2022 5th International Conference on Artificial Intelligence for Industries (AI4I) | 978-1-6654-5961-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/AI4I54798.2022.00022

Enhancing Zero-Shot Many to Many Voice Conversion via Self-Attention VAE with Structurally Regularized Layers

Ziang Long
Department of Mathematics
University of California
Irvine, CA, USA
zlong6@uci.edu

Yunling Zheng
Department of Mathematics
University of California
Irvine, CA, USA
yunliz1@uci.edu

Meng Yu
Tencent AI Lab
Tencent at Bellevue
Seattle, WA, USA
raymondmyu@tencent.com

Jack Xin
Department of Mathematics
University of California
Irvine, CA, USA
jack.xin@uci.edu

***Abstract*—Variational auto-encoder (VAE) is an effective neural network architecture to disentangle a speech utterance into speaker identity and linguistic content latent embeddings, then generate an utterance for a target speaker from that of a source speaker. This is possible by concatenating the identity embedding of the target speaker and the content embedding of the source speaker uttering a desired sentence. In this work, we propose to improve VAE models with self-attention and structural regularization (RGSM). Specifically, we found a suitable location of VAE's decoder to add a self-attention layer for incorporating non-local information in generating a converted utterance and hiding the source speaker's identity. We applied relaxed group-wise splitting method (RGSM) to regularize network weights and remarkably enhance generalization performance.**

In experiments of zero-shot many-to-many voice conversion task on VCTK data set, with the self-attention layer and relaxed group-wise splitting method, our model achieves a gain of speaker classification accuracy on unseen speakers by 28.3% while slightly improved conversion voice quality in terms of MOSNet scores. Our encouraging findings point to future research on integrating more variety of attention structures in VAE framework while controlling model size and overfitting for advancing zero-shot many-to-many voice conversions. [1]

I. Introduction

VC(voice conversion) is to convert speech of a source speaker to that of a target speaker while preserving its linguistic content. Recent works [1], [2] are able to do high quality End-to-End VC with parallel data, i.e. speech pairs of two speakers pronounce the same sentences. Yet, parallel training has the following obstacles: i) parallel data are expensive to collect, and utterance alignment takes even more effort, ii) VC to a different target requires re-training, iii) unable to do zero shot VC, i.e. conversion from/to the voice of an unseen speaker with only a few of his/her utterances.

To overcome these limitations above, non-parallel voice conversion models are proposed. Several studies use extra data [3], [4] or additional models [5] to facilitate training process, although it brings more cost of training. To avoid such disadvantages, recent studies introduced deep generative models, such as GANs [6], [7], and VAEs [6], [8]. Among them, CycleGAN-VC [9] (and the enhanced version [10]) is a GAN based model by configuring CycleGAN model (as widely used in image style transfer) with a gated CNN and cycle consistency loss. It produced promising results on parallel-VC without parallel corpus for training. However, it is only designed for one-to-one voice conversion. In contrast, StarGAN-VC [11] and its variants [12], [13] provided many-to-many voice conversion on non-parallel corpus by only single generator. Adaptations from StarGAN [14] include embedding loss and source-and-target conditional adversarial loss to enhance the models' generation accuracy.

However, the GAN based methods encounter saddle point problem that causes difficulties in training. Despite the good performance in computer vision, GAN based methods do not sound real [15], as the discriminator is easier to fool than human ears. In many-to-many VC task, the quality of converted voices are degraded as more speakers are trained simultaneously [16].

In another research direction, variational auto-encoders (VAE) based methods have a simple objective function, i.e. the maximization of ELBO (evidence lower bound) and its training strategy is suitable for self-supervised learning. Recent works [15], [17], [16] use autoencoder frameworks to disentangle input utterances into two embeddings which correspond to speaker and linguistic content information respectively. To generate a specific utterance from a target speaker, we use the concatenation of the speaker embedding of the target speaker and the content embedding of source speaker uttering the desired sentences. Previous works have showed that the transformer structure has advantages in speech recognition [18] and general speech applications compared to RNN's [19]. In speech tasks, attention mechanism was first adopted for speech recognition [20] and afterward used in VC tasks [21], [22], [23], [24], [25]. Prior works mainly used cross-attention between source and target in latent space.

