# Rhythm-controllable Attention with High Robustness for Long Sentence Speech Synthesis

*Dengfeng Ke*[1,†], *Yayue Deng*[2,1,†], *Yukang Jia*[1,†], *Jinlong Xue*[2,†], *Qi Luo*[1], *Ya Li*[2,*], *Jianqing Sun*[3], *Jiaen Liang*[3], *Binghuai Lin*[4]

[1] Beijing Language and Culture University, Beijing, China
[2] School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[3] Unisound AI Technology Co., Ltd, Beijing, China
[4] Smart Platform Product Department, Tencent Technology Co., Ltd, China

dengfeng.ke@blcu.edu.cn, dengyayue.stu@gmail.com, marshal_jia@foxmail.com, jinlong_xue@bupt.edu.cn, lqttkxljys@163.com, yli01@bupt.edu.cn, sunjianqing@unisound.com, liangjiaen@unisound.com, binghuailin@tencent.com

## Abstract

Regressive Text-to-Speech (TTS) system utilizes attention mechanism to generate alignment between text and acoustic feature sequence. Alignment determines synthesis robustness (e.g, the occurence of skipping, repeating, and collapse) and rhythm via duration control. However, current attention algorithms used in speech synthesis cannot control rhythm using external duration information to generate natural speech while ensuring robustness. In this study, we propose Rhythm-controllable Attention (RC-Attention) based on Tracotron2, which improves robustness and naturalness simultaneously. Proposed attention adopts a trainable scalar learned from four kinds of information to achieve rhythm control, which makes rhythm control more robust and natural, even when synthesized sentences are extremely longer than training corpus. We use word errors counting and AB preference test to measure robustness of proposed method and naturalness of synthesized speech, respectively. Results shows that RC-Attention has the lowest word error rate of nearly 0.6%, compared with 11.8% for baseline system. Moreover, nearly 60% subjects prefer to the speech synthesized with RC-Attention to that with Forward Attention, because the former has more natural rhythm.

**Index Terms**: attention mechanism, robust speech synthesis, rhythm control

## 1. Introduction

Current end-to-end speech synthesis systems [1, 2, 3, 4, 5] have the ability to generate high-quality and human-like speech. Hence, speech synthesis application scenarios are becoming more and more diversified. All of these scenarios have two basic demands on synthesis systems. Firstly, speech corresponding to text can be synthesized accurately and does not exist repetition, skipping, or gibberish. This demand focuses on the robustness of synthesis system. Secondly, in addition to accuracy, people would expect synthesized speech is capable of generating natural rhythm of human. This demand focuses on model's ability of modeling and controlling of rhythm. Both of demands can be satisfied by a well-designed attention mechanism.

The problems of robustness in speech synthesis arise from extremely unequal length between text sequence and acoustic feature sequence. In the process of upsampling, if mapping relationship from text sequences to acoustic feature sequences, named alignment matrix, is not learned well, two types of alignment problems are likely to occur. One is attention collapse which will lead to gibberish. The other is blurred alignment which means acoustic model fails to focus on a single input token in a decoding step, leading to skipping and repeating.

Some studies [6, 7, 8, 9] propose robust attention mechanisms in order to alleviate difficulty of alignment learning and accumulation of errors in the inference stage. Raffel et al. [10] propose Monotonic Attention. The core idea is that when decoding the current timestep, model only needs to decide whether to focus on the current phoneme or move the focus forward. However, completeness of alignment cannot be satisfied, which may lead to skipping. He et al. [11] propose Stepwise Monotonic Attention based on Monotonic Attention. By adding a constraint of forward stride to monotonic attention, the method ensures that each phoneme can be covered by at least one frame of mel spectrogram. Another novel attention is propoesd by Graves et al. [12], named GMM Attention which uses multiple mixed Gaussian distributions to model alignment. Meanwhile, in order to make sure monotonicity, the mean value of mixed Gaussian is constrained to increase as decoding time goes by. Many speech synthesis systems use these attentions to speed up converge of training and improve robustness of synthesis [13, 14].

Methods of rhythm control are investigated by a few studies. For example, Zhang et al. [15] propose a forward algorithm, named Forward Attention which contains a transition agent to achieve rhythm control. However, robustness may be damaged after adding the external rhythm control agent. In the inferring phase, adjusting value of transition agent by human may leads to accumulation of errors.

Therefore, in this work, we combine advantages of these attentions and propose a novel attention mechanism, called Rhythm-controllable Attention (RC-Attention), to satisfy demands of robustness and rhythm control. Our attention mechanism can improve robustness compared with other advanced attention mechanisms, especially in long sentences synthesis, meanwhile utilizing external duration information to synthesize speech with more natural rhythm.

