UniSyn: An End-to-End Unified Model for Text-to-Speech and Singing Voice Synthesis

Yi Lei[1], Shan Yang[2], Xinsheng Wang[1], Qicong Xie[1], Jixun Yao[1], Lei Xie[1*], Dan Su[2]

[1] Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University, Xi'an, China
[2] Tencent AI Lab, China
leiyi@npu-aslp.org, {xieqicong, yaojx}@mail.nwpu.edu.cn, lxie@nwpu.edu.cn,
w.xinshawn@gmail.com, {shaanyang, dansu}@tencent.com

Abstract

Text-to-speech (TTS) and singing voice synthesis (SVS) aim at generating high-quality speaking and singing voice according to textual input and music scores, respectively. Unifying TTS and SVS into a single system is crucial to the applications requiring both of them. Existing methods usually suffer from some limitations, which rely on either both singing and speaking data from the same person or cascaded models of multiple tasks. To address these problems, a simplified elegant framework for TTS and SVS, named *UniSyn*, is proposed in this paper. It is an end-to-end unified model that can make a voice speak and sing with only singing or speaking data from this person. To be specific, a multi-conditional variational autoencoder (MC-VAE), which constructs two independent latent sub-spaces with the speaker- and style-related (i.e. speak or sing) conditions for flexible control, is proposed in UniSyn. Moreover, supervised guided-VAE and timbre perturbation with the Wasserstein distance constraint are leveraged to further disentangle the speaker timbre and style. Experiments conducted on two speakers and two singers demonstrate that UniSyn can generate natural speaking and singing voice without corresponding training data. The proposed approach outperforms the state-of-the-art end-to-end voice generation work, which proves the effectiveness and advantages of UniSyn.

Introduction

Recent advances in text-to-speech (TTS) (Ren et al. 2019; Wang et al. 2017; Li et al. 2019; Chen et al. 2021; Liu et al. 2021a; Jeong et al. 2021; Weiss et al. 2021) and singing voice synthesis (SVS) (Nishimura et al. 2016; Kim et al. 2018; Blaauw and Bonada 2020; Liu et al. 2022; Gu et al. 2021; Lu et al. 2020; Ren et al. 2020) have greatly improved the quality and naturalness of the generated speech and singing voices, which results in the widespread applications of TTS and SVS. The increasing requirements in real-world applications necessitate the ability for versatile voices that can not only speak but also sing. Intuitively, training such a system needs high-quality speech and singing voice from the same speaker. However, obtaining such a corpus is more difficult than the corpus consisting of speaking and singing voice recorded by different persons, e.g., it is non-trivial to build a singing corpus for a person who is not good at singing. While the multi-task cascaded system consisting of TTS and singing voice conversion (SVC) (Zhang et al. 2020; Guo et al. 2022; Liu et al. 2021c) models can achieve this goal, it leads to a complex pipeline and inflexible generation depending on reference audio.

Both TTS and SVS share a similar pipeline that generates vocal voices from abstract information, i.e., textual input or musical scores, thus making it intuitively reasonable to realize the SVS and TTS tasks with a unified model that enables a target speaker timbre to speak and sing with the absence of speech or singing data. However, several challenges will rise to unify TTS and SVS into one model. First, as the important input information, musical scores in the SVS task are quite different from the textual input in the TTS task. Second, although both speaking and singing voices are generated from the same vocal articulation system, the acoustic outcomes have substantial differences in both timbre and prosodic aspects. Third, the speaker timbre and style (i.e. speak and sing) of source audio are heavily entangled. When only one style from each person is accessible in the training phase, this entanglement tends to result in the timbre leakage issue (Lee et al. 2020; Xue et al. 2021), which means the synthesized singing voice sounds like the timbre that provides the singing training data rather than the target speaker that only has speech training data and vice versa. Due to these challenges, previous efforts to build the unified TTS and SVS model did not obey the typical pipeline but utilized the explicit features (e.g. pitch contour and rhythm) extracted from the waveform as input to generate the desired speaking or singing voice for the target speaker timbre (Valle et al. 2020; Zhang et al. 2020; Xue et al. 2021). However, the reference signals are always essential for providing explicit features at inference time. Therefore, the generated singing voice relies on the reference and cannot synthesize arbitrary songs, which reduces the flexibility in practical applications.

In this paper, we develop *UniSyn*, which is an end-to-end unified system for both speech and singing voice synthesis with only speaking or singing data from each person. UniSyn consists of three components: 1) a variational autoencoder to encode the waveform into a latent space, 2) a prior model to estimate the latent distribution from the input

---

linguistic representations, and 3) a wave decoder to generate the waveform from sampled latent variables.

To overcome the challenges of unified modeling, we propose unified linguistic representations for the text content of TTS and the music score of SVS. To facilitate speaking/singing style transfer cross speakers in the unified model, we propose a multi-conditional variational autoencoder (MC-VAE) to lead the learned latent space more interpretable by being divided into two independent sub-spaces. One sub-space is for speaker conditioned on speaker identity, and the other is for the remaining information in the voice conditioned on other input representations except for speaker identity (Ding et al. 2020). To better decouple the speaker timbre and style (speaking/singing), we adopt the speaker identity and pitch contour as supervision to conduct disentanglement on the latent variables. Furthermore, the timbre perturbation strategy (Choi et al. 2021) with the Wasserstein distance constraint is also utilized to auxiliarily learn the remaining sub-space for improving the robustness of the disentangled speaker timbre. Note that different from the well-settled two-stage voice generation system that adopts an independently trained acoustic model and neural vocoder to output waveform, UniSyn is an *end-to-end* trained neural generation model that directly produces waveform.

The main contributions of our work are summarized as follows:

- We propose UniSyn, an end-to-end unified system for text-to-speech and singing voice synthesis. It leverages MC-VAE to learn an interpretable latent space for flexible control and utilizes guided-VAE and formant perturbation to conduct better disentanglement.