Noticing that the speaker information is global and has long range dependency, we add self-attention to gain a more effective many-to-many zero-shot style transfer, i.e. attention disentangled VAE. We find that *self-attention VAE with struc-*

[1] The work was partially supported by NSF grants DMS-1854434 and DMS-1952644 at UC Irvine.

978-1-6654-5961-7/22/$31.00 ©2022 IEEE
DOI 10.1109/AI4I54798.2022.00022

tured pruning has considerable advantages in voice conversion especially on unseen speakers. Merely adding self-attention to VAE is much less dramatic.

In our experimental section, we compared our method with Disentangled-VAE [16][2] and another recent method Fragment VC [26] on VCTK Corpus [27]. We employ group-ℓ_0 penalty and a splitting algorithm [28] in our implementation to remarkably improve model generalization on unseen speakers. Our converted utterances out-perform [16], [26] in both voice quality and conversion accuracy measured objectively by third party packages [29], [30].

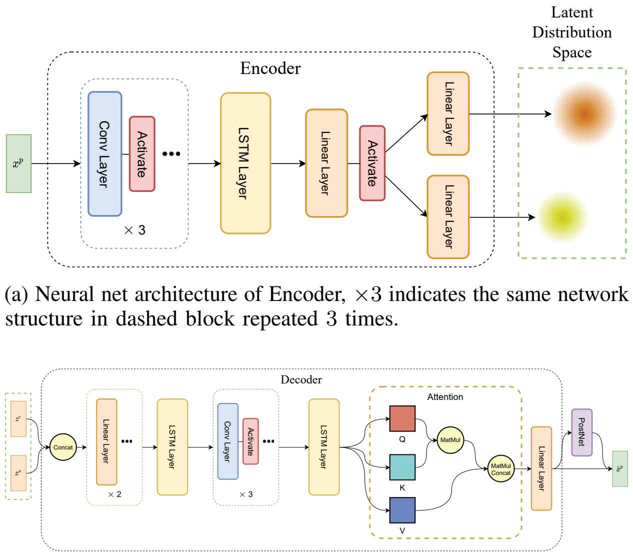
II. Related Works

A. Zero Shot Voice Conversion

IDE-VC [31] learned disentangled representation by introducing coarse lower/upper bounds of mutual information. Sequential AutoEncoder [32], [33] replaced standard normal prior with time-dependent learnable prior. Later works [16], [34], [15], [31] improved the performance of disentangled VAE to enable zero-shot voice conversion, i.e. the network has never listened to the voice of source/target speaker. Disentangled-VAE [16] used β-VAE [35], which modified the variational ELBO of VAE to encourage disentangled representations, and learn separate features for speaker and content respectively. AutoPST [17] furthermore disentangled prosody by re-sampling, which captures speaker's rhythm information.