† *Equal contribution.* ∗ *Corresponding author*

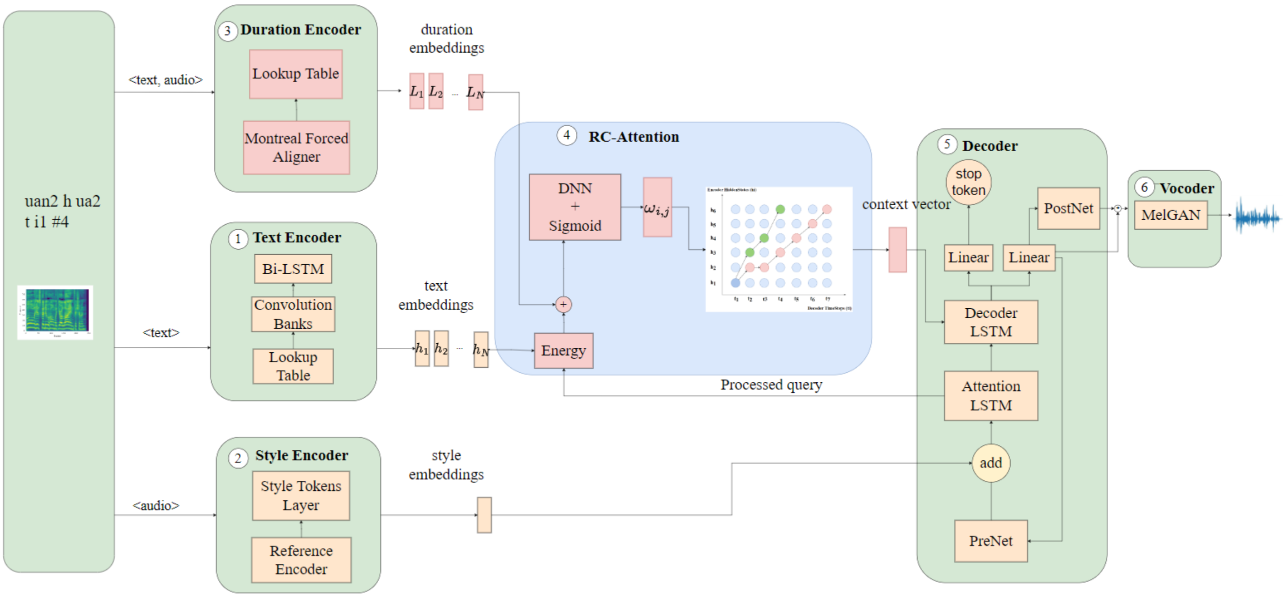


Figure 1: *The architecture of the whole speech synthesis model. Blue box shows computational procedure of RC-Attention. '$\omega_{i,j}$', '+' and 'add' represents a trainable scalar, concatenation and addition operation, respectively.*

## 2. Attention-based neural text-to-speech

Given an input sequence $X = [x_1, x_2, \ldots, x_N]$ with length N, text encoder produces a sequence of hidden state $H = [h_1, h_2, \ldots, h_N]$ which is convenient for attention mechanism to use, making the training more stable.

The mathematical representation of the first recurrent neural structure in the attention-based TTS is illustrated as follow:

$$s_i = LSTM\_Cell_{att}(s_{i-1}, c_{i-1}, y_{i-1}) \tag{1}$$

where $c_{i-1}$ and $y_{i-1}$ represent context vector and acoustic feature of previous decoding step, respectively. They are concatenated and sent into attention LSTM cell, producing current LSTM state $s_i$.

The general form of attention is shown in Eq. (2), where $e_{i,j}$ represents energy value of the $j$-th phoneme at the $i$-th decoding step.

$$e_{i,j} = Attention(s_i,\ h_j,\ \ldots) \tag{2}$$

$$a_{i,j} = Softmax(e_{i,j}) \tag{3}$$

After alignment vector $a_{i,j}$ is calculated as shown in Eq. (3), context vector $c_i$ is obtained by weighted sum of the hidden state sequences $H$.

$$c_i = \sum_{j=1}^{N} a_{i,j} h_j \tag{4}$$

## 3. Proposed method

Previous attention mechanisms or alignment tricks mostly focus on improving robustness of synthesis. Few alignment algorithms can achieve rhythm control by utilizing external duration information. Even if they can control rhythm, robustness would be damaged after adding rhythm controller. For example, after using Forward Attention with a transition agent, robust problems of synthesized speech would appear more frequently. Inspired by this, we integrate advantages of previous attention mechanisms and propose RC-Attention to achieve rhythm control in phoneme level without destroying robustness.

