FINE-GRAINED STYLE CONTROL IN TRANSFORMER-BASED TEXT-TO-SPEECH SYNTHESIS

Li-Wei Chen, Alexander Rudnicky

Language Technologies Institute, Carnegie Mellon University
{liweiche, air}@cs.cmu.edu

ABSTRACT

In this paper, we present a novel architecture to realize fine-grained style control on the transformer-based text-to-speech synthesis (TransformerTTS). Specifically, we model the speaking style by extracting a time sequence of local style tokens (LST) from the reference speech. The existing content encoder in TransformerTTS is then replaced by our designed cross-attention blocks for fusion and alignment between content and style. As the fusion is performed along with the skip connection, our cross-attention block provides a good inductive bias to gradually infuse the phoneme representation with a given style. Additionally, we prevent the style embedding from encoding linguistic content by randomly truncating LST during training and using wav2vec 2.0 features. Experiments show that with fine-grained style control, our system performs better in terms of naturalness, intelligibility, and style transferability. Our code and samples are publicly available. [1]

Index Terms— Text-to-speech synthesis, fine-grained style control, transformer, style transfer

1. INTRODUCTION

Human speech naturally consists of linguistic content and para-linguistic information. Para-linguistic information can be further categorized into speaker characteristics and speaking styles. Speaker characteristics correspond to the vocal properties of a specific speaker, such as accent and voice types. On the other hand, time-varying properties within a sentence such as speaking rate, loudness and prosody belong to speaking styles.

Separation of speaker characteristics is widely used in neural-based TTS systems [1, 2, 3, 4] to improve controllability and expressivity. In the global style token (GST) [5], a reference encoder is trained to capture speaker characteristics by representing the reference speech with a linear combination of style tokens. Conditioned on GST, TTS systems can synthesize speech with different speaker characteristics. Nonetheless, modeling of speaker characteristics and speaking styles often suffers from an issue termed "content-leakage" [6]. This is a situation where the system erroneously

extracts linguistic information from the reference speech, resulting in a content mismatch between the input text and the generated speech during inference. Content-leakage stems from minimizing a reconstruction loss of the reference speech during training. Without additional constraint, it relies solely on the design of the network architecture to prevent encoding linguistic information from the reference speech.

To make speaking styles controllable, Skerry-Ryan et al. [7] augment a style encoder that extracts a single prosody embedding from the reference audio. As fixed-length representation fails to model fine-grained styles, Lee et al. [8] extend the work to extract variable-length prosody embeddings with a soft attention mechanism for the fusion between text and prosody. In [9], the attention mechanism is further replaced by forced alignment of reference speech to aggregate prosody embeddings on a phoneme basis. Later, Tan et al. [10] introduce additional objectives using collaborative learning and adversarial training to mitigate content-leakage.

With the success of transformer based models in various domains [11, 12, 13], TransformerTTS [4] is developed as an alternative to seq2seq TTS systems, as it provides faster inference speed without noticeable quality degradation. Several studies have adapted prior works on modeling speaker and style properties to the TransformerTTS framework. MultiSpeech [14] extends TransformerTTS with an additional speaker module, resembling the approach of GST. UTTS [15] further adapts [7] to TransformerTTS and minimizes an additional Renyi divergence objective to alleviate content-leakage. However, fine-grained control of prosody, speaking rate, and other subtle speaking styles remains unexplored for TransformerTTS.