B. Attention Based Voice Conversion

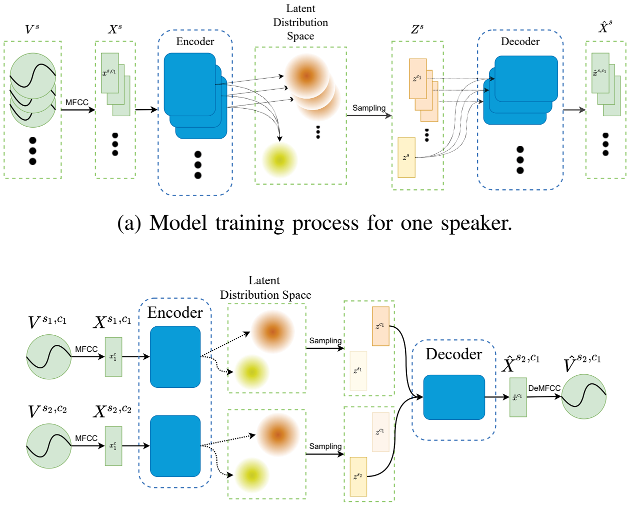
ATTS2S-VC [21] is a sequence-to-sequence model trained on parallel data with RNN appearing in encoder and decoder, and with guided cross-attention layer to do conversion in latent space. FastS2S-VC [22] used the same idea and improved the inference speed to real time voice conversion. Similarly, [23] also used cross-attention in latent space, but the encoder and decoder used CNN instead of RNN and one-shot voice conversion was supported. Apart from latent space, [24], [25] also tried attention to extract speaker identity information. FragmentVC [26] obtained the source utterance latent phonetic structure from Wav2Vec 2.0 and target utterance spectral features from log-mel-spectrograms. Aligning the hidden structures of the two different feature spaces with a two-stage training process, FragmentVC extracts fine-grained voice fragments from target speaker utterance(s) and fuse them into the desired utterance. Both self and cross attention mechanisms exist in the decoder. Different from previous works, *only self-attention is present in our VAE decoder to capture long-range dependency while keeping the model size increase due to adding attention moderate* (around 10%). Moreover, we performed *structured pruning on the resulting VAE to reduce its overfitting and generalization error*.



(a) Neural net architecture of Encoder, ×3 indicates the same network structure in dashed block repeated 3 times.

(b) Neural net architecture of Decoder, with self-attention mechanism, $\times n$ indicates the same network structure in dashed block repeated n times.

Fig. 1: Network structure of variational auto-encoder.



(a) Model training process for one speaker.

(b) An example of voice conversion process of generating voice $\hat{V}^{s_2,c_1}$ with speaker s_2 and linguistic content c_1.

Fig. 2: Model training and voice conversion process of our proposed method.

III. Proposed Approach

A. Preliminary

VAE [36] is an autoencoder that compresses the input into a regularized latent distribution in the encoder, and reconstructs the input back in the decoder. More precisely, let the encoder extract posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ of a group of latent variables $\boldsymbol{z}$ given an input data $\boldsymbol{x}$, while the decoder recovers

[2] The baseline code miscalculated the KL divergence term, see github discussion of this issue. We adopted the corrected version.

the conditional distribution $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ of input data $\boldsymbol{x}$ given a sample of the random variable $\boldsymbol{z}$. Since the marginal likelihood $p_\theta(\boldsymbol{x}) = \int p(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})d\boldsymbol{z}$ is intractable for a large dataset, it is hard to maximize directly. A variational lower bound(ELBO) is often optimized instead:

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right),$$

which is the same as minimizing the sum of the reconstruction error and Kullaback-Leibler divergence between the posterior and the prior. The β-VAE [35] is a modification of VAE with an emphasis on discovering disentangled latent factors. The objective function to maximize, different from the original, is

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \beta D_{KL}\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})\right),$$

where $\beta > 1$ is a trade-off between reconstruction quality and the extent of disentanglement.

B. Problem Formulation

We build on network structure and objective function in [35], [16] by adding a suitable attention mechanism [37] to the decoder and train with Group-ℓ_0 Splitting Method (RGSM, [28]) to alleviate over-fitting. Denote by $\boldsymbol{v}^{s_i,c_i}$ for speaker s_i to utter the linguistic content c_i, and by $\boldsymbol{x}^{s_i,c_i}$ the 80 dimensional log-MFCC of the corresponding utterance. The goal of this work is to generate $\boldsymbol{v}^{s_j,c_i}$ from $\boldsymbol{v}^{s_i,c_i}$ and $\boldsymbol{v}^{s_j,c_j}$ where s_i and s_j could be unseen during training, and $i \neq j$.

