

NNSPEECH: SPEAKER-GUIDED CONDITIONAL VARIATIONAL AUTOENCODER FOR ZERO-SHOT MULTI-SPEAKER TEXT-TO-SPEECH

Botao Zhao^{1,2}, Xulong Zhang¹, Jianzong Wang^{1*}, Ning Cheng¹, Jing Xiao¹

¹Ping An Technology (Shenzhen) Co., Ltd., China

²Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, China

ABSTRACT

Multi-speaker text-to-speech (TTS) using a few adaption data is a challenge in practical applications. To address that, we propose a zero-shot multi-speaker TTS, named nnSpeech, that could synthesis a new speaker voice without fine-tuning and using only one adaption utterance. Compared with using a speaker representation module to extract the characteristics of new speakers, our method bases on a speaker-guided conditional variational autoencoder and can generate a variable Z , which contains both speaker characteristics and content information. The latent variable Z distribution is approximated by another variable conditioned on reference mel-spectrogram and phoneme. Experiments on the English corpus, Mandarin corpus, and cross-dataset proves that our model could generate natural and similar speech with only one adaption speech.

Index Terms— zero-shot, multi-speaker text-to-speech, conditional variational autoencoder

1. INTRODUCTION

Recently, with the great success of neural network based text-to-speech (TTS) techniques, such as Tacotron [1], FastSpeech 1/2/s [2, 3], and Hifi-GAN [4] like neural vocoder, we can efficiently synthesize more natural voice. The single-speaker TTS model could be extended to multi-speaker scenarios using multi-speaker corpora [2, 5]. However, these models are only applied to the fixed set of speakers and the training requires a certain amount of human recordings.

Multi-speaker TTS for unseen speaker has many applications, such as news broadcast, personal assistant and meta-verse. Several studies have been done in this field and one popular solution is using few data to fine-tune the base model that was trained by a large amount of data [6, 7]. It achieves great voice quality when fine-tuning the whole model [8] or decoder [9]. There are also many studies only fine-tuning the speaker embedding [10]. However, fine-tuning based methods have two distinctive challenges: 1) Fine-tuning the model means the memory storage and serving cost increasing because that we have to train many new models for each speaker.

2). Fine-tuning cost a huge amount of computing when we have a large number of users. Zero-shot multi-speaker TTS has become a popular field in recent years [11, 12, 13], which can synthesize voices directly by the trained base model. In some studies, pre-trained speaker verification models are used to extract speaker information [14]. We can also optimize the speaker encoder and TTS system jointly [15] to make the speaker representation more optimal for the TTS system. But it is hard to extract the speaker representation to guide the TTS system and this could decrease voice quality or similarity. Zero-shot TTS needs training a generic model that could perform well on unseen data. However, due to the amount of data and the high dimension of the speaker vector, we can not estimate the distribution of speaker representation correctly. A solution is to use more adaption voices to get the averaged speaker representations [12]. But there are some applications requiring TTS model generate speech using a few adaption samples, such as personalized speech generation.

To solve the above problem, we introduce conditional variational autoencoder (CVAE) to this task. CVAE [16, 17] is one of the most popular conditional directed graphical models whose input observations modulate the prior on Gaussian latent variables that generate the outputs. In this paper, we propose a new method using both CVAE and the prior knowledge [17], speaker embeddings, for the zero-shot multi-speaker TTS. Our model is named of **nnSpeech** (no new speech) because our method can synthesize new voices directly instead of fine-tuning a new model. In our TTS system, the phoneme is the condition. We assume that the distribution of latent variable Z with the conditions of phoneme and speaker in multi-speaker TTS model could be approximated by the another variable conditioned on a reference mel-spectrogram and phoneme. To better training CVAE, we modified the standard CVAE and denotes it as speaker-guided CVAE (sgCVAE).

