ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446296

UNIFYING ONE-SHOT VOICE CONVERSION AND CLONING WITH DISENTANGLED SPEECH REPRESENTATIONS

Hui Lu[1,2], *Xixin Wu*[1,2,*], *Haohan Guo*[1], *Songxiang Liu*[4], *Zhiyong Wu*[1,3,*], *Helen Meng*[1,2,3]

[1]The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Center for Perceptual and Interactive Intelligence (CPII) Ltd, Hong Kong SAR, China
[3]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[4]Tencent AI Lab, Tencent, Shenzhen, China

ABSTRACT

We propose unifying one-shot voice conversion and cloning into a single model that can be end-to-end optimized. To achieve this, we introduce a novel extension to a speech variational auto-encoder (VAE) that disentangles speech into content and speaker representations. Instead of using a fixed Gaussian prior as in the vanilla VAE, we incorporate a learnable text-aware prior as an informative guide for learning the content representation. This results in a content representation with reduced speaker information and more accurate linguistic information. The proposed model can sample the content representation using either the posterior conditioned on speech or the text-aware prior with textual input, enabling one-shot voice conversion and cloning, respectively. Experiments show that the proposed method achieves better or comparable overall performance for one-shot voice conversion and cloning compared to state-of-the-art voice conversion and cloning methods.

Index Terms— Voice conversion, voice cloning, VAE, speech disentanglement

1. INTRODUCTION

Voice conversion and cloning are two techniques that are important for personalized speech generation. Voice conversion aims to modify speech from a source speaker to make it sound as though it was produced by a designated target speaker [1]. Voice cloning aims to generate speech from text for a speaker absent from the training data in a data-efficient manner [2]. The research community is especially interested in both tasks under one-shot scenarios, i.e., when only a few seconds of speech are available for the target speaker.

Traditionally, voice conversion [3] and cloning [2, 4] are studied separately. However, these two tasks share much in common regarding how information flows from the input to the output, as illustrated in Figure 1. Both tasks involve combining spoken content with speaker identity to generate the desired speech. The reference speech of the target speaker provides speaker identity information, while the spoken content is provided by the source speech for voice conversion and by the text for voice cloning. In real-world applications, it is desirable to model voice conversion and cloning jointly. For instance, a user may wish to convert either their speech or typed text into a designated character voice.

This motivates us to unify voice conversion and cloning into one framework. We propose a solution incorporating the disentangled content and speaker representations of speech, which serve as the key element to bridge voice conversion and cloning. With the desired disentanglement, one can map either source speech or text into the content representation, which can be combined with the speaker representation disentangled from the reference speech to generate the desired speech. Thus voice conversion and cloning can be respectively achieved.

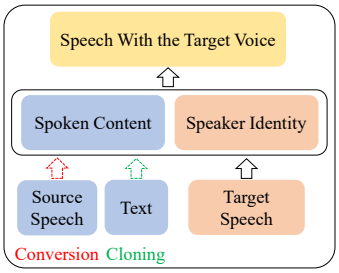

Fig. 1. Information flow for voice conversion and cloning

We adopt the variational auto-encoder (VAE) [5] as it is widely used in learning disentangled representations and voice conversion [6, 7, 8, 9, 10]. A naive voice conversion model based on VAE consists of a content posterior and a speaker posterior to respectively extract the content and speaker representations from speech. Two fixed Gaussian priors are imposed to regularize the two posteriors in a vanilla VAE. Recently proposed methods [9, 10] show that restricting two Kullback-Leibler (KL) divergence terms for two posteriors can limit the amounts of information captured by the two latent variables, thus facilitating the learning of disentangled representations and high-quality voice conversion. Based on this design, we propose a novel extension that incorporates textual information to improve voice conversion performance by inducing better disentanglement. Additionally, this extension allows for voice cloning, as textual input is also accepted.