- We conduct experiments on both TTS and SVS. The results demonstrate that UniSyn significantly outperforms the state-of-the-art end-to-end generation model on SVS for speakers without singing training data, as well as TTS for singers without speech training data.

- To the best of our knowledge, this is the first work on speech and singing voice synthesis unified system in the end-to-end way, which can not only teach the speaker without singing data to sing but also teach the singer without speech data to speak.

Background

Unified Speech & Singing Generation

Singing voice conversion (SVC) is a common solution for making a voice sing and speak without his/her singing or speaking data. Recent works (Saito et al. 2018; Polyak et al. 2020; Guo et al. 2022; Li et al. 2021) use the Phonetic PosteriorGrams (PPGs) extracted from audio to represent the linguistic content and eliminate the speaker characteristics. With PPGs and F0 modeling, these methods can convert the timbre of the singing voice while preserving the linguistic content. A framework (Zhang et al. 2020) is designed by unifying the features of speech and singing synthesis and hiring a speaker verification-based speaker encoder to disentangle the speaker timbre from the audio signal. The extracted speaker embedding is conditioned on the generation

model to synthesize speech and singing voice in the target timbre. For converting speaker timbre of the singing voice, VAE is employed in (Luo et al. 2020) to disentangle speaker timbre from audio in the latent space. The literature on the unified model for TTS and SVS without relying on speech or singing training data usually utilizes the explicit features extracted from the audio signal as the input representations. Mellotron (Valle et al. 2020) is proposed for both speech and singing voice synthesis by explicitly conditioning on the melodic information such as pitch and rhythm. However, reference audio is necessary for providing explicit features of speech or singing at inference time, which is inflexible for synthesizing arbitrary content. Instead of reference audio, approaches Xue et al. (2022); Liu et al. (2021b) are proposed to utilize the music score to produce singing voice conditioned on the speaker identity, whereas they cannot generate speech since the non-universal input. In this paper, we unify the textual content of speech and music score of singing and construct a unified model for TTS and SVS without leveraging any reference audio when inference.

VAE-Based Speech Generation

VAE is a generation model that extracts a latent distribution of the data and reconstructs the data from the sampled latent variables. The VAE family has been successfully applied to TTS and SVS. Recently, the conditional-VAE-based voice synthesis model, VITS (Kim, Kong, and Son 2021), has brought success to high-quality speech generation in a parallel end-to-end TTS framework. VITS adopts a variational autoencoder to approximate the latent distribution of waveform x and a text encoder followed by a stochastic duration predictor to encode the text sequence for producing a condition c. Then a normalizing flow-based decoder is used to generate waveform from the sampled latent variables z. VITS can be expressed in the CVAE process by optimizing the evidence lower bound (ELBO) of the interactable marginal log-likelihood of waveform $\log p\left(x|c\right)$:

$$\log p\left(x|c\right) \geq \mathbb{E}_{q(z|x)}\left[\log p\left(x|z\right) - \log \frac{q\left(z|x\right)}{p\left(z|c\right)}\right], \quad (1)$$

where $p\left(x|z\right)$ is the likelihood function that generate waveform x given the latent variables z, $q\left(z|x\right)$ is the posterior distribution approximated by the posterior encoder, $p\left(z|c\right)$ is the prior distribution of latent variables z given the condition c. Because of the success of VITS, its framework has been extended to SVS. By replacing the Monotonic Alignment Search (MAS), Zhang et al. (2022) introduce a length regulator with a duration predictor (Ren et al. 2019) and a frame prior network to predict a frame-level distribution for natural singing voice synthesis in an end-to-end way. However, these works leverage flow-decoder to disentangle the speaker information, which may reduce the robustness of the generation system and lack of interpretability of the latent space. In this paper, we improve the CVAE-based end-to-end synthesis framework by conducting disentanglement in the latent space for a more robust generation.
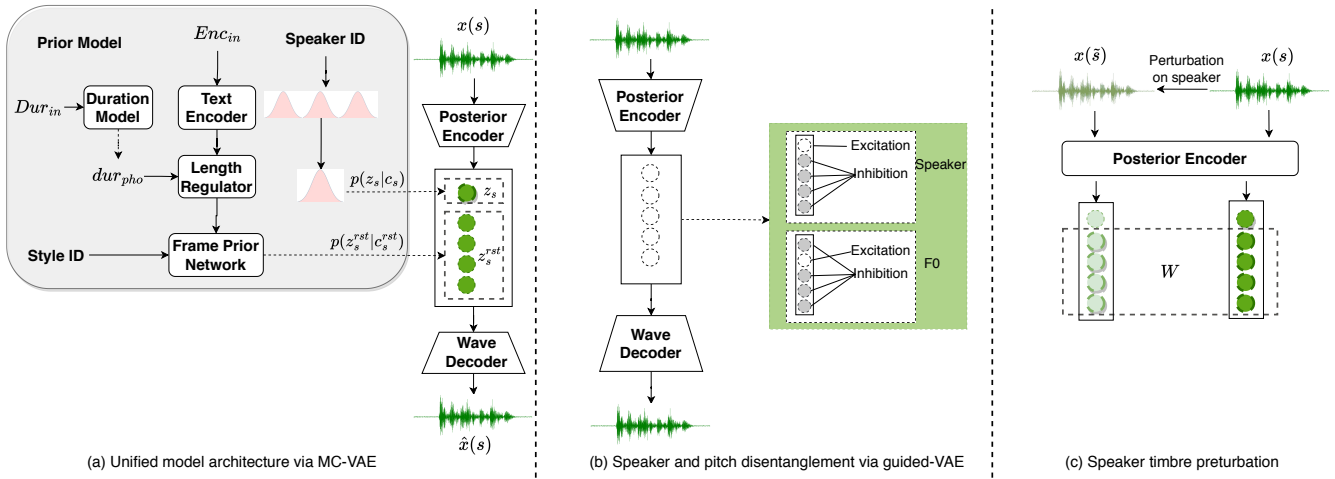
(a) Unified model architecture via MC-VAE   (b) Speaker and pitch disentanglement via guided-VAE   (c) Speaker timbre preturbation

Figure 1: The architecture of UniSyn. In subfigure (a), Enc_{in} is the unified input label for text encoder; Dur_{in} is the unified input for duration prediction; *Speaker ID* is the speaker identity; *style ID* is the style identity for speaking or singing; $x(s)$ is the waveform of the speaker s; $\hat{x}(s)$ is the reconstructed waveform; the dotted lines are the KL divergence. In subfigure (b), $Excitation$ and $Inhibition$ are the excitation function and inhibition function in the guided-VAE respectively. In subfigure (c), $x(\widetilde{s})$ is the speaker perturbed waveform; W is the Wasserstein distance loss between the latent distributions.

