Interspeech 2024
1-5 September 2024, Kos, Greece

Utilizing Adaptive Global Response Normalization and Cluster-Based Pseudo Labels for Zero-Shot Voice Conversion

Ji Sub Um, Hoirin Kim

School of Electrical Engineering, KAIST, Daejeon, Republic of Korea
twiz0311@kaist.ac.kr, hoirkim@kaist.ac.kr

Abstract

Recently, there has been an increase in research on zero-shot voice conversion. Many conventional studies use dynamic layers to conduct conversion for unseen speakers. Our aim is to extend dynamic methods to transmit content information as well. To achieve this, we propose AGRN-VC, which utilizes ConvNeXt V2 modules with adaptive global response normalization (AGRN) layers to convey content information. When conveying this information, it is crucial to ensure that the source speaker's information is not transmitted. So we adopt auxiliary learning with cluster-based pseudo labels. It helps the content encoder to focus on content information while excluding speaker information by performing a pseudo label classification task using its output. We conduct comparative experiments between various baseline models and the proposed model using subjective and objective metrics. Our proposed approach achieves better converted speech quality in terms of speaker similarity and naturalness.

Index Terms: zero-shot voice conversion, adaptive normalization layer, cluster-based pseudo label, auxiliary learning

1. Introduction

Customized voice generation has experienced a significant increase in interest due to its numerous applications in real-world scenarios such as singing voice generation, movie dubbing, and customer service. However, training a new model for each speaker is both time-consuming and expensive. Zero-shot voice conversion (VC) has been studied as a solution to this problem, enabling to generate customized voices. Zero-shot VC system converts the voice of a source speaker to that of an unseen target speaker using only a limited number of reference samples.

In recent years, many researchers have studied on zero-shot VC to improve the performance. These studies can be broadly classified into two main approaches based on their improvement strategies. The first approach focuses on effectively disentangling linguistic and speaker-related information. For instance, AutoVC [1, 2] introduces a bottleneck network to isolate linguistic information. VQMIVC [3] attempts disentanglement by using mutual information constraints [4]. It also utilizes vector quantization [5] and contrastive predictive coding loss to remove redundant information and focus on time-variant information. Some studies inject perturbation [6, 7, 8] into the input and conduct data augmentation [9] to extract robust linguistic information. The second approach involves research on adaptive layers that propagate disentangled information to the decoder to make speech sound more like the target speaker. For example, AdaIN-VC [10] and AGAIN-VC [11] introduce adaptive instance normalization (AdaIN) [12]. These adaptive layers remove global statistics from the normalization process. Affine transformations are performed in the de-normalization process using parameters obtained from speaker information. It improves the model's ability to adapt to unseen speakers. ACNN-VC [13] utilizes adaptive convolution neural networks [14] to adapt to speaker information while considering local structure on the time axis. TriAAN-VC [15] uses attention-based adaptive normalization [16] that carries out feature statistics transformation by generating attention-weighted mean and variance.

In particular, TriAAN-VC succeeds in generating high-quality converted samples in terms of naturalness and speaker similarity. However, there is still room for improvement. We aim to enhance naturalness by conveying linguistic information to the decoder through adaptive layers, in addition to speaker information. To achieve this, we start with TriAAN-VC as the backbone model and introduce the ConvNeXt V2 [17] module which includes the newly proposed adaptive global response normalization (AGRN). AGRN assists the model in generating speech with higher clarity by conveying linguistic information through channel selectivity process. Furthermore, during the adaptation of linguistic information, there may be issues with speaker adaptation due to interference from the source speaker information. To address this, it is necessary to extract robust linguistic information. Therefore, we adopt an auxiliary learning scheme with cluster-based pseudo labels and it imposes a constraint on the content encoder to perform disentanglement.

In this paper, we propose AGRN-VC, a new system that achieves better disentanglement in the encoder and adapts well to both speaker and content information in the decoder. We conduct comparative experiments between various baseline models and our proposed model using subjective and objective metrics. The results show that our proposed approach improves converted speech in terms of speaker similarity and naturalness. Additionally, we conduct an ablation study to analyze the effects of our proposed methods.