Specifically, we propose replacing the fixed isotropic Gaussian prior for the content representation with a trainable text-aware one. The text-aware prior is conditioned on the textual input to model the distribution of content representation. Compared to a fixed Gaussian prior, which provides a rather general regularization to the content representation learning, the text-aware prior is more informative of the linguistic content and can thus facilitate learning the more accurate and speaker-independent content representation. Besides, existing methods [9, 10] impose two weight parameters on the two KL divergence terms to facilitate disentanglement, the performance is very sensitive to the choices of these two weight parameters. In contrast, the proposed model is more robust to variations of weight parameters due to the strong regularization provided by the text-aware prior.

* Corresponding authors.

The improved disentanglement helps produce better voice conversion performance.

Furthermore, the text-aware prior makes it possible to sample the content representation from the textual input, which enables text-to-speech (TTS) and especially one-shot voice cloning. The TTS paradigm of the proposed model differs from the traditional multi-speaker TTS methods that directly map the textual embedding and speaker embedding to the speech acoustic features. Instead, the explicitly disentangled speech representations in the proposed model divide the TTS into two modeling phases: 1) to convert the text into a speaker-independent content representation using the text-aware prior; 2) to combine the speaker representation with the content representation and generate speech using the decoder. This two-phase modeling paradigm makes the learning more structured and efficient, thus helping produce good voice cloning performance.

We incorporate a modified evidence lower bound (ELBO) loss function to train the text-aware prior jointly with other parts of the proposed model end-to-end. After one training pass, the proposed model can be directly applied to new speaker voices for both voice conversion and cloning.

2. RELATED WORK

Our work is mainly related to recently proposed VAE-based speech disentanglement methods [9, 10] that incorporate a learning objective similar to β-VAE [6]. We propose to incorporate the trainable text-aware prior to better regularize the content representation learning, which improves voice conversion performance and enables one-shot voice cloning. Several other methods have been proposed to achieve voice conversion and cloning simultaneously. Some methods [11, 12] joint train a speech encoder and a text encoder to extract the speaker-independent content representation, but these models do not disentangle the speaker representation from speech, thus requiring further adaptation training for an unseen speaker. Other methods [13, 14, 15] adopt Glow [16] conditioned on the speaker embedding to extract the speaker-independent content representation. While the base distribution of Glow is set as the distribution of the textual embedding, it is encouraged to extract speaker-independent content representation. However, the content representation extracted using the conditional Glow remains significantly speaker-dependent, which reduces speaker similarity for voice conversion, as shown by the experimental results for SC-GlowTTS [13] in Section 5.3.

3. FORMULATION

3.1. VAE for speech disentanglement

We first introduce the VAE-based framework for speech disentanglement. Let $\mathcal{D}$ denote the speech corpus and y represent the speech acoustic feature; the goal is to disentangle y into two latent variables z_c and z_s to encode the information of linguistic content and speaker identity, respectively. To achieve this goal, we can adopt a VAE to model the distribution of speech y with latent variables z_c and z_s. We assume that the prior distributions for z_c and z_s are isotropic Gaussians, denoted as $p(z_c)$ and $p(z_s)$, respectively. Let the conditional distribution $p_\theta(y|z_c, z_s)$ parameterized by θ define the process of speech generation given the two latent variables, and let $q_\phi(z_c|y)$ and $q_\phi(z_s|y)$ denote respectively the posterior distributions of z_c and z_s, where ϕ represents the parameters. We refer to $p(z_c)$, $p(z_s)$, $q_\phi(z_c|y)$, $q_\phi(z_s|y)$ and $p_\theta(y|z_c, z_s)$ respectively as content prior, speaker prior, content posterior, speaker posterior and decoder. Recent works [9, 10] show the learning objective in Eqn. (1) can facilitate the disentanglement of content and speaker representations from speech, where β_c and β_s are two properly chosen hyper-parameters. Note that when $\beta_c = \beta_s = 1.0$, Eqn. (1) becomes the vanilla objective for VAE with two independent latent variables.

$$\begin{aligned}\mathbb{E}_{y\sim\mathcal{D}}[\log p(y)] \geq\ & \mathbb{E}_{y,q_\phi(z_c|y),q_\phi(z_s|y)}[\log p(y|z_c, z_s)] \\ & - \beta_c \cdot \mathbb{E}_y[D_{\mathrm{KL}}[q_\phi(z_c|y) \parallel p(z_c)]] \\ & - \beta_s \cdot \mathbb{E}_y[D_{\mathrm{KL}}[q_\phi(z_s|y) \parallel p(z_s)]] \end{aligned} \tag{1}$$

