

SEQUENCE-TO-SEQUENCE SINGING SYNTHESIS USING THE FEED-FORWARD TRANSFORMER

Merlijn Blaauw, Jordi Bonada

Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

merlijn.blaauw@upf.edu, jordi.bonada@upf.edu

ABSTRACT

We propose a sequence-to-sequence singing synthesizer, which avoids the need for training data with pre-aligned phonetic and acoustic features. Rather than the more common approach of a content-based attention mechanism combined with an autoregressive decoder, we use a different mechanism suitable for feed-forward synthesis. Given that phonetic timings in singing are highly constrained by the musical score, we derive an approximate initial alignment with the help of a simple duration model. Then, using a decoder based on a feed-forward variant of the Transformer model, a series of self-attention and convolutional layers refines the result of the initial alignment to reach the target acoustic features. Advantages of this approach include faster inference and avoiding the exposure bias issues that affect autoregressive models trained by teacher forcing. We evaluate the effectiveness of this model compared to an autoregressive baseline, the importance of self-attention, and the importance of the accuracy of the duration model.

Index Terms— Singing synthesis, sequence-to-sequence, self-attention, feed-forward, transformer

1. INTRODUCTION

In recent years, modern TTS systems have largely moved to sequence-to-sequence (Seq2Seq) models, e.g. [1, 2, 3], where the alignment between the phonetic or orthographic input sequence and the acoustic output sequence is learned during training and inferred during synthesis. One advantage of this approach is that it leads to a more end-to-end system, in the best case avoiding the need for pre-aligned training data, separate phonetic transcription, or a separate duration model at synthesis. For singing synthesis, not requiring pre-aligned training data is particularly attractive, as many existing tools (e.g. forced alignment with a HMM model) do not yield sufficiently accurate results on expressive singing, often requiring manual correction.

A common approach for Seq2Seq models in TTS is to use a content-based attention mechanism, e.g. [4], sometimes additionally using location-based information, e.g. [5]. As these mechanisms require access to acoustic information at inference, they are normally used in combination with an

autoregressive decoder. Recently, some systems have been proposed that use a feed-forward decoder and an alternative attention mechanism that does not rely on access to acoustic information [6, 7]. These notably provide faster, parallelizable inference, and are reported to produce more robust alignments with fewer mispronounced, repeated or skipped phonemes.

In the case of singing synthesis, this feed-forward approach is interesting as it avoids the exposure bias problem [8], caused by the discrepancy between teacher forced training and fully autoregressive inference. This problem can be especially noticeable in long sustained vowels where prediction errors tend to accumulate over time. Additionally, in our experience reaching similar quality results compared to non-Seq2Seq systems can be quite challenging with standard content-based attention mechanisms.

To facilitate evaluation of different systems, we only model timbre in this work, and assume F0 to be given. Although we use ground truth F0 extracted from recordings, it is feasible to predict F0 from the input score with an external model, or possibly predict it jointly. Related to this, we use WORLD vocoder features [9] as the output of our system, rather than the more commonly used mel-spectrogram features, as this allows exact control over the synthesized F0. For the best quality results, Seq2Seq systems are typically combined with a neural vocoder, e.g. [10, 11, 12], which can work well from both vocoder or mel-spectrogram features. However, in order to get a better idea of the performance of our model on its own, we do not use this approach in the experiments presented here.

The principal contributions of this paper are: 1. Propose a singing synthesizer based on the feed-forward Transformer with a practical Seq2Seq mechanism using an external duration model. 2. Evaluate the quality of this feed-forward model compared to a baseline autoregressive model. 3. Evaluate the importance of self-attention. 4. Evaluate the importance of the accuracy of the duration model used.

2. PROPOSED SYSTEM

In singing synthesis, the alignment between the input phonetic sequence and the output acoustic sequence is strongly constrained by the given musical score. This is a notable difference from TTS, which is generally only weakly constrained

by the (average) speech rate. Exploiting this fact, we propose to first generate an approximate initial alignment using note timings and a phoneme duration model. Once the input sequence is roughly aligned to the target output timesteps, we assume that the network is able to gradually refine the alignment through a series of transformations, until reaching something close to the target. Note that this approach is quite different from the approach using content-based attention, as here the initial alignment doesn't use any content at all.

An important point here is that we assume that the accuracy of the phoneme duration model is not critical to the end results. We assume that the decoder is powerful enough to be able to recover from errors in the initial alignment, to a certain degree. At the same time, the initial alignment can never hugely deviate from the true alignment, as it is heavily constrained by the note timings. To see if this assumption is correct, we purposely use a very simplistic duration model, based on average phoneme durations computed on a different dataset whose segmentation was corrected by hand. While language dependent, in this case the phoneme duration model is not singer dependent and the values could simply be copied from a table, without the need for any data with phonetic timings.