C. Encoder

We assume that the latent space $\boldsymbol{z}$ of any utterance $\boldsymbol{x}$ is the direct sum of two sub-latent spaces $\boldsymbol{z}_s$ and $\boldsymbol{z}_c$ that correspond to the speaker s and content c of that utterance respectively, where the distribution of $\boldsymbol{z} := (\boldsymbol{z}_s, \boldsymbol{z}_c)$ can be extracted by a well-designed encoder **Enc**.

Same as the standard VAE [36], we assume that $\boldsymbol{z}$ is normally distributed with a diagonal co-variance matrix and use re-parameterization trick to sample $\boldsymbol{z}$ during the training stage. In short, we assume that $(\boldsymbol{z}_s, \boldsymbol{z}_c) \sim p_\phi(\boldsymbol{z}_s, \boldsymbol{z}_c|\boldsymbol{x}) = \mathbf{Enc}(\boldsymbol{x};\phi)$, where $\boldsymbol{z}_s$ and $\boldsymbol{z}_c$ are assumed to be normally distributed $\boldsymbol{z}_s \sim \mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)$ and $\boldsymbol{z}_s \sim \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c)$[3] respectively.

D. Decoder

Since we have assumed that the speaker's sub-latent representation does not change much between different utterances of the same speaker, we adopt group based learned representation [38], where the decoder takes each content embedding $\boldsymbol{z}_c$ sampled from distribution of corresponding utterance and one speaker embedding $\boldsymbol{z}_s$ sampled from the "average distribution" of all utterances in the same group, i.e.

$$\boldsymbol{z}_s^{(i)} \sim \mathcal{N}\left(\frac{1}{n}\sum_{j=1}^{n}\boldsymbol{\mu}_s^{(j)}, \left(\bigodot_{j=1}^{n}\boldsymbol{\sigma}_s^{(j)}\right)^{1/n}\right)$$

for $i \in [n]$ where n is the group size. Decoder gives reconstructed input as follows:

$$\hat{\boldsymbol{x}} = \mathbf{Dec}\left(\boldsymbol{z}|\theta\right) = \mathbf{Dec}\left(\boldsymbol{z}_s, \boldsymbol{z}_c|\theta\right).$$

E. Objective Function

Same as in [16], we also adopt Post-Net [39], namely **Post**, to refine the reconstructed log-MFCC so that our reconstruction loss splits into two terms:

$$\mathcal{L}_{rec} = \|\hat{\boldsymbol{x}} - \boldsymbol{x}\| + \|\hat{\boldsymbol{x}} + \mathbf{Post}\left(\hat{\boldsymbol{x}}\right) - \boldsymbol{x}\|$$

where $\hat{\boldsymbol{x}} = \mathbf{Dec}\left(\boldsymbol{z}|\theta\right)$ is the decoder output.

Our loss function is a variant of **negative** Evidence Lower Bound (ELBO):

$$\mathcal{L} = \mathcal{L}_{rec} + \beta\ D_{KL}\left(p_\phi\left(\boldsymbol{z}|\boldsymbol{x}\right)||\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})\right),$$

for some $\beta \geq 1$ which was proposed in [35].

F. Relaxed Group-wise Splitting Method

Consider $\boldsymbol{W} = \{\cdots, w_g, \cdots\}$, $1 \leq g \leq G$ as grouped weights of a layer in VAE model, where G is the number of groups. The group ℓ_0 (Gl_0) penalty is: $||\boldsymbol{W}||_{Gl_0} := \Sigma_{g=1}^{G}\mathbf{1}_{||w_g||_2\neq 0}$, where $\mathbf{1}$ is the characteristic function, and its standard convex relaxation the group Lasso (GL) penalty [40] is: $||\boldsymbol{W}||_{GL} := \Sigma_{g=1}^{G}||w_g||_2$. By solving exactly the following Gℓ_0 proximal problem for positive parameter λ:

$$\arg\min_{z_g}\ \lambda\,\mathbf{1}_{\|z_g\|\neq 0} + \frac{1}{2}\,||z_g - w_g||_2$$

we deduce the proximal (*projection* or *shrinkage*) operator of Gℓ_0 penalty:

$$\mathbf{Prox}_{Gl_0,\lambda}(w_g) := w_g\mathbf{1}_{||w_g||_2>\sqrt{2\lambda}}.$$