It has been proved that $\mathbb{E}_y[D_{\mathrm{KL}}[q_\phi(z|y) \parallel p(z)]]$ is an upper bound of the mutual information between the latent variable and speech: $\mathbb{I}[y, z]$ [9], here z is a general variable name covering z_c and z_s. In this sense, properly chosen values for β_c and β_s can restrict the information flowing from z_c and z_s to be precisely the content and speaker identity, respectively. With the content and speaker representations being disentangled from speech, this model can achieve one-shot voice conversion by combining z_c extracted from the source speech with z_s extracted from the target speech.

3.2. VAE with text-aware prior

While the Gaussian prior $p(z_c)$ can serve as a general target to limit the amount of information captured by z_c, it does not specify how z_c should be, thus can cause content information loss when we impose a large β_c. When corpus $\mathcal{D}$ includes transcripts, i.e., we have the text x corresponding to each utterance y, it is desirable to incorporate the textual information to regularize the content representation learning further. We propose replacing the fixed Gaussian content prior with a trainable one, which we define as $p_\omega(z_c)$ with parameter ω. To make $p_\omega(z_c)$ text-aware, we model the conditional distribution of z_c given the text x, i.e., $p_\omega(z_c|x)$ with a neural network. For a given $\hat{z}_c$ that is sampled from $q_\phi(z_c|\hat{y})$, the prior probability of $\hat{z}_c$ is defined as $p_\omega(\hat{z}_c) = \int_x p_\omega(\hat{z}_c|x)p(x)dx \approx \frac{1}{|\mathcal{D}|}\sum_{x\sim\mathcal{D}} p_\omega(\hat{z}_c|x)$. In practice, we adopt a single-point approximation to the summation. Specifically, we let $p_\omega(\hat{z}_c) \approx \frac{1}{|\mathcal{D}|}p_\omega(\hat{z}_c|\hat{x})$, where $\hat{x}$ is the text corresponding to $\hat{y}$. This approximation is reasonable since $\hat{x}$ yields the largest $p_\omega(\hat{z}_c|\hat{x})$ and thus $\frac{1}{|\mathcal{D}|}p_\omega(\hat{z}_c|\hat{x})$ contributes most to the summation.

We substitute $p(z_c)$ in Eqn. (1) with $p_\omega(z_c)$, which is then approximated by $\frac{1}{|\mathcal{D}|}p_\omega(z_c|x)$ to derive the ELBO of the proposed model. This produces the resultant ELBO shown in Eqn. (2), in which we ignore a positive constant induced by $\frac{1}{|\mathcal{D}|}$.

$$\begin{aligned}\mathbb{E}_{(x,y)}[\log p(y)] \geq\ & \mathbb{E}_{y,q_\phi(z_c|y),q_\phi(z_s|y)}\log p(y|z_c, z_s) \\ & - \beta_c \cdot \mathbb{E}_{(x,y)}[D_{\mathrm{KL}}[q_\phi(z_c|y) \parallel p_\omega(z_c|x)]] \\ & - \beta_s \cdot \mathbb{E}_{(x,y)}[D_{\mathrm{KL}}[q_\phi(z_s|y) \parallel p(z_s)]] \end{aligned} \tag{2}$$

Since the text-aware prior is trainable, the optimization of the proposed model is slightly different from the vanilla VAE. For each batch of training data, we fix $p_\omega(\cdot)$ when updating other parts of the VAE. To update the text-aware prior, we fix other parts of the VAE and sample z_c from $q_\phi(\cdot|y)$ for different y to do maximum log-likelihood training over $p_\omega(\cdot)$, which can be approximated by the log-likelihood of $\hat{z}_c$ evaluated by $p_\omega(\cdot|x)$ as shown in Eqn. (3), where $\mathcal{C}$ is a constant.

