StylebookTTS: Zero-Shot Text-to-Speech Leveraging Unsupervised Style Representation

Juhwan Yoon, Hyungseob Lim, Hyeonjin Cha and Hong-Goo Kang
Yonsei University, Seoul, South Korea
E-mail: {jhyoon10, hyungseob.lim, hcha}@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr

***Abstract*—Zero-shot text-to-speech (ZS-TTS) is a TTS system capable of generating speech in voices it has not been explicitly trained on. While many recent ZS-TTS models effectively capture target speech styles using a single global style feature per speaker, they still face challenges in achieving high speaker similarity for voices that were not previously encountered. In this study, we propose StylebookTTS, a novel ZS-TTS framework that extracts and utilizes multiple target style embeddings based on the content. We begin by extracting style information from target speeches, leveraging linguistic content obtained through a self-supervised learning (SSL) model. The extracted style information is stored in a collection of embeddings called a stylebook, which represents styles in an unsupervised manner without the need for text transcriptions or speaker labeling. Simultaneously, the input text is transformed into content features using a transformer-based text-to-unit module, which links the text to the SSL representations of an utterance reading that text. The final target style is created by selecting embeddings from the stylebook that most closely align with the content features generated from the text. Finally, a diffusion-based decoder is employed to synthesize the mel-spectrogram by combining the final target style with the content features generated from the text. Experimental results demonstrate that StylebookTTS achieves greater speaker similarity compared to baseline models, while also being highly data-efficient, requiring significantly less paired text-audio data.**

I. Introduction

Text-to-speech (TTS) is a technology that converts written text into spoken language by passing it through various processing stages. In recent years, TTS systems have made substantial progress thanks to the advent of deep learning techniques [3]–[8], enabling their use in applications such as speech-based virtual assistants and accessibility tools for individuals with disabilities. Traditionally, TTS systems have been developed using text paired with the acoustic features of speakers encountered during training.

Recently, there has been increasing interest in generating voices for new speakers—who were not included in the training data—using only a few seconds of their speech samples. This approach is referred to as zero-shot multi-speaker TTS (ZS-TTS). Traditional ZS-TTS models [9]–[17] typically extract a global speaker embedding from the input speech using an external speaker encoder. This embedding is then used as a conditioning factor within the TTS framework. They demonstrate outstanding performance in synthesizing speech that closely resembles human-recorded utterances in terms of both naturalness and intelligibility. However, these networks continue to face challenges in bridging the similarity gap between observed and unobserved speakers during training. This issue arises because they rely on a single global vector to condition all the frames, which inadequately captures and transfers the target speaker's style to the generated speech from text. Although this issue can be mitigated by fine-tuning the pre-trained TTS network with a few seconds of speech samples from the desired speaker [12], this process is both time-consuming and memory-intensive for each new speaker.

In this work, we propose StylebookTTS, a zero-shot TTS framework that utilizes multiple target style representations during synthesis to enhance speaker similarity with the target speaker. StylebookTTS is built upon the stylebook framework [2], a speech analysis and synthesis method originally developed for any-to-any voice conversion. This framework captures diverse style representations based on linguistic content, without requiring text transcription or speaker labeling during training and inference. In the voice conversion framework, various style embeddings for different frames are extracted based on the source speech's content embeddings, which are derived from quantized self-supervised learning (SSL) features. These style and content embeddings are then integrated in a diffusion model [18] to produce the converted speech. To adapt the stylebook framework for TTS applications, we introduce a transformer-based [19] text-to-unit module that estimates the speech unit (i.e., quantized SSL feature) corresponding to the linguistic content of the speech from its text transcription. This module is trained independently and is then integrated with the pre-trained stylebook model to create the complete text-to-speech network. Our experiments demonstrate that StylebookTTS achieves enhanced speaker similarity with unseen target speakers and improved intelligibility, thanks to its use of fine-grained, content-dependent style embeddings.