### 3.1. The whole speech synthesis system architecture

The overall text-to-speech system is illustrated in Fig. 1. Our model uses Tacotron2 [16] as backbone. The whole synthesis model contains six parts: 1) text encoder to embed input sequence into hidden state sequence, 2) style encoder to extract style features from input audio, 3) duration encoder to map duration information to a fixed vector, 4) RC-Attention to align hidden state sequence with acoustic feature sequence, 5) decoder to decode current acoustic feature frame and 6) vocoder to synthesize waveform of speech from acoustic feature.

In the training phase, *(text, audio)* pairs are sent into the model. Text encoder takes phoneme id sequence as input to extract text hidden state sequence, namely text embeddings in Fig. 1. Style encoder structure is the same as global style tokens [17], generating style embedding from mel spectrograms. Inspired by [18], we inject style embedding into the attention LSTM by adding it to the output of PreNet. Duration encoder takes *(text, audio)* pairs as input and utilizes forced alignment tool to get duration of each phoneme. The duration sequence is mapped into duration embeddings by lookup table. In the inferring phase, duration sequence can be externally specified to achieve external control of rhythm. The addition of style embedding and previous acoustic frame processed by PreNet. The result is fed into the first recurrent neural structure as shown in Eq. (1). The energy function adopts additive attention mechanism which takes current state of Attention LSTM and text embeddings as query and key, respectively. RC-Attention takes the concatenation of energy value and duration embeddings as input and calculates context vector. The calculative process will be illustrated in the Section 3.2.

### 3.2. Rhythm-controllable attention

RC-Attention is proposed based on two assumptions which guarantee robustness and natural rhythm of synthesized speech. Inspired by Forward Attention [15], the first assumption is illustrated as follow: The text encoder hidden state noticed by the current decoding step can only be the encoder hidden state noticed by the previous decoding step or next one. As shown in the Fig. 2, there are only seven decoding steps. Take the decoding step $t_4$ as example, index of max value in the alignment vector at decoding step $t_4$ can only be the index of max value in decoding step $t_3$ or its following one, which is $h_2$ or $h_3$ respectively. Besides, each column only has a maximum value, which means at this decoding step acoustic model would generate acoustic feature frame according to the respective encoder hidden state. Hence, the first assumption guarantees monotonicity and completeness of alignment path.

The second assumption utilizes a trainable scalar $\omega$ which learns from four kinds of information: acoustic information, text information, style information and duration information. The text information refers to the hidden state sequences of encoder. Duration information is the most direct rhythm information, which can realize rhythm control in the phoneme level. Style information contains an average rhythm of the entire sentence. We perform MFA [19] on ground true audio to extract duration information of each phoneme. In order to achieve external rhythm control, when calculating the $j$-th alignment weight at current decoding step $i$, $\omega_{i,j}$ represents the probability of focused hidden state remain unchanged while $1-\omega_{i,j}$ represents the probability of focused hidden state move one step forward.

Applying the above two assumptions to basic procedure mentioned in Sec. 2, mathematical expressions of RC-Attention are formulated as follows:

$$e_{i,j} = Tanh(q_i + h_j) \tag{5}$$

The energy function is the same as Bahdanau attention [20] where $q_i$ is processed hidden state of attention LSTM in the $i$-th step, which contains style information provided by global style tokens and acoustic information.

$$\hat{\omega_{i,j}} = DNN(\, e_{i,j}\,,\, L_{i,j}\,) \tag{6}$$

where $L_{i,j}$ is the duration embedding. The shape of extracted duration embedding sequences is [Batch size, Max Length] after padding as part of the input of attention mechanism. Here, we simply ultilize a linear project as the DNN layer.

$$\omega_{i,j} = \delta\,(\,\hat{\omega_{i,j}}\,) \tag{7}$$

where $\delta$ represents sigmoid function which ensures value of $\omega$ is between 0 and 1.

After alignment vector $a_{i,j}$ is calculated as shown in Eq. (3), , we apply a trainable scalar $\omega_{i,j}$ in our Rhythm-controllable Attention to control rhythm.

$$a_{i,j} = (1-\omega_{i,j-1})a_{i-1,j-1} + \omega_{i,j}\, a_{i-1,j} \tag{8}$$

The bigger the value of $\omega_{i,j}$, the more difficult it is to transfer the weight from the previous encoder hidden state, and the slower rhythm of speech.

## 4. Experiments

### 4.1. Experimental setup

To verify effectiveness of the proposed method, we compare widely used alignment methods. This section will introduce training corpus, compared models and implementation details.

