ADVERSARIALLY LEARNING DISENTANGLED SPEECH REPRESENTATIONS FOR ROBUST MULTI-FACTOR VOICE CONVERSION

Jie Wang[1]*, Jingbei Li*[1]*, Xintao Zhao*[1]*, Zhiyong Wu*[1,2,*] *, Helen Meng*[1,2]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
{jie-wang19,lijb19,zxt20}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Factorizing speech as disentangled speech representations is vital to achieve highly controllable style transfer in voice conversion (VC). Conventional speech representation learning methods in VC only factorize speech as speaker and content, lacking controllability on other prosody-related factors. State-of-the-art speech representation learning methods for more speech factors are using primary disentangle algorithms such as random resampling and ad-hoc bottleneck layer size adjustment, which however is hard to ensure robust speech representation disentanglement. To increase the robustness of highly controllable style transfer on multiple factors in VC, we propose a disentangled speech representation learning framework based on adversarial learning. Four speech representations characterizing content, timbre, rhythm and pitch are extracted, and further disentangled by an adversarial network inspired by BERT. The adversarial network is used to minimize the correlations between the speech representations, by randomly masking and predicting one of the representations from the others. A word prediction network is also adopted to learn a more informative content representation. Experimental results show that the proposed speech representation learning framework significantly improves the robustness of VC on multiple factors by increasing conversion rate from 48.2% to 57.1% and ABX preference exceeding by 31.2% compared with state-of-the-art method.

Index Terms— disentangled speech representation learning, multi-factor voice conversion, prosody control in voice conversion, adversarial learning, gradient reverse layer

1. INTRODUCTION

Voice conversion (VC) aims at converting the input speech of a source speaker to sound as if uttered by a target speaker without altering the linguistic content [1]. Besides the conversion of timbre, the conversions can also be conducted in various domains such as prosody, pitch, rhythm or other non-linguistic domains. Representation learning methods for these speech factors have already been proposed and applied in many research fields in speech processing [2, 3, 4]. However, directly applying the speech representations extracted by these methods in VC may cause unexpected conversions of other speech factors as they may be not necessarily orthogonal. Therefore, disentangling the representations of intermingling various informative factors in speech signal is crucial to achieve highly controllable VC [5].

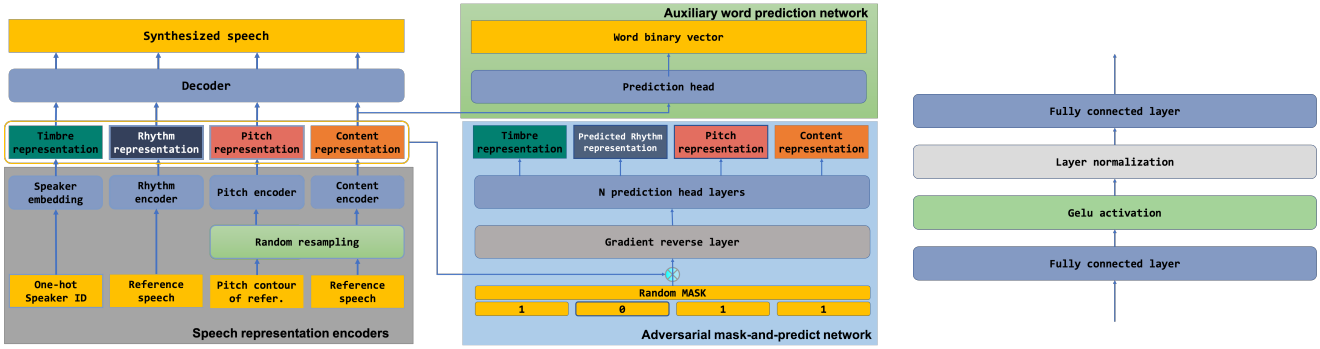
* Corresponding author.

Conventionally, only speaker and content information are factorized in VC. Auto-encoder which is composed of an encoder and a decoder is proposed and widely used for VC [6] . During training, the decoder reconstructs the speech from the speaker and content representations extracted from the encoder or other pretrained extractors. Variational autoencoder based methods [7] model the latent space of content information as Gaussian distributions to pursue the regularization property. Vector quantization based methods [8] are further proposed to model content information as discrete distributions which are more related to the distribution of phonetic information. An auxiliary adversarial speaker classifier is adopted [9] to encourage the encoder to cast away speaker information from content information by minimizing the mutual information between their representations [10].