$$\begin{aligned}\mathbb{E}_{y,q_\phi(z_c|y)}[\log p_\omega(z_c)] &\approx \mathbb{E}_{(x,y),q_\phi(z_c|y)}\left[\log \frac{1}{|\mathcal{D}|}p_\omega(z_c|x)\right] \\ &= \mathbb{E}_{(x,y),q_\phi(z_c|y)}[\log p_\omega(z_c|x)] + \mathcal{C} \end{aligned} \tag{3}$$

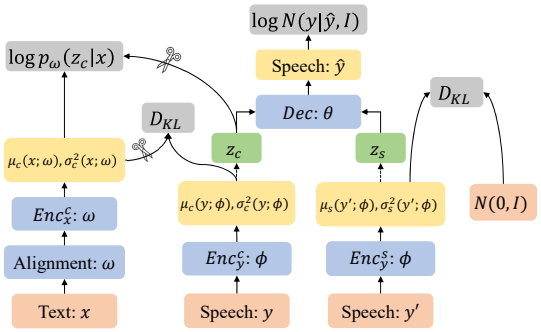

Fig. 2. Model architecture (scissors denote stop of gradient)

In summary, the learning objective to be maximized for the proposed model is shown in Eqn. (4), where $\text{sg}(\cdot)$ refers to the stop gradient operation. The first three terms form the VAE learning objective that treats the content prior as a fixed one. The last term guides the update of text-aware prior.

$$\begin{aligned}\mathcal{L}_T = &\mathbb{E}_{y,q_\phi(z_c|y),q_\phi(z_s|y)}[\log p(y|z_c, z_s)] \\ &- \beta_c \cdot \mathbb{E}_{(x,y)}[D_{\text{KL}}[q_\phi(z_c|y) \parallel \text{sg}(p_\omega(z_c|x))]] \\ &- \beta_s \cdot \mathbb{E}_{(x,y)}[D_{\text{KL}}[q_\phi(z_s|y) \parallel p(z_s)]] \\ &- \mathbb{E}_{(x,y)}[\mathbb{E}_{\text{sg}(q_\phi(z_c|y))}[\log p_\omega(z_c|x)]]\end{aligned} \tag{4}$$

4. IMPLEMENTATION

The overall architecture of the proposed model is shown in Figure 2. We adopt Enc_y^c and Enc_y^s to respectively model the content and speaker posteriors. These two modules take in the speech and output the mean and variance of the corresponding latent variable. Dec models $p_\theta(y|z_c, z_s)$ as a fixed-variance Gaussian where the mean is the predicted acoustic feature. The content prior consists of the Alignment module and Enc_x^c. The Alignment module aims to align the raw text onto a frame-level textual feature. Enc_x^c takes in the aligned textual feature and outputs the mean and variance of the content representation. During training, the input to Enc_y^c and Enc_y^s is the same, except that the input to Enc_y^s is chunked and shuffled along the time axis [17]. This operation can help learn a more content-independent speaker representation. The KL-divergence between the content posterior and the content prior is normalized by the time length.

To achieve voice conversion, we use Enc_y^c to extract the content representation from the source speech and obtain the speaker representation from the target speech using Enc_y^s; the two representations are fed into Dec to generate the desired speech. For voice cloning, the content representation is extracted from the text using Enc_x^c.

Enc_y^c consists of two layers of 1D 256-dimension convolution with a kernel size of 3 and 2 layers of 256-dimension self-attention. Enc_y^s contains 4 layers of 256-dimension 1D convolution with the 2-strided average pooling layer in between; the final output is averaged globally along the time axis to obtain a single vector, which is fed into a fully-connected layer to predict the mean and variance. The Alignment module combines the text encoder, length predictor, and length regulator from FastSpeech2 [18]. Enc_x^c and Dec have the same structure as Enc_y^c, except that for Dec the output of the self-attention is projected to the acoustic feature.

We train the proposed model on a single Tesla V100 GPU with a batch size of 32. we adopt the Adam optimizer with a constant learning rate of 10^{-4}. We heuristically search β_c over $[1.0, 10.0]$ and β_s over $[10^{-3}, 10^{-1}]$ using the validation set.

