USING MULTIPLE REFERENCE AUDIOS AND STYLE EMBEDDING CONSTRAINTS FOR SPEECH SYNTHESIS

Cheng Gong[1], *Longbiao Wang*[1*], *Zhenhua Ling*[2], *Ju Zhang*[3], *Jianwu Dang*[1,4]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]National Engineering Laboratory for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
[3]Huiyan Technology (Tianjin) Co., Ltd., China
[4]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{gongchengcheng, longbiao_wang}@tju.edu.cn, zhling@ustc.edu.cn

ABSTRACT

The end-to-end speech synthesis model can directly take an utterance as reference audio, and generate speech from the text with prosody and speaker characteristics similar to the reference audio. However, an appropriate acoustic embedding must be manually selected during inference. Due to the fact that only the matched text and speech are used in the training process, using unmatched text and speech for inference would cause the model to synthesize speech with low content quality. In this study, we propose to mitigate these two problems by using multiple reference audios and style embedding constraints rather than using only the target audio. Multiple reference audios are automatically selected using the sentence similarity determined by Bidirectional Encoder Representations from Transformers (BERT). In addition, we use "target" style embedding from a pre-trained encoder as a constraint by considering the mutual information between the predicted and "target" style embedding. The experimental results show that the proposed model can improve the speech naturalness and content quality with multiple reference audios and can also outperform the baseline model in ABX preference tests of style similarity.

Index Terms— Multiple references, style, naturalness, mutual information, content

1. INTRODUCTION

The goal of text-to-speech (TTS) is to synthesize intelligible and natural speech. Recent advances TTS have significantly improved the naturalness of synthetic speech [1, 2, 3, 4]. Naturalness largely depends on the expressiveness of the synthesized voice, which is determined by multiple characteristics, such as content, timbre, prosody, emotion, and style [5]. Despite having important applications, such as conversational assistants and long-form reading, the development of expressive TTS is still considered an important open problem.

To deliver true human-like speech, a TTS system must learn to model prosody. Researchers have addressed this problem by providing an additional reference speech signal to control the style of the generated speech [6, 7, 8, 9, 10]. But these methods still have two problems when synthesizing a sample: 1) one has to manually select an appropriate acoustic embedding, which can be challenging, and 2) during the training phase, the reference utterance and the input text are paired (i.e., the text is the transcription of the reference utterance). However, the input reference utterance and text are not paired during testing. Because the unpaired inputs are never seen during training, the generated voice has low speech naturalness and speech content quality.

One solution for prosody selection could be to remove the manually selection process by predicting the style from the text. In [11], the Text-Predicted Global Style Token (TP-GST) architecture learned to predict stylistic renderings from text alone, requiring neither explicit labels during training nor auxiliary inputs for inference. Similarly, Karlapati et al. [12] proposed a method to sample from this learned prosodic distribution using the contextual information available in the text. However, it is difficult to predict the latent style from text alone because the process of style embedding is an entangled representation of prosody and unknown underlying acoustic factors. Instead of using only text, Tyagi et al. [13] proposed an approach that leverages linguistic information to drive the style embedding selection of such systems. In this method, a sentence is selected as the reference audio if its linguistic similarity is similar to that of the target. However, only one audio, rather than multiple audios, was selected for style embedding.

Moreover, if the network architecture is not carefully designed, the generated voice is influenced by the content of the reference utterance. To tackle this content leakage problem, previous studies have proposed several methods that employ unpaired data training. For example, Liu et al. [14] proposed to mitigate the problem by using the unmatched text and speech during training, utilizing the automatic speech recognition (ASR) accuracy of an end-to-end ASR model to guide the training procedure. In addition, using the unpaired data, Ma et al. [15] introduced an end-to-end TTS model by combining a pairwise training procedure, an adversarial game, and a collaborative game into one training scheme.

In addition to methods that use unpaired data, a number of recently proposed methods improve the content quality of synthesized speech by adding constraints to the style embeddings. Ideally, the style embedding should not be able to reconstruct the content vector (i.e., there should be no information about the content in the style embedding). To this end, Hu et al. [16] minimized the mutual information (MI) between the style and content vectors. In addition, MI has been applied to other speech synthesis tasks, such as learning disentangled representations of the speakers and the languages they

* Corresponding Author.

speak [17], and maximizing the mutual information between the text and acoustic features to strengthen the dependency between them [18]. In this study, we also design a constraint for style embedding using mutual information.