Similarly, $\mathbf{Prox}_{GL,\lambda}(w_g) := w_g\max\{||w_g||_2 - \lambda, 0\}/||w_g||_2$. Both operators are group-wise operations. Applying group-wise splitting and proximal operators to gradient descent training, we implemented the relaxed group-wise splitting method (RGSM) [28], [41] summarized in Alg.1, where α is the learning rate of gradient descent, β is a positive (relaxation) parameter, $\delta = G\ell_0\ or\ GL$ refers to the type of penalty.

In our work here on voice conversion, RGSM is applied to *each VAE layer having width greater than* 128*, which turns out to be either a fully-connected layer or an LSTM layer* where *group means column of a weight matrix*. In our experiments, the group ℓ_0 penalty ($\delta = G\ell_0$) is found better than group Lasso and is adopted for all results reported below. Because the *group ℓ_0 penalty is discontinuous*, the splitting step in RGSM is absolutely necessary for the penalty to be integrated into the stochastic gradient descent training. The hyper-parameters we chose for RGSM are $\beta = 0.1$, $\lambda = 4 \times 10^{-2}$, $\lambda_l = 10^{-6}$. Due to its *built-in shrinkage mechanism, RGSM helps to reduce overfitting in network training and considerably improve the generalization capability* of our proposed self-attention VAE model as we shall see in section IV.

[3] Since we consider only Gaussians with diagonal co-variances, we specify a vector of variances in the second parameter to Gaussian distribution as convention.

Algorithm 1 Relaxed Group-wise Splitting Method

```
Hyper-parameters $\lambda, \lambda_l, \beta, \alpha$
Objective function: $f(w) = \text{loss}(w, x) + \lambda_l ||w||_{GL}$
Initialize $w^0$
for $e = 1, \cdots$, max-epoch, do
    for $g = 1, \cdots, G$, do
        $u^e = \text{Prox}_{\delta,\lambda}(w_{e-1})$;
    end for
    $w^{e+1} = w^e - \alpha \nabla f(w) - \alpha\beta(w^e - u^e)$;
end for
```

Block	Layer	
Encoder	Input layer	64 × 80
	1D Conv layer ×3	(5, 2, 512)
	BiLSTM layer	(512, 64, 2)
	FC layer	(8192, 2048)
	FC layer (content)	(2048, 56)
	FC layer (speaker)	(256, 8)
Decoder	FC layer	(32, 2048)
	FC layer	(2048, 8192)
	LSTM layer	(128, 512, 1)
	1D Conv Layer ×3	(5,2,512)
	LSTM layer	(512, 1024, 2)
	Multihead Self-Attention	(8, 128)
	FC layer	(1024, 80)

TABLE I: The architecture of our model. The parameters of the 1D convolution layer are denoted as kernel size, stride, output channel number. The parameters of the LSTM layer are input size, hidden layer width, and the number of hidden layers. The parameters of the multihead attention are number of head and feature dimension of each head.