5. EXPERIMENTS

5.1. Dataset

For evaluation, we utilize the multi-speaker English speech corpus VCTK [19]. It contains 109 speakers, from which we choose 8 for validation and 11 for testing. We remove any leading or trailing silence from all utterances and extract the mel-spectrograms using the same settings as HiFi-GAN [20]. We preprocess the text in a similar manner to FastSpeech-2 [18].

5.2. Baselines

We compare the proposed method with several strong baselines to show its effectiveness in learning disentangled speech representations to achieve one-shot voice conversion and cloning. For disentangled speech representation learning and voice conversion, we adopt VQMIVC [21] that applies vector quantization and mutual information minimization to disentangle speech and achieve one-shot voice conversion. The speech VAE with isotropic Gaussian content prior [9] is also included in the comparison – we set $\beta_c = 4.0$ and $\beta_s = 10^{-3}$ and denote it as VAE-GP. For voice cloning, we choose the recently proposed CDFSE [17] that utilizes an attention-based speaker encoder to capture the speaker characteristic. We also include SC-GlowTTS [13] that can achieve both one-shot voice conversion and cloning, and use the SC-GlowTTS-Trans variant as it achieves good overall results without the fine-tuned vocoder. We refer to the proposed speech VAE with text-aware prior as VAE-TP, for which we set $\beta_c = 5.0$ and $\beta_s = 10^{-3}$. The sizes of all compared models are shown in the second column of Table 2. We use the same pre-trained HiFi-GAN vocoder for all compared models to generate the waveform.

5.3. Disentanglement evaluation

We report the equal error rate (EER) of speaker verification (SV) using the content and speaker representations to demonstrate the performance of disentanglement. Ideally, the speaker identity information should be captured only in the speaker representation z_s, which is expected to yield good SV performance. On the other hand, the content representation z_c should produce relatively worse SV results. We extract the content and speaker representations from all utterances in the test set using all compared models except CDFSE, which does not explicitly disentangle speech into content and speaker representations. The content representation is averaged over time to obtain an utterance-level embedding. For SC-GlowTTS, the content representation is extracted using the Glow decoder conditioned on the speaker embedding of input speech. We randomly select 4 utterances from each speaker as anchors, while all the remaining utterances are taken as trials. The results yielded by a pre-trained speech recognition bottleneck feature [22] and a pre-trained SV model called Resemblyzer are also included for reference.

The results are shown in Table 1. As can be observed, the proposed method learns the best-disentangled representations for content and speaker in terms of the EER. VQMIVC also achieves good disentanglement, but the performance is worse than VAE-GP and VAE-TP. SC-GlowTTS performs well on SV when using speaker representation since the speaker encoder is pre-trained on the large-scale SV corpus, but the content representation remains largely speaker-dependent.

Resemblyzer: https://github.com/resemble-ai/Resemblyzer

Table 1. SV results on content and speaker representations

Model	EER (z_c) ↑	EER (z_s) ↓
Pre-trained	0.431	0.020
VQMIVC	0.361	0.076
SC-GlowTTS	0.131	0.030
VAE-GP	0.386	0.043
VAE-TP (ours)	**0.441**	**0.029**

5.4. Speech generation evaluation

We conduct objective and subjective evaluations to demonstrate the effectiveness of the proposed model in one-shot voice conversion and cloning. We randomly select five utterances from each speaker in the test set as voice conversion sources, and one utterance from each test speaker as the reference speech. In total, we obtain 550 converted utterances. For voice cloning, we select 50 sentences from the LibriTTS test set [23] and synthesize them using the voice of the 11 test speakers, resulting in 550 synthesized utterances in total. Part of the converted and synthesized samples are available online. We refer to voice conversion as CV and voice cloning as CL for ease of writing. We include the evaluation results on samples copy-synthesized by Hifi-GAN for reference, denoted as Copy-Syn.

For objective evaluations, we use Resemblyzer to extract speaker embeddings from converted and synthesized utterances. We report the cosine similarities (CS) between the generated utterances and the reference speech as an indicator of speaker similarity. Additionally, we employ a pre-trained speech recognition model [24] to transcribe the generated utterances. A lower character error rate (CER) of the transcription indicates more accurate content representation and more intelligible generated speech.