UniSyn

We will introduce the proposed unified model for cross-speaker speech and singing voice generation as illustrated in Figure 1, including 1) the unified textual information modeling for TTS and SVS, 2) the proposed MC-VAE for flexible multi-speaker voice generation with multiple conditions, 3) the supervised guided-VAE to disentangle speaker timbre and pitch contour from the latent representations of audio signals, and 4) the timbre perturbation strategy with Wasserstein distance constraint to further enhance the disentanglement ability on the timbre of speech.

Unified Textual Features for TTS and SVS

Previous work (Valle et al. 2020) has shown that both speaking and singing voice could be simultaneously decomposed into explicit acoustic features like rhythm, pitch, speaking style, and text content. To achieve the unified voice generation, we design a unified textual representation for speaking and singing. In TTS, the textual information often can be factorized as *phoneme*, *tone* (for tonal language), and *phoneme duration*. While in SVS, the music score usually includes *phoneme*, *note pitchID*, *phoneme duration*, and *note duration*. Aiming at building a unified voice synthesis model, it's necessary to factorize the common and different aspects between the text and music score inputs.

For the common part, 1) the linguistic content can be defined as the *phoneme* attribute for both speech and singing, which is marked as pho in our model; 2) the melody variations of speaking and singing voice are defined as an attribute named tp merged by the *note pitchID* of the music score and the *tone* of input text, since both they determine the melody acoustic feature. Noted that for non-tonal languages, there is no explicit input feature representing speech melody, so the *note pitchID* should be treated as a different feature in unified textual representations.

With the basic pronunciation skeleton, the phoneme duration dur_{pho} defines the rhythm of both speech and singing. But unlike speech, the duration of a phoneme in singing voice is bound by the music note duartion dur_{note} in the music scores. To form a unified representation, we set a placeholder of dur_{note} for speaking voice. To unify speaking and singing prosody rhythm, a shared duration predictor is leveraged for modeling the phoneme duration of both speech and singing voices. Intuitively, dur_{pho} is strong affected by pho, dur_{note}, and the style tag. Besides, we observe that in Mandarin Chinese, the vowel always has a longer duration than the consonant in the same syllable, especially in singing voices. Therefore, we introduce a *relative position* attribute pos for a more accurate estimation of phoneme duration in the duration model, which is obtained based on the total number of phonemes and the rank of the current phoneme in the current music note or speaking syllable.

As a summary, the attribute set $Enc_{in} = \{pho, tp\}$ is involved to model the linguistic information in the text encoder, while the duration predictor takes the $Dur_{in} = \{pho, dur_{note}, pos, style\}$ to predict the phoneme duration for both speech and singing, which is then applied in the Length Regulator to match the time resolution of acoustic latent features.

MC-VAE for Voice Generation

For building an end-to-end generation system, we follow the CVAE-based skeleton in VITS (Kim, Kong, and Son 2021) to produce waveform from the latent variables. In this work, we separate the latent space into two sub-spaces to make the latent variables learned from VAE more interpretable. Ideally, we tend to decompose the audio signal into two independent attributes, named speaker timbre and remaining attributes. To achieve this goal, we propose the multi-conditional variational autoencoder (MC-VAE).

Given input waveform x, we can learn its latent distribution $q(z|x)$ through the VAE encoder. In MC-VAE, we try to divide the latent z into two independent spaces $z = (z_s, z_s^{rst})$, where z_s only defines a speaker subspace and $z_s^{rst})$ denotes the remaining information in waveform such as linguistic content and style. To make the z_s and z_s^{rst} independent, we utilize two independent conditions $c = (c_s, c_s^{rst})$ to restrict the sub-spaces, where c_s is the categorical speaker label $speakerID$ for modeling z_s and c_s^{rst} is the remaining information generated from the frame-level speaker-independent prior network for modeling z_s^{rst}.

We define the prior distribution of speaker follows the conditional Gaussian distribution $p(z_s|c_s) = N(\mu_{c_s}, \sigma_{c_s})$, where μ_{c_s} and σ_{c_s} are the prior mean and variance and only determined by the speaker ids. Meanwhile, the remaining information z_s^{rst} of audio can be approximated from the unified textual features and the style embedding, which we denote as c_s^{rst}. With the independent conditions $c = (c_s, c_s^{rst})$ and independent latent variables $z = (z_s, z_s^{rst})$, by the conditionally independent property of joint probability distributions, the variational lower bound of MC-VAE in our model can be written as:

$$
\begin{aligned}
&\textbf{ELBO}\,(p, q; x, c) \\
&= \mathbb{E}_{q(z|x)}\left[\log p(x|z) - \log \frac{q(z|x)}{p(z|c)}\right] \\
&= \mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathbb{E}_{q(z_s, z_r|x)}\left[\log \frac{q(z_s, z_s^{rst}|x)}{p(z_s, z_s^{rst}|c)}\right] \\
&= \mathbb{E}_{q(z|x)}[\log p(x|z)] \\
&\quad - \mathbb{E}_{q(z_s|x)q(z_s^{rst}|x)}\left[\log \frac{q(z_s|x)\,q(z_s^{rst}|x)}{p(z_s|c_s)\,p(z_s^{rst}|c_s^{rst})}\right] \\
&= \mathbb{E}_{q(z|x)}[\log p(x|z)] \\
&\quad - KL\left(q(z_s|x)\,||\,p(z_s|c_s)\right) \\
&\quad - KL\left(q(z_s^{rst}|x)\,||\,p(z_s^{rst}|c_s^{rst})\right),
\end{aligned}
\tag{2}
$$