In addition to improved speaker similarity, our proposed model efficiently utilizes small amounts of paired data. In contrast to previous models that need to learn both the pronunciation of sounds and speaker-specific style factors from paired text-audio data, our framework delegates the style adaptation to the pre-trained stylebook model. Instead, it focuses on establishing the relationship between text and its pronunciation, as captured by the quantized SSL features. Ultimately, StylebookTTS requires significantly less paired data to achieve near-optimal performance. These results indicate that this content-based target style modeling methodology can deliver excellent performance in speech synthesis tasks, even with limited paired

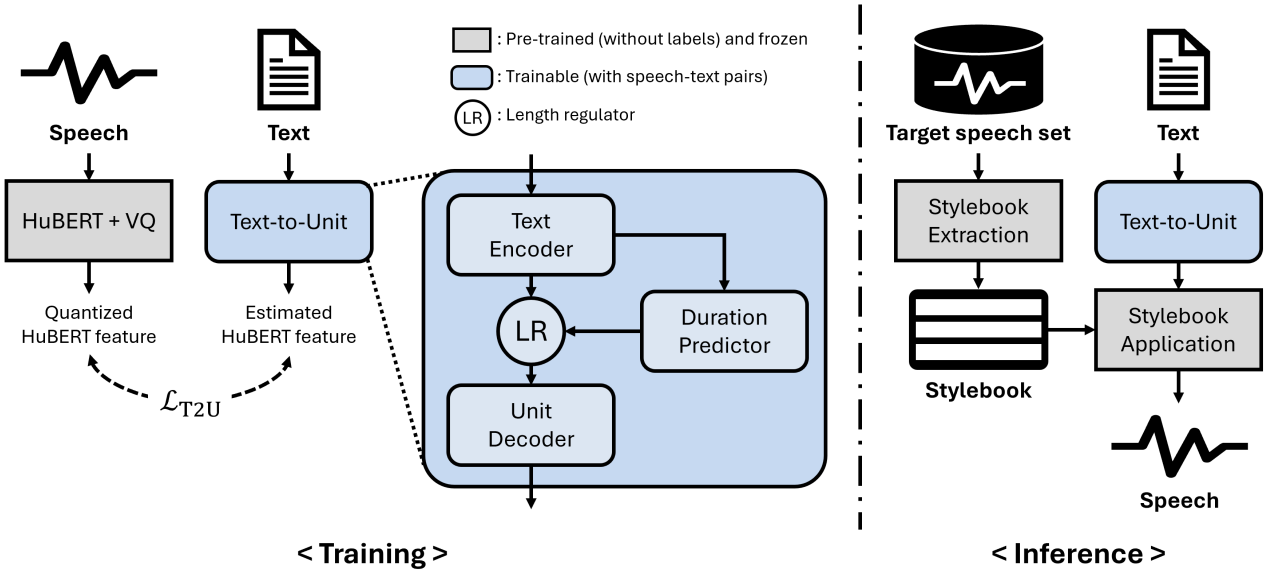


Fig. 1. Block diagram of the proposed text-to-speech framework utilizing pre-trained HuBERT [1] and a style extraction and application model [2]. During training, a text-to-unit module is trained to estimate the quantized HuBERT feature from a given text. After trainig, the HuBERT model originally included in the style application module is replaced with the trained text-to-unit module so that a speech can be generated from a text instead of a HuBERT feature.

data.

Our contributions are summarized as follows[1]:

- We extend the content-dependent stylebook network [2] to the TTS domain, achieving high speaker similarity and intelligibility with unobserved speakers, while maintaining comparable overall speech quality.
- We demonstrate that our proposed StylebookTTS can be efficiently optimized even with small amounts of paired data by relying on external stylebook module, which has been pre-trained in an unsupervised manner to handle the complex style modeling.

II. Related works

A. Zero-shot text-to-speech

Zero-shot text-to-speech aims to synthesize speech from a text transcript and reference audio (or target speech) of an unobserved speaker, ensuring that the characteristics of the generated speech closely match those of the target speaker. The goal is typically achieved by extracting style-related information from the target speaker's utterance and incorporating it as a conditional feature into the TTS model during synthesis. For style-related information, many previous works [9]–[12] have commonly used a speaker embedding vector extracted by a speaker verification model, which may be either pre-trained or jointly trained with a classification loss. However, compressing all style-related information into a single vector can lead to information loss, which limits the style transfer capability of the ZS-TTS system.

