

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

Jiatong Shi^{1*}, Shuai Guo^{2*}, Nan Huo¹, Yuekai Zhang¹, 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 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 mainstream sequence-to-sequence models, including the RNN-based, transformer-based, and conformer-based models. Our experiments show that the PE loss can mitigate the over-fitting problem and significantly improve the synthesized singing quality reflected in objective and subjective evaluations.

Index Terms— Sequence-to-Sequence Singing Voice Synthesis, Perceptual Loss, Perceptual Entropy

1. INTRODUCTION

The singing voice synthesis (SVS) system utilizes both the musical (i.e., music score) and lyric information to synthesize human singing voices. Similar to the Text-to-Speech (TTS) task, the SVS task focuses on signal generation. However, SVS task has more requirements on the score and pitch, while the TTS has relatively loose restrictions on these factors. Conventional methods for the SVS task include concatenative methods [1–3] and statistical parametric methods [4, 5]. The concatenative methods generate waveform signals through concatenating specific singing units. Though these methods have shown high sound quality, they require large corpora to support the synthesis process. They also lack flexibility because they cannot generate sounds that are not in their stock corpora. The statistical parametric methods tackle the problem by modeling the singing signal from a statistical point of view. The most popular model among these methods is the Hidden Markov Model (HMM). It enables a flexible synthesizer but leads to loss of the naturalness [6].

In recent years, Deep Neural Network (DNN) based systems have achieved great performance in the synthesis field and showed their superiority over traditional HMM. The usage of the neural networks in SVS is similar to that in TTS. Firstly, the DNN model is proposed to predict the spectral information and begin to outperform conventional HMMs significantly [7, 8]. Later on, variations of the neural networks, including Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN), also demonstrate their power on acoustic modeling [6, 9–11]. Other architectures, such as

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 TTS with sufficient transcribed data, SVS suffers from data limitation due to its high data annotation cost and more strict copyright issues in the music domain. As speech is similar to singing, some works investigate to transfer knowledge from speech to singing. In [24], Valle et al. extract style information (e.t., rhythm, pitch, and other style tokens) from speech and extend them to SVS tasks. However, the method needs speech in advance to perform the style transfer. In [25], transfer learning from speech allows the synthesizer to generate singing voices with higher quality. However, the method requires parallel speech & singing corpora, which is difficult to obtain. Given the deep structure in Seq2Seq models, over-fitting would be a severe challenge to adapt the model for practical usage.

In this work, we propose a perceptual entropy (PE) loss that acts as a regularization term for the Seq2Seq SVS architecture. The PE is derived from the masking theory in the psycho-acoustic model of speech coding [26]. It is an inherent attribute of audio signals. As it models the audio signal’s perceptible information, the loss is designed to maximize the PE to regularize the networks. Additionally, we investigate different architectures for the Seq2Seq SVS, including the RNN-based model, the transformer-based model [18], and the conformer-based model. Our experiment results on the public Japanese singing voice database, Kiritan, show that the PE loss can significantly improve the synthesized singing quality in all three mainstream models, which is proved in objective and subjective evaluations. We also find that PE loss can help models produce better F0-contour and high-frequency-band spectrum prediction.

2. PERCEPTUAL ENTROPY LOSS

Compared to other speech synthesis tasks, the dataset for singing voice synthesis is small. Training using an end-to-end network architecture with a large number of parameters may lead to the over-fitting problem. We propose a perceptual entropy (PE) based loss as a regularization factor to alleviate the problem in network training.

As defined in [26], the perceptual entropy (PE) applies a psycho-acoustic model to compute the maximal perceptible information of an audio wave. Intuitively, the PE indicates how much frequency information humans can perceive from a given audio signal with a specific quantization strategy.

*Equal Contribution.

†Corresponding Author.