2. Proposed methods

In this section, we introduce details about the architecture and our proposed methods: ConvNeXt block and auxiliary learning strategy. The basic architecture of our proposed model is similar to [15] and an overview of the process is shown in Figure 1.

2.1. Architecture

Feature Extraction We utilize contrastive predictive coding (CPC) features [18] as inputs of the content and speaker encoders. CPC features, $x_{cpc} \in \mathbb{R}^{D \times T}$, are extracted from a pre-trained model [19]. D and T are channel size and time length, respectively. Also, we extract pitch information, log-normalized fundamental frequency (F_0), $x_{f0} \in \mathbb{R}^{T}$, to capture changes in intonation and use the cluster-based pseudo labels

10.21437/Interspeech.2024-539

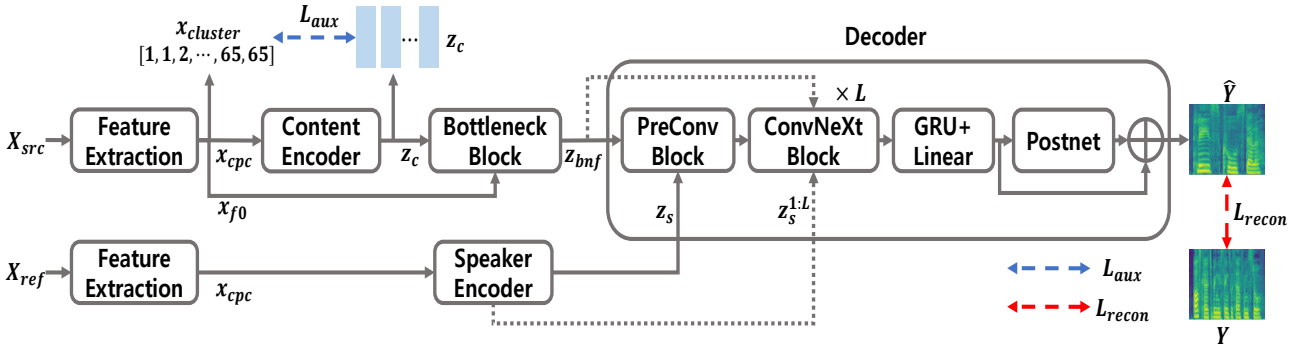

Figure 1: *Overview of our proposed AGRN-VC system.*

to train the content encoder. The labels, $x_{\text{cluster}} \in \mathbb{R}^N$, are extracted from a pre-trained model that applies k-means clustering to the output of the seventh transformer layer of HuBERT-base. The selection of the layer is based on prior studies on the phonetic relations across layers [20]. N denotes the label length. The auxiliary learning with $x_{cluster}$ is explained in Section 2.3.

Encoder The content and speaker encoders receive CPC features x_{cpc} to output the content features, z_c, and the speaker features, z_s, respectively. The content encoder comprises a total of L convolution blocks and instance normalization (IN) layers. Each Convolution block consists of two convolution layers with residual sum. The content features, z_c, are concatenated with the pitch information, x_{f0}, and passed through the bottleneck block, which includes two gated recurrent unit (GRU) layers and a linear layer, to extract the bottleneck feature, z_{bnf}. The speaker encoder comprises a total of L convolution blocks, attention blocks and IN layers. The mid-level output of each attention module, denoted as z_s^l, is then used in the decoder's attentive layer normalization (AttLN). l means the l^{th} output of the attention module and $l = 1, 2, \cdots, L$.

Decoder The decoder comprises PreConv block, a total of L ConvNeXt blocks, GRU layers and a linear layer. The Postnet is added to refine the predicted mel-spectrogram. PreConv block consists of AttLN and a convolution layer, where z_s is used as a condition for AttLN. ConvNeXt blocks utilize z_s^l as conditional inputs and further details are explained in Section 2.2.

2.2. ConvNeXt block

The ConvNeXt block is adopted in the decoder to convey both speaker and content information. This block consists of two ConvNext V2 modules. Figure 2 shows that this module comprises three convolution layers and two normalization layers: AttLN and AGRN. AttLN is used to propagate speaker information, while AGRN propagates content information.

