GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech

Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao

Zhejiang University　Sea AI Lab

Demo Page: https://generspeech.github.io

Code Release: https://github.com/Rongjiehuang/GenerSpeech

Motivation

A growing number of applications, such as voice assistant services and long-form reading, have been actively developed and deployed to real-world speech platforms.

Unlike typically controllable speech synthesis, style transfer for generalizable out-of-domain (OOD) text-to-speech aims to generate high-quality speech samples with unseen styles (e.g., timbre, emotion, and prosody) derived from an acoustic reference (i.e., custom voice), which is hampered by two major challenges:

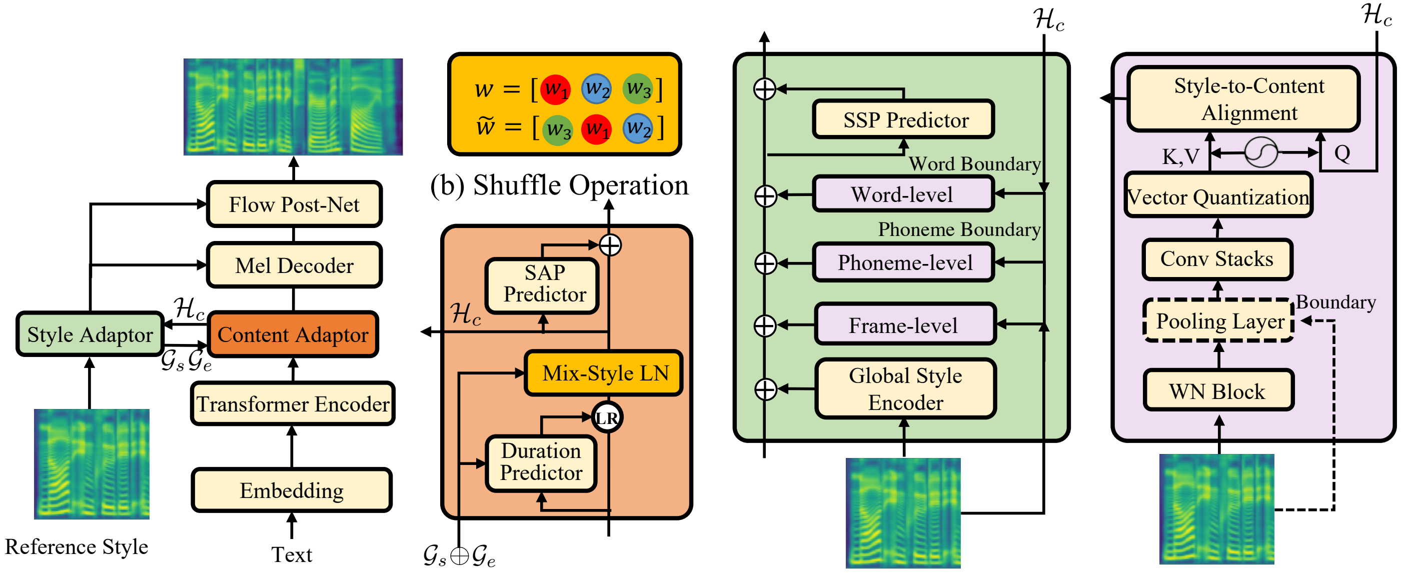
1) style modeling and transferring: the high dynamic range in expressive voice is difficult to control and transfer.

2) model generalization: when the distributions of style attributes in custom voice differ from training data, the quality and similarity of synthesized speech often deteriorate due to distribution gaps.

GenerSpeech

Decompose a model into the domain-agnostic and domain-specific parts via disentangled representation learning. We model and control the style-agnostic (linguistic content) and style-specific (speaker identity, emotion, and prosody) variations in speech separately:

1. To improve model generalization, we propose mix-style layer normalization(MSLN) to eliminate the style information in the linguistic content representation.
2. To enhance modeling and transferring style attributes, we introduce a multi-level style adaptor consisting of a global encoder for speaker and emotion feature embeddings and three differential (frame, phoneme, and word-level) local encoders for prosodic style representations.
3. To reconstruct details in these expressive speech samples, we include a flow-based post-net to refine the transform decoder output and generate fine-grained mel-spectrograms.



