7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414043

LITESING: TOWARDS FAST, LIGHTWEIGHT AND EXPRESSIVE SINGING VOICE SYNTHESIS

Xiaobin Zhuang, Tao Jiang, Szu-Yu Chou, Bin Wu, Peng Hu, Simon Lui

Tencent Music Entertainment, Shenzhen, China

ABSTRACT

LiteSing proposed in this paper is a high-quality singing voice synthesis (SVS) system, which is fast, lightweight and expressive. This model mainly stacks several non-autoregressive WaveNet blocks in the encoder and decoder under a generative adversarial architecture, predicts full conditions from the musical score, and generates acoustic features from these conditions. The full conditions in this paper consist of dynamic spectrogram energy, voiced/unvoiced (V/UV) decision

In recent years, some non-autoregressive sequence-to-sequence(Seq2Seq) models have been proposed to speed up the model inference. In [10], convolutional neural networks (CNNs) were used to model long-term dependencies of singing voices and fully convolutional networks (FCN) were employed to generate acoustic feature sequences in a segment-by-segment manner. WGANSing [11] introduced an adversarial singing synthesis approach based on U-Net architecture, optimized the network using the Wasserstein-GAN (WGAN) loss function [12]. XiaoiceSing [13] adopted