Model	VAL SET MCD(dB)	TEST SET			
		MCD(dB)	F0_RMSE(Hz)	VUV_ERROR(%)	F0_CORR
RNN	3.41	7.19	52.05	9.40	0.79
RNN*	3.39	5.99	45.53	5.26	0.85
Transformer	3.38	6.48	44.99	4.95	0.85
Transformer*	3.52	6.60	48.07	4.60	0.84
Conformer	3.49	7.01	47.82	6.69	0.83
Conformer*	3.77	6.47	40.79	4.53	0.89

Table 1. Evaluation of the synthesized singing quality based on different models with and without PE loss on the Kiritan dataset. * means that the model is trained with the PE loss. We apply four metrics including Mel-cepstrum distortion (MCD), F0 root mean square error (F0_RMSE), Voice/unvoiced error rate (VUV_ERROR), and Correlation coefficients of F0 measures (F0_CORR). Among them, larger F0_CORR means better performance, while for other metrics smaller means the better.

4. EXPERIMENTS

4.1. Dataset

We carry out experiments on an open-source Japanese singing voice database “Kiritan” [33]. It consists of 65 minutes of singing from a female singer. There are 50 songs in total. As there is no official split of the “Kiritan” database, we use 48 songs for training, 1 for validation, and 1 for testing. Follow previous works [6, 18], we split each song of several minutes of singing into phrases, resulting in 467 phrases for training, 18 for validation, and 10 for testing. The splitting is based on the silence between lyrics. We down-sample the songs to a sampling rate of 22050 in data pre-processing. The labels (i.e., phones), pitches, and beats information are quantized to 30ms to align with the resolution of our spectrogram.

4.2. Experiment Settings

We compare three network architectures, including RNN, transformer, and conformer-based models, respectively. We conduct a series of comparison experiments to investigate the impact of the PE loss with different network architectures. All the models follow the same framework as shown in Fig. 1, but with different types of encoders and decoders. The hyper-parameters in our experiments are selected based on the validation set. Detailed settings are as follows:

RNN: we use four bidirectional LSTM modules to process the embedding of phone, position, beats, and pitch at each time frame. The hidden size of these LSTM modules is 256, and the layer number is three. In training with the PE loss, We set λ defined in Eq. (8) as 0.01 and use the Adam optimizer with a 0.001 learning rate.

Transformer: For the transformer-based model, the encoder module consists of a single 3x1 GLU block with 256 channels. The decoder module consists of six layers with 4 heads self-attention and 3x1 GLU blocks with 256 channels. In the training stage with the PE loss, the ratio of PE loss λ defined in Eq. (8) is set as 0.01. The Adam optimizer with 0.001 learning rate and noam warm-up policy [34] are utilized in the training stage.

Conformer: The conformer encoder has ten blocks of encoder layers. Each encoder layer consists of a 256-dimension, four-heads self-attention layer with relative position representations. The size of linear units in feed-forward module is 1024 and the kernel size in each convolutional module is set as 7. The decoder module employs the six blocks that include stacked MHSA layers and Gated Linear Units (GLU) layer. In the training stage with the PE loss, the weight of the PE loss λ defined in Eq. (8) is set as 0.02. The Adam optimizer with 0.001 learning rate and OneCycle policy are utilized in the training stage [35].

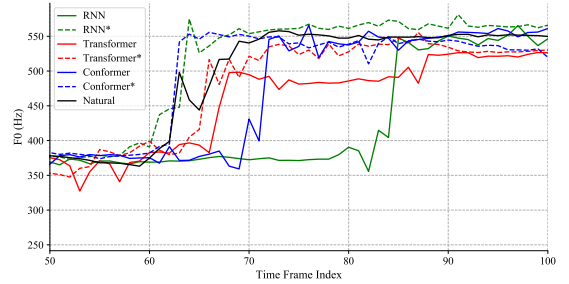


Fig. 2. F0 Contour prediction based on different models. * stands for models trained using PE loss. Natural is the ground truth singing.

The following settings are shared in all the experiments: the ground-truth Mel-spectrum uses 80 Mel-bins. The embedding size of the phone, pitch, and beats are 256. Global mean normalization is applied for outputs. The dropout rate in encoder and decoder is set as 0.1. We train these models for 300 epochs and choose the model with the lowest $\mathcal{L}_1^{\text{linear}}$ on the validation set.

