NAT	SAC	WER	SIM	NAT
AdaINVC [10]	0.957	0.594	3.45±0.02	3.23±0.02	0.964	0.506	3.42±0.02	2.89±0.02	0.927	0.620	3.18±0.02	2.69±0.02
AdaINVC w LoIN	0.961	0.528	**3.47±0.04**	3.37±0.03	**0.969**	0.489	3.43±0.04	3.02±0.04	0.942	0.542	**3.30±0.02**	2.89±0.02
AgaINVC [11]	0.860	0.532	3.22±0.05	3.21±0.02	0.857	0.545	3.27±0.02	2.81±0.02	0.837	0.573	2.87±0.03	2.72±0.03
AgaINVC w LoIN	0.928	**0.478**	3.41±0.02	3.41±0.02	0.902	0.526	3.39±0.04	2.95±0.02	0.891	0.512	3.10±0.01	3.01±0.04
MediumVC [26]	0.952	0.502	3.36±0.02	3.49±0.04	0.952	0.513	3.35±0.01	3.31±0.04	0.921	0.491	3.09±0.04	3.09±0.02
MediumVC w LoIN	**0.972**	0.482	3.43±0.02	**3.51±0.02**	0.968	**0.473**	**3.47±0.04**	**3.36±0.04**	**0.957**	**0.422**	3.28±0.02	**3.18±0.04**

proximately an 8% improvement in SACs. The decline in SAC can be attributed to overfitting, as models with more parameters are prone to overfitting when the structures are similar. Similarly, the BN line also experiences a decline. It can be inferred that as the training continues, EAM_36_9_0.5 will eventually fall behind EAM_36_6_0.5. Additionally, even with overfitting, LoIN maintains a slight advantage because the random strategy is also a data augmentation method against overfitting.

Evaluation on Frame-selection Ratio. Experiments are conducted on the EAM_36_3_θ using LoIN. Fig.3(c) shows that setting $\theta = 0.5$ leads to significant improvements in SACs compared to IN ($\theta = 1$). However, when $\theta = 0.2$, the performance of SACs slightly lags behind that of IN, suggesting that as θ decreases, the fluctuation of local feature statistics increases, making it more difficult to achieve local consistency. Furthermore, the consistency of the content remains consistently high, likely due to the relatively wide bottleneck ($bot = 36$). Overall, setting $\theta = 0.5$ provides both the flexibility of style representation and feasibility of training.

4.2. Evaluation on EB

To further confirm the validity of LoIN, comparative experiments are performed on three IN-driven VC methods: AdaINVC [10], a VAE-based model that employs dense cascaded INs in content encoders; AgaINVC [11], a Unet-based model [12] where channel-wise feature statistics are directly used as speaker styles; MediumVC [26], an autoencoder model that employs predefined speaker style representations and 2 INs in the content encoder. As shown in Table 1, in terms of SACs, AdaINVC fails to obtain the significant boost. We observe that the number of cascaded INs in AdaINVC had reached 11 (without overfitting), leaving little room for further improvement in similarity. On the other hand, AgaINVC benefited more from LoIN due to its unique Unet structure, where feature statistics computed in LoIN are directly used as style representations in a feed-forward manner. It shows the more accurate the style representation is, the more significant gains on performance.

A point worth considering is the steady improvement in WERs by LoIN, which has not been observed in EAM. For control experiments on EAM, it requires lightweight structure to provide more space for performance improvement. Therefore, the feature decoupling of EAM is not sufficient, as content embeddings often contain residual source speaker style information due to the relatively wide bottleneck in the content encoder (bot=36). This makes the local consistency brought by LoIN more conducive to the elimination of style information. However, for the present three methods, to obtain higher speaker similarity, content embedding tends to suffer from over-compression (bot=4 in AgaINVC). Therefore, the more accurate style representation computed in LoIN does not continue to improve style significantly but rather better maintains content information. In all, the differences depend on the inherent decoupling mechanism of IN.