2.1. Model architecture

The input to our system is a musical score, consisting of a sequence of notes. Each note consists of an onset, duration, pitch, and a sequence of phonemes, typically corresponding to a syllable. In this work, we define the note onset as the vowel onset, and note end as the onset of the following vowel or silence. Additionally, we provide an external F0 to our system, in order to capture the affect of pitch on timbre. The output of our system is sequence of harmonic and aperiodic vocoder features, which in this case are simply concatenated.

The main components of our proposed system, as depicted in Fig. 1, are the encoder, the aligner and the decoder. The encoder takes the input phonetic sequence and computes a sequence of hidden states corresponding to each phoneme and their local context. The aligner provides a hard alignment by repeating these states according to the predicted phoneme durations, obtaining a sequence of the same length as the output acoustic sequence. Next, some additional conditioning signals derived from F0 and position are added. The decoder, based on the Transformer model [13], finally transforms the sequence of encoder hidden states to the target output sequence, through a series of self-attention and convolutional layers.

2.2. Encoder

Our encoder is based on the encoder proposed in [3]. First, embeddings are computed from each input phoneme. Then, a series of convolutional blocks with gated linear units (GLUs) [14] allows encoding information about the phonetic context of each phoneme, e.g. corresponding to triphones or pentaphones.

Finally a residual shortcut connection from the monophone embeddings is added to the local context output of the convolutional blocks.

2.3. Duration model

As noted, we purposely choose to use a very simplistic phoneme duration model in this work. It consists of a simple look-up table, populated with average phoneme durations computed from a dataset of a different singer with manually corrected phonetic segmentation. A simple heuristic is used to ensure that the sum of predicted phoneme durations matches the target note duration.

As we assume the note onset to correspond to the vowel onset, we first shift all onset consonants of each note to the preceding note (or silence). Then, we look up the sequence of average phoneme durations for each note, $[d_1, d_2, \dots, d_N]$, where N is the corresponding number of phonemes, and d_1 corresponds to the duration of the vowel. In order to match the target note duration, d_n , we use the predicted consonant durations and fill the remaining duration with the vowel. However, we also ensure at least half of the note's duration is occupied by the vowel by fixing $r_v = 0.5$. The scaling factor for the consonants, r_c , then becomes,

$$r_c = \begin{cases} 1 & \text{for } N = 1, \\ \min \left(1, \frac{d_n - \lfloor r_v d_n \rfloor}{\sum_{i=2}^N d_i} \right) & \text{otherwise.} \end{cases} \quad (1)$$

And, the adjusted phoneme durations, $[\hat{d}_1, \hat{d}_2, \dots, \hat{d}_N]$,

$$\hat{d}_i = \begin{cases} d_n - \sum_{j=2}^N \max(1, \lfloor r_c d_j \rfloor) & \text{for } i = 1, \\ \max(1, \lfloor r_c d_i \rfloor) & \text{for } i = 2, 3, \dots, N. \end{cases} \quad (2)$$

Note that all durations here are in integer number of frames, and that there are corrections for rounding errors and zero frame durations.

2.4. F0 and position conditioning

Continuous log F0 is encoded as a low dimensional vector between zero and one by evaluating several triangular basis functions whose centers are placed at frequencies appropriate for the training data's pitch range.

Transformers typically use an additive trigonometric positional encoding to give the self-attention blocks a sense of the position of their inputs, and provide a linear inductive bias along early on in training. However, in our case we found that a simple K -dimensional cyclical encoding of the normalized frame position within each note, $p \in [0, 1] \subset \mathbb{R}$, gave slightly better results,

$$v = \frac{1}{2} \cos \left(2\pi p - 2\pi \frac{k-1}{K} \right) + \frac{1}{2} \quad \text{for } k = 1 \dots K. \quad (3)$$