AttLN AttLN is a case where adaptive attention normalization is applied to a layer normalization (LN). We use the output of the previous layer, f, as the query and the features z_s^l as the key, and the value. Attention operation is then performed using these values to obtain mean, μ and variance, σ^2. The details are explained by the following equations.

$$\begin{aligned} &Q = \text{TIN}(f)W_q, K = \text{TIN}(z_s^l)W_k, V = z_s^l W_v \\ &A = \text{softmax}(QK^T/\sqrt{d}) \\ &\mu = AV, \sigma^2 = A(V * V) - \mu * \mu, \end{aligned} \tag{1}$$

where, $\text{TIN}(\cdot)$ denotes time-wise instance normalization [15], $W_{q,k,v}$ represent learnable parameters, and d denotes the channel size of the query. The obtained mean and variance values are averaged along the time axis, and then used in the denormalization process as shown in a following equation.

$$\tilde{f} = \sigma * \text{LN}(f) + \mu \tag{2}$$

By utilizing AttLN, the model can adapt to speaker information and mimic the characteristics of the target speaker.

AGRN AGRN is a modified version of GRN and designed to effectively transfer linguistic information. There are four steps in the process of adapting to the information. Firstly, it aggregates global features, $g(h) = \{||h_1||, ||h_2||, \cdots, ||h_C||\}$, for each channel using L2-norm when the input is h and C denotes the channel size. Then, it measures the relative importance, $N(g(h)_i)$, by comparing it with all other channels. Subsequently, the initial input responses are adjusted based on the calculated feature normalization scores. In the end, additional adjustments are made through affine parameters that vary on the bottleneck features, z_{bnf}. These steps are represented by the following equations.

$$\begin{aligned} &\tilde{h}_i = \gamma_i * h_i * N(g(h)_i) + \omega_i + h_i \\ &N(g(h)_i) = \frac{||h_i||}{\sum_{j=1,\cdots,C} ||h_j||}, \end{aligned} \tag{3}$$

where $||h_i||$ is the L2-norm of the i-th channel. Also, γ_i and ω_i means the scaling and shift factor, respectively. By calculating and applying the relative importance along the channel axis, channel competition and selectivity are achieved, and at the same time, affine transformation is conducted according to content information. It enables channel selectivity to occur based on content information. Through this process, the model can adapt to content information and generate more intelligible speech.

2.3. Training strategy with auxiliary loss

To facilitate adaptation to linguistic information without loss of adaptability to the target speaker, it is crucial to eliminate any distractions caused by the source speaker information. Therefore, we need to conduct better disentanglement. To do this, we utilize auxiliary learning with cluster-based pseudo labels.

Some VC studies [20, 21, 22] have started to use features from speech self-supervised learning (SSL) models such as HuBERT [23] or WavLM [24]. In particular, [20] applies k-means clustering to the features to eliminate speaker-related information and uses the resulting discrete features, the cluster index sequences, as content information. However, this process also

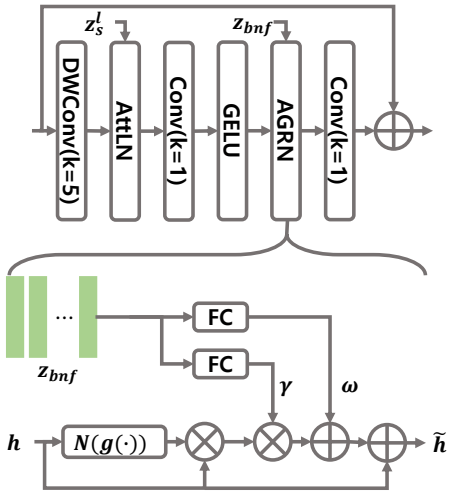

Figure 2: *Architecture of the ConvNeXt V2 module and AGRN. FC denotes a fully-connected layer. DWConv denotes a depth-wise convolution layer, and k means a kernel size.*

loses some important phoneme information. To address this, [21] refines the speech SSL model to obtain soft features by utilizing the discrete features as pseudo labels. Inspired by these works, we adopt auxiliary learning to classify cluster-based pseudo labels for reducing unnecessary information.

