Towards Multi-Scale Style Control for Expressive Speech Synthesis

Xiang Li[1,†], Changhe Song[1,†], Jingbei Li[1], Zhiyong Wu[1,2,∗], Jia Jia[1,3], Helen Meng[1,2]

[1]Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2]Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China

{xiang-li20, sch19, lijb19}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk,
jjia@tsinghua.edu.cn

Abstract

This paper introduces a multi-scale speech style modeling method for end-to-end expressive speech synthesis. The proposed method employs a multi-scale reference encoder to extract both the global-scale utterance-level and the local-scale quasi-phoneme-level style features of the target speech, which are then fed into the speech synthesis model as an extension to the input phoneme sequence. During training time, the multi-scale style model could be jointly trained with the speech synthesis model in an end-to-end fashion. By applying the proposed method to style transfer task, experimental results indicate that the controllability of the multi-scale speech style model and the expressiveness of the synthesized speech are greatly improved. Moreover, by assigning different reference speeches to extraction of style on each scale, the flexibility of the proposed method is further revealed.

Index Terms: text-to-speech, expressive speech synthesis, prosody, multi-scale, speech style

1. Introduction

Due to the prosperous development of deep learning, statistical parametric speech synthesis models built with artificial neural networks have achieved profound advancement on both the quality and naturalness of the synthesized speech [1, 2]. Popular end-to-end text-to-speech (TTS) models like Tacotron are now embodied with the ability to generate almost human-like audios with the support of high fidelity neural vocoder models [3, 4]. But the limited expressiveness of synthesized audio persists as one of the major gaps between synthesized speech and real human speech, which draws growing attention to expressive speech synthesis studies.

Frequently mentioned expressive speech synthesis methods like reference encoder introduce the idea of extracting a global style embedding from the given reference audio [5]. Global style token and their derivatives further manage to memorize and reproduce the global-scale style features of speech [6, 7]. Other more recently proposed methods seek to replace the global-scale latent style embedding in previous works with fine-grained latent prosody embedding sequence. [8] incorporates an additional reference attention mechanism to aligned the extracted prosody embedding sequence to phoneme sequence. Other researches like [9] achieve the same goal by resorting to force-alignment tools. [10] further proposes phoneme-scale content-independent style extracting module based on force-alignment.

However, most of these existing expressive speech synthesis methods focus on only style features of one single scale. On one hand, reference encoder and global style tokens accomplish the modeling of global-scale speech style, but suffer from the absence of local-scale style modeling, and are thus incapable of controlling the fine-grained prosody of the synthesized speech. On the other hand, fine-grained models usually find it quite challenging to recognize the utterance-level style context due to the removal of global-scale style features, and are consequently not provided with the ability to control the global style of the synthesized speech directly.

Meanwhile, the actual expressiveness of human speech can be perceived as a compound of multi-scale acoustic factors. One is the global-scale speech style, which includes, but not limited to the timbre and emotion of the speaker. Style of this level are supposed to be consistent throughout one single utterance. The other is the local-scale speech style, which comprises the speed, energy, pitch, pause and other acoustic features of the speech. Fluctuation of these features is frequently observed even among different locations within the same utterance.

Some latest researches are observed to devote effort to performing a multi-scale style modeling on some specific tasks like emotional speech synthesis. For example, [11] employs a learnt ranking function [12] to extract local-scale emotion intensity scalars of each phoneme in the utterance. The extracted emotion intensity sequence and a global-scale emotion label are then together fed into the TTS system in order to generate speech with desired strength fluctuation, as well as matching the target emotion category. But in these methods, style modules of different scales are not accommodated into the end-to-end train process of the whole TTS system, or heavily relies on auxiliary labels. This implies that the local-scale style modules are not aware of the utterance-level style context, and the learnt global-scale latent style embedding might eventually reach an average of different samples with the same label, rather than matching the global-scale style features of given utterance specifically.

With all the listed imperfections of previous works taken into consideration, this paper proposes a multi-scale speech style modeling method for end-to-end expressive speech synthesis, which is composed by a multi-scale reference encoder, a reference attention mechanism and a vanilla encoder-decoder-based TTS backbone.