5. RESULTS AND DISCUSSION

We carry out both objective and subjective evaluations to verify the effectiveness of our proposed model.¹

5.1. Objective Evaluations

For the objective evaluation, We utilize four metrics, including Mel-cepstrum distortion (MCD), F0 root mean square error (F0_RMSE), Voice/unvoiced error rate (VUV_ERROR), and Correlation coefficients of F0 measures (F0_CORR). We also report the MCD value on the validation set. The results are shown in Table 1. For RNN and conformer, the model with the PE loss achieves better performance on all metrics. While for transformer, PE loss only shows its benefit on V/VUV predictions. As for all the models, the transformer reaches the best MCD value, and the conformer with the PE loss shows favorable results on other metrics.

As shown in Table 1, there are gaps between the MCD value on the validation sets and test sets. This phenomenon shows that our training suffers from the over-fitting problem. The regularization from the PE loss, however, can alleviate the problem by reducing the MCD gap between the two sets for the RNN model and the conformer model. PE is a measure of the acoustic information that could be perceived by a human. For spectrogram prediction,

¹The synthesised sound examples can be found at https://peterguoruc.github.io/SVS_pe.github.io/

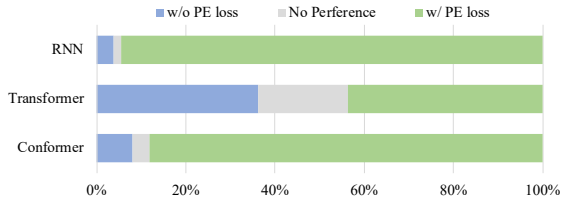


Fig. 3. The first set of Subjective Evaluation on the impact of the PE loss. We carry out pairwise A/B tests among the models with and without the PE Loss. The grey part in the middle means listeners have no preference between these two models.

maximizing PE suggests a way of presenting more human perceptible details, which, in other words, adds penalties to non-perceptive acoustic information in the signal. Those penalties might increase the training errors but “regularize” the model to focus less on learning non-perceptive acoustic details. Fig. 2 illustrates the F0 contours predicted from all models using 1sec in a test sample. The dashed line represents F0 contours predicted by models with the PE loss. The solid black line represents the ground truth F0 contour, and other solid lines represent F0 predicted by models without PE loss training. Models without PE loss training tend to deviate from the ground truth duration. The PE loss, however, shows to stabilize the decoding process and achieves better musical duration representation.

5.2. Subjective Evaluations

For the subjective evaluation, two sets of A/B tests are conducted with different models discussed in Section 4.2. In the first set, listeners are asked to compare two samples of the same song synthesized by a model with and without PE loss for all the three models.² In the second set, listeners compare two synthesized singing samples of the same song from two random models among RNN, transformer, and conformer. All the models in the second set are trained with the PE loss. Eighteen listeners in total participate in the subjective evaluation. As shown in Fig. 3, 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 ($p < 0.001$). We use the one-side 2-sample z-test to calculate the p-values, which sample size is 180.

Noted that the transformer model with the PE loss has worse objective performance than that without the PE loss, as shown in Table 1. However, the listeners prefer the synthesized singing from the model with the PE loss in the subjective evaluation. One possible reason is that the PE loss aims to maximize the human-perceptible information using masking theory. It selectively ignores some information that is not intelligible to humans. However, objective metrics focus on the whole spectrum. The ignored information due to the PE loss will lead to mismatches for the objective evaluation. As illustrated in Fig. 4, the RNN model achieves the best synthetic performance. The transformer model shows its superiority over the conformer model. All the results are significant ($p < 0.001$).

5.3. Further Discussion

Fig. 5 visualizes the time-frequency spectrum of one example synthesized singing from the six models described in Section 4. We can see from the figure that RNN demonstrates a strong ability to capture and reconstruct the formants as well as the harmonics. The better prediction offers smoother and less-jittery singing from the RNN-based model. It is such property that leads to the winning of

²The A/B pairs and models that generated the pairs are randomly shuffled for testing. Names of models are hidden during tests.

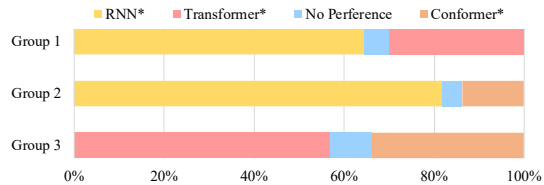


Fig. 4. The second set of Subjective Evaluation on the impact of different model architectures. We carry out pairwise A/B tests among the three models RNN, Transformer, and Conformer. All the three models are trained with the PE loss.