Unlike previous methods that fine-tune speech SSL models or employ them to derive prior distribution parameters [22], we employ auxiliary learning as a constraint to train the content encoder to focus on linguistic information, and this approach carries out a classification over pseudo labels by utilizing content features, z_c. Thus, we train the model to minimize cross entropy loss between the distribution over cluster indices and target pseudo labels, α. The equation of the auxiliary loss is as follows:

$$\begin{aligned} L_{aux} &= -\sum_{t=1}^{T} \log P(\alpha_t | z_{c,t}) \\ P(\alpha_t = i | z_{c,t}) &= \frac{\exp(e_i^T W z_{c,t}/\tau)}{\sum_{k=1}^{K} \exp(e_k^T W z_{c,t}/\tau)}, \end{aligned} \tag{4}$$

where i is the cluster index, W is a linear transformation and K is the number of clustering. e_i means the learnable embedding corresponding to the i^{th} cluster, and $z_{c,t}$ refers to the content feature of the t^{th} time frame. Also τ is a temperature and we set as a square root on the channel dimension of the embedding.

We combine this loss with the reconstruction loss to train our proposed model. During the training stage, the same speech sample is used as inputs of the content and speaker encoder, denoted as X_{src} and X_{ref} in Figure 1. From this utterance, we extract speaker, content, and pitch vectors, and the decoder predicts the mel-spectrograms, $\hat{Y}$ using these features. When we represent Y as the ground truth mel-spectrograms, the L1 distance based reconstruction loss is as follows:

$$L_{rec} = ||\hat{Y} - Y||_1 \tag{5}$$

Therefore, the total loss is as follows:

$$L_{total} = L_{rec} + \lambda L_{aux}, \tag{6}$$

where λ is a constant weight for auxiliary learning loss.

2.4. Inference

After training, we utilize an utterance from the target speaker to extract speaker features. Simultaneously, pitch and content vectors are derived from an utterance of the source speaker. These extracted features are then fed into the decoder to generate the converted mel-spectrograms. Subsequently, Parallel WaveGAN vocoder [25] is employed to generate raw waveforms from these mel-spectrograms. The generated samples is available at https://twiz0311.github.io/AGRN-VC/

3. Experiments

3.1. Dataset

We utilize the VCTK dataset [26] to train our model. The dataset consists of approximately 400 utterances per speaker, for a total of 109 speakers. Of these, 89 speakers are used in the training process while the remaining 20 speakers are used in the inference step as unseen scenarios. Among the 89 speakers, the utterances are divided into 80%, 10%, and 10% to serve as train, valid, and test datasets, respectively. The raw waveforms are down-sampled to 16 kHz and 80-dimensional mel-spectrograms are extracted with stride size of 160 and window size of 400. The cluster labels are extracted with stride size of 320 by a pre-trained model provided in open repository.[1] The number of clusters is 100. Finally, F0 is obtained through World Vocoder [27] with stride size of 160.

3.2. Training

We train the model using the Adam optimizer [28] and a learning rate with 1e-4. We train the model for a total of 300 epochs with a batch size of 128. We set the weight of auxiliary loss to 0.01. We set the content and speaker encoders to be the same as the TriAAN-VC model [15], and the decoder uses the same parameters except for the ConvNeXt block. Each convolution layer in the ConvNeXt block use a channel size of 512, 2048, 512, respectively. The number of blocks, L, is 6. The dimension of the learnable cluster embedding is 256.

3.3. Evaluation metric

We conduct subjective and objective measures to show the effectiveness of our proposed model. These metrics are measured with three baseline models for comparison: AGAIN-VC[2], VQMIVC[3], and TriAAN-VC[4]. These baseline models are trained using publicly available official codes.

As a subjective measure, we utilize mean opinion score (MOS), which is a metric that assesses speaker similarity or naturalness based on evaluation by native listeners. Naturalness MOS (NMOS) and Speaker similarity MOS (SMOS) are scored on a 9-point scale ranging from 1 to 5. A score of 1 indicates poor sound quality or a speech that sounds different to the target speaker. A score closer to 5 indicates a speech that is more natural and similar to the human voice, or a speech that sounds more like the target speaker.

