GENERSPEECH

AUTHORS

Rongjie Huang, Ti Ren, Jinglin Liu, Chenye Cui, Zhou Zhao

AFFILIATIONS

Zhejiang University, Sea AI Lab

GenerSpeech: Towards Style Transfer for Generalizable Out-Of-Domain Text-to-Speech
Generating speech samples with unseen style like speaker identity, emotion and prosody face the challenges of
a) modelling and trasferring of highly dynamic features and b) robustness in handling diverse OOD conditions.
We propose an approach that decomposes the speech variation into style-specific and style-agnostic parts that can tackle the above challenges effectively.

Introduction

TTS is a widely used technology with numerous applications, such as navigation systems, accessibility tools, and virtual assistants. Today's times seek personalisation in everything. Increasing demand for personalized speech generation challenges TTS models especially in unseen scenarios regarding domain shifts.

Objective

To generate high-quality and similar speech samples with unseen style (e.g., speaker identity, emotion, and prosody) derived from a reference utterance, even in diverse OOD conditions that differ from the source data.

CHALLENGES

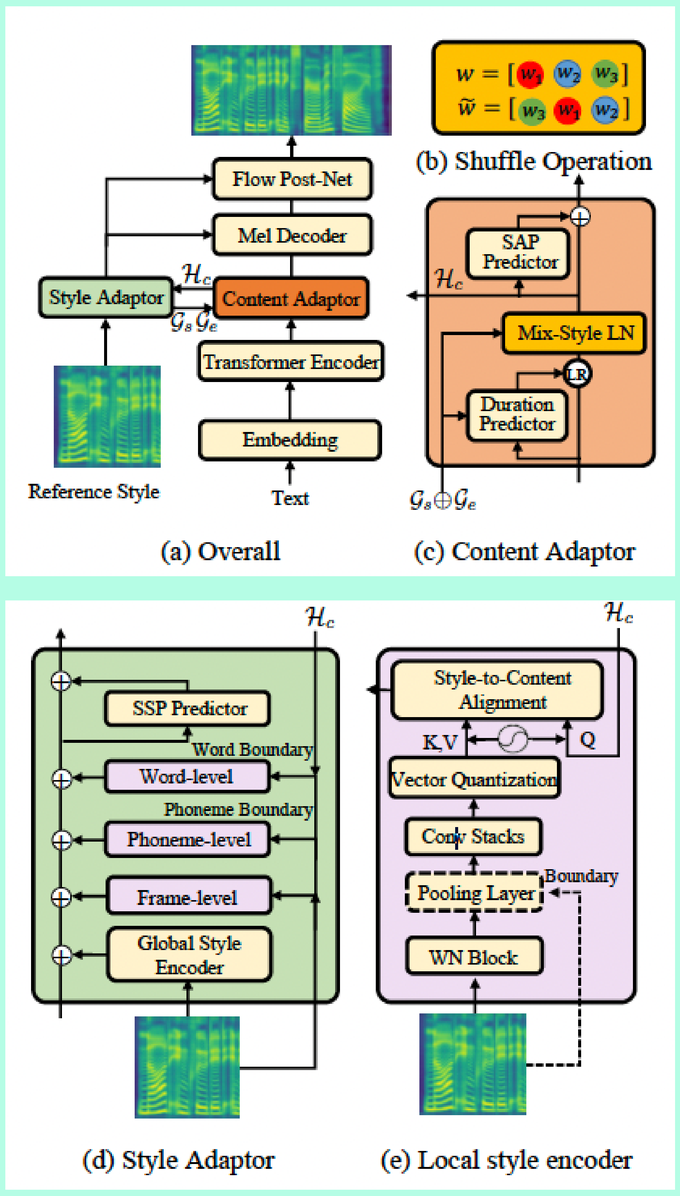
Model Generalisation

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.

Style Modeling and Transferring

The high dynamic range in expressive voice is difficult to control and transfer. Many TTS models only learn an averaged distribution over input data and lack the ability to fine-grained control style in speech sample.

Architecture



Generalizable Content Adaptor

It disentangles linguistic content-related variation from global style attributes, such as speaker and emotion deriving style-agnostic representation initially. It utilises Mix-Style Layer Normalisation is that uses learnable scale and bias vectors to adaptively perform scaling and shifting of the normalised input features based on the style embedding.

$$\gamma(w) = E^{\gamma} * w, \quad \beta(w) = E^{\delta} * w$$

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

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

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

Multi-level Style Adaptor

It consists of a global encoder for speaker and emotion feature embeddings and three differential local encoders for prosodic style representations at the frame, phoneme, and word levels. The multi-level style adaptor allows GenerSpeech to generate high-quality speech samples with unseen styles by capturing both global and fine-grained prosodic variations in speech.

Results

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**

Conclusion

We proposed GenerSpeech, a text-to-speech model towards high-fidelity zero-shot style transfer of out-of-domain custom voices. T

We demonstrated that GenerSpeech achieved new **state-of-the-art** for the task through several experiments. We also did ablation study and presented the results, showing effectiveness of each of the newly proposed components in our architecture

Experiments with differeing amounts of data for style adaption proved the model's robustness in the few-shot data setting

PRESENTED BY

Adarsh Kumar: 19D180003
Shreya Illindra: 190050050
Manan Agarwal: 190050065