On the other hand, prompt-based approaches [13]–[17] allow the model to directly access the acoustic features of the target speech through the attention mechanism [19]. In contrast to approaches that rely on a global speaker embedding, prompt-based models can locally adapt the style at each time frame. This results in better speaker similarity, regardless of whether the model undergoes a discriminative learning process with speaker labels [13]–[17]. However, these models inevitably face increased computational costs as the length of the target speech grows, since the attention layer must reference the entire target utterance at each attention step.

Unlike the previous approaches, we employ stylebook [2]—a collection of style embeddings each implicitly linked to different linguistic content—as the source of style-related information for conditioning a ZS-TTS model. By adopting this approach, we achieve better speaker similarity compared to models that use global speaker embeddings, and we further enhance performance with longer target speech without increasing computational costs during inference.

B. Speech analysis and synthesis using stylebook

The stylebook [2] model is an autoencoder-like speech analysis and synthesis network originally designed for any-to-any voice conversion. In this framework, target speeches are first processed by the mel and style encoders, which transform them into style embeddings. Simultaneously, the target speeches are converted into HuBERT [1] representations and undergo vector quantization to extract features related solely to linguistic content, as done in [20] and [21]. The obtained style and contents embeddings are fed into a multi-head attention mechanism with a learned query set, resulting in a set of embeddings—referred to as the stylebook— for the given target speaker. Since the query set consists of embeddings representing different content-related features, the resulting stylebook forms a collection of embeddings that capture various speaking styles of the target speaker. Similarly,

[1]Audio samples are available at http://pineville17.github.io/StylebookTTS

source speeches are converted into content embeddings using a HuBERT network and vector quantization. The model then selects the appropriate style embeddings from the stylebook by comparing the source speech's content embeddings with the stylebook entries using attention mechanisms. The obtained embeddings are combined in a diffusion-based U-Net model [22] to generate the mel-spectrogram of the converted speech. Finally, a pre-trained HiFi-GAN [23] vocoder is used to synthesize speech from the obtained mel-spectrogram. During training, the same speech is used as both the source and the target. Here, the model is trained to effectively represent the missing information lost during the quantization of HuBERT features—typically, the detailed speaking style associated with specific linguistic content.

The main contribution of the stylebook model is its use of transposed dual attention to generate a stylebook that represents the speaking style of a target speaker and to adapt the style embedding at each time frame based on the content of the source speeches. The content-dependent style embeddings can be extracted without the need for text transcription or speaker labeling, as the model is trained to reconstruct the input itself following decomposition with HuBERT and vector quantization. Additionally, since the number of embeddings in the stylebook remains fixed regardless of the length of the target speech, the stylebook model can leverage longer target speeches without increasing computational costs once the stylebook is extracted.

III. Proposed Method

In this work, we propose StylebookTTS designed to extend the stylebook framework to a zero-shot text-to-speech (ZS-TTS) system. StylebookTTS consists of two distinct components: the stylebook network and the text-to-unit module. The stylebook network, which follows the architecture outlined in [2], handles both the extraction and application of the stylebook, as shown in Fig. 1.

A. Text-to-unit estimation

The core component of our proposed method is the text-to-unit (T2U) module, which estimates vector-quantized HuBERT representations from the input text. This enables the network to extract content embeddings from the input text, as vector-quantization of SSL features effectively captures content information from the input speech.

The architecture of the text-to-unit module is inspired by FastSpeech2 [4], a non-autoregressive TTS model known for its high generation speed and excellent speech quality. FastSpeech2 is designed to train pitch, energy, and duration predictors during training, and incorporates the estimated pitch, energy, and duration as conditions during inference, thereby providing style information. In our text-to-unit module, we omit the pitch and energy prediction components from FastSpeech2 to avoid confusion in capturing contextual information. We retain the duration predictor to accurately estimate phoneme durations, ensuring that the input text aligns well with the HuBERT units. During inference, the text is processed through the text encoder to generate phoneme embeddings. The length regulator adjusts the duration of the input sequence based on the predicted duration of each phoneme, indicating how many discrete HuBERT unit frames correspond to each phoneme. Subsequently, the unit decoder generates discrete HuBERT units from the length-adjusted sequence.