To overcome the situation that prosody is also converted while replacing the speaker representation in conventional VC, different information bottlenecks are applied to decompose the speaker information into timbre and other prosody-related factors such as rhythm and pitch [11]. To improve disentanglement, restricted sizes of bottleneck layers encourage the encoders to discard the information which can be learnt from other bottlenecks. Random resampling is also proposed to use in the information bottlenecks to remove rhythm information from content and pitch representations.

However, without explicit disentanglement modeling, random resampling [12] and restricting the sizes of bottleneck layers can only gain limited disentanglement of speech representations. Random resampling which is usually implemented as dividing and resampling speech segment using linear interpolation on time dimension can only be used in removing time-related information such as rhythm. Moreover, random resampling is proved as a partial disentanglement algorithm that can only contaminate a random portion of the rhythm information [11]. Besides, the sizes of bottlenecks layer need to be carefully designed to extract disentangled speech representations which are ad-hoc and may not be suitable for other datasets. And the content encoder actually is a residual encoder which cannot ensure that the content information is only modeled in the content representation.

In this paper, to achieve robust and highly controllable style transfer for multiple factors VC, we propose a disentangled speech representation learning framework based on adversarial learning. The proposed framework explicitly removes the correlations between the speech representations which characterize different factors of speech by an adversarial network inspired by BERT [13]. The speech is firstly decomposed into four speech representations which represent content, timbre and another two prosody-related factors,



(a) Overall architecture (b) Prediction head layer

Fig. 1. Architecture of the proposed multiple factor voice conversion system with adversarially disentangled speech representation learning.

rhythm and pitch. During training, one of the speech representations will be randomly masked and inferred from the remaining representations by the adversarial mask-and-predict (MAP) network. The MAP network is trained to maximize the correlations between the masked and the remaining representations, while the speech representation encoders are trained to minimize the correlations by taking the reversed gradient of the MAP network. In this way, the representation learning framework is trained in the adversarial manner, with speech representation encoders trying to disentangle the representations, while MAP network trying to maximize the representation correlations. A word prediction network is employed to predict word existence vector from content representations, which indicate whether each vocabulary exists in the reference speech. The decoder reconstructs the speech from the representations during training and can achieve VC on multiple factors by replacing the corresponding speech representations.

Experimental results show that the proposed speech representation learning framework significantly improves the robustness of VC on multiple factors, increasing conversion rate from 48.2% to 57.1% and ABX preference exceeding by 31.2% compared to state-of-the-art speech representation learning methods for multiple factors. Furthermore, the proposed framework also eschews the laborious manual effort for sophisticated bottleneck tuning.

2. METHODOLOGY

Our proposed disentangled speech representation learning framework, shown in Figure 1, is composed of three sub networks: (i) multiple speech representation encoders which encode speech into different speech representations characterising content, timbre, rhythm and pitch, (ii) an adversarial MAP network that is trained to capture the correlations between different speech representations based on the mask-and-predict operations, (iii) an auxiliary word prediction network which predicts a binary word existence vector indicating whether the content representation contains corresponding vocabulary words. Finally, a decoder is employed to synthesize speech from these disentangled speech representations.

2.1. Speech representation learning

Three encoders in SpeechFlow [11] are fine-tuned to extract rhythm, pitch and content representations from reference speech at frame-level. One-hot speaker labels(ID) are embedded at utterance-level and used as the timbre representations.

2.2. Adversarial learning for speech representation disentanglement

An adversarial MAP network inspired by BERT [13] is designed to explicitly disentangle the extracted speech representations. During training, one of these four speech representations is randomly masked and the adversarial network infers the masked representation from other representations. The adversarial network is composed of a gradient reverse layer [14] and a stack of prediction head layers [15] which has also been used in masked acoustic modeling. Each prediction head layer is composed of a fully-connected layer, GeLU activation [16], layer normalization [17] and another fully-connected layer demonstrated in Figure 1(b). The gradient of the adversarial network is reversed by a gradient reversal layer [14] before backward propagated to the speech representation encoders. $L1$ loss is adopted here to measure the adversarial loss demonstrated in the following equations:

$$Z = (Z_r, Z_c, Z_f, Z_u) \tag{1}$$

$$M \in \{(0,1,1,1),(1,0,1,1),(1,1,0,1),(1,1,1,0)\} \tag{2}$$

$$Z_{input} = Z \cdot M \tag{3}$$

$$Z_{predicted} = Adversarial\,(Z_{input}) \tag{4}$$

$$Z'_{predicted} = (1-M)\cdot Z_{predicted} + M \cdot Z \tag{5}$$

$$L_{adversarial} = \left|\left|Z - Z'_{predicted}\right|\right| \tag{6}$$

where $L_{adversarial}$ is adversarial loss, Z is the concatenation of Z_r, Z_c, Z_f, Z_u denoting rhythm, content, pitch and timbre representations respectively, M is a randomly selected mask, Z_{input} is the representations after mask operation in which masked representation is set as 0 if the corresponding indicator value in M is 0 and unmasked representations remain the original value, $Z_{predicted}$ is the speech representations predicted by the adversarial MAP network, $Z'_{predicted}$ is the results of keeping unmasked representations the same as in the Z and replacing the masked representation with the predicted one in the $Z_{predicted}$.

The MAP network is trained to predict the masked representation as accurate as possible by minimizing the adversarial loss, while in the backward propagation, the gradient is reversed which encourages the representations learned by the encoder contain as little mutual information as possible.

2.3. Auxiliary word prediction network

To avoid that the content information is encoded into other representations, an auxiliary word prediction network is designed to predict the existences of each vocabulary from the content representation. The word prediction network is a stack of prediction head layer which is to produce a binary vocabulary-size vector where each dimension indicates whether the corresponding vocabulary word exists in this sentence. The word existence vector is denoted as $V_{word} = [\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_n]$ where $v_i = 1$ if word i is in speech, otherwise $v_i = 0$. Cross entropy loss is applied here to force the content prediction as accurate as possible:

$$L_{word} = -\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} v_i \\ 1 - v_i \end{bmatrix}^T \begin{bmatrix} v_i' \\ 1 - v_i' \end{bmatrix} \tag{7}$$

where the v_i' is the predicted word exist indicator, n is the size of vocabulary. $v_i' = 1$ if the word i is predicted present other wise $v_i' = 0$.

2.4. VC with disentangled speech representations

The decoder in SpeechFlow [11] is employed to generate mel spectrogram from the disentangled speech representations. During training, four speech representations are extracted from the same utterance and the decoder is trained to reconstruct the mel spectrogram from the speech representations with a loss function defined as the following equation:

$$L_{reconstruct} = \|\boldsymbol{S} - \hat{\boldsymbol{S}}\|_2^2 \tag{8}$$

where S and $\hat{S}$ is the mel spectrogram of the input and reconstructed speeches respectively. The entire model is trained with a loss defined as the following equation:

$$loss = \alpha * L_{adversarial} + \beta * L_{word} + \gamma * L_{reconstruct} \tag{9}$$

where α, β, γ are the loss weights for adversarial loss, word prediction loss and reconstruction loss respectively. To improve the robustness of our proposed framework, the loss weight for the reconstruction loss is designed to be exponential decaying.

3. EXPERIMENT

3.1. Training setup

The experiments are performed on the CSTR VCTK corpus [18], which contains audio data produced by 109 speakers in English. We randomly select a subset of 10 females and 10 males. After preprocessing, the corpus for experiment contains 6471 sentences in total, 5176 sentences for training, 647 sentences for validation and 285 sentences for testing.

All the audios are down-sampled to 16000Hz. Mel spectrograms are computed through a short time Fourier transform (STFT) using a 50 ms frame size, 12.5 ms frame hop, and a Hann window function. We transform the STFT magnitude to the mel scale using an 80 channel mel filterbank spanning 125 Hz to 7.6 kHz, followed by log dynamic range compression. The filterbank output magnitudes are clipped to a minimum value of 0.01. The weights of adversarial loss and word prediction loss are fixed to 10^{-1} and 10^{-2} respectively. The weight of reconstruction loss γ applies an initial weight of 1 with decay factor of 0.9 every 200,000 steps. We train a vanilla SpeechFlow [11] as the baseline approach on the same training and validation sets.

We program all neural networks used in the experiments based on an open source pytorch implemention of SpeechFlow [11]. We train all models with a batch size of 16 for 500,000 steps using the ADAM optimizer with learning rate fixed to 10^{-4} on a NVIDIA 2080Ti GPU. We use a pretrained wavenet vocoder on VCTK corpus [19] to synthesize the audios from the spectrogram. The demo is available https://thuhcsi.github.io/icassp2021-multi-factor-vc/.