IV. Experiments

A. Dataset and experimental setting

We use VCTK Corpus [27] which includes hundreds of utterances of 109 speakers respectively. We split the data set into two parts: a training set and a testing set. The test set contains all utterances of 6 speakers (3 males and 3 females) and 10 utterances of the rest speakers. The six unseen speakers are p225 (female), p226 (male), p227 (male), p228 (female), p229 (female), and p232 (male), while the seen speakers to evaluate are p286 (male), p287 (male) p288 (female), p292 (male), p293 (female) and p294 (female). The feature we fed to and the output of our network are both 80 dimensional log MFCC (Mel-frequency cepstral coefficients), for which the STFT (short-term Fourier transform) is performed with window size 1024 and hop length 256. We chose the vocoder from wavnet [42]. The Table I presents the detail of our model parameter settings.

B. Evaluation

To evaluate our method, we consider two objective metrics: i) voice quality and ii) speaker classification accuracy. Both metrics are measured with third-party pre-trained network and applied to conversion between both seen and unseen speakers. We picked 6 seen and 6 unseen speakers and do 10 cross utterance conversions for each pair.

For each converted utterance, the voice quality is measured by MOSNet [29] that predicts human ratings of converted speech. MOSnet score ranges from 1 to 5, with lowest score of 1 and highest score of 5. Table II compares MOSNET scores of converted utternance of seen/unseen speakers from VAE models with/without attention and RGSM in training, showing that *the objective speech quality is improved by having RGSM*.

	seen	unseen
Baseline	3.59	3.35
FragmentVC	3.01	3.07
Attention	3.49	3.39
RGSM	**3.74**	**3.58**
Attention + RGSM	3.63	3.51

TABLE II: Average MOSNet scores of converted utternance of seen/unseen speaker and with/without attention.

Given any audio file of a speech, Resemblyzer [30] creates a summary vector of 256 values summarizing the characteristics of the voice spoken. Given multiple speech wav files of a speaker, the Resemblyzer speaker embedding is the mean of multiple Resemblyzer speech embeddings. From the full VCTK dataset [27], we generated six speaker embeddings of seen and unseen speakers respectively. For each converted utterance, we generate the utterance embedding and select the speaker with closest speaker embedding among the six candidates. The prediction is correct if the selected speaker is the target speaker. Since our classifier picks the closest embedding among the six known speaker embeddings, no threshold is involved in our classification. Table III compares the average speaker classification accuracy of converted utterances of seen/unseen speakers with/without attention and with/without RGSM in training, showing that VAE model with attention layer and trained by RGSM reaches the highest classification accuracy.

	seen	unseen
Baseline	78.6%	36.7%
FragmentVC	67.3%	63.3%
Attention	84.0%	37.7%
RGSM	80.0%	34.3%
Attention + RGSM	**93.3%**	**65.0%**

TABLE III: Average speaker classification accuracy of converted utterances of seen/unseen speakers and with/without attention.

V. Conclusions

We successfully integrated a self-attention layer into the decoder of a VAE framework to enhance the zero-shot many-to-many voice conversion task. To improve generalization due to the ensuing larger model capacity, an efficient group-wise splitting and thresholding algorithm has been found efficient in maintaining the generated voice quality of VAE while significantly increasing speaker classification accuracy of converted utterance of seen/unseen speakers. In future work, we plan to explore some of the feature extraction and attention structures in [26] to further reduce generalization error.

References

[1] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanevsky, and Y. Jia, "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation," 2019.

[2] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," 2019.

[3] C.-H. Lee and C.-H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in Ninth International Conference on Spoken Language Processing, 2006.

[4] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in Twelfth Annual Conference of the International Speech Communication Association, 2011.

[5] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN-based approach to voice conversion without parallel training sentences." in Interspeech, 2016, pp. 287–291.

[6] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," arXiv preprint arXiv:1704.00849, 2017.

[7] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," arXiv preprint arXiv:1711.11293, 2017.

[8] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Acvae-vc: Non-parallel voice conversion with auxiliary classifier variational autoencoder," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 9, pp. 1432–1443, 2019.

[9] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in 2018 26th European Signal Processing Conference (EUSIPCO). IEEE, 2018, pp. 2100–2104.

[10] 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.

[11] 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.