The multi-scale reference encoder is trained to extract a global-scale latent style embedding (GSE) vector and a local-scale latent prosody embedding (LPE) sequence from the mel-spectrogram of the given reference speech. And Following [8], a reference attention is assigned to learn the alignment between LPE sequence and the phoneme sequence. Both the multi-scale
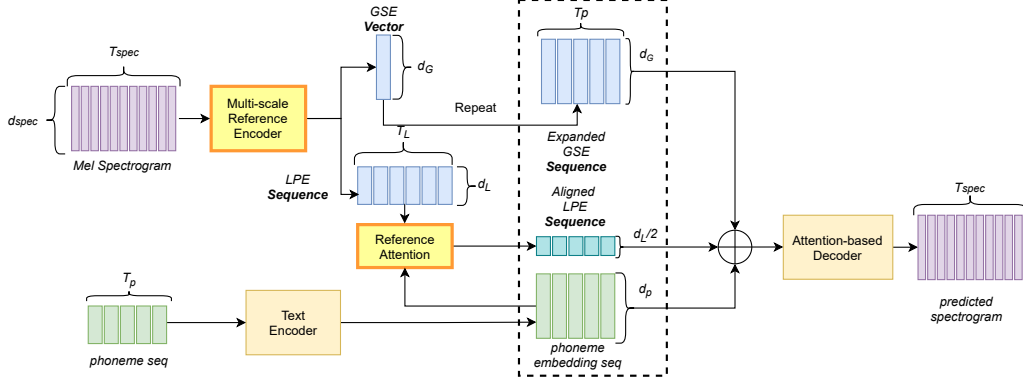
---

† Equal contribution.
∗ Corresponding author.

Figure 1: *Schematic diagram of the proposed model*

reference encoder and the reference attention are jointly trained with the TTS system in an end-to-end fashion. Moreover, instead of inheriting the frame-level granularity from the input spectrogram, downsampling is introduced in our multi-scale reference encoder to obtain intermediate latent features with the granularity closer to human spoken phoneme.

Benefiting from these design features, our method is able to control the global-scale and local-scale speech styles simultaneously by assigning reference audios. We prove the advantages of the proposed method by conducting style transfer experiment on an emotional dataset. Through additional cross-emotion and non-parallel transfer experiments, the flexibility and potentiality of our model is further revealed upon a closer inspection.

2. Methodology

Our multi-scale style modeling method consists of two major components: (i) A multi-scale reference encoder that extracts the global-scale style embedding (GSE) vector and the local-scale prosody embedding (LPE) sequence from the mel-spectrogram of the given reference speech; (ii) A reference attention mechanism which aligns the extracted LPE sequence to the phoneme embedding sequence produced by the text encoder. Based on above components, an end-to-end encoder-decoder based TTS backbone is constructed to achieve controllability and improved expressiveness.

2.1. Multi-scale expressive TTS system

As shown in Figure 1, given an input reference spectrogram with d_{spec} bands and T_{spec} frames, the multi-scale reference encoder generates a GSE vector with dimensionality of d_G, as well as an LPE sequence with dimensionality of d_L and length of T_L. The reference attention, which aims to align the LPE sequence to the phoneme embedding sequence, outputs an aligned LPE sequence. On the other hand, the GSE vector is repeated T_p times along time dimension into an extended GSE sequence $(d_G \times T_p)$. With the same length T_p, the aligned LPE sequence and the expanded GSE sequence are concatenated altogether with the phoneme embedding sequence, which produces a $(d_p + d_L/2 + d_G) \times T_p$ latent feature sequence further fed to the attention-based decoder.

2.2. Multi-scale reference encoder

To extract both global and local scale style features from the given reference speech, we introduce a specifically designed multi-scale reference encoder, which is made up by a stack of 6
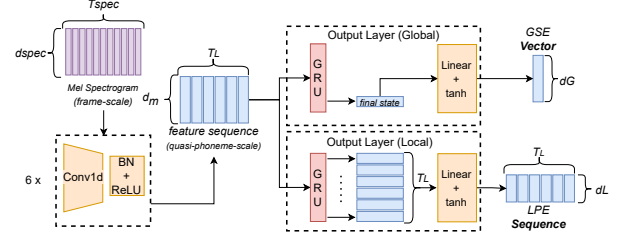

Figure 2: *Flowchart of the Multi-scale Reference Encoder*

convolution layers and 2 output layers, as shown in Figure 2.

2.2.1. 1D-Convolution downsampling layers

For the convolution layers, 1D-convolution along temporal dimension is adopted to help the reference attention to learn the alignment between the output local-scale prosody embedding sequence and the phoneme embedding sequence. Each of the convolution layers is composed by 3x1 filters, ReLU activation and batch normalization [13].