3.2. Objective evaluation

Mel-cepstral distortion (MCD) is calculated on a sub set of the testing set which consists 300 parallel conversion pairs of 155 sentences including inter-gender and intra-gender converison. The audios in the test set are perceptually distinct in pitch and rhythm. MCD is defined as the Euclidean distance between the predicted mel spectrogram and the that of target speech. The MCD compariosn is shown in Table 1. The proposed voice conversion system outperforms the baseline with decreasing the MCD from 4.00 to 3.94.

Table 1. MCD comparison between different approaches.

	Baseline	**Proposed**
MCD	4.00	3.94

Table 2. ABX comparison between proposed and baseline approaches. PR refers to preference rate.

	Baseline	**Proposed**	**Neutral**
PR	20.6%	51.8%	27.6%

3.3. Subjective evaluation

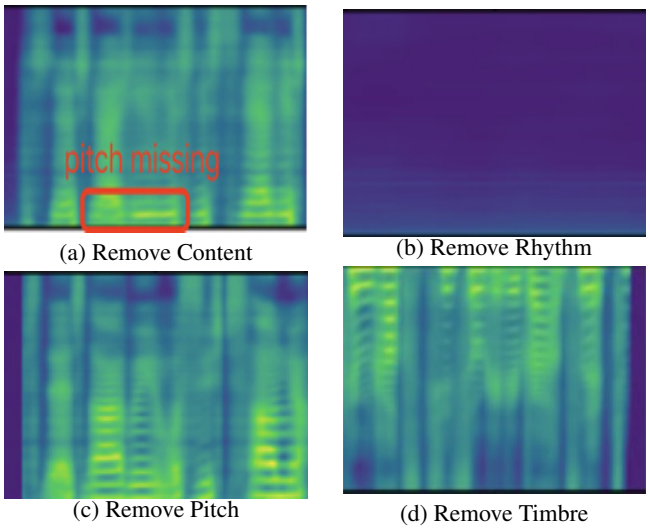
We perform the ABX test on 20 utterances selected from the testing set in terms of similarity between the converted and reference speech when different factors of speech are converted. The listeners are presented with the target utterance and the factors which are converted and asked to select the most similar speech from the ones synthesized from different approaches in random order. As shown in table 2, our proposed model outperforms the baseline with 31.2% higher than baseline on average. It means that while converting the same aspect, the proposed framework endows the voice conversion system a strong disentanglement and conversion ability. It also improves the interpretability as promising a distinct outstanding conversion results.

We conduct another subjective evaluation to measure the conversion rate of different approaches. The listeners are presented with both the source and target utterances in random order and a random synthesized speech. The listeners are asked to select the converted speech is more similar to the source or the target utterance for each speech factor converted in the synthesized speech. For each speech factor, listeners are asked to choose whether the converted speech is more similar to the source or target utterance individually. It means that the conversion rates of different speech factors are evaluated independently and not influenced by each other. The conversion rate is defined as the percentage of answers that choose the target utterance [11].

As shown in Table 3, our proposed model outperforms the baseline in the most conversion conditions which means a highly controllable voice conversion.

Table 3. Conversion rate comparison between proposed and baseline approaches.

	Baseline	Proposed
conversion rate	48.2%	57.1%



(a) Remove Content (b) Remove Rhythm

(c) Remove Pitch (d) Remove Timbre

Fig. 2. Reconstructed Mel spectrogram when one component is removed of the sentence "I must do something about it." of the Baseline system.

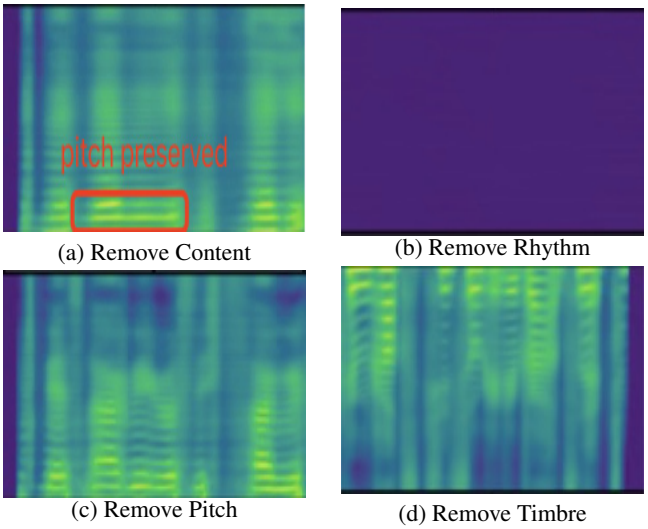
3.4. Analysis and discussion

To further show the disentanglement performance of our proposed framework, we generate mel spectrograms with four speech factors removed by set the corresponding input as zero [11] as shown in Figure 2 and Figure3. Take content removed as an example as shown in Figure 2(a) and 3(a), after the content information is removed, the spectrogram of the proposed system is composed of more uninformative blanks. It can be observed that the proposed system removes the content information more thoroughly than the baseline which means that in the proposed system, the amount of content information leaking into other encoder is less than baseline system. The pitch information is preserved more than as less flat than baseline approach as annotated in Figure 2 and 3.