where the first term is the reconstruction loss $\mathcal{L}_{rec}$. The second and third terms are KL divergences, marked as $\mathcal{L}_{kl}^{z_s}$ and $\mathcal{L}_{kl}^{z_s^{rst}}$ respectively. We use the L_1 loss between the mel-spectrogram mel_x from ground-truth waveform x and $mel_{\hat{x}}$ from the predicted waveform $\hat{x}$ in frequency domain to reconstruct x:

$$
\mathcal{L}_{rec} = \|mel_x - mel_{\hat{x}}\|_1.
\tag{3}
$$

$\mathcal{L}_{kl}^{z_s}$ is a reverse-KL divergence to estimate the true distribution $p(z_s|c_s) \sim N(\mu_{c_s}, \sigma_{c_s})$ with our approximate distribution $q(z_s|x)$. The reverse-KL divergence makes objective *mode seeking* (Vaswani et al. 2022; Mei et al. 2019) that helps the approximate distribution $q(z_s|x)$ to find an accurate speaker with high probability and mimic it exactly. On the contrary, $\mathcal{L}_{kl}^{z_s^{rst}}$ is a forward-KL divergence to estimate the posterior distribution $q(z_s^{rst}|x)$ with the approximate prior distribution $p(z_s^{rst}|c_s^{rst})$, which is a *mode covering* (Agarwal et al. 2019) instance that is suitable for learning the timbre-independent remaining information.

Since the timbre of speech is almost invariant over time, the variance of the speaker distribution should be relatively smaller. We directly treat the speaker id c_s as μ_{c_s} and set σ_{c_s} is equal to 0.01, where the smaller variance is also benefit to distinguish speaker with the Gaussian distribution.

Supervised Guided-VAE for Disentanglement

To disentangle the desired attribute, such as speaker timbre, from the latent variables z, we utilize a supervised guided-VAE (GVAE) (Ding et al. 2020), as shown in Figure 1(b). For an attribute f that needs to be disentangled from waveform, GVAE defines the latent variable from the posterior encoder as $z = (z_f, z_f^{rst})$, where z_f is a scalar variable deciding this attribute and z_f^{rst} represents the remaining latent variables. The objective of GVAE contains an adversarial excitation and inhibition method as:

$$
\mathcal{L}_{excitation}^f = L_{pred}(\text{Pred}_f(z_f), f_x)
\tag{4}
$$

$$
\mathcal{L}_{inhibition}^f = L_{pred}(\text{Pred}_f(z_f^{rst}), f_x)
\tag{5}
$$

$$
\mathcal{L}_{gvae}^f = \mathcal{L}_{excitation}^f + 1 \Big/ \mathcal{L}_{inhibition}^f,
\tag{6}
$$

where f_x is the ground-truth value of the attribute f of waveform x, Pred_f refers to the network to predict the attribute value from latent z, and L_{pred} denotes the loss function to optimize prediction results. By minimizing the Eq. 6, the excitation process encourages z_f containing the attribute information f, and the inhibition process can be treated as an adversarial term that makes z_f^{rst} as uninformative to f as possible.

In Unisyn, we utilize the GVAE to disentangle both speaker timbre and pitch information, since the pitch contour is critical to singing voice generation for the speaker without singing data. Specifically, the latent z can be presented as $z = (z_s, z_s^{rst})$ and $z = (z_p, z_p^{rst})$ respectively. For speaker disentanglement, we optimize the shared latent variable $z = (z_s, z_s^{rst})$ from MC-VAE, where the *cross entropy* is treated as L_{pred} to predict the speaker identity. As for pitch disentanglement, $z = (z_p, z_p^{rst})$ is optimized to make sure the z_p only contains pitch-related information, where the *MSE* loss is used to predict pitch values.

Speaker Timbre Perturbation

With the above MC-VAE and GVAE, the timbre and pitch contour can be disentangled from the waveform, and the vocal timbre also can be flexibly controlled by a manual label or sampling from reference audio. To further improve the robustness of generating speaking or singing voice with a target timbre without corresponding training data, we further conduct information perturbation (Choi et al. 2021) on formant of waveform to obtain speaker-independent augmented training data.

In detail, we utilize a formant shifting function fs to distort the timbre of the audio at a random range, since the formant is highly related to vocal timbre. Given the waveform $x(s)$ from speaker s, we apply the function fs on it during each training step to obtain $x(\tilde{s})$, where only formant is randomly shifted and other information is preserved. In this way, we assume the latent variables z_s^{rst} and $z_{\tilde{s}}^{rst}$ in

MC-VAE should follow the same distribution. We then utilize a Wasserstein distance (Zhao et al. 2018) constraint between z_s^{rst} and $z_{\tilde{s}}^{rst}$ to encourage them to learn the speaker-independent information:

$$\mathcal{L}_{pert} = W\left(q\left(z_s^{rst}|x(s)\right)||q\left(z_{\tilde{s}}^{rst}|x(\tilde{s})\right)\right), \qquad (7)$$

where W represents the Wasserstein distance between two distributions. We utilize the Wasserstein distance since it has been shown good performance on the text sequences (Zhao et al. 2018).

Training

To further improve the quality of generated audio, we also apply the adversarial training strategy during training (Kong, Kim, and Bae 2020; Kim, Kong, and Son 2021). Following the state-of-the-art vocoder HiFi-GAN (Kim, Kong, and Son 2021), we utilize several discriminators, including a Multi-Period Discriminator (MPD) and a Multi-Scale Discriminator (MSD), to discriminate the real and fake waveform on different scales. The feature-mapping loss $\mathcal{L}_{fm}$ is also applied to the generator to constrain its output in each hidden layer of the discriminators (Kong, Kim, and Bae 2020).