In addition, in order to regulate the temporal granularity of the convolution output closer to human vocal perception, the filter strides of 6 convolution layers are set as [2, 1, 2, 1, 2, 2]. Through this, the input spectrogram $(d_{spec} \times T_{spec})$ is down-sampled to $(d_m \times T_L)$, where d_m is the number of filters in the last convolution layer, and T_L is equal to $\lceil T_{spec}/16 \rceil$. For any input spectrogram with a common frame shift like 12.5 msec, the temporal granularity of the outcome features is approximately 200 msec, which is in line with the common duration range of consonant-vowel syllables from 150 msec to 200 msec [14]. The downsampling operation ensures that after the convolution stack, the temporal granularity of the intermediate feature sequence is properly reformed to a quasi-phoneme-scale. In this way, the robustness of reference attention is enhanced.

2.2.2. Scale-specific output layers

Output layers of local-scale and global-scale share the same structure, which is made up of consecutive Gated Recurrent Unit (GRU) [15] and fully-connected layer, as well as tanh activation. Both output layers take the above quasi-phoneme-scale feature sequence $(d_m \times T_L)$ as input. However, inside the global-scale output layer, only the final state of GRU is fed to subsequent layers. The outcome of global-scale output layer is forced to be a latent style vector $(d_G \times 1)$, while that of local-scale output layer is still a quasi-phoneme-scale prosody em-

Table 1: *MOS Evaluation Result (Average score and 95% confidence interval)*

| | Emotion (Global-scale style) | | | Fine-grained Prosody (Local-scale style) | | |
	Proposed	Base-G	Base-L	Proposed	Base-G	Base-L
Overall	**4.155 ± 0.042**	4.149 ± 0.041	2.955 ± 0.071	**4.136 ± 0.048**	3.471 ± 0.067	3.625 ± 0.054
Neutral	**4.425 ± 0.087**	4.308 ± 0.103	2.742 ± 0.207	**4.375 ± 0.098**	3.958 ± 0.143	3.825 ± 0.123
Angry	3.883 ± 0.138	**4.192 ± 0.111**	3.200 ± 0.151	**4.075 ± 0.138**	3.583 ± 0.179	3.558 ± 0.155
Fear	**4.225 ± 0.102**	4.167 ± 0.092	3.767 ± 0.146	**3.975 ± 0.149**	3.275 ± 0.168	3.575 ± 0.153
Disgust	**3.942 ± 0.126**	3.892 ± 0.125	2.325 ± 0.161	**3.983 ± 0.130**	3.350 ± 0.190	3.542 ± 0.142
Happy	**4.300 ± 0.091**	4.200 ± 0.099	2.192 ± 0.173	**4.208 ± 0.110**	3.200 ± 0.174	3.475 ± 0.150
Sad	**4.267 ± 0.095**	4.133 ± 0.105	3.833 ± 0.130	**4.200 ± 0.124**	3.492 ± 0.180	3.758 ± 0.135
Surprised	4.042 ± 0.113	**4.150 ± 0.108**	2.625 ± 0.178	**4.133 ± 0.119**	3.442 ± 0.193	3.642 ± 0.136

bedding sequence $(d_L \times T_L)$.

2.3. Reference attention mechanism

In terms of expressive speech synthesis, fine-grained style embedding sequence is usually reformed into sequence with the same length as phoneme sequence, in order that it can be used as an extension by the speech synthesis backbone.

In our model, scaled dot-product attention [16] is employed as the reference attention mechanism, which learns to find the alignment between the quasi-phoneme-scale LPE sequence $(d_L \times T_L)$ from the reference encoder and the target phoneme embedding sequence. The quasi-phoneme-scale LPE sequence is splitted along feature dimension into two sub-sequences $(d_L/2 \times T_L)$, which are fed as the key and value of the reference attention mechanism respectively. Meanwhile, the phoneme embedding sequence is fed as the query. Finally, the reference attention outputs an aligned LPE sequence $(d_L/2 \times T_p)$ with the same length as the phoneme embedding sequence.

2.4. Model usage

During training time, the whole end-to-end system are jointly trained. And the reference speech is exactly the target speech of the TTS system. Our first attempt on training the whole model altogether in one single session is deterred by the misalignment of reference attention. Thus, the training process is divided into 2 stages. In the first stage, the global-scale output layer of the reference encoder is not inserted into the model so that the local-scale style model and the reference attention is trained without disturbance. While in the second stage, the training of global-scale output layer is involved, and the parameters of text encoder, reference attention, as well as reference encoder (excluding the global-scale output layer) are frozen.

During inference time, by accepting input text and reference speech, the proposed model can synthesize speech with style close to the reference speech. Normally, the global and local scale share the same reference speech. But in our model, unique reference speech for each scale can be accepted respectively. Here, the reference speech for local scale is often required to have the same content with the input text, while the content of the reference speech for global scale is not restricted.