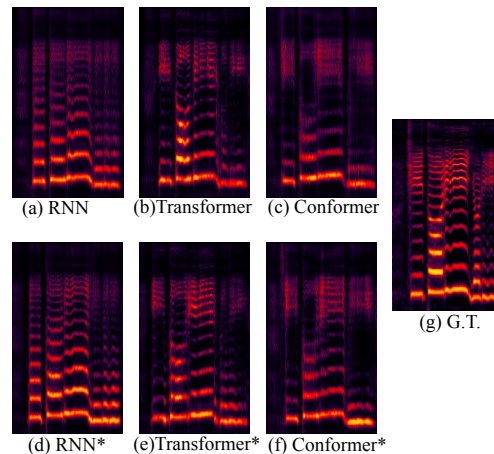


Fig. 5. Time-Frequency spectrum of an example synthesized singing from different models. * stands for models trained using the PE loss. G.T. refers to the ground truth singing voice.

the RNN-based model in the subjective evaluation. However, RNN-based model achieves the worst performance in the objective evaluation. This is due to its less effective phone envelope estimation compared to other models, as shown in Fig. 5a and Fig. 5b.

It is also worth mentioning that the PE loss is shown to be helpful for the network to reconstruct the information in the high-frequency band. From Fig. 5, we notice that there is more detailed formant information in the high-frequency-band of the spectrum for models using the PE loss. This is also a possible reason for significant performance improvement after applying the PE loss. This observation is interesting because the PE loss is computed on Bark-scale, which does not pay more attention to the high-frequency-band spectrum. We will further investigate this in our future works.

6. CONCLUSION

This paper presents the perceptual entropy-based loss as the regularization term to alleviate the over-fitting problem in training the singing voice synthesis networks. We explore the effectiveness of the PE loss with various mainstream Seq2Seq models, including the RNN-based model, transformer-based model, and conformer-based model. The PE loss brings significant performance improvement for all the models, which is verified in both objective and subjective evaluations. The PE loss benefits the F0-contour and high-frequency-band spectrum prediction as well.

7. 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).