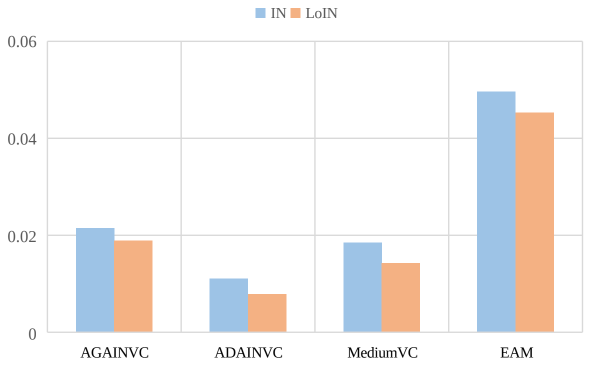


Figure 5: *The standard deviation of the frame-wise means in latent feature maps with extensive test samples.*

4.3. Quantification on Local Consistency

Fig.5 provides a visual representation of the quantification of local consistency in four different methods. The test samples used in this analysis are taken from VCC2020, consisting of 4 females and 4 males, with each speaker providing 10 utterances. For each method, feature maps $A \in \mathbb{R}^{128\times128}$ $(D \times L)$ are used, where each map corresponds to a segment with a length of $L = 128$. To represent the local statistics, the means of $A[:, 0 : X]$ are computed, with X taking on the values of 16, 32, ..., and 128, respectively. The standard deviation of these means is then plotted in Fig.5 to reflect the stability of local consistency. Smaller values indicate better consistency.

5. Conclusion

We conduct an investigation into the shortcomings of IN when it comes to decoupling speech features. Our findings suggest that the inconsistency between local and global channel feature statistics can lead to inadequate decoupling. To address it, we propose LoIN, which incorporates consistency constraints by utilizing randomly selected local feature statistics to normalize feature maps during training, rather than relying on global feature maps like IN. The experiments show that LoIN is a straightforward yet powerful module that can achieve robust content-style tradeoffs. We anticipate that future research will concentrate on exploring the consistency of speech styles further.

6. Acknowledgements

This work was supported by National Key Technology Research and Development Program under 2020AAA0140000.

7. References

[1] 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 *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[2] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5939–5943.

[3] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Acvae-vc: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *arXiv preprint arXiv:1808.05092*, 2018.

[4] Y. Gu, X. Zhao, X. Yi, and J. Xiao, "Voice conversion using learnable similarity-guided masked autoencoder," in *Digital Forensics and Watermarking: 21st International Workshop, IWDW 2022, Guilin, China, November 18-19, 2022, Revised Selected Papers*. Springer, 2023, pp. 53–67.

[5] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," *arXiv preprint arXiv:2010.11672*, 2020.

[6] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[7] S.-H. Lee, J.-H. Kim, H. Chung, and S.-W. Lee, "Voicemixer: Adversarial voice style mixup," *Advances in Neural Information Processing Systems*, vol. 34, pp. 294–308, 2021.

[8] Y. A. Li, A. Zare, and N. Mesgarani, "Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," *arXiv preprint arXiv:2107.10394*, 2021.

[9] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[10] J.-c. Chou, C.-c. Yeh, and H.-y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," *arXiv preprint arXiv:1904.05742*, 2019.

[11] Y.-H. Chen, D.-Y. Wu, T.-H. Wu, and H.-y. Lee, "Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5954–5958.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[14] M. Chen, Y. Shi, and T. Hain, "Towards low-resource stargan voice conversion using weight adaptive instance normalization," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5949–5953.

[15] T. Dutoit, *An introduction to text-to-speech synthesis*. Springer Science & Business Media, 1997, vol. 3.

[16] 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.

[17] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 6, pp. 1411–1418, 2014.

[18] J. Lian, C. Zhang, and D. Yu, "Robust disentangled variational speech representation learning for zero-shot voice conversion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6572–6576.

[19] 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.

[20] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, *Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit*. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 1997.

[21] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," *arXiv preprint arXiv:2008.12527*, 2020.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[25] I. Rec, "P. 800.1, mean opinion score (mos) terminology," *International Telecommunication Union, Geneva*, 2006.

[26] Y. Gu, Z. Zhang, X. Yi, and X. Zhao, "Mediumvc: Any-to-any voice conversion using synthetic specific-speaker speeches as intermedium features," *arXiv preprint arXiv:2110.02500*, 2021.

[27] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5210–5219.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.