We use character error rate (CER), word error rate (WER), speaker verification accuracy (SV), and speaker embedding cosine similarity (SECS) as objective measures. CER and WER are measured using a pre-trained wav2vec 2.0 [29] based au-

[1] https://github.com/bshall/hubert
[2] https://github.com/KimythAnly/AGAIN-VC
[3] https://github.com/Wendison/VQMIVC
[4] https://github.com/winddori2002/TriAAN-VC

Table 1: *Objective scores of converted samples in the seen & the unseen scenarios: CER, WER, SV and SECS. GT denotes the results of ground-truth. GT (mel+vocoder) refers the results of samples converted to mel-spectrograms and reconstructed by the vocoder.*

Methods	Seen-to-Seen				Unseen-to-Unseen			
	CER (%, ↓)	WER (%, ↓)	SV (%, ↑)	SECS (↑)	CER (%, ↓)	WER (%, ↓)	SV (%, ↑)	SECS (↑)
GT	1.64	4.95	-	-	1.84	5.21	-	-
GT (mel+vocoder)	2.10	5.87	-	-	2.32	6.24	-	-
AGAIN-VC	29.99	50.31	30.43	0.652	25.93	44.03	39.57	0.670
VQMIVC	12.31	24.42	89.43	0.775	12.83	24.99	39.50	0.668
TriAAN-VC	13.19	24.80	93.97	0.788	12.00	23.16	91.77	0.777
AGRN-VC	**6.94**	**14.79**	**95.13**	**0.795**	**6.97**	**14.51**	**93.63**	**0.786**

Table 2: *Subjective scores in the unseen scenarios: NMOS and SMOS. ± denotes 95% confidence intervals.*

	NMOS	SMOS
GT	4.45 ± 0.07	4.39 ± 0.07
VQMIVC	$\mathbf{3.67 \pm 0.09}$	3.05 ± 0.11
TriAAN-VC	3.47 ± 0.10	3.63 ± 0.09
AGRN-VC	$\mathbf{3.67 \pm 0.09}$	$\mathbf{3.78 \pm 0.09}$

Table 3: *Objective scores of ablation study in the unseen scenarios: CER, WER, SV, and SECS.*

	CER (%)	WER (%)	SV (%)	SECS
AGRN-VC	6.97	14.51	93.63	0.786
-L_{aux}	7.02	14.87	86.43	0.760
-AGRN	8.16	16.57	94.20	0.788
-L_{aux} -AGRN	7.72	16.06	90.50	0.770

tomatic speech recognizer (ASR) model[5]. These metrics show how accurately the converted speech is pronounced. SV and SECS are measured by utilizing a pre-trained speaker verification model provided by the resemblyzer library[6] [30]. They indicate that the higher the score, the closer the generated speech is to the target speaker.

4. Results

4.1. Objective and subjective evaluation

In this section, we measure objective and subjective metrics for comparing our proposed model with the baseline models. For objective evaluation, we conduct experiments for two scenarios: the seen scenario and the unseen scenario. In each scenario, we produce 3000 converted speech samples to measure the metrics. Also, we evaluate subjective metrics: NMOS and SMOS. We conduct four types of voice conversion in unseen domains: female-to-female, female-to-male, male-to-female, and male-to-male. We randomly choose ten samples for each case, resulting in a total of 40 samples per system. Each speech sample is assessed by ten native listeners, and we recruit participants through Amazon mechanical turk as done in [31].

As shown in Table 1, our proposed model shows lower CER and WER in the seen and unseen scenario. It means converted speech pronounce characters and words more accurately than the baseline models, resulting in a more clear and intelligible sound. Also, SV and SECS improves in both cases. It indicates that converted samples produced by our model have a higher similarity to the target speaker's voice. In addition, as shown in Table 2, our proposed model has higher or similar NMOS and higher SMOS scores than the baseline models. We achieve similar NMOS to VQMIVC, but SMOS significantly improves. Compared to TriAAN-VC, we can observe an improvement of 0.2 in SMOS and 0.15 in NMOS. Based on these results, we confirm that performing disentanglement through auxiliary learning and propagating content information to the model via AGRN improves naturalness and speaker similarity. Therefore, our proposed model can produce more natural and intelligible speech and better mimics characteristics of the target speaker.