In this paper, we incorporate fine-grained style control into TransformerTTS. We condition the speaking style on reference speech, making our model infer the style factors where the perceptual difference may be hard to quantify manually. Motivated by [8], we develop a cross-attention module to fuse linguistic content and style information into a mixed representation. We show that our designed model architecture and training procedure suffer less from content-leakage and preserves style information without additional regularization objective. Moreover, by separately controlling fine-grained style factors, our system generates higher quality speech.
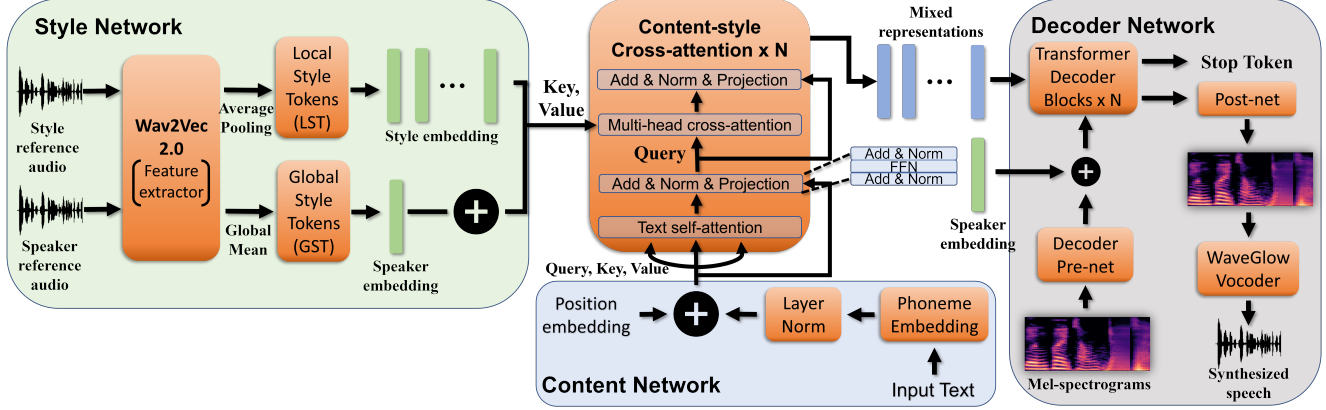
---

[1] https://github.com/b04901014/FG-transformer-TTS

Fig. 1: Overview of our LST-TTS system.

2. METHOD

Figure 1 shows the overview of our LST-TTS system. We decompose our system into four sub-modules: content network, style network, decoder network, and the content-style cross-attention blocks. The decoder network follows an architecture similar to MultiSpeech [14], while the content network is a simplified version containing only phoneme embedding and layer normalization. Here we focus on the design of the style network and content-style cross-attention blocks.

2.1. Style network

The purpose of the style network is to extract and disentangle speaker characteristics and fine-grained speaking styles from the reference speech.

Wav2vec 2.0 features. Instead of using spectral based features or pretrained speaker embeddings, we adopt wav2vec 2.0 [12] to extract meaningful representations from the audio. While wav2vec 2.0 was originally meant to be fine-tuned on downstream tasks, here we use it as a feature extractor. The benefit this brings is twofold. The content leakage is mitigated since the reconstruction of reference speech from wav2vec 2.0 features is more difficult compared to raw waveform or spectrum-based inputs. Moreover, since wav2vec 2.0 is already a transformer based model, its output features are naturally suitable to be processed by the succeeding attention blocks.

Speaker embedding. For the global style embedding, we adopt a framework similar to GST [5]. Specifically, the output features from wav2vec 2.0 are processed by a position-wise linear layer with ReLU activation. A single layer LSTM [16] is applied for further contextualization. Finally we aggregate a mean representation across time-steps to form a global style query vector. To obtain the speaker embedding, the query vector will be fed to a multi-head attention [17], with a trainable codebook of global style tokens as both key and value.

Style embedding. Similarly, the output features from

wav2vec 2.0 are also processed by a LSTM for fine-grained style embeddings. However, instead of taking the mean across the time dimension, we adopt average pooling with stride 4 and kernel size 8 to smooth out the representation. Based on this representation, each time steps will be fed as a query to a multi-head attention with another trainable codebook as key and value to produce a sequence of style embeddings. We term this embedding codebook as "local style tokens" (LST), as it is constructed in a similar way with GST but focuses on frame-level properties. The final style representation is obtained by broadcast-adding speaker embedding with style embedding across the time dimension.

2.2. Content-style cross-attention blocks

In fine-grained style control, one persistent issue is to align variable-length prosody features with the phonetic representations. In [8], a scaled dot-product attention is applied to learn the alignment between the two modalities. The prosody features are then aggregated per phone to concatenate with the phone representation. However, simple concatenation is a rather naive approach for the fusion of style and content information and is more vulnerable to content-leakage.

