Paper Number: 4187

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

Jiatong Shi[1*], *Shuai Guo*[2*], *Nan Huo*[1], *Yuekai Zhang*[1], *Qin Jin2*[†]

[1]Johns Hopkins University, USA [2]Renmin University of China, P.R.China
jiatong_shi@jhu.edu, {shuaiguo, qjin}@ruc.edu.cn

Introduction

The neural network (NN) based singing voice synthesis (SVS) systems require sufficient data to train well and 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.**

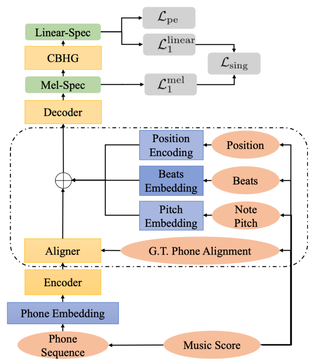
In this work, we propose a Perceptual Entropy (PE) loss

- Derived from the masking theory in the **psycho-acoustic** model of speech coding
- As **regularization term**, mitigate the over-fitting problem
- Significantly improve the synthesized singing quality reflected in objective and subjective evaluations

Model Structure

The Seq2Seq SVS System framework consists of four modules: an encoder, an encoder post-net, a decoder, and a decoder post-net.

- The model first accepts the phone sequence and encodes them into the phone embedding sequence. Next, the encoder converts the phone embedding into hidden states by considering the context in the sequence.
- The encoder post-net aligns the hidden states into pseudo acoustic segments, then applies encoded fundamental frequency, beats information, and positional encoding to the segments.
- The outputs are then fed into the decoder, which decodes the pseudo acoustic segments into Mel-spectrogram.
- The decoder post-net then converts the mel-spectrogram into a linear spectrogram. We use Griffin-Lim vocoder to generate waveforms from the linear spectrogram.



Perceptual Entropy Loss

➢ Motivation

We propose a perceptual entropy (PE) based loss as a **regularization factor** to alleviate the problem in network training.

The perceptual entropy (PE)

- Applies a psycho-acoustic model to compute the maximal perceptible information of an audio wave.
- A measure of the acoustic information that could be perceived by a human.
- Maximizing PE suggests a way of presenting more **human perceptive details**, which, in other words, adds penalties to non-perceptive acoustic information in the signal.
- Might increase the training errors but "regularize" the model to focus less on learning non-perceptive acoustic details.

➢ Calculation

The final PE at time t is defined as formulation 1, it indicates how much frequency information humans can perceive from a given audio signal with a specific quantization strategy intuitively.

To combine the PE with other loss functions, we define the PE loss as formulation 2, where PE is the mean perceptual entropy over time.

In training, the PE loss is interpolated with the synthesis network loss via a scaling hyper parameter λ in formulation 3.

$$PE(t) = \sum_{i=1}^{n} \sum_{\omega=l_{ti}}^{h_{ti}} \left[\log_2\left(2 \cdot \left|\frac{Re(\omega)}{\sqrt{6T_i'/k_i}}\right| + 1\right) + \log_2\left(2 \cdot \left|\frac{Im(\omega)}{\sqrt{6T_i'/k_i}}\right| + 1\right)\right] \quad (1)$$

$$\mathcal{L}_{pe} = \frac{1}{1 + PE} \quad (2)$$

$$\mathcal{L} = \mathcal{L}_{sing} + \lambda \cdot \mathcal{L}_{pe} \quad (3)$$

where i is the index of critical band; n is the number of bands given the sampling rate of the system; l_{ti} and h_{ti} are the upper and lower bounds of band i at time t; k_i is the number of frequency components in band i, T_i' is the masking threshold in band i. Re(ω) and Im(ω) denote the real and imaginary parts of the predicted spectrum.

Experiments

➢ Dataset – kiritan

- 65 mins in total, down-sample to 22050 HZ
- 467 phrases training, 18 validation, 10 testing
- labels (i.e., phones), pitches, and beats, 30 ms

➢ Model Settings

RNN:
Four bidirectional LSTM modules. The hidden size: 256, the layer number: three.
Weight of the PE loss λ : 0.01

Transformer:
Encoder module, a single 3x1 GLU block with 256 channels. Decoder module, six layers with 4 heads self-attention and 3x1 GLU blocks with 256 channels.
Weight of the PE loss λ : 0.01

Conformer:
Encoder, ten blocks of encoder layers, 256-dimension, four-heads self-attention layer . Decoder module employs the six blocks that include stacked MHSA layers and Gated Linear Units (GLU) layer.
Weight of the PE loss λ : 0.02

Objective Evaluation

Model	MCD(dB) – VAL SET	MCD(dB)	F0_CORR
RNN	3.41	7.19	0.79
RNN + PE	3.39	**5.99**	0.85
Transformer	3.38	6.48	0.85
Transformer + PE	3.52	6.60	0.84
Conformer	3.49	7.01	0.83
Conformer + PE	3.77	6.47	**0.89**

please refer to the paper for more metrics

- Huge gaps between the MCD value on the validation sets and test sets → over-fitting problem
- RNN and conformer, PE loss achieves better performance on all metrics
- RNN with the PE achieves the best MCD value, and the conformer with the PE loss shows favorable results on other metrics

Subjective Evaluation

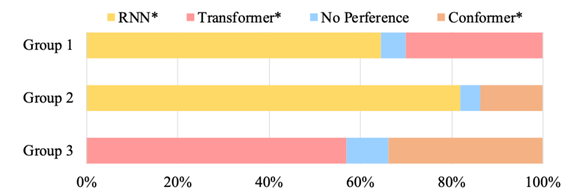
Two sets of A/B tests are conducted with different models

- randomly shuffled
- names of models are hidden during tests.
- 18 listeners in total

➢ First test
same song synthesized by model with and without PE loss

Result:
all listeners prefer the singing generated by a model trained with the PE loss, which indicates that the PE loss significantly improves the singing quality for RNN and Conformer architectures

➢ Second test
same song from two random models among RNN, transformer, and conformer(All with the PE loss).



Result:
performance: RNN model > Transformer > Conformer

✓ The synthesized sound examples can be found at https://peterguoruc.github.io/SVS_pe.github.io/
✓ Detailed further discussion can be found in our paper

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (No. 62072462) and the National Key R&D Program of China (No. 2020AAA0108600).