Combining the aforementioned ideas of using text information to predict style embeddings and using unpaired data to improve speech content quality, we propose multi-reference TTS (MRTTS). This model is trained using multiple automatically selected and weighted reference audio to generate expressive speech. The contributions of this paper are as follows. *I)* We present an approach that uses automatically selected and weighted multi-reference audios for speech style modeling. *II)* For the proposed method, we design a constraint for learned style embeddings using mutual information. *III)* To the best of our knowledge, this work is the first attempt to improve expressive speech quality using multiple reference audios rather than only one target audio. The objective and subjective evaluation results demonstrate that MRTTS exhibits superior performance compared to the baseline model in terms of speech naturalness, speech content quality, and style similarity.

2. PROPOSED MRTTS MODEL

The proposed method, shown in Fig. 1, is based on the end-to-end TTS architecture Tacotron2. For the text encoder input x, we used the character sequence of the normalized text for training. All the style encoders learned to model prosody from one or multiple reference audio with the same settings as the Global Style Token (GST) module. Multiple N reference audios were selected based on the corresponding text similarity as determined by BERT. For multiple style embeddings obtained from multiple reference audios, we used the scaled attention method to obtain the final style embedding.

2.1. Linguistics-driven multiple references selection

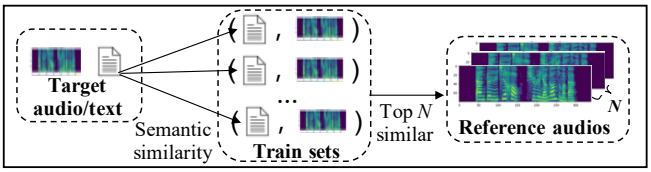
The objective of this selection step is to select reference audios from training sets for training and inference. Unlike GST, which uses the target audio as the reference audio, we selected the reference audio to be the audio that has a semantic similar to the target, as in [13]. We defined the semantic distances as the cosine similarity between the two sentence-level embeddings, and we selected the reference audio to be the audio for which the semantic distance is closest to the target audio, as shown in Fig. 1(a).

To generate sentence embedding, we used BERT [19, 20] because it is one of the best pre-trained models and it produces state-of-the-art results on a large number of NLP tasks. We used word representations from the uncased base (12-layer) model without fine tuning. Sentence-level representations were achieved by averaging the second to last hidden layers for each token in the sentence. We did not use "[CLS]" because it acts as an "aggregate representation" for classification tasks and it is not the best choice for quality sentence embeddings.

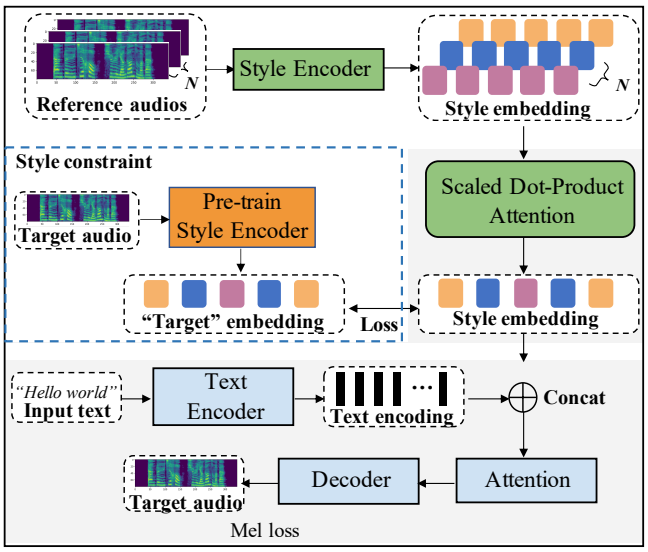
2.2. Multi-references based acoustic modeling

From the selected N reference audios, the style encoder obtains N style embedding $S = [s_1, s_2, ..., s_n]$. The attention query Q, key K, value V, and final style embedding E are calculated as

$$\boldsymbol{Q} = \boldsymbol{Q}^{'}\boldsymbol{W_q}, \quad \boldsymbol{K} = \boldsymbol{S}\boldsymbol{W_k}, \quad \boldsymbol{V} = \boldsymbol{S}\boldsymbol{W_v},$$
$$\boldsymbol{E} = \mathrm{A}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{softmax}(\frac{f(\boldsymbol{Q})f(\boldsymbol{K})^T}{\sqrt{d_m}})f(\boldsymbol{V}), \quad (1)$$



(a) Linguistics-driven multiple reference selection.