With the transformer architecture, we can embed the style information into the linguistic content in a learnable and natural manner. Specifically, we replace the original content encoder blocks in TransformerTTS with cross-attention blocks. As shown in Figure 1, phone representations are first processed by a typical transformer encoder block. The output is then used as the query for a multi-head cross-attention module with style representation as both key and value. The resulting aligned style representation will be added back to the content representation with the residual design in transformer blocks. Similar to the transformer encoder, this block can be stacked multiple times for better performance. Since the same style representation is fed as key and value for multiple blocks within the skip connections, one can view the functionality of this module as gradually refining the content representation

based on the style representation.

2.3. Training and inference procedure

Our system is trained end-to-end with essentially the same loss as TransformerTTS [4]. We denote the style network as S, content network as T, content-style cross-attention as A, and decoder network as D. During training, we random sample a pair of speech and text $(\mathbf{x}^s_{sty}, \mathbf{c})$ from speaker s, along with an additional utterance $\mathbf{x}^s_{spk}$ from the same speaker as the speaker reference audio. The system is then trained to minimize the following loss:

$$\mathcal{L}_{tts} = \|D(A(S(\mathbf{x}^s_{sty}, \mathbf{x}^s_{spk}), \mathbf{c}), \mathbf{x}^s_{sty}) - \text{Mel}(\mathbf{x}^s_{sty})\|_1$$

Additionally, a post-net is trained to further refine the output mel-spectrogram, and a stop token is jointly trained to predict the end of the speech.

Random truncation of style embedding. At training time, we truncate the style embedding to a random length $\tilde{l}_{sty}$ chosen from $[\alpha, l_{sty}]$, where l_{sty} is the length of the original embedding and α is a hyper-parameter controlling the minimum length after truncation. We set $\alpha = 15$ in our experiments. This benefits our system in two aspects. First, at inference time, $\mathbf{x}_{sty}$ may be much shorter than expected synthetic speech as the content is different. Our designed training procedure simulates this situation. The model is trained to match the style of synthetic and reference speech up to $\tilde{l}_{sty}$, and fill out the remaining styles based on the given style information before $\tilde{l}_{sty}$. One can also observe this behavior from the synthesized speech samples we provided. Furthermore, content-leakage is alleviated as only a partial segment of reference speech is given to the decoder, and the remaining linguistic information is forced to be extracted from the phone sequence.

During inference, our system supports style reference audio $\mathbf{x}_{sty}$ and speaker reference audio $\mathbf{x}_{spk}$ from different speakers and with different content. Moreover, we can adopt a similar sampling method in GST [5] to sample style embeddings without the style reference speech $\mathbf{x}_{sty}$. In particular, we sample uniformly from the LST codebook independently for each time step of the style embedding. We then multiply the style embedding by a scale factor $\beta = 0.25$. The length of the style embedding is chosen randomly from $[80, 160]$.

3. EXPERIMENT SETUP

We conduct our experiments under both single-speaker and multi-speaker settings, each designed to evaluate different aspects of our system. We use two datasets, the LJ Speech dataset [18] and the VCTK Corpus [19] respectively. The LJ Speech dataset [18] is a commonly used dataset for single-speaker TTS systems. It contains 13,100 English utterances from a single female speaker with total duration of about 24 hours. The VCTK Corpus consists of 110 English speakers

each reading about 400 sentences. As various accents are included, it is often adopted for evaluation of voice conversion systems. For VCTK, 44 male and female speakers were randomly selected for the training set, with the remaining speakers as test set. Additionally, we use the ESD database [20], an emotional speech synthesis corpus to evaluate style transferability in our system. We only use the English corpus in ESD which consists of 13 hours of speech from 10 speakers with 5 emotions.