(a) Overall (c) Content Adaptor (d) Style Adaptor (e) Local style encoder

Generalizable Content Adaptor

Leveraging recent progress on domain generalization, in this work, we design the Mix-Style Layer Normalization for regularizing TTS model training by perturbing the style information in training samples:

$$\gamma_{\text{mix}}(w) = \lambda\gamma(w) + (1-\lambda)\gamma(\tilde{w}) \qquad \beta_{\text{mix}}(w) = \lambda\beta(w) + (1-\lambda)\beta(\tilde{w})$$

$$\text{Mix-StyleLN}(x, w) = \gamma_{\text{mix}}(w)\frac{x-\mu}{\sigma} + \beta_{\text{mix}}(w)$$

By utilizing the Mix-Style Layer Normalization in the generalizable content adaptor, the linguistic content-related variation could be disentangled from the global style attributes (i.e., speaker and emotion), which promotes the generalization of TTS model towards out-of-domain custom style.

Multi-level Style Adaptor

Global Representation

We use a generalizable wav2vec 2.0 model to capture the global style characteristics, including the speaker and emotion acoustic conditions. In practice, we add an average pooling layer and fully-connected layers on the top of the wav2vec 2.0 encoder, which allows for fine-tuning the model on speaker and emotion classification tasks. The AM-softmax criteria is employed as the loss function for downstream classification.

To sum up, the fine-tuned wav2vec 2.0 model generates discriminative global representations $\mathcal{G}_s$, and $\mathcal{G}_e$ to model the speaker and emotion characteristics, respectively.

Multi-level Style Adaptor

浙江大学 ZHEJIANG UNIVERSITY

Local Representation

1. Frame level. To catch the frame-level latent representation $\mathcal{S}_u$, we remove the optional pooling layer in the local style encoder.
2. Phoneme level. To catch the phoneme-level style latent representation $\mathcal{S}_p$ from speech, we take the phoneme boundary as an extra input and apply pooling on the refined sequences before feeding into the vector quantization layer.
3. Word level. To catch the word-level style latent representation $\mathcal{S}_w$ from speech, we take the word boundary as an extra input and apply pooling to refine the sequences.

Style-To-Content Alignment Layer

To align the variable-length local style representations with the phonetic representation, we introduce the Style-To-Content Alignment Layer for learning the alignment between the two modalities of style and content. In practice, we adopt the popular Scaled Dot-Product Attention as the attention module.

$$\text{Attention}(Q, K, V) = \text{Attention}(\mathcal{H}_c, \mathcal{S}_u, \mathcal{S}_u) = \text{Softmax}(\frac{\mathcal{H}_c \mathcal{S}_u^T}{\sqrt{d}})\mathcal{S}_u$$

Flow-based Post-Net

To further improve the quality and similarity of synthesized mel-spectrograms, we introduce a flow-based post-net to refine the coarse-grained outputs of the mel-spectrogram decoder.

During training, the flow post-net converts the synthesized mel-spectrogram into the gaussian prior distribution and calculates the exact log-likelihood of the data. During inference, we sample the latent variables from the prior distribution and pass them into the post-net reversely to generate the expressive mel-spectrogram.

Performance: Parallel Transfer

Table 1: Quality and style similarity of parallel customization samples when generalized to out-of-domain VCTK and ESD testsets. The evaluation is conducted on a server with 1 NVIDIA 2080Ti GPU and batch size 1. The mel-spectrograms are converted to waveforms using Hifi-GAN (V1).

Method	VCTK				ESD			
	MOS	SMOS	Cos	FFE	MOS	SMOS	Cos	FFE
Reference	4.40 ± 0.09	/	/	/	4.47 ± 0.08	/	/	/
Reference(voc.)	4.37 ± 0.09	4.30 ± 0.09	0.96	0.05	4.40 ± 0.09	4.47 ± 0.10	0.99	0.07
Mellotron	3.91 ± 0.08	3.88 ± 0.08	0.74	0.32	3.92 ± 0.07	4.01 ± 0.08	0.80	0.27
FG-TransformerTTS	3.95 ± 0.1	3.90 ± 0.09	0.86	**0.30**	3.90 ± 0.10	3.94 ± 0.08	0.67	0.43
Expressive FS2	3.85 ± 0.08	3.87 ± 0.10	0.85	0.41	4.04 ± 0.08	3.93 ± 0.09	0.93	0.41
Meta-StyleSpeech	3.90 ± 0.07	3.95 ± 0.08	0.83	0.38	4.02 ± 0.10	3.97 ± 0.10	0.86	0.41
Styler	3.89 ± 0.09	3.82 ± 0.08	0.76	0.38	3.76 ± 0.08	4.05 ± 0.08	0.68	0.39
GenerSpeech	$\mathbf{4.06 \pm 0.08}$	$\mathbf{4.01 \pm 0.09}$	**0.88**	0.35	$\mathbf{4.11 \pm 0.10}$	$\mathbf{4.20 \pm 0.09}$	**0.97**	**0.26**