Combining the above MC-VAE, GVAE, speaker perturbation and adversarial training, we optimize the model by minimizing the loss function as

$$\begin{aligned}\mathcal{L}_G =& \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{kl}^{z_s} + \gamma\mathcal{L}_{kl}^{z_s^{rst}} + \lambda\mathcal{L}_{gvae} \\ &+ \mu\mathcal{L}_{pert} + \eta\mathcal{L}_{fm} + \theta\mathcal{L}_{adv}(G) + \phi\mathcal{L}_{dur},\end{aligned} \qquad (8)$$

where $\mathcal{L}_{gvae}$ is the sum of $\mathcal{L}_{gvae}^s$ and $\mathcal{L}_{gvae}^p$, $\mathcal{L}_{dur}$ is the L_1 loss between the predicted duration and the ground-truth duration in the log domain, and $\mathcal{L}_{adv}(G)$ is the adversarial loss for the generator. The hyper-parameters α, β, γ, λ, μ, η, θ, and ϕ are the weights of the losses. We set $\alpha = 60$, $\beta = 12$, $\gamma = 1.5$, $\lambda = 10$, $\mu = 0.02$, $\eta = 2$, $\theta = 2$, and $\phi = 1.5$ in our model empirically.

Inference

The inference procedure of UniSyn is shown in Figure 2, which supports both TTS and SVS. The speaker-related latent variable z_s is sampled from the speaker prior distribution conditioned on speaker ID, and the latent z_s^{rst} is encoded from the prior model with the unified textual features. With $z = (z_s, z_s^{rst})$, the wave decoder can generate corresponding speech and singing voice of the target speaker.
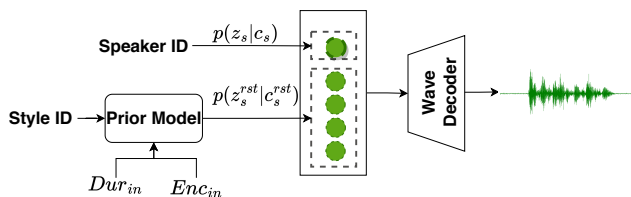

Figure 2: The inference procedure of UniSyn.

Model Architecture

Prior Model. The prior model aims at providing the speaker and remaining prior distributions. It consists of several components: 1) a **text encoder** followed by a length regulator to encode the unified textual inputs into frame-level representations, which contains 6 Feed-Forward Transformer blocks (Ren et al. 2019); 2) a **duration predictor** including 3 convolution layers with dropout to provide each phoneme duration during inference; 3) a **frame prior network**, including 6 Transformer blocks, to produce the frame-level prior latent variables z_s^{rst}; 4) a speaker prior procedure to produce the $p(z_s|c_s)$.

Posterior Encoder. The posterior encoder contains a linear-spectrogram extractor, several WaveNet (Shen et al. 2018) residual blocks, and a linear projection layer, aiming at extracting the mean and variance of the posterior distribution $q(z|x)$ from waveform.

Wave Decoder. Given the sampled $z \sim p(z|x)$ with reparametrization trick, the wave decoder tends to reconstruct x from z, where z is sliced to fixed length for more efficient training. The wave decoder consists of a stack of transposed convolutions with the multireceptive field fusion module (MRF) (Kong, Kim, and Bae 2020) to match the resolution of audio samples.

Experimental Setup

Dataset

To evaluate the performance of UniSyn, we conduct experiments with a mixture of Mandarin speech and singing corpora. The singing corpora contain data from two female singers: 1) the Opencpop [1] (Wang et al. 2022), an open-source singing corpus with 100 pop songs recorded by a professional singer, which we denote as *Singer-1*; and 2) an internal singing corpus with 100 songs from another female singer, denoted as *Singer-2*. Both of the singing datasets have about 5-hour audio, and we split them into sentence pieces for training. The speech corpora consist of two speakers: 1) an open-source Mandarin TTS dataset [2] recorded from a female speaker, which contains about 10-hour neutral speech, denoted as *Speaker-1*; and 2) an internal Mandarin corpus recorded from another female speaker, totally about 5 hours, which we denote as *Speaker-2*.

To balance the amount of data for speech and singing, we randomly select about 2 hours of audio from each speaker for training. For validation and evaluation, 100 utterances from the rest data and two preserved songs from each singing corpus are involved. We down-sample all the speech and singing audios into 24k Hz, and set the frame size and hop size to 1200 and 300 respectively when extracting optional auxiliary acoustic features like pitch and spectrogram. The auxiliary pitch contour is extracted through WORLD (Morise, Yokomori, and Ozawa 2016), and the implementation of formant shifting follows the NANSY (Choi et al. 2021) model using Praat (Boersma and Van Heuven

---

[1]https://wenet.org.cn/opencpop/
[2]https://www.data-baker.com/open_source.html

2001). The phoneme duration is obtained through an HMM-based force alignment model (Sjölander 2003).

Benchmark Systems

We choose the state-of-the-art end-to-end speech synthesis framework VITS for comparison since VITS can conduct the cross-speaker generation by the flow-decoder to disentangle the speaker timbre. Because there is rare research to make a singer speak, we use Learn2Sing 2.0 (Xue et al. 2022) to conduct a comparison on the SVS task additionally. We separately measure the performance of TTS and SVS for the unified model. We conduct experiments on the following systems: 1) GT, the ground-truth recordings; 2) VITS_{TTS}, where we only train the system on the speaking data as the top-line of the unified model; 3) VITS_{SVS}, which is only trained with the singing data as the top-line of VITS for SVS; 4) $\text{VITS}_{\text{unify}}$, the unified model constructed on VITS using the flow-decoder for generating speaking and singing voice; 5) $\text{UniSyn}_{\text{TTS}}$, the proposed system trained on the speech data as the top-line of the proposed model; 6) $\text{UniSyn}_{\text{SVS}}$, which is only trained with the singing data as the top-line of UniSyn for SVS; 7) $\text{UniSyn}_{\text{unify}}$, the proposed model for both TTS and SVS; 8) Learn2sing (Xue et al. 2022), the system to teach speakers to sing with a HiFi-GAN vocoder (Kong, Kim, and Bae 2020) to synthesize audio from mel-spectrogram. All VITS family systems utilize the same length regulator and duration predictor with our proposed methods instead of Monotonic Alignment Search (MAS). Because the duration of the singing voice is highly related to the music score and MAS will lead to the failure of duration prediction. All the above models are trained with 4 NVIDIA V100 GPUs for fair comparison.