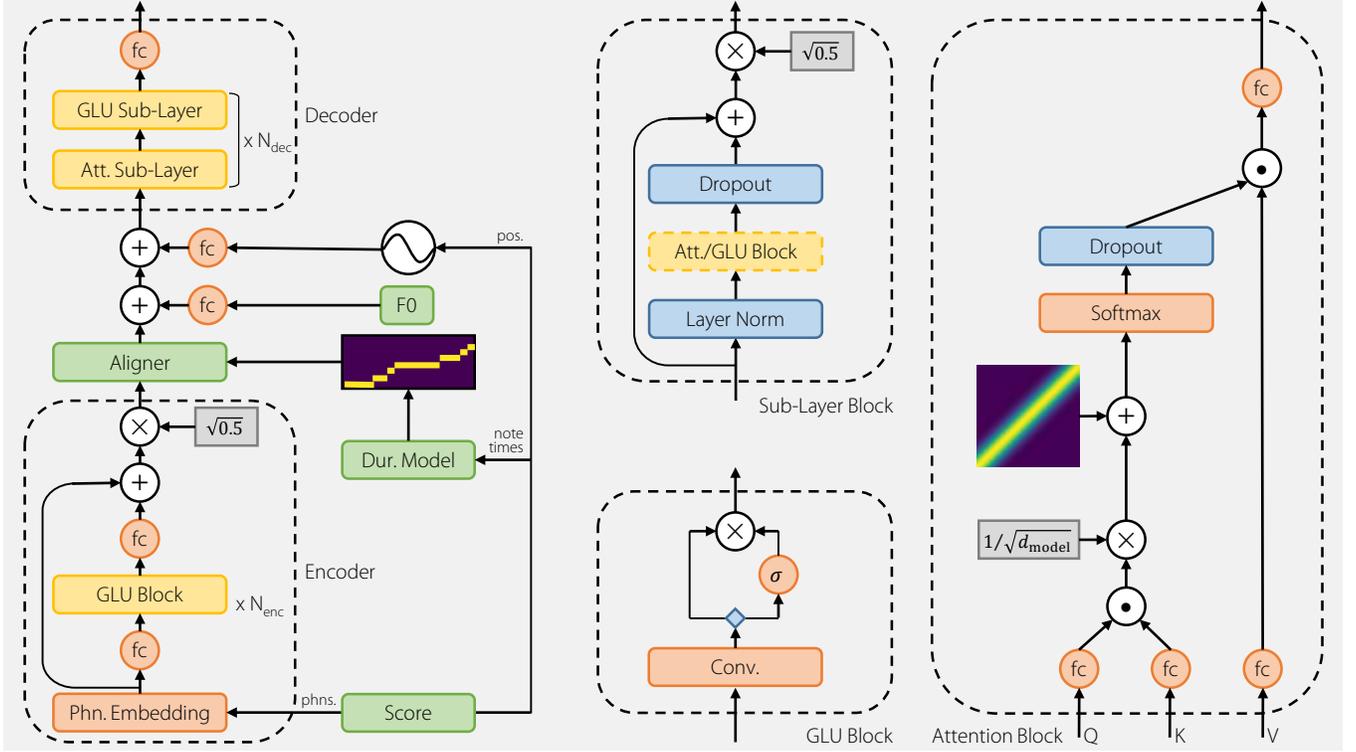


Fig. 1. A diagram of the complete model architecture. On the left is the full system, composed of encoder, aligner and decoder, which themselves are composed of different higher level blocks. On the right, these higher level blocks (sub-layer, gated linear unit (GLU) and attention) are shown in detail.

2.5. Decoder

Our decoder is based on a feed-forward variant of the Transformer model [13], similar to [7]. Each layer consists of a self-attention sub-layer block and a convolutional sub-layer block. Both sub-layers blocks have layer normalization [15], dropout and a residual shortcut connection.

Following [13], our self-attention blocks use the scaled dot product as a scoring function. Additionally, similar to [16], we bias the scores with a Gaussian along the diagonal to favor a more localized self-attention,

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_{\text{model}}}} + M \right) V, \quad (4)$$

$$M_{j,k} = \frac{-(j-k)^2}{2\sigma^2}, \quad (5)$$

where d_{model} is the dimensionality of the input vectors, $M \in \mathbb{R}^{T \times T}$ for sequence length T , and σ is a learned scale parameter. To reduce memory and computational requirements for the self-attention layers, we may use a reduction factor $r \geq 1$, which means r frames are predicted per output timestep [1, 3]. While the use of multi-head attention is typical for NLP applications, we did not find this improved results in our case.

For the convolutional blocks we use GLUs, which for our case outperform the 2-layer convolutional network with central ReLU activation typically used in Transformer architectures.

3. EXPERIMENTS

For the experiments in this work, we train a model on a proprietary dataset of 41 pop songs performed by a professional English male singer. From this dataset 35 songs were used for training (1 h 26 total), 4 for validation and 2 for testing.

