A Unified Model for Zero-shot Singing Voice Conversion and Synthesis

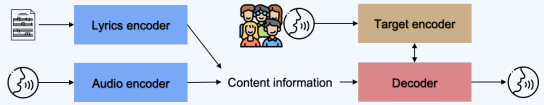
Jui-Te Wu[1], Jun-You Wang[2], Jyh-Shing Roger Jang[1,2], Li Su[1,3]
[1] *NTU-AS Data Science Degree Program, National Taiwan University, Taiwan*
[2] *Department of Computer Science and Information Engineering, National Taiwan University, Taiwan*
[3] *Institute of Information Science, Academia Sinica, Taiwan*

Introduction

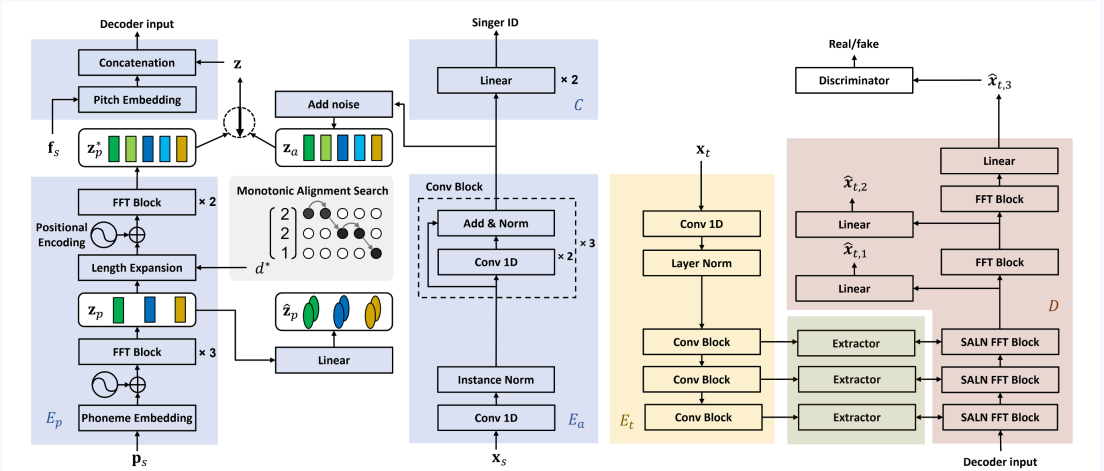
- Most recent works on singing voice generation only support the generation of preset singers' voice.
- Zero-shot singing voice generation is challenging due to the lack of publicity multi-singer dataset.
- Current state-of-the-art zero-shot voice conversion and text-to-speech show potential in improving zero-shot singing voice generation.

Overview

- The source encoders map the audio and text input respectively into a similar content latent space.
- The target encoder takes target singing segments as references.



Methods



Results

Task	Model	Unseen singers		Seen singers	
		Similarity	Naturalness	Similarity	Naturalness
SVC	Baseline (SVC)	3.14 ± 0.17	3.29 ± 0.16	3.20 ± 0.18	3.21 ± 0.16
	Proposed (S)	$\mathbf{3.61 \pm 0.16}$	3.27 ± 0.17	3.63 ± 0.16	3.27 ± 0.16
	Proposed (C)	3.56 ± 0.17	$\mathbf{3.53 \pm 0.17}$	$\mathbf{3.70 \pm 0.16}$	$\mathbf{3.46 \pm 0.16}$
SVS	Baseline (SVS)	3.14 ± 0.18	2.98 ± 0.16	$\mathbf{3.88 \pm 0.15}$	$\mathbf{3.39 \pm 0.16}$
	Proposed (S) w/o Musdb-V	$\mathbf{3.18 \pm 0.17}$	$\mathbf{3.06 \pm 0.17}$	3.87 ± 0.15	3.32 ± 0.17

Table 1. Subjective evaluation MOS and the 95% confidence interval are shown

Datasets	SVC	SVS
All	0.290	0.233
w/o MPOP600	0.292	0.225
w/o Musdb	0.178	0.145
w/o MPOP600 and w/o Musdb	0.194	0.149

Table 2. Ablation study Objective similarity scores with reduced training data. All four settings are trained with Proposed (S) under the unseen-to-unseen scenario.

Conclusion

- Proposed unified model jointly supports the zero-shot SVC and SVS tasks, and has achieved state-of-the-art performance.
- Different design of the attention mechanism determines the trade-off between perceptual similarity and naturalness.
- Using a dataset containing a large number of singers for training is critical in improving zero-shot singing voice generation.