Evaluation Metrics

We conduct Mean Opinion Score (MOS) and Comparative Mean Opinion Score (CMOS) tests to evaluate the perceptual audio quality and speaker similarity of the synthetic speech and singing samples. The score of MOS test ranges from 1 to 5 with an interval of 0.5, in which 1 means very bad and 5 means excellent. The rating score of CMOS ranges from -3 to 3, in which a positive score means the first model is better and vice versa. In evaluations, we generate 20 speech and singing samples for each speaker/singer, which are listened to by at least 20 listeners. For an additional subjective evaluation, we calculate the note-level pitch RMSE and Pearson correlation between generated voice and the ground-truth recordings. Note that in the speaker similarity evaluation, the real speech of the two speakers is used as reference for both TTS and SVS tasks as the real singing voices of the speakers are not available. Likewise, the ground truth singing of the two singers are used as reference for both TTS and SVS evaluation.

Results

Subjective Evaluation

We first conduct MOS evaluation on both TTS and SVS tasks separately in terms of naturalness and speaker similarity, as shown in Table 1 and Table 2 respectively.

Evaluation on Text-to-speech. According to the MOS results of TTS shown in Table 1, given the target speech data, VITS_{TTS}, $\text{UniSyn}_{\text{TTS}}$, $\text{VITS}_{\text{unify}}$ and $\text{UniSyn}_{\text{unify}}$ achieve similar scores on both audio naturalness and speaker similarity on Speakers-1 and Speaker-2. It demonstrates that the proposed UniSyn with interpretable latent distribution has the equivalent ability in speech generation with VITS. When generating speaking voice for singers, the proposed $\text{UniSyn}_{\text{unify}}$ significantly outperforms $\text{VITS}_{\text{unify}}$ on speech naturalness and speaker similarity. Specifically, the scores of generated speech of singers from $\text{VITS}_{\text{unify}}$ drop significantly compared to the generated speech of speakers, and $\text{UniSyn}_{\text{unify}}$ can still maintain relatively high MOS scores. This result shows the superiority of the proposed method in producing synthetic speaking voice of the target singer without speech training data. [3]

	Speaker-1	Speaker-2	Singer-1	Singer-2
	Naturalness MOS (↑)			
GT	4.67±0.08	4.62±0.06	-	-
VITS $_{\text{TTS}}$	4.17±0.09	4.13±0.08	-	-
VITS $_{\text{unify}}$	4.15±0.09	4.14±0.10	3.33±0.11	3.36±0.12
UniSyn $_{\text{TTS}}$	4.18±0.08	4.16±0.12	-	-
UniSyn $_{\text{Unify}}$	4.19±0.08	4.15±0.09	**3.79±0.06**	**3.81±0.10**
	Similarity MOS (↑)			
GT	4.78±0.05	4.75±0.07	-	-
VITS $_{\text{TTS}}$	4.26±0.13	4.25±0.11	-	-
VITS $_{\text{unify}}$	4.23±0.10	4.23±0.12	3.45±0.11	3.48±0.09
UniSyn $_{\text{TTS}}$	4.27±0.11	4.25±0.10	-	-
UniSyn $_{\text{Unify}}$	4.24±0.09	4.22±0.12	**3.71±0.11**	**3.76±0.12**

Table 1: Speech naturalness and speaker similarity MOS of TTS with 95% confidence interval

Evaluation on Singing Voice Synthesis. We then investigate the performance of singing voice synthesis for all the testing speakers and singers, as shown in Table 2. For the singer singing task, the UniSyn family can be on par with the VITS family, and they are much better than the baseline Learn2Sing. For the speaker singing task, the naturalness scores of $\text{VITS}_{\text{unify}}$ and Leanr2Sing are much lower than the singer singing task, where the produced singing voice of the speakers has obvious artifacts. $\text{UniSyn}_{\text{unify}}$ surpasses VITS and Leanr2Sing by a large margin when generating speaker's singing voice. As for speaker similarity, the scores of Learn2Sing, VITS, and Unisyn are close, which means the three models have similar ability to maintain speaker timbre. The results of naturalness and speaker similarity indicate that the UniSyn can generate natural singing voices for the speakers, even if they do not have any singing training data. It is worth noticing that the speaker similarity scores of speakers are much lower than the singers in the SVS task. This is mainly because there is no ground-truth singing voice for the speakers, he listeners only can judge

---

[3]Audio samples are available at: https://leiyi420.github.io/UniSyn

the similarity with the speaking data, which has obvious differences from their singing speech.

	Speaker-1	Speaker-2	Singer-1	Singer-2
	Naturalness MOS (↑)			
GT	-	-	4.65±0.08	4.78±0.05
Learn2Sing	3.07±0.16	3.03±0.15	3.85±0.08	3.87±0.09
VITS$_{SVS}$	-	-	3.95±0.08	3.97±0.09
VITS$_{unify}$	3.15±0.10	3.05±0.09	3.93±0.11	3.95±0.12
UniSyn$_{SVS}$	-	-	3.95±0.06	3.96±0.07
UniSyn$_{unify}$	**3.73±0.11**	**3.78±0.10**	3.95±0.06	3.97±0.08
	Speaker similarity MOS (↑)			
GT	-	-	4.57±0.06	4.61±0.04
Learn2Sing	3.75±0.12	3.76±0.15	3.89±0.12	3.93±0.08
VITS$_{SVS}$	-	-	4.15±0.08	4.17±0.07
VITS$_{unify}$	3.75±0.11	**3.79±0.13**	4.15±0.13	4.16±0.11
UniSyn$_{SVS}$	-	-	4.17±0.07	4.18±0.09
UniSyn$_{unify}$	**3.76±0.10**	3.78±0.12	4.14±0.08	4.18±0.07

Table 2: Speech naturalness and speaker similarity MOS of SVS with 95% confidence interval.