[12] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion," arXiv preprint arXiv:1907.12279, 2019.

[13] Z. Zhang, B. He, and Z. Zhang, "Gazev: Gan-based zero-shot voice conversion over non-parallel speech corpus," arXiv preprint arXiv:2010.12788, 2020.

[14] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8789–8797.

[15] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," 2019.

[16] M. Luong and V. A. Tran, "Many-to-many voice conversion based feature disentanglement using variational autoencoder," 2021.

[17] K. Qian, Y. Zhang, S. Chang, J. Xiong, C. Gan, D. Cox, and M. Hasegawa-Johnson, "Global rhythm style transfer without text transcriptions," 2021.

[18] A. Zeyer, P. Bahar, K. Irie, R. Schlüter, and H. Ney, "A comparison of transformer and lstm encoder decoder models for asr," in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 8–15.

[19] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs rnn in speech applications," in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2019, pp. 449–456.

[20] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4960–4964.

[21] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," 11 2018.

[22] H. Kameoka, K. Tanaka, and T. Kaneko, "Fasts2s-vc: Streaming non-autoregressive sequence-to-sequence voice conversion," CoRR, vol. abs/2104.06900, 2021. [Online]. Available: https://arxiv.org/abs/2104.06900

[23] T. Ishihara and D. Saito, "Attention-based speaker embeddings for one-shot voice conversion," 10 2020, pp. 806–810.

[24] Y. Zhang, H. Che, J. Li, C. Li, X. Wang, and Z. Wang, "One-shot voice conversion based on speaker aware module," in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 5959–5963.

[25] S. Liu, J. Zhong, L. Sun, X. Wu, X. Liu, and H. Meng, "Voice conversion across arbitrary speakers based on a single target-speaker utterance," 09 2018, pp. 496–500.

[26] 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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 5939–5943.

[27] C. Veaux and a. K. M. Junichi Yamagishi, "Superseded - CSTR VCTK corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2017. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/2651

[28] B. Yang, J. Lyu, S. Zhang, Y.-Y. Qi, and J. Xin, "Channel pruning for deep neural networks via a relaxed group-wise splitting method," In Proc. of International Conference on AI for Industries, Laguna Hills, CA, 2019.

[29] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," 2019.

[30] G. Louppe, "Resemblyzer," https://github.com/resemble-ai/Resemblyzer, 2019.

[31] S. Yuan, P. Cheng, R. Zhang, W. Hao, Z. Gan, and L. Carin, "Improving zero-shot voice style transfer via disentangled representation learning," in International Conference on Learning Representations, 2021. [Online]. Available: https://openreview.net/forum?id=TgSVWXw22FQ

[32] Y. Li and S. Mandt, "Disentangled sequential autoencoder," 2018.

[33] Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf, "S3vae: Self-supervised sequential VAE for representation disentanglement and data generation," 2020.

[34] J. Chou, C. Yeh, and H. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," 2019.

[35] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "β-vae: Learning basic visual concepts with a constrained variational framework," in International Conference on Learning Representations, 2017. [Online]. Available: https://openreview.net/forum?id=Sy2fzU9gl

[36] D. Kingma and M. Welling, "Auto-encoding variational Bayes," 2014.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[38] H. Hosoya, "A simple probabilistic deep generative model for learning generalizable disentangled representations from grouped data," CoRR, vol. abs/1809.02383, 2018. [Online]. Available: http://arxiv.org/abs/1809.02383

[39] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," CoRR, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[40] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 68, no. 1, pp. 49–67, 2006.

[41] T. Dinh, B. Wang, A. Bertozzi, S. Osher, and J. Xin, "Sparsity meets robustness: Channel pruning for the Feynman-Kac formalism principled robust deep neural nets," in Proc. of International Conference on Machine Learning, Optimization, and Data Science; and LNCS, vol. 12566, 2020, pp. 362–381.

[42] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," CoRR, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499