Abstract:

Phonetic Posteriorgrams based Many-to-Many Singing Voice Conversion via Adversarial Training

Haohan Guo1,2*, Heng Lu2, Na Hu2, Chunlei Zhang2, Shan Yang1,2, Lei Xie1, Dan Su2, Dong Yu2 1 School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
2 Tecent AI Lab, Beijing, China hhguo,lxie@nwpu-aslp.org, bearlu@tencent.com

arXiv:2012.01837v1 [cs.SD] 3 Dec 2020

Abstract

This paper describes an end-to-end adversarial singing voice conversion (EA-SVC) approach. It can directly generate arbitrary singing waveform by given phonetic posterior- gram (PPG) representing content, F0 representing pitch, and speaker embedding representing timbre, respectively. Pro- posed system is composed of three modules: generator G, the audio generation discriminator DA, and the feature disentan- glement discriminator DF . The generator G encodes the features in parallel and inversely transforms them into the target waveform. In order to make timbre conversion more stable and controllable, speaker embedding is further decomposed to the weighted sum of a group of trainable vectors represent- ing different timbre clusters. Further, to realize more robust and accurate singing conversion, disentanglement discrimina- torDF isproposedtoremovepitchandtimbrerelatedinformation that remains in the encoded PPG. Finally, a two-stage training is conducted to keep a stable and effective adversarial training process. Subjective evaluation results demonstrate the effectiveness of our proposed methods. Proposed system outperforms conventional cascade approach and the WaveNet based end-to-end approach in terms of both singing quality and singer similarity. Further objective analysis reveals that the model trained with the proposed two-stage training strat- egy can produce a smoother and sharper formant which leads to higher audio quality.

Conclusion

This paper proposes an adversarial training based end-to-end singing voice conversion approach. In the generator, encoder uses two CNN-BLSTM modules to encode PPG and F0 re- spectively, and decomposes the speaker embedding into a weight distribution of a group of trainable vectors represent- ing different timbre components. The adversarial training is applied in two aspects, audio generation and feature disen- tanglement. The multi-scale discriminator is use to adversarially train the generator to produce high-fidelity audio. Another discriminator aiming to map the encoded PPG to the encoded F0 and SE is proposed to remove overlapped information remained in PPG. A two-stage training strategy combining MR-STFT loss and adversarial loss is employed to keep a more stable and effective adversarial training process. We conduct MOS tests to evaluate the proposed meth- ods. The results show that EA-SVC achieves the best performance in both quality and similarity over the conventional cascade approach and the WaveNet based end-to-end apporach. Objective anlysis is also conducted to further demon- strate the effectiveness of our proposed modules.