Once the quantized SSL features are estimated from the input text, they are fed into the content encoder and then processed through the multi-head attention layer during the stylebook application stage. This process facilitates the generation of speech that preserves the content of the given text while closely resembling the style of the target speaker.

B. Training Method

The training process for the text-to-unit module involves several key steps to ensure accurate prediction of the SSL features and high speech quality. During training, both speech and its corresponding text transcription are provided as inputs to the text-to-unit module. The target for estimation is the discrete HuBERT representation, which captures the content-related information from the input speech. A duration predictor, similar to the one used in FastSpeech2, estimates phoneme durations. This predictor utilizes the Montreal Forced Alignment (MFA) tool [24] to obtain target durations for each phoneme, ensuring precise alignment between the input text and the target HuBERT representations.

As the module generates outputs, two loss functions are utilized: duration loss and HuBERT loss. For the duration loss, mean-squared error (MSE) is computed between the target and predicted durations provided by the duration predictor, ensuring that the predicted durations closely match the target phoneme durations. For the HuBERT loss, L1 loss is used to optimize feature prediction accuracy by comparing the predicted and target feature representations, which helps maintain high fidelity in the synthesized speech. The overall loss function for the T2U network is the sum of these two loss components, combining their advantages to enhance performance, as follows:

$$\mathcal{L}_{\mathrm{dur}} = \frac{1}{n}\Sigma_{i=1}^{n}(d_{target,i} - d_{pred,i})^2, \tag{1}$$

$$\mathcal{L}_{\mathrm{hub}} = \Sigma_{i=1}^{n}|\mathrm{hvq}_{target,i} - \mathrm{hvq}_{pred,i}|, \tag{2}$$

$$\mathcal{L}_{\mathrm{T2U}} = \mathcal{L}_{\mathrm{dur}} + \mathcal{L}_{\mathrm{hub}}, \tag{3}$$

where d indicates the phoneme duration and hvq indicates vector quantized HuBERT representations and n indicates the number of data.

IV. Experiments

A. Experiment settings

1) ***Dataset****:* We conducted experiments using the LibriTTS dataset [25]. For training the T2U module, we randomly selected 115,183 samples from 900 speakers, totaling 186 hours, from the LibriTTS-train-clean-360 subset. For evaluating the proposed StylebookTTS network, we chose 27 speakers from the LibriTTS-test-clean subset and created target speeches with varying lengths, ranging from 5 to 180 seconds.

TABLE I
NATURALNESS EVALUATION RESULTS IN MOS RATINGS. SUBJECTIVE MOS VALUES WITH 95% CONFIDENCE INTERVALS ARE SHOWN.

	MOS ↑	
Target Length	5 sec	180 sec
MetaStyleSpeech	2.72±0.13	2.93±0.13
YourTTS	2.69±0.13	2.63±0.13
StylebookTTS	**3.57±0.12**	**4.00±0.09**
Ground Truth	4.55±0.09	

*2) **Model configuration**:* The model configuration of the stylebook framework follows that of [2], specifically using 128 embeddings in the stylebook, and is pre-trained on the same LibriTTS dataset. For the T2U module, the details are as follows. The text encoder consists of 4 feed-forward transformer (FFT) blocks, each with 2 attention heads and a hidden size of 256. The unit decoder is comprised of 6 FFT blocks, each with 2 attention heads and a hidden size of 256. The duration predictor consists of 2 convolution layers with a filter size of 256, a kernel size of 3, followed by a linear layer. The number of parameters of StylebookTTS including the T2U module and the stylebook framework is 77M.