Our proposed system uses 64-dimensional input features similar to [17], extracted with a 10 ms hop time. A reduction factor, $r = 2$, is used. We use 256-dimensional phoneme embeddings, and an encoder with a single 3x1 GLU block with 64 channels. F0 is coarse coded to a 4-dimensional vector, as is the position within the note, albeit with a cyclical encoding. The decoder consists of 6 layers with (single head) self-attention and 3x1 GLU blocks, all with 256 channels. The final output projection is to $64r$ channels. Dropout probability is set to 0.1 throughout the model. The learned standard deviation of the Gaussian bias of the self-attention blocks is initialized to 30. Initialization of convolutional layers follows [4]. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1 \times 10^{-9}$, and a batch size of 32. We follow the learning rate schedule from [13], with a 4000 step warm-up, a base learning rate of 1×10^{-3} , and a total of 50k updates. Additionally, we use Polyak averaging with a decay of 0.995 for validation and testing. The objective that we optimize is a simple L1 loss between output and target features.

We compare our proposed feed-forward model, which we

Table 1. Mean Opinion Score (MOS) ratings on a 1–5 scale with their respective 95% confidence intervals.

System	Mean Opinion Score
Hidden reference	4.49 ± 0.09
AR-NPSS	2.56 ± 0.10
FFT-NPSS (proposed)	2.79 ± 0.11
FFT-NPSS w/ ground truth dur.	2.87 ± 0.11
FFT-NPSS w/o self-attention	2.48 ± 0.10

label FFT-NPSS, to an autoregressive baseline model roughly following [17], labeled AR-NPSS. To study the importance of the accuracy of the approximate initial alignment, we train a version of our model, which uses ground truth phonetic durations rather than predictions by the simple averages duration model. Note that the baseline AR-NPSS is a non-Seq2Seq model, so it is also trained on ground truth phonetic durations. To study the importance of self-attention in the model we train a version of our model without attention blocks as well.

We ran a MOS listening test with 18 participants, which each rated a random subset of 12 out of 20 phrases. Per test 6 stimuli were presented; the 4 systems mentioned previously, and visible and hidden references consisting of a WORLD re-synthesis of the target recording. All systems are presented and rated together to encourage a comparison between them.

The results of our listening test are shown in Table 1. We can see that the FFT-NPSS system using ground truth phoneme durations performs best, but it is closely followed by the proposed Seq2Seq variant using a simple averages duration model. This shows that the initial alignment provided by the duration model has some importance, but it is not critical. Additionally, our proposed system outperforms the baseline autoregressive AR-NPSS system, possibly due to avoiding issues related to exposure bias. Finally, the variant of the FFT-NPSS system without self-attention layers performed worst overall, showing that self-attention is an important component for this kind of Seq2Seq mechanism, in our observations especially in terms of providing a coherent timbre over time. While all systems are still rated considerably below the reference WORLD re-synthesis, we expect that this gap would be reduced if we combine our system with a neural vocoder. Some sound examples, both with and without neural vocoder, are available online¹.

4. RELATION TO PRIOR WORK

Our work is most closely related to the recently proposed Fast-Speech model for TTS [7]. This model is also based on the feed-forward Transformer and an initial alignment obtained from a duration model. However, in this case the duration model is trained with the help of a teacher model based on an autoregressive Transformer [18], which is also used for generating the target mel-spectrogram features. We wanted to avoid

the need to train an autoregressive teacher model, as we found this generally challenging for the case of singing voice. Additionally, we apply some modifications to the architecture, such as the use of GLU convolutional blocks, alternative positional encoding and a Gaussian bias for the self-attention layers.

The ParaNet model [6] proposes a different approach to feed-forward TTS. Here, standard content-based encoder-decoder attention is used, but the model is trained with the help of attention distillation with an autoregressive teacher model based on [3]. Besides the reasons mentioned above, we found that the hard alignment used in our approach makes it easier to obtain a quality similar to non-Seq2Seq models, compared to the soft alignment of encoder-decoder attention.

In singing synthesis, the only Seq2Seq system we are aware of is [19]. This model is based on the DCTTS model [20], using content-based encoder-decoder attention, with autoregressive decoder. Similar to our approach, there is an initial alignment of the input states to the output timesteps. However, relying on the fact that the Korean syllable structure has at most one onset and one coda consonant, the first and last frame of the note are assigned to each consonant respectively, and the remaining frames are assigned to the vowel. After which, learning the attention alignment can be facilitated by using diagonally guided attention [20].

Non-Seq2Seq singing synthesizers include those based on autoregressive models [17, 21, 22], feed-forward CNN [23], and feed-forward GAN-based approaches [24, 25].