3. Experiments

3.1. Dataset and model details

An internal single-speaker emotional corpus on Mandarin is employed in our experiment. The emotions of all speeches are classified by 7 categories (neutral, angry, fear, disgust, happy, sad, surprised). A total of 28.79 hours speech data are formed

as 22,500 utterances, among which 10,500 belong to the neutral emotion, and the other part is evenly divided by the rest emotions. Additionally, about 15% of the whole dataset is parallel: for any text content in this part, there are 7 speech recordings corresponding to 7 different emotions respectively.

For the backbone of our expressive speech synthesis system, Tacotron 2 [4] with decoder reduction factor set to 3 is adopted. In the mutli-scale reference encoder, the number of GRU hidden units in both global and local output layer is set to 128. And in terms of dimensionalities of multi-scale style features, d_L is set to 6 as information bottleneck [8], while d_G is kept as 128. To obtain GSE vectors that are more distinguishable among different emotions, a vanilla emotion classifier is appended to the tail of the reference encoder module, which takes the extracted GSE vector as input and makes prediction on emotion category label. The classifier consists of 2 linear layers, and is jointly trained with other modules by a cross-entropy loss. As for waveform generation, HiFi-GAN [17] is chosen as our neural vocoder.

3.2. Mutli-scale style transfer

3.2.1. Ablation study

In order to demonstrate the influence of GSE and LPE in our proposed multi-scale speech style modeling method, two mono-scale style models are established as our baselines.

Base-G: The global-scale baseline model, which shares the same TTS backbone and multi-scale reference encoder, but excludes the reference attention mechanism and local output layer in reference encoder. It is an adapted version of [5] with the additional emotion classifier as training assistant.

Base-L: The local-scale baseline model, where the global output layer in reference encoder is removed. It is a slightly modified version of [8] with convolution layers in the reference encoder replaced by our 1D-convolution downsampling layers.

The style modeling abilities of the proposed model and the two baseline models are investigated by applying them to parallel style transfer task, where each of the tested model is given a reference audio with the same content as input text, to generate speech with style as close to the reference as possible.

In Figure 3, the 8-th character in the reference speech of test case 1 is followed by a pause (which are noted by the red rectangles). The pause in the speech generated by Base-G is mistakenly located in front of the 8-th character, revealing that the Base-G model is troubled with inaccurate transfer result on fined-grain speech prosody. On the other hand, as depicted in test case 2 of Figure 3, the formant characteristic of mel-spectrogram predicted by Base-L model is apparently different from the give reference speech. Which indicates that due to
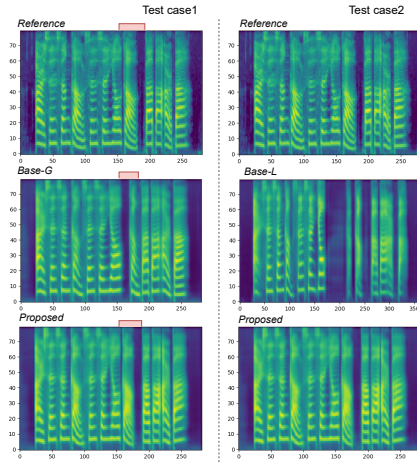
Figure 3: *Comparison on parallel transfer result. (Upper row: reference speech, middle row: speech produced by the Base-G (Test case 1) / Base-L (Test case 2) model, bottom row: speech produced by the proposed model.)*
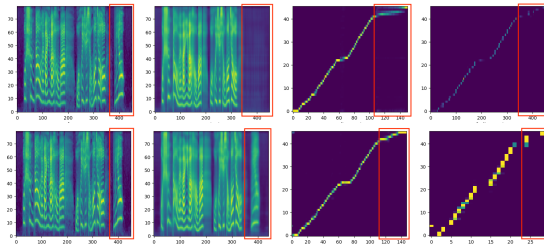

Figure 4: *Comparison on the robustness of reference attention (Upper row: test result on Base-FS model, lower row: test result on proposed quasi-phoneme-scale model. From left to right, each column is: reference speech, predicted speech, alignment of decoder attention, alignment of reference attention.)*

the absence of global style context, the global emotion styles of speeches generated by the Base-L model fail to match the original emotion. While none of these mistakes described above are observed in the speech generated by our proposed multi-scale model. And the enhancement on the similarity between synthesized speech and reference speech also shows the promoted controllability of the proposed model.

