7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414348

SEQUENCE-TO-SEQUENCE SINGING VOICE SYNTHESIS WITH PERCEPTUAL ENTROPY LOSS

Jiatong Shi[1*], *Shuai Guo*[2*], *Nan Huo*[1], *Yuekai Zhang*[1], *Qin Jin*[2†]

[1] Johns Hopkins University, USA
[2] Renmin University of China, P.R.China

{jiatong_shi, nhuo1, yzhan400}@jhu.edu, {shuaiguo, qjin}@ruc.edu.cn

ABSTRACT

The neural network (NN) based singing voice synthesis (SVS) systems require sufficient data to train well and are are prone to over-fitting due to data scarcity. However, we often encounter data limitation problem in building SVS systems because of high data acquisition and annotation cost, . In this work, we propose a Perceptual Entropy (PE) loss derived from a psycho-acoustic hearing model to regularize the network. With a one-hour open-source singing voice database, we explore the impact of the PE loss on various main-

the generative adversarial network (GAN), are also shown to improve the synthesized singing quality [12–17].

As sequence-to-sequence (Seq2Seq) models have become the predominant architectures in neural-based TTS, state-of-the-art SVS systems have also adopted the encoder-decoder methods and showed improved performance over simple network structure (e.g., DNN, CNN, RNN) [17–23]. In these methods, the encoders and decoders vary from bi-directional Long-Short-Term Memory units (LSTM) to multi-head self-attention (MHSA) based blocks. However, unlike