The results are shown in Table 2. We can observe that the proposed VAE-TP model achieves the best speaker similarity for voice conversion and cloning in terms of CS. VAE-TP also achieves better results for CER of voice conversion than VAE-GP and VQMIVC. SC-GlowTTS obtains the best conversion CER performance thanks to the information preservation capability of the Glow decoder. Both VAE-TP and CDFSE surpass SC-GlowTTS in cloning CER. CDFSE achieves better cloning CER than VAE-TP by adopting a phoneme classifier to regularize the content representation learning, while VAE-TP relies only on the VAE objective without extra supervision.

Table 2. Speech generation objective evaluation results

Model	Size	CV-CS ↑	CV-CER ↓	CL-CS ↑	CL-CER ↓
Copy-Syn	-	0.827	0.12%	0.827	0.12%
VQMIVC	336M	0.700	8.43%	-	-
VAE-GP	188M	0.715	7.93%	-	-
CDFSE	592M	-	-	0.758	**1.96%**
SC-GlowTTS	384M	0.719	**0.58%**	0.722	6.99%
VAE-TP (ours)	399M	**0.740**	1.49%	**0.783**	2.82%

We randomly select 15 converted and 15 synthesized utterances for subjective evaluation. We ask 19 subjects who understand English sufficiently to listen to these samples and evaluate their naturalness and similarity to their corresponding reference speech. The speech naturalness and speaker similarity are evaluated with a 5-scale mean opinion score (MOS). The proposed model achieves overall comparable or better performance than all baselines for all metrics, as shown in Table 3. While SC-GlowTTS achieves relatively better performance in conversion naturalness, its similarity and cloning naturalness are worse than other methods.

Samples: https://light1726.github.io/voice_conversion_and_cloning/

Table 3. Speech generation subjective evaluation results

Models	CV MOS (95% CI)		CL MOS (95% CI)	
	Naturalness	Similarity	Naturalness	Similarity
Copy-Syn	4.36±0.09	4.51±0.09	4.36±0.09	4.51±0.09
VQMIVC	3.69±0.10	3.63±0.11	-	-
VAE-GP	3.66±0.10	3.67±0.10	-	-
CDFSE	-	-	3.82±0.05	3.60±0.11
SC-GlowTTS	**3.92±0.12**	3.66±0.14	3.56±0.13	3.38±0.10
VAE-TP (ours)	3.73±0.10	**3.71±0.10**	3.81±0.08	**3.66±0.09**

5.5. Ablation study

In the ablation study, we aim to examine the effect of the text-aware prior by comparing the proposed VAE-TP with VAE-GP, which has a fixed Gaussian prior. We fix the weight parameter $\beta_s = 10^{-3}$ and vary β_c. We report SV EER using z_c and transcription CER of converted speech to indicate the disentanglement and voice conversion performance, respectively. Table 4 shows that the performance of VAE-GP is very sensitive to the value of β_c. Although an increase in β_c leads to a more speaker-independent z_c, it causes more loss of linguistic information and transcription errors. While VAE-TP yields a similar trend, it produces overall better performance and more robustness to varying β_c than VAE-GP. This demonstrates the superiority of the proposed text-aware prior in facilitating better speech disentanglement and voice conversion.

Table 4. The effects of varying β_c on VAE-GP and VAE-TP

Model	Tasks	$\beta_c = 1.0$	$\beta_c = 5.0$	$\beta_c = 10.0$
VAE-GP	EER (z_c) ↑	0.232	0.374	0.418
	CV-CER ↓	1.10%	11.11%	32.21%
VAE-TP (ours)	EER (z_c) ↑	0.325	0.443	0.469
	CV-CER ↓	0.27%	1.49%	3.42%

6. CONCLUSION

We propose a unified framework that can perform both one-shot voice conversion and cloning. We incorporate a learnable text-aware prior into a speech VAE to disentangle speech into content and speaker representations. Compared to the vanilla fixed Gaussian prior, the text-aware prior is more informative regarding the linguistic content and can aid the speech VAE in learning better-disentangled representations. Additionally, the text-aware prior allows for the flexibility to sample content representation from text, enabling the proposed model to achieve both one-shot voice conversion and cloning. Experimental results demonstrate that the proposed model can achieve better or comparable voice conversion and cloning performance compared to existing methods.