When the content is removed, both the reconstructed mel spectrograms of the two systems are blank except that there is a bright line in the Figure 2(b) indicating that partial rhythm information is encoded by other encoders. When the pitch is removed, the pitch contour of the reconstructed speech generated by the proposed system retains the curve but is flatter than that of baseline. When the timbre is removed, both the formant position shift indicates the speaker identity changes. When one of the four speech factors is set zero, the proposed system not only removes the corresponding information more thoroughly but also keeps other information undamaged which shows that the proposed system achieves a better disentanglement .

3.5. Ablation study

Ablation studies are conducted to validate the effectiveness of the word prediction network. For investigating the effects, we train the proposed model but without the word prediction network. As shown in Table 4, the reconstruction loss decreases from 21.5 to 12.8 and the adversarial loss decreases from 0.016 to 0.015 on training set after applying the word prediction network.



(a) Remove Content (b) Remove Rhythm

(c) Remove Pitch (d) Remove Timbre

Fig. 3. Reconstructed Mel spectrogram when one component is removed of the sentence "I must do something about it." of the Proposed system.

Table 4. Effect of Word prediction network on reconstruction loss and adversarial loss reduction.

	Proposed-without word prediction	Proposed with word prediction
$L_{reconstruct}$	21.5	**12.8**
$L_{adversarial}$	0.016	**0.015**

The decrease of cost functions demonstrates the contributions of word prediction network for enabling a more robust disentangled speech representation learning voice conversion system. The results show that the word prediction network boosts the performance of the voice conversion system.

4. CONCLUSION

In order to increase the robustness of highly controllable style transfer on multiple factors in VC, we propose a disentangled speech representation learning framework based on adversarial learning. We extract four speech representations which characterizing content, timbre, rhythm and pitch, and we employ an adversarial network inspired by BERT to further disentangle the speech representations. We employ a word prediction network to learn a more informative content representation. Experimental results show that the proposed speech representation learning framework significantly improves the robustness of VC on multiple factors.

5. REFERENCES

[1] Jing-Xuan Zhang, Zhen-Hua Ling, Yuan Jiang, Li-Juan Liu, Chen Liang, and Li-Rong Dai, "Improving sequence-to-

sequence voice conversion by adding text-supervision," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6785–6789.

[2] Zhen Zeng, Jianzong Wang, Ning Cheng, and Jing Xiao, "Prosody learning mechanism for speech synthesis system without text length limit," *arXiv preprint arXiv:2008.05656*, 2020.

[3] Hossein Zeinali, Shuai Wang, Anna Silnova, Pavel Matějka, and Oldřich Plchot, "But system description to voxceleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[4] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[5] Lantian Li, Dong Wang, Yixiang Chen, Ying Shi, Zhiyuan Tang, and Thomas Fang Zheng, "Deep factorization for speech signal," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5094–5098.

[6] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min 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.

[7] Wen-Chin Huang, Hsin-Te Hwang, Yu-Huai Peng, Yu Tsao, and Hsin-Min Wang, "Voice conversion based on cross-domain features using variational auto encoders," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 51–55.

[8] Da-Yi Wu and Hung-yi Lee, "One-shot voice conversion by vector quantization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7734–7738.

[9] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.

[10] Orhan Ocal, Oguz H Elibol, Gokce Keskin, Cory Stephenson, Anil Thomas, and Kannan Ramchandran, "Adversarially trained autoencoders for parallel-data-free voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2777–2781.

[11] Kaizhi Qian, Yang Zhang, Shiyu Chang, David Cox, and Mark Hasegawa-Johnson, "Unsupervised speech decomposition via triple information bottleneck," *arXiv preprint arXiv:2004.11284*, 2020.

[12] Adam Polyak and Lior Wolf, "Attention-based wavenet autoencoder for universal voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6800–6804.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[14] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[15] Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[16] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[18] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al., "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[19] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," *arXiv preprint arXiv:1905.05879*, 2019.