3.2.2. Subjective evaluation result

The described advantages of the proposed model is further verified by conducting Mean Opinion Score (MOS) test on the parallel style transfer result of the proposed and baseline models. An amount of 10 test reference audio groups are established, each of which is made up of 7 speech audios with the same text content but unique emotion category and various prosody. The participants are asked to give two subjective scores for the similarity of the synthesized speech to the reference speech, in terms of global emotion and local prosody respectively.

As presented in Table 1, the proposed model outperforms the baselines on the precision of both global-scale and local-scale style transfer. Between the baseline models, the global-scale Base-G model possesses noticeable outperformance on global emotion transfer, while the speech produced by local-scale Base-L model tends to be better controlled on reproducing fined-grained prosody. This indicates that since each mono-scale baseline style model possesses its own expertise in style

modeling on the corresponding scale, by integrating them into a whole multi-scale style model, significant improvement on the overall performance can be achieved.

3.3. Quasi-phoneme-scale v.s. Frame-scale

To show the impact of regulating the temporal granularity of style feature to quasi-phoneme-scale, another multi-scale model **Base-FS** is trained with the strides of all convolution layers in the reference encoder being 1. Thus, the temporal granularity of the intermediate feature sequence in the reference encoder and the extracted LSE sequence is rolled back to frame scale.

In Figure 4, given the same reference audio, Base-FS fails to synthesize the final character in the sentence (encompassed by the red rectangles), while the proposed quasi-phoneme-scale model is not encountered with such problem. And the alignment of the reference attention in Base-FS is evidently less bright in the area of the last character, which arouses confusion in the decoding process, and eventually damages the quality of synthesized speech. These phenomena reveal the robustness of the reference attention received considerable degradation due to the removal of downsampling.

3.4. Multi-reference style transfer

Different from previous mono-scale style models which only support accepting one single reference speech, our multi-scale model is innately adaptable to multiple reference speeches with considerable diversities. As long as the text content of the local reference is correspondent with the input text, arbitrary global reference speech could be adopted to transfer the embedded global-scale style to the synthesized speech, regardless of its local-scale style and content.

Experiment evidences have confirmed that, the transfer result preserves the local-scale prosody factors of the local-scale reference, while its global-scale emotion style is altered to match the global-scale reference. Since the effectiveness of global emotion transfer typically varies among the subjective perception of each and every individual, checking our demo page[1] for detailed samples is strongly recommended.

4. Conclusions

In this work, a novel multi-scale style modeling method for expressive speech synthesis is introduced. Experimental results support that the designed multi-scale scheme improves the effectiveness of style modeling on both global and local scales. Further attempts on cross-emotion and partially non-parallel style transfer brings an insight to the practical flexibility of our method, which indicates the significance of the proposed model towards multi-scale style control of the generated speech. In the future, dedicated discussion on the consistency of style transfer results among various emotions or speakers is believed to be promising research topics.

5. Acknowledgement

---

[1] https://thuhcsi.github.io/interspeech2021-multi-scale-style-control

6. References

[1] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2013, pp. 7962–7966.

[2] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2014, pp. 3829–3833.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech 2017*, 2017, pp. 4006–4010.

[4] 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.* IEEE, 2018, pp. 4779–4783.

[5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *International Conference on Machine Learning.* PMLR, 2018, pp. 4693–4702.

[6] 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.

[7] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference.* IEEE, 2019, pp. 623–627.

[8] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2019, pp. 5911–5915.

[9] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-Grained Robust Prosody Transfer for Single-Speaker Neural Text-To-Speech," in *Proc. Interspeech 2019*, 2019, pp. 4440–4444.

[10] T. Daxin and L. Tan, "Fine-grained style modelling and transfer in text-to-speech synthesis via content-style disentanglement," *arXiv preprint arXiv:2011.03943*, 2020.

[11] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *2021 IEEE Spoken Language Technology Workshop.* IEEE, 2021, pp. 423–430.

[12] X. Zhu, S. Yang, G. Yang, and L. Xie, "Controlling emotion strength with relative attribute for end-to-end speech synthesis," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop.* IEEE, 2019, pp. 192–199.

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning.* PMLR, 2015, pp. 448–456.

[14] M. Steinschneider, K. V. Nourski, and Y. I. Fishman, "Representation of speech in human auditory cortex: is it special?" *Hearing research*, vol. 305, pp. 57–73, 2013.

[15] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, A. Moschitti, B. Pang, and W. Daelemans, Eds. ACL, 2014, pp. 1724–1734.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[17] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.