2. METHOD

2.1. VAE and CVAE

VAEs [18] are one of the auto-encoder neural network architectures. The encoder network denotes the conditional

* Corresponding author: Jianzong Wang, jzwang@188.com

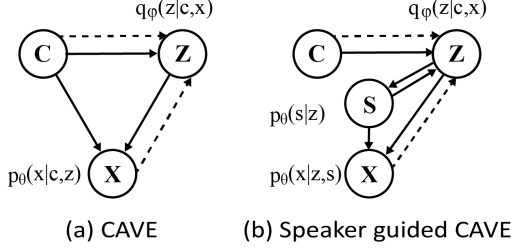


Fig. 1. Graphical models of CVAE and Speaker-Guided CVAE (sgCVAE).

distribution $q_\phi(Z|X)$ for the latent space variable Z given input data X and the decoder network means the distribution $p_\theta(X|Z)$ for the generated data X given input vector Z . VAEs can learn the parameters from training data to make the $q_\phi(Z|X)$ consistent with the $p_\theta(Z|X) \propto p_\theta(X|Z)p(Z)$. The log marginal distribution of data X can be lower-bounded by the evidence lower bound (ELBO) as followed:

$$\mathcal{L}(\theta, \phi; X) = \mathbb{E}_{q_\phi(Z|X)}[\log(p_\theta(X|Z))] - KL[q_\phi(Z|X)||p(Z)], \quad (1)$$

where the $KL(q_\phi||p_\theta)$ denotes the Kullback-Leibler (KL) divergence that the minimum value is zero when the two distributions are consistent. So we can minimize the KL divergence between $p_\theta(Z|X)$ and $q_\phi(Z|X)$ by maximizing the ELBO, which can be further reduced to formula Eq.1. One of the typical way to modeling $p_\theta(Z|X)$ and $q_\phi(Z|X)$ is to assume normal distributions with a diagonal covariance matrix. As for the prior distribution $p(Z)$, it can be designed as a specific form according to the assumption that we need. The first term of Eq.1 is the autoencoder reconstruction error, and the second term is the KL divergence $q_\phi(Z|X) = \mathcal{N}(Z|\mu_\phi(x), \sigma_\phi(x))$ and $p(Z) = \mathcal{N}(Z|0, I)$. Using a reparameterization $z = \mu_\phi(x) + \sigma_\phi \odot \epsilon, \epsilon \sim \mathcal{N}(\epsilon|0, I)$, we can sample ϵ from its distribution to generate z instead of sampling from $q_\phi(Z|X)$ directly, so that we can compute the gradient of the lower bound with respect to θ .

Conditional VAEs are the extended version of VAEs and are trained to maximize the conditional log-likelihood of X given C . As shown in Fig. 1(a), the variational lower bound to be maximized becomes

$$\begin{aligned} \mathcal{L}(\theta, \phi; X, C) &= \mathbb{E}_{q_\phi(Z|X, C)}[\log(p_\theta(X|Z, C))] \\ &\quad - KL[q_\phi(Z|X, C)||p_\theta(Z|C)] \\ &\leq \log p_\theta(X|C). \end{aligned} \quad (2)$$

Be similar to VAEs, the prior distribution $p(Z|C)$ can be designed for the requirements. When we assume the latent Z is independent with C , the $p_\theta(Z|C) = p(Z)$. As for the neural network architecture, the differences between VAE and CVAE are that the encoder and decoder networks can take an auxiliary variable C as an additional input.

2.2. Speaker-guided CVAE for Zero-shot Multi-speaker TTS

Most multi-speaker TTS model consists of a phoneme encoder to encode the phoneme information to a latent variable Z , a decoder to generate the mel-spectrogram from Z and a speaker representation added to Z to control the timbre. Speaker representation can be a jointly-optimized looked-up embedding table or pretrained speaker verification system.

Based on the multi-speaker model, we model the multi-speaker TTS using CVAE. The CVAE for multi-speaker TTS is composed of mel-spectrogram, phoneme, and the latent variable. To better fuse the features, we define X, C simply as the output of mel encoder and phoneme encoder. The mel encoder utilizes the speaker encoder of AdaIN-VC [19]. The phoneme encoder and mel decoder are based on FastSpeech2 [3].