7. ACKNOWLEDGEMENT

This research is supported by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission's InnoHK Scheme, as well as the National Natural Science Foundation of China (62076144), Shenzhen Science and Technology Program (WDZC20220816140515001).

8. REFERENCES

[1] Yannis Stylianou, Olivier Cappé, and Eric Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] Sercan Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, "Neural voice cloning with a few samples," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[3] Ju-Chieh Chou and Hung-yi Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *Interspeech 2019, Graz, Austria, 15-19 September 2019*. 2019, pp. 664–668, ISCA.

[4] Dongyang Dai, Yuanzhe Chen, Li Chen, Ming Tu, Lu Liu, Rui Xia, Qiao Tian, Yuping Wang, and Yuxuan Wang, "Cloning one's voice using very limited data in the wild," in *ICASSP 2022*. IEEE, 2022, pp. 8322–8326.

[5] Diederik P. Kingma and Max Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[6] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.

[7] Wei-Ning Hsu, Yu Zhang, and James R. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," in *Advances in Neural Information Processing Systems, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 1878–1889.

[8] Yuki Saito, Yusuke Ijima, Kyosuke Nishida, and Shinnosuke Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5274–5278.

[9] Hui Lu, Disong Wang, Xixin Wu, Zhiyong Wu, Xunying Liu, and Helen Meng, "Disentangled speech representation learning for one-shot cross-lingual voice conversion using ß-vae," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 814–821.

[10] Jiachen Lian, Chunlei Zhang, and Dong Yu, "Robust disentangled variational speech representation learning for zero-shot voice conversion," in *IEEE ICASSP*. IEEE, 2022.

[11] Hieu-Thi Luong and Junichi Yamagishi, "NAUTILUS: A versatile voice cloning system," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2967–2981, 2020.

[12] Tasnima Sadekova, Vladimir Gogoryan, Ivan Vovk, Vadim Popov, Mikhail Kudinov, and Jiansheng Wei, "A Unified System for Voice Cloning and Voice Conversion through Diffusion Probabilistic Modeling," in *Proc. Interspeech 2022*, 2022, pp. 3003–3007.

[13] Edresson Casanova, Christopher Shulby, Eren Gölge, Nicolas Michael Müller, Frederico Santos de Oliveira, Arnaldo Candido Jr., Anderson da Silva Soares, Sandra Maria Aluisio, and Moacir Antonelli Ponti, "SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model," in *Proc. Interspeech 2021*, 2021, pp. 3645–3649.

[14] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021, pp. 5530–5540.

[15] Yi Lei, Shan Yang, Jian Cong, Lei Xie, and Dan Su, "Glow-wavegan 2: High-quality zero-shot text-to-speech synthesis and any-to-any voice conversion," in *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, Hanseok Ko and John H. L. Hansen, Eds. 2022, pp. 2563–2567, ISCA.

[16] Durk P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *Advances in neural information processing systems*, vol. 31, 2018.

[17] Yixuan Zhou, Changhe Song, Xiang Li, Luwen Zhang, Zhiyong Wu, Yanyao Bian, Dan Su, and Helen Meng, "Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis," in *Interspeech 2022*. 2022, pp. 2573–2577, ISCA.

[18] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[19] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[20] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[21] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng, "VQMIVC: vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Interspeech 2021, Brno, Czechia, 30 August - 3 September 2021*. 2021, pp. 1344–1348, ISCA.

[22] Songxiang Liu, Yuewen Cao, Disong Wang, Xixin Wu, Xunying Liu, and Helen Meng, "Any-to-many voice conversion with location-relative sequence-to-sequence modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1717–1728, 2021.

[23] Heiga Zen, Rob Clark, Ron J. Weiss, Viet Dang, Ye Jia, Yonghui Wu, Yu Zhang, and Zhifeng Chen, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech*, 2019.

[24] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al., "Nemo: a toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.