Objective Evaluation

Since pitch is critical to the singing voice, we further conduct objective evaluations to measure the pitch accuracy of the synthesized singing voice, as shown in Table 3. The measurements include root mean squared error (RMSE) and Pearson correlation (Corr) with the ground-truth recordings. For Singer-1 and Singer-2, all the models achieve a similar level of RMSE and Corr on pitch. While for the singing voice of Speaker-1 and Speaker-2, the proposed UniSyn$_{unify}$ achieves the lowest RMSE and highest Corr results, which reaches the same level as those of the singer's singing voice. But for Learn2Sing and VITS, RMSE remains very high while Corr is obviously at a lower level. The objective evaluation results are consistent with the subjective tests and further prove that our unified model can synthesize natural singing voices for the speakers without singing training data.

	Speaker-1	Speaker-2	Singer-1	Singer-2
Learn2Sing				
RMSE (↓)	29.559	33.554	10.868	10.280
Corr (↑)	0.837	0.813	0.922	0.925
VITS$_{unify}$				
RMSE (↓)	24.651	29.458	9.012	9.389
Corr (↑)	0.866	0.835	0.964	0.963
UniSyn$_{unify}$				
RMSE (↓)	**8.672**	**9.954**	**8.698**	**8.738**
Corr (↑)	**0.967**	**0.957**	**0.967**	**0.967**

Table 3: Objective evaluation on pitch, where "RMSE" and "Corr" denote the RMSE and Pearson correlation of pitch.

Ablation Studies

The CMOS results of ablation studies are illustrated in Table 4, where "-pert" and "-GVAE" denote removing speaker

perturbation and supervised guided-VAE from UniSyn respectively. A more negative CMOS score means UniSyn is much better. From the TTS of speakers and SVS of singers results, we find removing the perturbation or GVAE would not have much impact on the synthetic voices, which indicates the proposed MC-VAE can effectively conduct voice generation. From the cross-over TTS and SVS, it can be seen that: 1) removing the speaker perturbation strategy leads to a significant decline in speaker similarity, which indicates that speaker perturbation mainly contributes to the timbre during cross-speaker TTS; 2) removing GVAE results in a significant drop on naturalness and also noticeable influence on speaker similarity, which demonstrates that GVAE plays an important role in keeping naturalness and speaker similarity; 3) removing both components have medium scores compared with individually removing them.

We also notice that removing the two strategies leads to different degrees of performance degradation. To be specific, removing speaker timbre perturbation noticeably leads to different degree of degradation on speaker similarity for TTS and SVS. We also find that GVAE is more important to the naturalness of the TTS task although it is necessary for both tasks and metrics.

		Naturalness (↑)	Similarity (↑)
TTS			
-pert	Speakers	-0.006	-0.012
	Singers	0.002	-0.356
-GVAE	Speakers	-0.015	-0.023
	Singers	-0.430	-0.164
-pert-GVAE	Speakers	-0.015	-0.014
	Singers	-0.347	-0.224
SVS			
-pert	Speakers	-0.004	-0.235
	Singers	-0.021	-0.018
-GVAE	Speakers	-0.163	-0.102
	Singers	-0.012	-0.007
-pert-GVAE	Speskers	-0.104	-0.106
	Singers	-0.021	-0.036

Table 4: CMOS values for ablation studies of UniSyn

Conclusions

In this work, we propose *UniSyn* to conduct TTS and SVS in an end-to-end unified trained neural model. UniSyn also has the advantage of synthesizing both speaking and singing voice with only speaking or only singing training data of a target person. Based on a specifically designed unified textual representation, we propose MC-VAE to create a more interpretable latent space for speaking/singing style transfer across speakers. We further introduce GVAE and timbre perturbation into UniSyn to improve the speaker similarity and naturalness of synthetic speech and singing. Experiments and ablation studies show that UniSyn outperforms the state-of-the-art end-to-end synthesis framework, VITS, on both speaker singing and singer speaking tasks, which indicates the effectiveness of the proposed approach in the unified modeling of TTS and SVS.

Acknowledgments

References

Agarwal, R.; Liang, C.; Schuurmans, D.; and Norouzi, M. 2019. Learning to generalize from sparse and underspecified rewards. In *Proceedings of International conference on machine learning*, 130–140. PMLR.

Blaauw, M.; and Bonada, J. 2020. Sequence-to-sequence singing synthesis using the feed-forward transformer. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7229–7233. IEEE.

Boersma, P.; and Van Heuven, V. 2001. Speak and unSpeak with PRAAT. *Glot International*, 5(9/10): 341–347.

Chen, M.; Tan, X.; Li, B.; Liu, Y.; Qin, T.; Zhao, S.; and Liu, T.-Y. 2021. Adaspeech: Adaptive text to speech for custom voice. arXiv:2203.16408.

Choi, H.-S.; Lee, J.; Kim, W.; Lee, J.; Heo, H.; and Lee, K. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34: 16251–16265.

Ding, Z.; Xu, Y.; Xu, W.; Parmar, G.; Yang, Y.; Welling, M.; and Tu, Z. 2020. Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7920–7929.

Gu, Y.; Yin, X.; Rao, Y.; Wan, Y.; Tang, B.; Zhang, Y.; Chen, J.; Wang, Y.; and Ma, Z. 2021. Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders. In *Proceedings of International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5. IEEE.

Guo, H.; Zhou, Z.; Meng, F.; and Liu, K. 2022. Improving Adversarial Waveform Generation Based Singing Voice Conversion with Harmonic Signals. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6657–6661. IEEE.