We define the conditional distribution $p(X, Z|C) = p_\theta(X|C, Z)p_\theta(Z|C)$, and $p_\theta(X|C, Z)$ and $p_\theta(Z|C)$ are modeled by the mel decoder network and the prior encoder network. As shown in Fig.1(a), we use a $q_\phi(Z|C, X)$ to approximate the true posterior $p_\theta(Z|C)$. But this assumption is counterintuitive, because the latent variable Z should consist the speaker information if we want reconstruct the mel-spectrogram. So we propose the speaker-guided CVAE (as shown in Fig.1(b) and Fig.2(b)) that assume the prior of latent Z is based on the speaker representation S , $p_\theta(Z|C) = p_\theta(Z|C, S)$, and then use $q_\phi(Z|C, X)$ to approximate $p_\theta(Z|C, S)$. We assume that the latent Z contains the whole information of C , so $p_\theta(X|C, Z, S) = p_\theta(X|Z, S)$. In speaker-guided CVAE, we modify the conditional distribution to

$$p(X, Z|C, S) = p_\theta(X|Z, S)p_\theta(Z|C, S).$$

We hypothesise the latent variable z follows the Gaussian distribution with a diagonal covariance matrix, so $q_\phi(Z|C, X) \sim \mathcal{N}(\mu_1, \sigma_1^2 \mathbf{I})$ and $p_\theta(Z|C, S) \sim \mathcal{N}(\mu_2, \sigma_2^2 \mathbf{I})$. Then the $q_\phi(Z|C, X)$, named recognition network, and the $p_\theta(Z|C, S)$ named prior network are represented as:

$$\begin{aligned} \begin{bmatrix} \mu_1 \\ \ln \log \sigma_1^2 \end{bmatrix} &= \mathbf{MLP}_{q_\phi} \left(\begin{bmatrix} X \\ C \end{bmatrix} \right), \\ \begin{bmatrix} \mu_2 \\ \ln \log \sigma_2^2 \end{bmatrix} &= \mathbf{MLP}_{p_\theta} \left(\begin{bmatrix} C \\ S \end{bmatrix} \right). \end{aligned} \quad (3)$$

The reparameterization trick is used to sample Z_1, Z_2 from the distribution $\mathcal{N}(\mu_1, \sigma_1^2 \mathbf{I})$ and $\mathcal{N}(\mu_2, \sigma_2^2 \mathbf{I})$. Then, we can generate mel-spectrogram X based on $p_\theta(X|Z, S)$, which is represent by mel decoder network. Additionally, We utilize a network to predict $\hat{S} = \mathbf{MLP}_S(Z)$. In the inference step, the speaker embedding S will be replaced by predicted \hat{S} . Finally, we can use Hifi-GAN to generate voice [4].

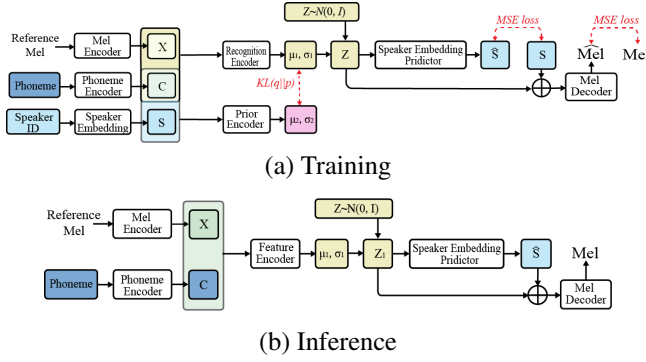


Fig. 2. Neural network architectures of nnSpeech.

2.3. Loss Functions

nnSpeech model is trained by maximizing the modified ELBO:

$$\begin{aligned} \mathcal{L}(\theta, \phi; X, C, S) = & \mathbb{E}_{q_\phi(Z|X, C)}[\log(p_\theta(X|Z, S))] \\ & + \mathbb{E}_{q_\phi(Z|X, C)}[\log(p_\theta(S|Z))] \\ & - KL[q_\phi(Z|X, C) || p_\theta(Z|C, S)], \end{aligned} \quad (4)$$

where the first two terms are the expectation of generated mel-spectrogram and speaker embedding. Maximizing them means minimizing the mean square error of the estimations and the coefficient of the third term, KL divergence, is negative. Therefore, Maximizing the ELBO can be implemented by minimizing three loss functions, mel loss, speaker loss, and KL loss. The mel loss and speaker loss are defined as:

$$\begin{aligned} \mathcal{L}_{mel} &= \mathbb{E}[\|X - \hat{X}\|_2], \\ \mathcal{L}_{spk} &= \mathbb{E}[\|S - \hat{S}\|_2], \end{aligned} \quad (5)$$