[5]https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

[6]https://github.com/resemble-ai/Resemblyzer

4.2. Ablation study

In this section, we conduct an objective evaluation for the ablation study. We generate 3000 converted samples to measure metrics in the unseen scenario. -AGRN means that we use only vanilla GRNs. -L_{aux} denotes that we train the model using only a reconstruction loss without the auxiliary loss.

As shown in Table 3, when the auxiliary loss is not used, SV and SECS significantly decrease. These results demonstrate that training with auxiliary learning helps the content model to output robust linguistic features with reducing source speaker-related information. When AGRN is not used, both WER and CER increase significantly while SECS and SV increase slightly. This indicates the importance of AGRN in improving the intelligibility of the speech. Additionally, we observe a trade-off where the model slightly sacrifices speaker similarity when adapting to linguistic information via AGRN. When examining the results obtained without utilizing both the auxiliary loss and AGRN, it becomes apparent that our proposed model, which combines both methods, can mitigate the sacrifice of speaker similarity caused by the use of AGRN. Consequently, our proposed model leads to improved quality in terms of speaker similarity and naturalness.

5. Conclusions

In this paper, we propose a new system that introduces AGRN and utilizes auxiliary learning with cluster-based pseudo labels. The role of an auxiliary learning is to extract more robust content features while reducing source speaker information. This information is effectively conveyed throughout the model via AGRN, resulting in a more accurate pronunciation sound. The effectiveness of the methods is demonstrated through various metrics and ablation studies. Our proposed model generates higher quality speech, but it relies on prosody such as pitch variation and rhythm from the source speaker. As prosody also play a crucial role in customizing various voices, we aim to develop systems capable of controlling prosody in the future work.

6. Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C1014044).

7. References

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

[2] K. Qian, Z. Jin, M. Hasegawa-Johnson, and G. J. Mysore, "F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6284–6288.

[3] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Proc. of Interspeech*, 2021, pp. 1344–1348.

[4] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *Proc. of International Conference on Machine Learning (ICML)*, 2020, pp. 1779–1788.

[5] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," in *Proc. of Neural Information Processing Systems (NeurIPS)*, 2017.

[6] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, "Neural analysis and synthesis: Reconstructing speech from self-supervised representations," *Proc. of Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 16 251–16 265, 2021.

[7] H.-S. Choi, J. Yang, J. Lee, and H. Kim, "Nansy++: Unified voice synthesis with neural analysis and synthesis," in *Proc. of International Conference on Learning Representations (ICLR)*, 2022.

[8] S. H. Lee, H. Y. Choi, H. S. Oh, and S. W. Lee, "Hiervst: Hierarchical adaptive zero-shot voice style transfer," in *Proc. of Interspeech*, 2023, pp. 4439–4443.

[9] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[10] J.-c. Chou and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *Proc. of Interspeech*, 2019, pp. 664–668.

[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 *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5954–5958.

[12] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1501–1510.

[13] J. Um, Y. Choi, and H.-R. Kim, "Acnn-vc: Utilizing adaptive convolution neural network for one-shot voice conversion," in *Proc. of Interspeech*, 2022, pp. 2998–3002.

[14] P. Chandran, G. Zoss, P. Gotardo, M. Gross, and D. Bradley, "Adaptive convolutions for structure-aware style transfer," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7972–7981.

[15] H. J. Park, S. W. Yang, J. S. Kim, W. Shin, and S. W. Han, "Triaan-vc: Triple adaptive attention normalization for any-to-any voice conversion," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, "Adaattn: Revisit attention mechanism in arbitrary neural style transfer," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6649–6658.

[17] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 133–16 142.

[18] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[19] M. Riviere, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7414–7418.

[20] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Proc. of Interspeech*, 2021.

[21] B. van Niekerk, M.-A. Carbonneau, J. Zaïdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6562–6566.

[22] J. Lian, C. Zhang, G. K. Anumanchipalli, and D. Yu, "Towards improved zero-shot voice conversion with conditional dsvae," in *Proc. of Interspeech*, 2022.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[24] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[25] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[26] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[27] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

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

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

[31] Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin *et al.*, "Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias," *arXiv preprint arXiv:2306.03509*, 2023.