(b) Model diagram. The blue dashed box means style constraint module which uses a pre-train style encoder to constrain the style embedding during the training process.

Fig. 1. Proposed MRTTS model.

respectively, where $\boldsymbol{Q}^{'}$ is an embedding in which the values are randomly initialized and automatically learned by backpropagation, and all W are linear projection matrices.

Here, attention is not used to learn the alignment. Instead, it learns a similarity measure between the $\boldsymbol{Q}^{'}$ vector and each style embedding s_i in S. The weighted sum of the style embedding S, which we call the final style embedding E, is passed onto the text encoder outputs for conditioning in every time step.

Without the style embedding constraints, as shown in Fig. 1(b), the style encoder is jointly trained with the rest of the model, driven only by the Mel reconstruction loss $\mathcal{L}_{mel}$ from the Tacotron2 decoder, as in [7]. Thus, the style encoder does not require any explicit style or prosody labels.

2.3. Model training with style embedding constraints

The modeling style in TTS is somewhat underdetermined, and training models with reconstruction loss alone are insufficient to disentangle content and style from other factors of variation.

The pathway of the style embedding constraints, which are represented by the blue dashed box in Fig. 1 (b), is trained using a loss between the predicted and "target" style embedding. An additional pre-trained encoder is used to extract the "target" embedding in the training process. Therefore, the first step is the style encoder pre-training, which can be simply treated as a neural GST training process. The pre-trained style encoder is trained using the same training sets. As in the case of TPCE-GST [11], we stop the gradient flow to ensure that the style prediction error does not backpropagate through the pre-trained encoder.

Algorithm 1 Pseudocode for MI estimator training

Input: Pairs of predicted and target embeddings $(E_i, E_i^{'})$.
Output: M

$M \leftarrow$ initialization with random weights.
while M not converged **do**
  Sample a mini-batch of $(E_i, E_i^{'}), i = 1, 2, ..b$.
  $E_i^{'}$ = random permutation of $E_i^{'}$.
  $\mathcal{L}_{mi} = \frac{1}{b}\sum_{i=1}^{b} M(E_i, E_i^{'}) - log(\frac{1}{b}\sum_{i=1}^{b} e^{M(E_i, E_i^{'})})$
  $M = M + \epsilon\Delta_M\mathcal{L}_{mi}$
end while

The total loss with style embedding constraints is defined as

$$\begin{aligned} \mathcal{L}_{total} &= \mathcal{L}_{mel} + \mathcal{L}_s, \\ \mathcal{L}_s &= \text{MSE}(\boldsymbol{E}, \boldsymbol{E}^{'}) - \text{MI}(\boldsymbol{E}, \boldsymbol{E}^{'}), \end{aligned} \tag{2}$$

where $\mathcal{L}_s$ is the loss between the predicted style embedding E and the target embedding $E^{'}$. Not only do predict and "target" tend to have the same value with low MSE$(\boldsymbol{E}, \boldsymbol{E}^{'})$, but they also need to have strong mutual information MI$(\boldsymbol{E}, \boldsymbol{E}^{'})$.

Mutual information (MI) measures the dependence of two random variables from the perspective of information [16, 21]. Given two random variables X and Y, the MI $I(X;Y)$ between them is equivalent to the Kullback–Leibler (KL) divergence between their joint distribution, $P_{X,Y}$, and the product of marginals, $P_X P_Y$.