where the X means mel-spectrogram and S denotes speaker embedding vector. The KL loss is calculated by

$$\begin{aligned} \mathcal{L}_{kl} &= \frac{1}{2} \log \frac{|\Sigma_2|}{|\Sigma_1|} - n + \text{tr}(\Sigma_2^{-1} \Sigma_1) \\ &+ (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2), \end{aligned} \quad (6)$$

where the μ_1, Σ_1 and μ_2, Σ_2 denotes the Gaussian distribution p_θ and q_ϕ . Then the total loss is:

$$\mathcal{L} = \alpha * \mathcal{L}_{mel} + \beta * \mathcal{L}_{spk} + \gamma * \mathcal{L}_{kl} \quad (7)$$

α and β are set as 1, and γ is the hyper-parameter of our model that will be analyzed in the following experiments. Besides, our model has duration loss, energy loss, and pitch loss like FastSpeech2.

3. EXPERIMENTS

3.1. Experimental Setup

We trained our model on a multi-speaker English corpus (903 speakers, 191.29 hours, a subset 'train-clean-360' of the

whole LibriTTS [20]), and a multi-speaker Mandarin corpus (218 speakers, 85 hours, AISHELL-3 [21]). We randomly chose several male and female speakers from both the LibriTTS and AISHELL3 for evaluation. The only single speaker dataset LJSpeech [22] was used to do a cross-dataset evaluation.

The speech data were resampled to 22050Hz and then extracted the mel-spectrogram using 256 hop size and 1024 window size for all datasets. We also calculated the duration, pitch, and energy for variance adaptor modular as the same as FastSpeech2 [3].

In the training stage, we first trained the model without speaker embedding predictor module 100,000 steps, and then trained the whole model without speaker embedding module 500,000 steps on NVIDIA V100 GPU with the batch size 16. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$. In the testing stage, one speech was randomly selected as the reference for each speaker.

3.2. Evaluation

We evaluated our model both on subjective and objective tests. For subjective test, we focused on the adaption voice quality, which consists of the voice naturalness mean opinion score (MOS), and the voice similarity score (VSS), as shown in Table 1. Fourteen listeners judged each sentence, and the averaged MOS scores and VSS score were the final result. In addition, we also calculated the parameters that needs to be stored and the consumed time of fine-tune based methods. As for the adaption utterances, 20 sentences were used for retraining 2,000 steps for fine-tune-based methods, and one same reference utterance was selected randomly for the one-shot multi-speaker methods. Besides, we calculated the mel-cepstral distortion (MCD) score as the objective test.

We compared our method with fine-tuning the different parts of a multi-speaker TTS model. It could be found that fine-tuning the decoder could achieve a much better performance than fine-tuning the speaker embedding and zero-shot methods. For zero-shot multi-speaker methods, we compared our model with two FastSpeech2-based methods: 1). Using the x-vector [23] as the speaker embedding in FastSpeech2; 2). StyleSpeech [13]. Our method outperformed the x-vector in three datasets and got better performance than StyleSpeech both on the MOS and VSS on AISHELL3 and LibriTTS. However, our model performed poorly under cross-dataset LJSpeech. Both fine-tuning decoder and using the style adaptive layer normalization (StyleSpeech) perform better than fine-tuning speaker embedding and using x-vector.

3.3. Method Analysis

In this section, we first studied the voice quality with different adaption data. Next, we analyzed the hyper-parameter γ , the weight of KL divergence loss. In addition, we conduct the ablation studies.

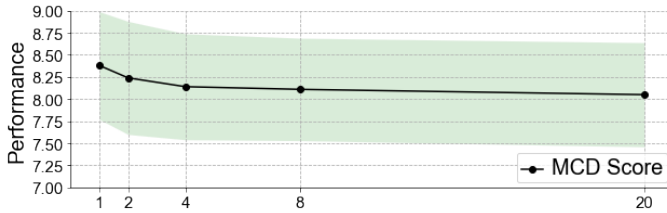
Table 1. Subjective Comparison of the fine-tune based TTS methods