Jeong, M.; Kim, H.; Cheon, S. J.; Choi, B. J.; and Kim, N. S. 2021. Diff-tts: A denoising diffusion model for text-to-speech. arXiv:2104.01409.

Kim, J.; Choi, H.; Park, J.; Hahn, M.; Kim, S.; and Kim, J.-J. 2018. Korean singing voice synthesis system based on an LSTM recurrent neural network. In *Proceedings of INTERSPEECH*, 1551–1555.

Kim, J.; Kong, J.; and Son, J. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of International Conference on Machine Learning*, 5530–5540. PMLR.

Kong, J.; Kim, J.; and Bae, J. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. 17022–17033.

Lee, J.; Choi, H.-S.; Koo, J.; and Lee, K. 2020. Disentangling timbre and singing style with multi-singer singing synthesis system. In *the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 7224–7228. IEEE.

Li, N.; Liu, S.; Liu, Y.; Zhao, S.; and Liu, M. 2019. Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6706–6713.

Li, Z.; Tang, B.; Yin, X.; Wan, Y.; Xu, L.; Shen, C.; and Ma, Z. 2021. Ppg-based singing voice conversion with adversarial representation learning. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7073–7077. IEEE.

Liu, J.; Li, C.; Ren, Y.; Chen, F.; and Zhao, Z. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11020–11028.

Liu, P.; Cao, Y.; Liu, S.; Hu, N.; Li, G.; Weng, C.; and Su, D. 2021a. Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention. arXiv:2102.06431.

Liu, R.; Wen, X.; Lu, C.; Song, L.; and Sung, J. S. 2021b. Vibrato Learning in Multi-Singer Singing Voice Synthesis. In *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 773–779. IEEE.

Liu, S.; Cao, Y.; Hu, N.; Su, D.; and Meng, H. 2021c. Fastsvc: Fast cross-domain singing voice conversion with feature-wise linear modulation. In *Proceedings of International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.

Lu, P.; Wu, J.; Luan, J.; Tan, X.; and Zhou, L. 2020. Xiaoicesing: A high-quality and integrated singing voice synthesis system. arXiv:2006.06261.

Luo, Y.-J.; Hsu, C.-C.; Agres, K.; and Herremans, D. 2020. Singing voice conversion with disentangled representations of singer and vocal technique using variational autoencoders. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3277–3281. IEEE.

Mei, J.; Xiao, C.; Huang, R.; Schuurmans, D.; and Müller, M. 2019. On principled entropy exploration in policy optimization. In *Proceedings of International Joint Conference on Artificial Intelligence*, 3130–3136.

Morise, M.; Yokomori, F.; and Ozawa, K. 2016. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7): 1877–1884.

Nishimura, M.; Hashimoto, K.; Oura, K.; Nankaku, Y.; and Tokuda, K. 2016. Singing Voice Synthesis Based on Deep Neural Networks. In *Interspeech*, 2478–2482.

Polyak, A.; Wolf, L.; Adi, Y.; and Taigman, Y. 2020. Unsupervised Cross-Domain Singing Voice Conversion. In *Proceedings of INTERSPEECH*, 801–805.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Fastspeech: Fast, robust and controllable text to speech. 3165–3174.

Ren, Y.; Tan, X.; Qin, T.; Luan, J.; Zhao, Z.; and Liu, T.-Y. 2020. Deepsinger: Singing voice synthesis with data mined from the web. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1979–1989.

Saito, Y.; Ijima, Y.; Nishida, K.; and Takamichi, S. 2018. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5274–5278. IEEE.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4779–4783. IEEE.

Sjölander, K. 2003. An HMM-based system for automatic segmentation and alignment of speech. In *Proceedings of fonetik*, volume 2003, 93–96. Citeseer.

Valle, R.; Li, J.; Prenger, R.; and Catanzaro, B. 2020. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6189–6193. IEEE.

Vaswani, S.; Bachem, O.; Totaro, S.; Müller, R.; Garg, S.; Geist, M.; Machado, M. C.; Castro, P. S.; and Le Roux, N. 2022. A general class of surrogate functions for stable and efficient reinforcement learning. In *Proceedings of AISTATS*, 8619–8649.

Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; et al. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proceedings of INTERSPEECH*, 4006–4010.

Wang, Y.; Wang, X.; Zhu, P.; Wu, J.; Li, H.; Xue, H.; Zhang, Y.; Xie, L.; and Bi, M. 2022. Opencpop: A High-Quality Open Source Chinese Popular Song Corpus for Singing Voice Synthesis. arXiv:2201.07429.

Weiss, R. J.; Skerry-Ryan, R.; Battenberg, E.; Mariooryad, S.; and Kingma, D. P. 2021. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5679–5683. IEEE.

Xue, H.; Wang, X.; Zhang, Y.; Xie, L.; Zhu, P.; and Bi, M. 2022. Learn2Sing 2.0: Diffusion and Mutual Information-Based Target Speaker SVS by Learning from Singing Teacher. In *Proc. Interspeech 2022*, 4267–4271.

Xue, H.; Yang, S.; Lei, Y.; Xie, L.; and Li, X. 2021. Learn2sing: Target speaker singing voice synthesis by learning from a singing teacher. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, 522–529. IEEE.

Zhang, L.; Yu, C.; Lu, H.; Weng, C.; Zhang, C.; Wu, Y.; Xie, X.; Li, Z.; and Yu, D. 2020. DurIAN-SC: Duration Informed Attention Network Based Singing Voice Conversion System. In *Proceedings of INTERSPEECH*, 1231–1235.

Zhang, Y.; Cong, J.; Xue, H.; Xie, L.; Zhu, P.; and Bi, M. 2022. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7237–7241. IEEE.

Zhao, J.; Kim, Y.; Zhang, K.; Rush, A.; and LeCun, Y. 2018. Adversarially regularized autoencoders. In *Proceedings of International conference on machine learning*, 5902–5911. PMLR.