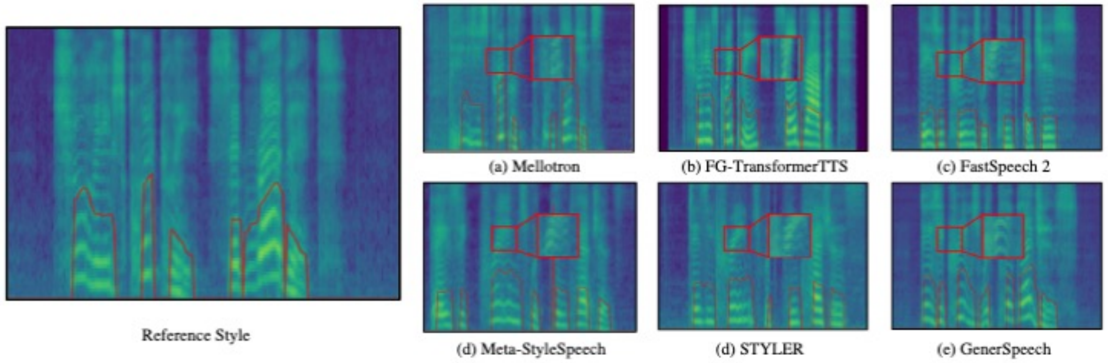


Figure 2: Visualizations of the reference and generated mel-spectrograms in Non-Parallel style transfer. The corresponding texts of reference and generated speech samples are "*Daisy creams with pink edges.*" and "*Chew leaves quickly, said rabbit.*", respectively.

Audio quality: GenerSpeech has achieved the highest MOS with scores of 4.06 (VCTK) and 4.11 (ESD) compared with the baseline models, especially in ESD dataset.

Style similarity: GenerSpeech score the highest overall SMOS of 4.01 (VCTK) and 4.20 (ESD). The objective results of both Cos and FFE further show that GenerSpeech surpasses the state-of-the-art models in transferring the style of custom voices.

Performance: Non-Parallel Transfer

Table 2: The AXY preference test results for parallel and non-parallel style transfer. We select 20 samples from VCTK and ESD testing sets for evaluation. For each reference (A), the listeners are asked to choose a preferred one among the samples synthesized by baseline models (X) and proposed GenerSpeech (Y), from which AXY preference rates are calculated. The scale ranges of 7-point are from “X is much closer" to “Both are about the same distance" to “Y is much closer", and can naturally be mapped on the integers from -3 to 3.

Baseline	Parallel				Non-Parallel			
	7-point score	Perference (%)			7-point score	Perference (%)		
		X	Neutral	Y		X	Neutral	Y
Mellotron	1.51 ± 0.10	26%	14%	40%	1.62 ± 0.09	6%	28%	66%
FG-TransformerTTS	1.07 ± 0.14	22%	30%	48%	1.29 ± 0.10	34%	20%	46%
Expressive FS2	1.22 ± 0.12	30%	20%	50%	1.42 ± 0.11	24%	16%	60%
Meta-StyleSpeech	1.13 ± 0.09	26%	26%	48%	1.18 ± 0.12	14%	26%	60%
Styler	1.49 ± 0.10	18%	24%	58%	1.27 ± 0.09	20%	22%	58%

GenerSpeech can generate mel-spectrograms with rich details in frequency bins between two adjacent harmonics, unvoiced frames, and high-frequency parts, which results in more natural sounds. However, some baseline models (especially Mellotron) fail to generate high-fidelity mel-spectrograms in Non-Parallel style transfer;

GenerSpeech can resemble the prosodic style of the reference signal and demonstrates its precise style transfer, which is nearly time-aligned in pitch contours. However, most baseline models failed to match the prosodic style. They generated the “average” distribution over their input data, generating less expressive speech, especially for long-form phrases.

Thanks !