5. CONCLUSIONS

We presented a singing synthesizer based on the Transformer model, with a practical Seq2Seq mechanism allowing feed-forward operation. This approach allows training models without the need for pre-aligned training data, which can be cumbersome to prepare for singing data. Compared to a baseline autoregressive model, the proposed model allows for faster inference, avoids issues related to exposure bias, and rates as good or slightly better in listening tests. The use of self-attention resulted to be a key factor in obtaining good quality results, especially in terms of producing coherent timbre. As our model relies on an initial alignment provided by a duration model, we compared a very simplistic duration model to ground truth durations, to see the importance of the initial alignment’s accuracy. In listening tests, using ground truth durations was rated highest, but the difference was relatively small. While not shown due to lack of space, in our observations the model can recover from errors in the initial alignment, most likely thanks to the decoder’s non-causal convolutions and self-attention layers. For example, while the duration of phrase-final consonants tends to be systematically underpredicted by average durations, in the output of the synthesizer these phonemes have durations close to the target.

¹<https://mtg.github.io/singing-synthesis-demos/transformer/>

This work was funded by TROMPA H2020 No 770376.

6. REFERENCES

- [1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyriannakis, Rob Clark, and Rif A. Saurous, “Tacotron: A fully end-to-end text-to-speech synthesis model,” in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 4006–4010.
- [2] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyriannakis, and Yonghui Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.
- [3] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ö. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, “Deep Voice 3: Scaling text-to-speech with convolutional sequence learning,” in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.
- [4] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin, “Convolutional sequence to sequence learning,” in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1243–1252.
- [5] Alex Graves, “Generating sequences with recurrent neural networks,” *CoRR*, vol. abs/1308.0850, 2013.
- [6] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao, “Parallel neural text-to-speech,” *CoRR*, vol. abs/1905.08459, 2019.
- [7] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, “FastSpeech: Fast, robust and controllable text to speech,” in *Advances in Neural Information Processing Systems 32 (NIPS)*, 2019, pp. 3165–3174.
- [8] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, “Sequence level training with recurrent neural networks,” in *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016.
- [9] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” *IEICE Transactions on Information and Systems*, vol. E99.D, pp. 1877–1884, 2016.
- [10] Akira Tamamori, Tomoki Hayashi, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, “Speaker-dependent WaveNet vocoder,” in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech)*, 2017, pp. 1118–1122.
- [11] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 3617–3621.
- [12] Xin Wang, Shinji Takaki, and Junichi Yamagishi, “Neural source-filter-based waveform model for statistical parametric speech synthesis,” in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 5916–5920.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017, pp. 6000–6010.
- [14] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier, “Language modeling with gated convolutional networks,” in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 933–941.
- [15] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton, “Layer normalization,” *CoRR*, vol. abs/1607.06450, 2016.
- [16] Matthias Sperber, Jan Niehues, Graham Neubig, Sebastian Stüker, and Alex Waibel, “Self-attentional acoustic models,” in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech)*, 2018, pp. 3723–3727.
- [17] Merlijn Blaauw and Jordi Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” *Applied Sciences*, vol. 7, no. 12, 12 2017.
- [18] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, “Neural speech synthesis with transformer network,” in *The 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019, pp. 6706–6713.
- [19] Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, and Kyogu Lee, “Adversarially trained end-to-end Korean singing voice synthesis system,” in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2588–2592.
- [20] Hideyuki Tachibana, Katsuya Uenoyama, and Shunsuke Aihara, “Efficient trainable text-to-speech system based on deep convolutional networks with guided attention,” in *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 4784–4788.
- [21] Frederik Bous and Axel Roebel, “Analysing deep learning-spectral envelope prediction methods for singing synthesis,” in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, 2019.
- [22] Yuan-Hao Yi, Yang Ai, Zhen-Hua Ling, and Li-Rong Dai, “Singing voice synthesis using deep autoregressive neural networks for acoustic modeling,” in *Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech)*, 2019, pp. 2593–2597.
- [23] Kazuhiro Nakamura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, “Singing voice synthesis based on convolutional neural networks,” *CoRR*, vol. abs/1904.06868, 2019.
- [24] Yukiya Hono, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, “Singing voice synthesis based on generative adversarial networks,” in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 6955–6959.
- [25] Pritish Chandna, Merlijn Blaauw, Jordi Bonada, and Emilia Gómez, “WGANSing: A multi-voice singing voice synthesizer based on the Wasserstein-GAN,” in *Proceedings of the 27th European Signal Processing Conference (EUSIPCO)*, 2019.