8. REFERENCES

- [1] M. Macon, L. Jensen-Link, E. B. George, et al., “Concatenation-based midi-to-singing voice synthesis,” in *Proc. Audio Engineering Society Convention*. Audio Engineering Society, 1997.
- [2] H. Kenmochi and H. Ohshita, “Vocaloid-commercial singing synthesizer based on sample concatenation,” in *Proc. Interspeech*, 2007.
- [3] J. Bonada, M. Umberto M., and M. Blaauw, “Expressive singing synthesis based on unit selection for the singing synthesis challenge 2016,” in *Proc. Interspeech*, 2016.
- [4] K. Oura, A. Mase, T. Yamada, et al., “Recent development of the hmm-based singing voice synthesis system—sinsy,” in *Proc. Seventh ISCA Workshop on Speech Synthesis*, 2010.
- [5] K. Saino, H. Zen, Y. Nankaku, et al., “An hmm-based singing voice synthesis system,” in *Proc. Ninth International Conference on SLP*, 2006.
- [6] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” *Applied Sciences*, vol. 7, no. 12, pp. 1313, 2017.
- [7] M. Nishimura, K. Hashimoto, K. Oura, et al., “Singing voice synthesis based on deep neural networks,” in *Proc. Interspeech*, 2016, pp. 2478–2482.
- [8] Y. Hono, S. Murata, K. Nakamura, et al., “Recent development of the dnn-based singing voice synthesis system—sinsy,” in *Proc. APSIPA ASC*, 2018, pp. 1003–1009.
- [9] J. Kim, H. Choi, J. Park, et al., “Korean singing voice synthesis system based on an lstm recurrent neural network,” in *Proc. Interspeech*, 2018, pp. 1551–1555.
- [10] K. Nakamura, K. Hashimoto, K. Oura, et al., “Singing voice synthesis based on convolutional neural networks,” *arXiv preprint arXiv:1904.06868*, 2019.
- [11] K. Nakamura, S. Takaki, K. Hashimoto, et al., “Fast and high-quality singing voice synthesis system based on convolutional neural networks,” in *Proc. ICASSP*, 2020, pp. 7239–7243.
- [12] Y. Hono, K. Hashimoto, K. Oura, et al., “Singing voice synthesis based on generative adversarial networks,” in *Proc. ICASSP*, 2019, pp. 6955–6959.
- [13] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, “Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan,” in *Proc. EUSIPCO*, 2019, pp. 1–5.
- [14] O. Angelini, A. Moinet, K. Yanagisawa, and T. Drugman, “Singing synthesis: with a little help from my attention,” *arXiv preprint arXiv:1912.05881*, 2019.
- [15] J. Liu, Y. Chen, Y. Yeh, and Y. Yang, “Score and lyrics-free singing voice generation,” *arXiv preprint arXiv:1912.11747*, 2019.
- [16] S. Choi, W. Kim, S. Park, S. Yong, and J. Nam, “Korean singing voice synthesis based on auto-regressive boundary equilibrium gan,” in *Proc. ICASSP*, 2020, pp. 7234–7238.
- [17] J. Chen, X. Tan, J. Luan, et al., “Hifisinger: Towards high-fidelity neural singing voice synthesis,” *arXiv preprint arXiv:2009.01776*, 2020.
- [18] M. Blaauw and J. Bonada, “Sequence-to-sequence singing synthesis using the feed-forward transformer,” in *Proc. ICASSP*, 2020, pp. 7229–7233.
- [19] L. Zhang, C. Yu, H. Lu, et al., “Durian-sc: Duration informed attention network based singing voice conversion system,” *arXiv preprint arXiv:2008.03009*, 2020.
- [20] Y. Wu, S. Li, C. Yu, et al., “Peking opera synthesis via duration informed attention network,” *arXiv preprint arXiv:2008.03029*, 2020.
- [21] Y. Gu, X. Yin, Y. Rao, et al., “Bytesing: A chinese singing voice synthesis system using duration allocated encoder-decoder acoustic models and wavernn vocoders,” *arXiv preprint arXiv:2004.11012*, 2020.
- [22] P. Lu, J. Wu, J. Luan, et al., “Xiaoicesing: A high-quality and integrated singing voice synthesis system,” *arXiv preprint arXiv:2006.06261*, 2020.
- [23] Y. Ren, X. Tan, T. Qin, et al., “Deepsinger: Singing voice synthesis with data mined from the web,” in *Proc. ACM SIGKDD*, 2020, pp. 1979–1989.
- [24] R. Valle, J. Li, R. Prenger, and B. Catanzaro, “Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens,” in *Proc. ICASSP*, 2020, pp. 6189–6193.
- [25] L. Zhang, C. Yu, H. Lu, C. Weng, Y. Wu, X. Xie, Z. Li, and D. Yu, “Learning singing from speech,” *arXiv preprint arXiv:1912.10128*, 2019.
- [26] J. D. Johnston, “Transform coding of audio signals using perceptual noise criteria,” *IEEE Journal on selected areas in communications*, vol. 6, no. 2, pp. 314–323, 1988.
- [27] T. Painter and A. Spanias, “Perceptual coding of digital audio,” in *Proc. IEEE*, 2000, pp. 451–515.
- [28] N. S. Jayant and P. Noll, “Digital coding of waveforms: principles and applications to speech and video,” *Englewood Cliffs, NJ*, pp. 115–251, 1984.
- [29] N. Jayant, J. Johnston, and R. Safranek, “Signal compression based on models of human perception,” *Proc. the IEEE*, vol. 81, no. 10, pp. 1385–1422, 1993.
- [30] H. Fletcher, “Auditory patterns,” *Reviews of modern physics*, vol. 12, no. 1, pp. 47, 1940.
- [31] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *Proc. ICML*, 2017, pp. 933–941.
- [32] A. Gulati, J. Qin, C. Chiu, et al., “Conformer: Convolution-augmented transformer for speech recognition,” *arXiv preprint arXiv:2005.08100*, 2020.
- [33] Moris M., “Tohoku kiritan singing voice corpus,” <https://zunko.jp/kiridev/login.php>, Accessed: 2020.10.15.
- [34] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 5998–6008.
- [35] L. N. Smith, “A disciplined approach to neural network hyperparameters: Part 1—learning rate, batch size, momentum, and weight decay,” *arXiv preprint arXiv:1803.09820*, 2018.