**Training corpus:** Considered distribution of rhythm, we take an self-built emotional corpus as training corpus in which distribution of duration is more diverse. The training corpus contains text, audio, emotion label pairs, covering three different emotion categories (neutral, happy and angry). Each sentence has an average length of 14 words. The whole corpus consists of 7000 utterances, including 2334 happy utterances, 2332 angry utterances and 2334 neutral utterances. It has a total length of approximately 8.4 hours uttered by a young female. Audios are sampled at 48 kHz, but all utterances are downsampled to 22 kHz and are represented as 80 dimensional mel spectrograms in the experiments.

**Compared models:** To evaluate the effects of rhythm-controllable attention, models on which we conduct experiments for comparison include:

1. Baseline: Location Sensitive Attention [6].
2. GMM: GMMv2b [12] with five GMM mixture number.
3. FA: Forward Attention [15] with transition agent.
4. Proposed: Proposed model describes in Sec. 3

**Implementation details:** The parameter settings of text encoder, decoder and style encoder structure are the same as original Tacotron2 [16] and Global Style Tokens [17], respectively. For style encoder, style embedding channel and number of style tokens is equal to 256 and ten, respectively. A MelGAN vocoder [21] is used in our experiments as the vocoder. For GMMv2b Attention, we set GMM mixture number, delta bias and sigma bias to 5, 0.2 and 2.0, respectively. In our experiment, training corpus is randomly split into a training set (6900) and validation set (100). All models are trained for at least $100k$ steps with a batch size of 32 using the Adam optimizer [22]. In the inferring phrase, a reference utterance from each of styles is used.

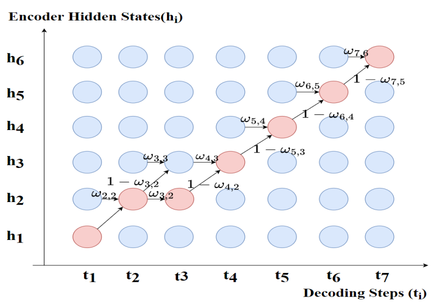


Figure 2: *Calculation process of alignment in RC-Attention. Vertical axis represents encoder hidden state sequence generated by text encoder. Horizontal axis represents decoding step.*

### 4.2. Robustness evaluation

To measure robustness of proposed method, each model synthesizes 40 speech samples for each emotion. Two subjects randomly choose ten samples from 120 speech synthesized by different models. All texts synthesized are out-of-domain sentences which have an average length of 147 words. Results are summarized in Table 1. Even though sentences are nearly ten times longer than sentences in training corpus, proposed attention can still achieve the best robustness.

Moreover, in order to intuitively demonstrate robustness of different attention mechanisms, we visualize alignment diagrams as shown in Fig. 3. For baseline, a blurred horizontal

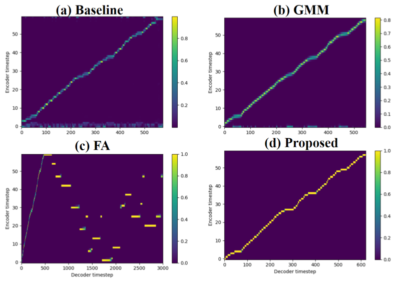


Figure 3: *Alignment visualization of different attentions.*

Table 1: *Number of word errors from selected models. The lower number means the better robustness performance and the bold indicates the best performance in all the models. *(Total 1426 words)**

| Model | Skipping | Repeating | Collapse |
|---|---|---|---|
| Baseline | 60 | 10 | 98 |
| GMM | 7 | **0** | 5 |
| FA | 10 | 4 | 4 |
| Propoesd | **6** | **0** | **3** |

line exists at the bottom of alignment diagram, which indicates that baseline would pay attention to the first few encoder hidden states instead of only focusing on one encoder hidden state at each decoding step. Even though alignment of GMMv2b attention shows a more clear diagonal trend than baseline and Forward Attention, there still exist blurred situations associated with pauses, which tends to generate speech with gibberish. Nevertheless, the proposed attention creates a clearer alignment without dispersing weights on unrelated hidden states at each time step.

### 4.3. Rhythm naturalness evaluation

To figure out which method produces audio with more natural rhythm, subjects are asked to select preferred samples according to overall impression on the naturalness of rhythm. For each model, we randomly generate 60 samples in which number of samples for every emotion is equal. Totally, there are 240 utterances judged by nine subjects in this experiment.