Methods	Params/spk	Time/spk	MOS			VSS		
			LibriTTS	AISHELL3	LJSpeech	LibriTTS	AISHELL3	LJSpeech
GT	–	–	4.73 \pm 0.49	4.60 \pm 0.60	4.75 \pm 0.60	–	–	–
GT+Vocoder	–	–	4.48 \pm 0.69	4.52 \pm 0.68	4.71 \pm 0.54	4.84 \pm 0.68	4.79 \pm 0.57	4.92 \pm 0.43
Finetune spk emb	256	1410 \pm 396	3.45 \pm 0.92	4.10 \pm 0.87	3.33 \pm 0.90	3.56 \pm 1.03	3.91 \pm 0.97	2.88 \pm 1.00
Finetune decoder	14 M	1690 \pm 526	3.48 \pm 0.88	4.25 \pm 0.83	3.39 \pm 0.91	4.03 \pm 0.98	4.68 \pm 0.77	4.00 \pm 0.87
X-vector [23]	–	–	3.27 \pm 0.85	3.95 \pm 0.84	3.03 \pm 0.94	3.30 \pm 0.81	3.86 \pm 0.82	2.21 \pm 0.88
StyleSpeech [13]	–	–	3.33 \pm 0.91	4.07 \pm 0.91	3.37 \pm 0.90	3.36 \pm 0.98	3.97 \pm 0.76	3.24 \pm 0.78
nnSpeech	–	–	3.42 \pm 0.92	4.13 \pm 0.86	3.13 \pm 0.98	3.46 \pm 0.93	4.17 \pm 0.73	2.65 \pm 0.89

Table 2. Objective Comparison of the fine-tune based methods (MCD)

Metric	LibriTTS	AISHELL3	LJSpeech
Finetune spk emb	7.55 \pm 0.64	8.29 \pm 0.65	7.65 \pm 0.66
Finetune decoder	7.18 \pm 0.57	7.74 \pm 0.57	7.05 \pm 0.50
X-vector	7.95 \pm 0.65	8.89 \pm 0.54	8.73 \pm 0.68
StyleSpeech	7.82 \pm 0.68	8.64 \pm 0.67	7.38 \pm 0.51
nnSpeech	7.44 \pm 0.59	8.38 \pm 0.53	8.27 \pm 0.54

MCD under vary adaption data. We studied the MCD score with different adaption data on the AISHELL3. For our model, we replaced the output of mel encoder X and predicted \hat{S} by their averaged vectors. It can be seen that the MCD score decreases slightly with the increase of the adaption data, as shown in Fig.3. Our method could get better performance when using four adaption utterances, which proves that our method could extracts the speaker information from a few adaption data.

**Fig. 3.** Performance with varying number of adaption voices.

Hyper-parameter analysis. As shown in Table 3, we analyzed the hyper-parameter γ on the dataset AISHELL3, and the model got the best performance when the γ was set as 0.0005 compared with the set as 0.005 and 0.05. That is because the KL divergence is always large than MSE loss, such as \mathcal{L}_{mel} and \mathcal{L}_{spk} . So the model will be hard to learn something from other losses if the weight of KL loss is large.

Analyses on the model architecture. We further compared the model architecture with two other settings: 1). Add the content information C to the latent variable Z , so our model will become the standard CVAE; 2). remove the speaker embedding predictor module and not add the pre-

Table 3. The MCD with different γ setting on AISHELL3

Metric	$\gamma=0.05$	$\gamma=0.005$	$\gamma=0.0005$
MCD	8.90 \pm 0.49	8.50 \pm 0.51	8.38 \pm 0.53

Table 4. The ablation experiment on AISHELL3

Metric	Content add	w/o spk pred	nnSpeech
MCD	8.83 \pm 0.45	8.41 \pm 0.63	8.38 \pm 0.53

dicted speaker embedding C to the latent Z . As shown in Table 4, we found that both settings get the higher MCD than our model. For setting 1, we believe that the latent variable Z involves the content information and speaker information, so adding with the content C again is not necessary and maybe cover up the speaker information. For setting 2, the result proves that the speaker predictor module is helpful.