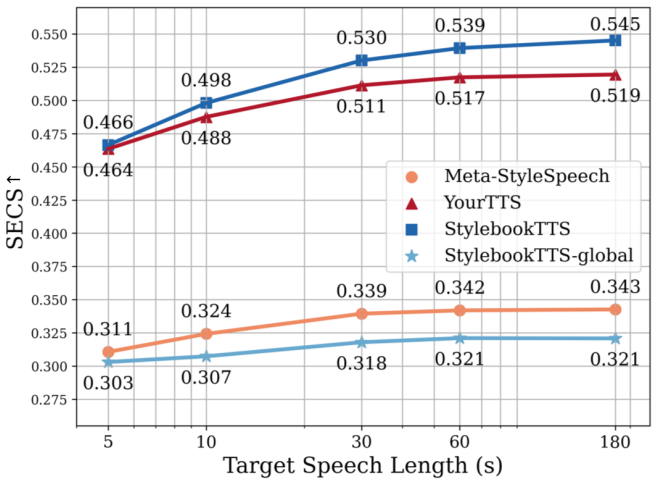
B. Evaluation Methods

We analyze and evaluate our proposed system using four metrics: mean opinion score (MOS) [26], ABX test preference score, speaker embedding cosine similarity (SECS), and character error rate (CER). The MOS test assesses the perceived quality of the generated speech through a subjective quality evaluation. The ABX test compares two synthesized speeches to determine which one exhibits higher speaker similarity to the real target speech. MOS and ABX results were obtained from evaluations involving 13 participants. SECS measures how closely the speaker features of the generated speech align with those of the target speech. We computed SECS by extracting speaker embeddings from both the synthesized and target speeches using the pre-trained ECAPA-TDNN [27] network[2] and calculated the cosine similarity between these embeddings. CER indicates the intelligibility of the generated speech. We computed CER by comparing the text of the source speech with the text of the converted speech using a HuBERT [1]-based automatic speech recognition (ASR) network[3].

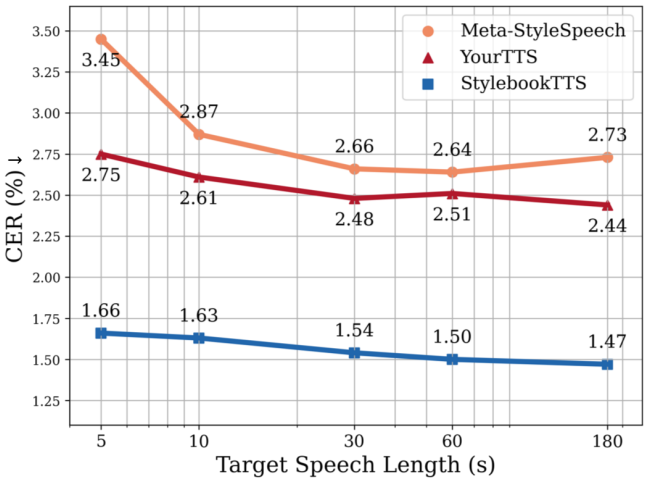
To evaluate performance based on the length of the target speech, we tested various target speeches with lengths of 5, 10, 30, 60, 120, 180, and 300 seconds. For comparison, we selected YourTTS [10] and Meta-StyleSpeech [11] as baseline models for non-finetuning zero-shot TTS frameworks and used their official implementations for synthesis.

C. Experiment results

*1) **Naturalness**:* Table I shows the MOS ratings for synthesized speeches generated by both the baseline and proposed networks for short (5 seconds) and long (180 seconds) input target speech lengths. The results show that our proposed StylebookTTS outperforms the baseline networks in terms of naturalness, regardless of the target speech length. Notably, as the target speech length increases, StylebookTTS generates speech that is nearly as natural as human utterances.



(a)



(b)

Fig. 2. Evaluation results of the proposed StyleBookTTS and the baseline models. (a) illustrates the SECS results and (b) shows CER (%) of various ZS-TTS networks.