Result, summarized in Table 2, demonstrates that different alignment methods lead to different rhythm performance. Our proposed model performs better than baseline in all three emotions. Although rhythm generated by GMMv2b attention is slightly better than the proposed attention in term of angry, our proposed model can generate more natural utterances in most cases. Among these compared models, only Forward Attention and proposed attention are capable of controlling rhythm. It is noted that 60 percent of subjects prefer audio synthesized by proposed method rather than Forward Attention, which illustrates our proposed method achieves better control of rhythm. RC-Attention makes use of four kinds of information to study how to control rhythm. Moreover, it indirectly controls rhythm by influencing energy value. By contrast, Forward Attention try to learn the transition agent from only two kinds of information. Besides, it directly operates transition probability when controls rhythm. The latter method increases probability of occurrence of non-robustness problems, meanwhile rhythm of synthesized speech can be damaged.

Table 2: *AB preference test on naturalness, where "N/P" stands for no preference.*

| Emotion | Baseline | GMM | FA | Proposed | N/P |
|---|---|---|---|---|---|
| **neutral** | 7% | - | - | **55%** | 38% |
| | - | 12% | - | 32% | **56%** |
| | - | - | 3% | **86%** | 11% |
| **happy** | 21% | - | - | **44%** | 35% |
| | - | 11% | - | **60%** | 29% |
| | - | - | 29% | 32% | **39%** |
| **angry** | 8% | - | - | **54%** | 38% |
| | - | **41%** | - | 26% | 33% |
| | - | - | 13% | **62%** | 25% |

Table 3: *The results of subjective MOS tests for emotion expression.*

| Model | Happy | Angry | Avg |
|---|---|---|---|
| Baseline | 3.60 | 3.68 | 3.64 |
| GMM | 3.67 | 3.71 | 3.69 |
| FA | 3.55 | 3.63 | 3.59 |
| Propoesd | **3.70** | **3.79** | **3.75** |

### 4.4. Emotion expression evaluation

Different emotions have different patterns to control rhythm. Good rhythm control can promote emotion expression. Therefore, we evaluate emotional expression by subjective mean opinion score (MOS). Subjects are asked to give a score between 1 to 5 points in terms of emotion expression. Table 3 demonstrates results that RC-Attention produces best results in two emotions. Score of GMMv2b is the closest to RC-Attention, followed by the baseline and Forward Attention. We attribute Forward Attention's performance in emotional expressiveness to its blunt rhythmic control. The results in Section 4.3 show that poor control of rhythm in Forward Attention would lead to unnatural rhythm, furthermore, it would damage emotional expressiveness of synthesized speech.

## 5. Conclusions

In this work, we propose a novel attention mechanism, called RC-Attention, which integrates advantages of previous attentions. RC-attention achieves rhythm control in phoneme level without destroying robustness. We conduct objective and subjective experiments to evaluate robustness and naturalness of rhythm, respectively. Results demonstrate that RC-Attention can enhance robustness of synthesis compared with other attention mechanisms. Furthermore, our attention mechanism achieves rhythm control with more naturalness.

## 6. Acknowledgements

This work is supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (202200042) and New Talent Project of Beijing University of Posts and Telecommunications (2021RC37) and the research fund of China testing International Co., Ltd. on speech synthesis of Chinese listening text (HT-202011-374).

# 7. References


[1] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, p. 125, 2016.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[3] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," *arXiv preprint arXiv:2103.00993*, 2021.

[4] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[5] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-tts: A non-autoregressive network for text to speech based on flow," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7209–7213.

[6] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[7] Q. Tian, Z. Zhang, C. Liu, H. Lu, L. Chen, B. Wei, P. He, and S. Liu, "Feathertts: Robust and efficient attention based neural tts," *arXiv preprint arXiv:2011.00935*, 2020.

[8] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, T. Qin, and T.-Y. Liu, "Multispeech: Multi-speaker text to speech with transformer," *arXiv preprint arXiv:2006.04664*, 2020.

[9] C.-C. Chiu and C. Raffel, "Monotonic chunkwise attention," *arXiv preprint arXiv:1712.05382*, 2017.

[10] C. Raffel, M.-T. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2837–2846.

[11] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts," *arXiv preprint arXiv:1906.00672*, 2019.

[12] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[13] R. Liu, B. Sisman, and H. Li, "Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability," *arXiv preprint arXiv:2104.01408*, 2021.

[14] F. Yang, J. Luan, and Y. Wang, "Improving emotional speech synthesis by using sus-constrained vae and text encoder aggregation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8302–8306.

[15] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4789–4793.

[16] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[17] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[18] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447*, 2017.

[19] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, 2017, pp. 498–502.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, 2019, pp. 14 881–14 892.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.