We extract mel-spectrograms from audio following the convention of Tacotron2 [3]. WaveGlow [21] is used as the vocoder to transform mel-spectrogram to synthetic speech. We use 64 GSTs and 32 LSTs and set the embedding and hidden size of all modules to 256 except the bottleneck layer of the decoder pre-net (32) and the FFN in attention blocks (1024). We set the number of content-style cross attention blocks and transformer decoder blocks to $N = 5$. The system is trained using Adam optimizer [22] with learning rate 2×10^{-4}. We first train our system on LJ Speech for 250k steps with a batch size of 128. The resulting model is used for single-speaker evaluation. Since the speaker embedding is not useful for the single-speaker system, we simply use the average speaker embedding of the training examples at inference time to eliminate the need of speaker reference audio. Bootstrapping from the pretrained model on LJ Speech, we further train our system for 150k steps on VCTK and ESD with a batch size of 64. Additional implementation details are in our github repository.

Comparing methods. We evaluate four different variants of our LST-TTS system. SPK is a version which only uses speaker embedding without style embedding, this is similar to MultiSpeech [14] in VCTK and TransformerTTS [4] in LJ Speech. LST-A refers to the situation where the speaker and style reference speech are from the same speaker, while in LST-S they are shuffled and may be from different speakers. In LST-R, random sampling using the method described in Section 2.3 is applied for LST without conditioning on any reference speech. Additionally, we denote ground truth speech in both dataset as GT.

4. RESULTS AND DISCUSSIONS

4.1. Objective evaluation

We first evaluate the systems using the word error rate (WER) obtained from an ASR model to measure content integrity. We use the ASR model pretrained by ESPNet [23] on the LibriSpeech dataset [24]. Table 1 presents the WER of competing methods on both datasets. In LJ Speech, our methods LST-A and LST-R both yield a low WER, which verifies that content-leakage is minimized in our system and ensures the intelligibility of the synthetic speech. We also show that randomly sample the LST without sequential dependency (LST-R) does not affect the intelligibility of the generated speech. This enables our system to sample different styles without

Table 1: Objective evaluation on LJ Speech and VCTK

	LJ Speech					VCTK				
	GT	LST-A	Tan et al. [10]	SPK	LST-R	GT	LST-A	UTTS [15]	SPK	LST-S
WER(%)	3.6	9.5	21.4	34.5	8.7	3.0	12.5	18.3	40.5	12.3

the presence of $\mathbf{x}_{sty}$. In addition, we observe that the SPK variant of our system tends to suffer from deletion and repetition, which causes the high WER. We conjecture that with our style network and content-style cross-attention module, the content representation is adjusted to contain the prosody information. As a result, our decoder network can focus on learning the alignment without modeling the speaking style, leading to better robustness.

We further show on VCTK the ability of our system to scale to multi-speaker settings. Table 1 shows that both LST-A and LST-S are able achieve low WER, and that passing different speakers for $\mathbf{x}_{sty}$ and $\mathbf{x}_{spk}$ does not affect intelligibility. While not directly comparable due to the difference in ASR models, we include the WER of recent TTS systems with style modeling [10, 15] on each dataset for reference.

Emotional TTS. We assess the style transferability of our method objectively using a pretrained speech emotion recognition (SER) classifier with 95.8% accuracy on the test set of ESD. We label synthesized audio of our model with the emotion label of the style reference speech $\mathbf{x}_{sty}$ and calculate the accuracy of the SER classifier over 1000 randomly sampled synthetic speech. Under the LST-A setup, we achieved 68.4% accuracy, which demonstrates emotion patterns can be transferred from $\mathbf{x}_{sty}$ with our system. While being lower in performance compared to recent Emotional TTS (ETTS) systems [25, 26, 27], our system can realize ETTS without any supervision from the emotion labels.

4.2. Subjective evaluation

We evaluate the naturalness and style transferability of our system in terms of a listening test. All tests were conducted using Amazon Mechanical Turk. For naturalness, each competing method generated 20 utterances with text and reference speech randomly sampled from the test set. Workers were asked to rate the naturalness of the speech on a 5-scale from 1 (Bad) to 5 (Excellent). Each audio was rated 20 times by different workers and we report the mean opinion score (MOS) along with the 95% confidence interval. For style similarity, 20 pairs of style reference speech and synthesized speech was presented to the workers for each system. Similarly, the workers rated the speaking style similarity between the two audios from 1 (not similar at all) to 5 (highly similar).