4. CONCLUSION

We proposed the speaker-guided CVAE for zero-shot multi-speaker TTS named nnSpeech. Our method could improve the generalization of the zero-shot TTS model. It performs excellent when only using one adaption voice both in the English and Chinese corpus. We have done some model analysis to find the best hyper-parameter and proves the our model is robust and each module is helpful. However, our model performs poorly on the cross-dataset, and there is still a gap between zero-shot methods and fine-tune methods. For future work, we will focus on the cross-dataset problem. The result in section 3.2 shows using adaptive layer normalization to modulate the decoder is a possible solution.

5. ACKNOWLEDGEMENT

This paper is supported by the Key Research and Development Program of Guangdong Province No. 2021B0101400003 and the National Key Research and Development Program of China under grant No. 2018YFB0204403.

6. REFERENCES

- [1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., “Tacotron: Towards end-to-end speech synthesis,” *INTERSPEECH*, pp. 4006–4010, 2017.
- [2] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “Fastspeech: Fast, robust and controllable text to speech,” in *Advances in Neural Information Processing Systems*, 2019, pp. 3165–3174.
- [3] 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 *ICLR*, 2020.
- [4] 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*, 2020.
- [5] Xulong Zhang, Jianzong Wang, Ning Cheng, Edward Xiao, and Jing Xiao, “CycleGEAN: cycle generative enhanced adversarial network for voice conversion,” in *ASRU*. 2021, pp. 1–6, IEEE.
- [6] Jian Cong, Shan Yang, Lei Xie, Guoqiao Yu, and Guanglu Wan, “Data efficient voice cloning from noisy samples with domain adversarial training,” in *INTERSPEECH*, 2020, pp. 811–815.
- [7] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu, “Adaspeech: Adaptive text to speech for custom voice,” in *ICLR*, 2021.
- [8] Zvi Kons, Slava Shechtman, Alex Sorin, Carmel Rabinovitz, and Ron Hoory, “High quality, lightweight and adaptable tts using lpcnet,” in *INTERSPEECH*, 2019, pp. 176–180.
- [9] Henry B Moss, Vatsal Aggarwal, Nishant Prateek, Javier González, and Roberto Barra-Chicote, “Boffin tts: Few-shot speaker adaptation by bayesian optimization,” in *ICASSP*, 2020, pp. 7639–7643.
- [10] Serkan O Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou, “Neural voice cloning with a few samples,” in *Advances in Neural Information Processing Systems*, 2018.
- [11] Tao Wang, Jianhua Tao, Ruibo Fu, Jiangyan Yi, Zhengqi Wen, and Chunyu Qiang, “Bi-level speaker supervision for one-shot speech synthesis,” in *INTERSPEECH*, 2020, pp. 3989–3993.
- [12] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi, “Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings,” in *ICASSP*, 2020, pp. 6184–6188.
- [13] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang, “Meta-stylespeech: Multi-speaker adaptive text-to-speech generation,” *ICML*, pp. 7748–7759, 2021.
- [14] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al., “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” in *Advances in Neural Information Processing Systems*, 2018.
- [15] Mengnan Chen, Minchuan Chen, Shuang Liang, Jun Ma, Lei Chen, Shaojun Wang, and Jing Xiao, “Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding,” in *INTERSPEECH*, 2019, pp. 2105–2109.
- [16] Kihyuk Sohn, Honglak Lee, and Xinchen Yan, “Learning structured output representation using deep conditional generative models,” in *Advances in neural information processing systems*, 2015.
- [17] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi, “Learning discourse-level diversity for neural dialog models using conditional variational autoencoders,” in *ACL*, 2017, pp. 654–664.
- [18] Diederik P Kingma and Max Welling, “Stochastic gradient vb and the variational auto-encoder,” in *ICLR*, 2017, p. 121.
- [19] Ju-chieh Chou, Cheng-chieh Yeh, and Hung-yi Lee, “One-shot voice conversion by separating speaker and content representations with instance normalization,” in *INTERSPEECH*, 2019, pp. 664–668.
- [20] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, “Libri-tts: A corpus derived from librispeech for text-to-speech,” in *INTERSPEECH*, 2019, pp. 1526–1530.
- [21] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, “Aishell-3: A multi-speaker mandarin tts corpus and the baselines,” *arXiv preprint arXiv:2010.11567*, 2020.
- [22] Keith Ito and Linda Johnson, “The lj speech dataset,” 2017.
- [23] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *ICASSP*, 2018, pp. 5329–5333.