*2) **Speaker similarity**:* StylebookTTS aims to enhance the speaker similarity of the synthesized voice to the original target voice by collecting and applying target style embeddings in a content-dependent manner. Fig. 2 (a) illustrates the SECS trends for synthetic speeches generated by the baseline and proposed models across various target lengths. In addition to the two baseline models and the proposed StylebookTTS, which uses multiple style embeddings (128 in this case) to form the stylebook, we also present results for another StylebookTTS network with a single style embed-

[2]https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
[3]https://huggingface.co/facebook/hubert-large-ls960-ft

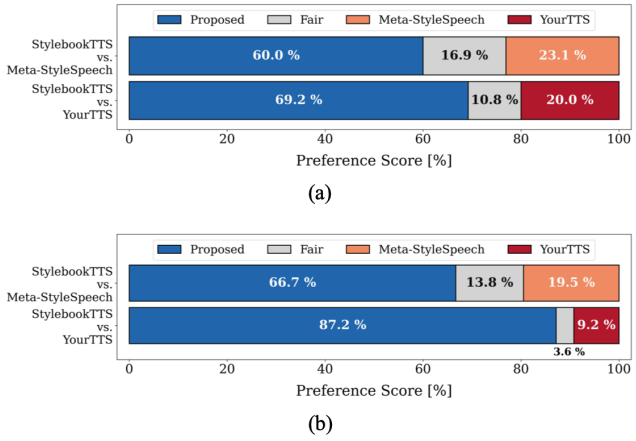


Fig. 3. ABX test result between the proposed StyleBookTTS and the baseline models. (a) illustrates the result when 5 seconds of target speech is introduced. (b) shows the result when 180 seconds of target speech is introduced.

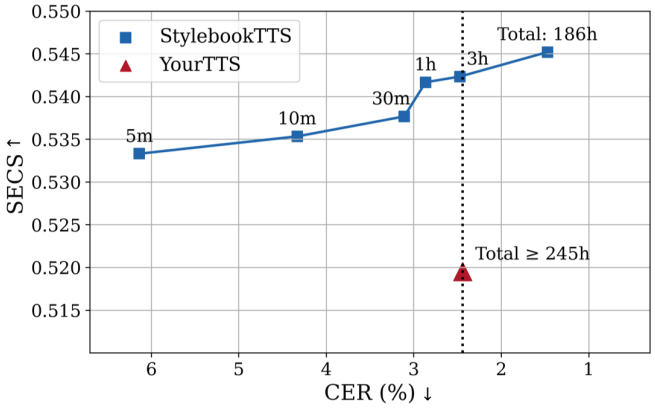


Fig. 4. Evaluation results (SECS and CER) regarding the length of the total speech paired with the text transcription used to train the text-to-unit module.

ding (StylebookTTS-global). The StylebookTTS-global can be viewed as a global style embedding within the stylebook framework. The figure demonstrates that, in terms of speaker similarity, StylebookTTS outperforms both the baseline networks and the StylebookTTS framework using a global style embedding to represent the target voice.

This trend is also evident in the ABX results. As shown in Fig. 3, StylebookTTS demonstrates superior speaker similarity compared to the baseline networks ($p < 0.01$ in all cases). When only 5 seconds of target voice are used, 60% of listeners preferred StylebookTTS over the baseline models for higher speaker similarity. This proportion increases significantly as the target speech length is extended to 180 seconds. Over 87% of listeners indicated that the proposed model generates speech that is more similar to the target speaker compared to YourTTS. These results demonstrate that our content-dependent style extraction approach enhances speaker similarity by extracting multiple content-dependent target styles and applying the most appropriate styles to match the content of the text.

*3) **Intelligibility**:* In addition to enhanced speaker similarity, our approach achieves remarkable intelligibility as shown in Fig. 2 (b). StylebookTTS achieves a CER of 1.66% with just 5 seconds of target speech, whereas Meta-StyleSpeech and YourTTS show CERs of 3.45% and 2.75%, respectively. As the target length increases, StylebookTTS continues to outperform other networks in terms of intelligibility.