The MI neural estimation (MINE) [21] method constructs a lower bound of MI based on the Donsker-Varadhan representation of KL divergence via

$$I(X;Y) \geq \hat{I}_M(X;Y) = \sup_M E_{P_{X,Y}}[M] - log(E_{P_X P_Y}[e^M]), \tag{3}$$

where M can be any function that forces the two expectations in the above equation to be finite. The authors in [16] proposed the use of a deep neural network for M, which enables the MI between X and Y to be estimated by maximizing the lower bound in Eq. 3 with respect to M using gradient descent.

The MI estimator function, M, is updated in each step of the training. The predicted style embeddings, E, and the "target" embeddings, E', are used to train the MI estimator M. Then, the MI of E and $E^{'}$, MI$(E, E^{'})$, can be estimated using the MI estimator M. The MI estimator training process is summarized in Algorithm 1.

3. EXPERIMENTS

3.1. Experimental Setup

Datasets. Experiments were performed on a high-quality dataset of English audiobook recordings featuring the voice of Catherine Byers (the speaker from the 2013 Blizzard Challenge). Some books contain very expressive character voices with high dynamic ranges, which are challenging to model. We selected five fairy tale books, and approximately 13,000 utterances were used for training and validation. All speech data were downsampled to 22050 Hz. We trained an open-source WaveRNN [22] vocoder using the same data to reconstruct waveforms from the Mel spectrogram.

Method. For the experiment, we built several TTS systems, as listed in Table 1. We chose the vanilla Tacotron2 and TPSE-GST as our baseline systems (B1-B4) because GST-Tacotron requires either a reference signal or a manual selection of style token weights at the time of inference. Because the original TPSE-GST B2 only uses the mean square error (MSE) constraint, we also modified the TPSE model and added MI loss to it, which is B3-B4.

Table 1. TTS systems used for our analysis. TPSE and Tacotron2 are baseline models, U-MRTTS and C-MRTTS represent the without or with style embedding constraint respectively. N means the number of multiple reference audios

Method	Architecture	Attention	Constraint		N
			MSE	MI	
B1	Tacotron2	N/A	N/A	N/A	N/A
B2	TPSE-GST	N/A	✓	N/A	N/A
B3		N/A	N/A	✓	N/A
B4		N/A	✓	✓	N/A
P1	U-MRTTS	N/A	N/A	N/A	3
P2		✓	N/A	N/A	1
P3		✓	N/A	N/A	3
P4	C-MRTTS	N/A	✓	N/A	3
P5		N/A	N/A	✓	3
P6		N/A	✓	✓	3
P7		✓	✓	N/A	3
P8		✓	N/A	✓	3
P9		✓	✓	✓	1
P10		✓	✓	✓	3

In addition, we included style embedding constrains C-MRTTS with and without MI loss to verify the effectiveness of the mutual information constraints. The MRTTS was also built with and without attention to investigate whether or not the attention model was necessary. The dimensionality of the style embeddings of all systems was 256.

Evaluation metrics. In terms of subjective evaluation, the mean opinion score (MOS) was calculated on a scale from 1 to 5 with 0.5-point increments. We also conducted an ABX test. The rating criterion was determined by answering the question "Which one's speaking style is closer to the target audio style?" with one of three choices: (1) the first is better, (2) the second is better, and (3) neutral. In all tests, 25 native listeners were asked to rate the performance of 50 randomly selected synthesized utterances from the test set.

Because another main objective of the MRTTS algorithm is to reduce the content leakage of the generated speech and to improve the speech content quality, we objectively evaluated the performance by measuring the content quality using an ASR algorithm like [16]. We adopted iFlytek's online API [1] as the ASR system, and computed the word error rate (WER) as a metric for the content preservation ability of the model. We used a PyPi package called jiwer [23] to calculate the WER.

3.2. Result and analysis

Objective evaluation of speech content quality. We present our WER results in the last column of Table 2, which shows that the proposed method produced a better WER than the baseline methods (a smaller WER indicates less content leakage). For the TPSE-GST model, the direct and only use of text to predict the style embedding easily caused the content information to be entangled into the style embedding. Because we used multiple related reference audios as input instead of only target audio during training, style embedding is unlikely to contain more textual information compared to the baseline. Thus, our P3 and P10 methods can synthesize high-content

[1] https://www.xfyun.cn/services/voicedictation

Table 2. Mean opinion score (MOS) evaluations with 95% confidence intervals computed from the t-distribution and WER for various systems.