Table 2 presents the evaluation result on LJ Speech. With the TransformerTTS architecture, SPK already achieves a high naturalness score. Adding the style control with our method does not degrade but improves the naturalness, sug-

gesting the success of the fusion between linguistic content and speaking style by cross-attention module. Compared to [10] which applied fine-grained style control on Tacotron2, our method outperforms in both naturalness and style similarity. Table 3 indicates the speech quality of our method is not impaired under multi-speaker configuration. While LST-S lowers the similarity score, it could be attributed to the increase in difficulty of comparing speaking styles from different speakers. On the other hand, for LST-A, the style similarity score in VCTK remains close to the that of LJ Speech. These results demonstrate that our system is able to perform speaking style transfer within each speaker.

Table 2: MOS with 95% confidence interval on LJ Speech

	Naturalness	Style similarity
GT	4.07 ± 0.07	-
LST-A	$\mathbf{3.62} \pm 0.09$	$\mathbf{3.28} \pm 0.10$
Tan et al. [10]	3.19 ± 0.11	3.23 ± 0.11
SPK	3.50 ± 0.11	-

Table 3: MOS with 95% confidence interval on VCTK

	Naturalness	Style similarity
GT	3.87 ± 0.10	-
LST-A	$\mathbf{3.75} \pm 0.10$	$\mathbf{3.25} \pm 0.12$
LST-S	3.73 ± 0.10	3.09 ± 0.10
SPK	3.47 ± 0.10	3.07 ± 0.13

5. CONCLUSION

We developed a transformer-based TTS system that incorporates fine-grained style control. We evaluate the intelligibility, naturalness, and style transferability of the synthetic speech and show that our system can generalize to various settings. We believe that the techniques we described are capable of abstracting key properties of speech and allowing these to be recombined to generate novel outputs. Based on this work, we intend to inspect the learned fine-grained style embedding of our system and its potential downstream applications.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. Saurous, "Tacotron: Towards end-to-end speech synthesis," 08 2017, pp. 4006–4010.

[2] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*. 2019, vol. 32, Curran Associates, Inc.

[3] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[4] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *AAAI*, 2019.

[5] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *ICML*, 2018.

[6] T. Hu, A. Shrivastava, O. Tuzel, and C. Dhir, "Unsupervised style and content separation by minimizing mutual information for speech synthesis," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2020, pp. 3267–3271.

[7] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," in *Proceedings of the 35th International Conference on Machine Learning*. 2018, vol. 80, pp. 4693–4702, PMLR.

[8] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5911–5915, 2019.

[9] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-Grained Robust Prosody Transfer for Single-Speaker Neural Text-To-Speech," in *Proc. Interspeech 2019*, 2019, pp. 4440–4444.

[10] D. Tan and T. Lee, "Fine-Grained Style Modeling, Transfer and Prediction in Text-to-Speech Synthesis via Phone-Level Content-Style Disentanglement," in *Proc. Interspeech 2021*, 2021, pp. 4683–4687.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[13] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "What does BERT with vision look at?," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 5265–5275, Association for Computational Linguistics.

[14] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "Multispeech: Multi-speaker text to speech with transformer," in *INTERSPEECH*, 2020.

[15] D. Paul, S. Mukherjee, Y. Pantazis, and Y. Stylianou, "A Universal Multi-Speaker Multi-Style Text-to-Speech via Disentangled Representation Learning Based on Rényi Divergence Minimization," in *Proc. Interspeech 2021*, 2021, pp. 3625–3629.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*. 2017, vol. 30, Curran Associates, Inc.

[18] K. Ito and L. Johnson, "The LJ Speech dataset," 2017.

[19] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[20] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 920–924.

[21] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," 05 2019, pp. 3617–3621.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[23] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[25] R. Liu, B. Sisman, and H. Li, "Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability," in *Proc. Interspeech 2021*, 2021, pp. 4648–4652.

[26] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," 06 2021, pp. 5734–5738.

[27] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5, 2021.