*4) **Data efficiency**:* In Fig. 4, we illustrate the objective performance (SECS, CER) of our proposed model relative to the size of the paired data used for T2U training. For inference, 3 minutes of target speech were provided. Even with only 5 minutes of paired data, our proposed model achieves higher SECS compared to YourTTS trained on the full dataset. This demonstrates that StylebookTTS effectively utilizes the pre-trained stylebook model for style adaptation. Regarding intelligibility, StylebookTTS trained with 3 hours of paired data achieves performance comparable to YourTTS, which is trained on over 245 hours of text-audio paired data. These results indicate that our content-based target style modeling approach has the potential to deliver exceptional performance in various speech synthesis tasks, including multilingual and multi-modal synthesis, even with limited paired data.

V. Limitations and Future Works

Despite the improved performance in speaker similarity and intelligibility demonstrated by StylebookTTS, there are still some limitations to address. Since StylebookTTS models the target style based on the content of the target speech, its performance may be less reliable when the input target speech is short and lacks sufficient content information. This issue could be mitigated by conditionally combining global style embeddings with content-dependent style embeddings to address content scarcity. Another limitation is that StylebookTTS-generated speeches tend to have relatively monotonic prosody compared to ground truth speech, despite the high speaker similarity scores. Future research could address this by incorporating a separate variance adaptor to provide additional pitch information for more refined prosody control.

VI. Conclusion

In this work, we present StylebookTTS, a zero-shot text-to-speech (TTS) framework that synthesizes speech using multiple content-dependent target style representations derived from the given text. We developed a text-to-unit module that establishes the relationship between text and its pronunciation by utilizing quantized SSL features. The text-generated SSL features are then fed into the stylebook network for style application and synthesis. Our experimental results demonstrate that this approach achieves superior speaker similarity and intelligibility compared to other zero-shot TTS frameworks that rely on a single global style embedding. Furthermore, StylebookTTS outperforms baseline models even when trained with a relatively small amount of paired text and audio data. This work highlights the effectiveness of content-based target style modeling in speech synthesis applications with limited paired data, suggesting its potential for expansion into research areas such as multilingual or multimodal synthesis.

REFERENCES

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[2] H. Lim, K. Byun, S. Moon, and E. Visser, "Stylebook: Content-dependent speaking style modeling for any-to-any voice conversion using only speech data," *arXiv preprint arXiv:2309.02730*, 2023.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017.

[4] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[5] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018.

[6] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI conference on artificial intelligence*, 2019.

[7] Y. Ren, Y. Ruan, X. Tan, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. Advances in neural information processing systems (NeurIPS)*, 2019.

[8] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. International Conference on Machine Learning (ICML)*, 2021.

[9] E. Casanova, C. Shulby, E. Gölge, *et al.*, "Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model," in *Proc. Interspeech*, 2021.

[10] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *Proc. International Conference on Machine Learning (ICML)*, 2022.

[11] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *Proc. International Conference on Machine Learning (ICML)*, 2021, pp. 7748–7759.

[12] H. Kim, S. Kim, J. Yeom, and S. Yoon, "Unitspeech: Speaker-adaptive speech synthesis with untranscribed data," in *Proc. Interspeech*, 2023.

[13] Y. Zhou, C. Song, X. Li, *et al.*, "Content-dependent fine-grained speaker embedding for zero-shot speaker adaptation in text-to-speech synthesis," in *Proc. Interspeech*, 2022.

[14] C. Wang, S. Chen, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[15] S. Kim, K. Shih, J. F. Santos, *et al.*, "P-flow: A fast and data-efficient zero-shot tts through speech prompting," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[16] C. Du, Y. Guo, F. Shen, *et al.*, "UniCATS: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding," in *Proc. AAAI Conference on Artificial Intelligence*, 2024.

[17] K. Shen, Z. Ju, X. Tan, *et al.*, "NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[19] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Proc. Advances in neural information processing systems (NIPS)*, 2017.

[20] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Proc. Interspeech*, 2021.

[21] B. van Niekerk, M.-A. Carbonneau, J. Zaıdi, M. Baas, H. Seuté, and H. Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[22] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[23] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," 2020.

[24] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Proc. Interspeech*, 2017.

[25] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019.

[26] *Method for the subjective assessment of small impairments in audio systems*. Rec. ITU-R BS.1116-3, 2015.

[27] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Proc. Interspeech*, 2020.