Method	Architecture	MOS	WER
B1	Tacotron2	4.015 ± 0.023	28.2%
B2	TPSE-GST	4.175 ± 0.016	26.7%
B3		4.177 ± 0.022	26.3%
B4		4.183 ± 0.063	26.6%
P1	U-MRTTS	4.011 ± 0.074	27.1%
P2		4.103 ± 0.105	21.2%
P3		4.292 ± 0.035	18.2%
P4	C-MRTTS	3.911 ± 0.025	30.2%
P5		3.985 ± 0.025	29.3%
P6		3.997 ± 0.031	29.5%
P7		4.123 ± 0.015	18.7%
P8		4.285 ± 0.103	18.4%
P9		4.109 ± 0.127	19.1%
P10		$\mathbf{4.313 \pm 0.024}$	**17.9%**
GT	N/A	4.674 ± 0.013	15.6%

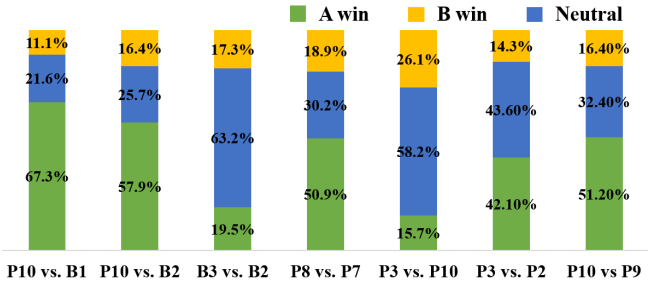


Fig. 2. ABX test results for style similarity.

quality speech and achieve a small WER in the ASR system. The results also show that there is no significant difference between the WER produced by P3 and P10. Therefore, we infer that the improvement in speech content quality is mainly due to unpaired data rather than style embedding constraints.

Subjective evaluation of speech naturalness. The results of the MOS test are presented in the third column of Table2. The proposed model P10 demonstrated a significantly better naturalness compared to the baseline models Tacotron2 and TPSE. This result shows the advantage of predicting style embeddings from multiple reference audios rather than from text. Without the attention mechanism, the synthesized speech quality was significantly lower, especially for the C-MRTTS style embedding constraint models (P4-P6). P7, P8, and P10 yielded better results. However, for the baseline TPSE model, B3 and B4 are similar to B2, and there is no significant improvement. On the one hand, MI is a better constraint for style embedding in our proposed model. On the other hand, MI does not bring significant gains to the TPSE model because MI constraints is limited by the style modeling ability of the TPSE model. Furthermore, we also found that if only one audio was used as a reference, as in the cases of P2 and P9, the result was worse than when multiple audios were used, as in the cases of P3 and P10.

Subjective evaluation of ABX test for style similarity. Fig. 2 shows the results of ABX test. As expected, a gap between our proposed model P10 and the baseline models B1 and B2 is visible. This shows that the proposed MRTTS model can produce better latent style representations, which results in better style similarity. The results also show that raters had a strong preference for P3 and P10.

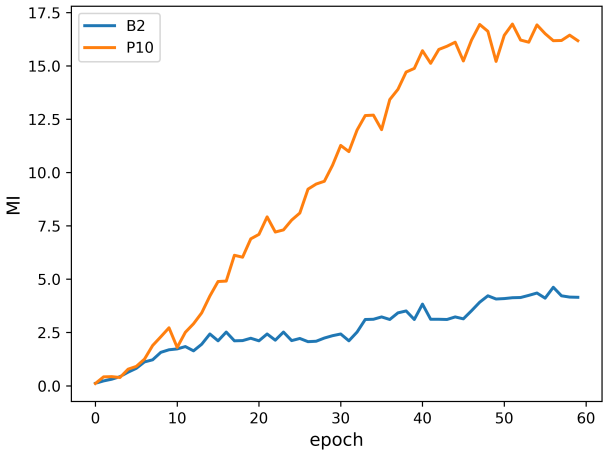


Fig. 3. Estimated MI value in the training process.

That means that the style constraint methods we proposed can perform well in style models. The results of the ABX test for the TPSE baseline model show that B3 was similar to B2, and there is no significant difference. However, comparing P7 and P8, the style embedding constraint with mutual information improved the style model performance. This proves once again that the use of MI cannot compensate for the shortcomings of the TPSE model. In addition, we also found that if only one audio was used as the reference, as in the cases of P2 and P9, the style similarity was worse than when multiple audios were used, as in the cases of P3 and P10. In other words, multiple audios contain richer style information than a single audio.

Mutual Information Evaluation. We estimate the mutual information between the two random variables (i.e. the predict style embedding and the "target" embedding) from our trained model (with frozen weights) using the MINE algorithm, and the result is illustrated in Fig. 3. As expected, P10 with style embedding constraint have much high MI value than the B2. We assume this is because the style embedding constraints in MI model may improve the mutual information between predict and "target" style embedding and consequently increase the MI value.

In summary, using multiple reference audios (rather than only the target audio) as input reference can more accurately model latent style representations. In addition, the style constraints of MI can improve the speech style similarity. Both objective and subjective evaluations showed that our proposed MRTTS can synthesize more intelligible and natural speech. We present synthetic samples at `https://gongchenghhu.github.io/ICASSP2022_demo/`.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a MRTTS model that uses multiple reference audios and style embedding constraints for the synthesis of expressive speech. The multiple reference audios were automatically selected using the sentence similarity determined by BERT. In addition, we considered the MI between style embeddings as constraints. The experimental results showed that the proposed model can improve speech naturalness and content quality, and that it can outperform the baseline model according to ABX preference tests of style similarity. In the future, we aim to build a fine-grained model that learns variable-length style information from multi-reference audios.

5. ACKNOWLEDGEMENTS

This work was supported by the the National Key R&D Program of China (Grant No. 2018YFB1305200) and the National Natural Science Foundation of China (Grant No. 62176182).

6. REFERENCES

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Zongheng Yang Jaitly, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, et al., "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6706–6713.

[4] Cheng Gong, Longbiao Wang, Ju Zhang, Shaotong Guo, Yuguang Wang, and Jianwu Dang, "TacoLPCNet: Fast and Stable TTS by Conditioning LPCNet on Mel Spectrogram Predictions," in *Proc. Interspeech 2021*, 2021, pp. 111–115.

[5] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[6] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[7] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[8] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018.

[9] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.

[10] Shuang Ma, Daniel Mcduff, and Yale Song, "A generative adversarial network for style modeling in a text-to-speech system," in *International Conference on Learning Representations*, 2019, vol. 2.

[11] Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[12] Sri Karlapati, Ammar Abbas, Zack Hodari, Alexis Moinet, Arnaud Joly, Penny Karanasou, and Thomas Drugman, "Prosodic representation learning and contextual sampling for neural text-to-speech," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6573–6577.

[13] Shubhi Tyagi, Marco Nicolis, Jonas Rohnke, Thomas Drugman, and Jaime Lorenzo-Trueba, "Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020.

[14] Da-Rong Liu, Chi-Yu Yang, Szu-Lin Wu, and Hung-Yi Lee, "Improving unsupervised style transfer in end-to-end speech synthesis with end-to-end speech recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 640–647.

[15] Shuang Ma, Daniel Mcduff, and Yale Song, "Neural tts stylization with adversarial and collaborative games," in *International Conference on Learning Representations*, 2018.

[16] Ting-Yao Hu, Ashish Shrivastava, Oncel Tuzel, and Chandra Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3267–3271.

[17] Detai Xin, Tatsuya Komatsu, Shinnosuke Takamichi, and Hiroshi Saruwatari, "Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual tts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6608–6612.

[18] Peng Liu, Xixin Wu, Shiyin Kang, Guangzhi Li, Dan Su, and Dong Yu, "Maximizing mutual information for tacotron," *arXiv preprint arXiv:1909.01145*, 2019.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT (1)*, 2019.

[20] Han Xiao, "bert-as-service," `https://github.com/hanxiao/bert-as-service`, 2018.

[21] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 531–540.

[22] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, Jennifer Dy and Andreas Krause, Eds. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 2410–2419, PMLR.

[23] jitsi, "Jiwer: Similarity measures for automatic speech recognition evaluation," `https://github.com/jitsi/jiwer`, 2019.