Spoken Style Learning with Multi-modal Hierarchical Context Encoding for Conversational Text-to-Speech Synthesis

Jingbei Li[1], *Yi Meng*[1], *Chenyi Li*[1], *Zhiyong Wu*[1,2,∗], *Helen Meng*[2], *Chao Weng*[3], *Dan Su*[3]

[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]AI Lab, Tencent, Shenzhen, China

{lijb19,my20,licy20}@mails.tsinghua.edu.cn, {zywu,hmmeng}@se.cuhk.edu.hk,
{cweng,dansu}@tencent.com

Abstract

For conversational text-to-speech (TTS) systems, it is vital that the systems can adjust the spoken styles of synthesized speech according to different content and spoken styles in historical conversations. However, the study about learning spoken styles from historical conversations is still in its infancy. Only the transcripts of the historical conversations are considered, which neglects the spoken styles in historical speeches. Moreover, only the interactions of the global aspect between speakers are modeled, missing the party aspect self interactions inside each speaker. In this paper, to achieve better spoken style learning for conversational TTS, we propose a spoken style learning approach with multi-modal hierarchical context encoding. The textual information and spoken styles in the historical conversations are processed through multiple hierarchical recurrent neural networks to learn the spoken style related features in global and party aspects. The attention mechanism is further employed to summarize these features into a conversational context encoding. Experimental results demonstrate the effectiveness of our proposed approach, which outperform a baseline method using context encoding learnt only from the transcripts in global aspects, with MOS score on the naturalness of synthesized speech increasing from 3.138 to 3.408 and ABX preference rate exceeding the baseline method by 36.45%.

Index Terms: conversational text-to-speech synthesis, spoken style learning, human-computer speech interaction

1. Introduction

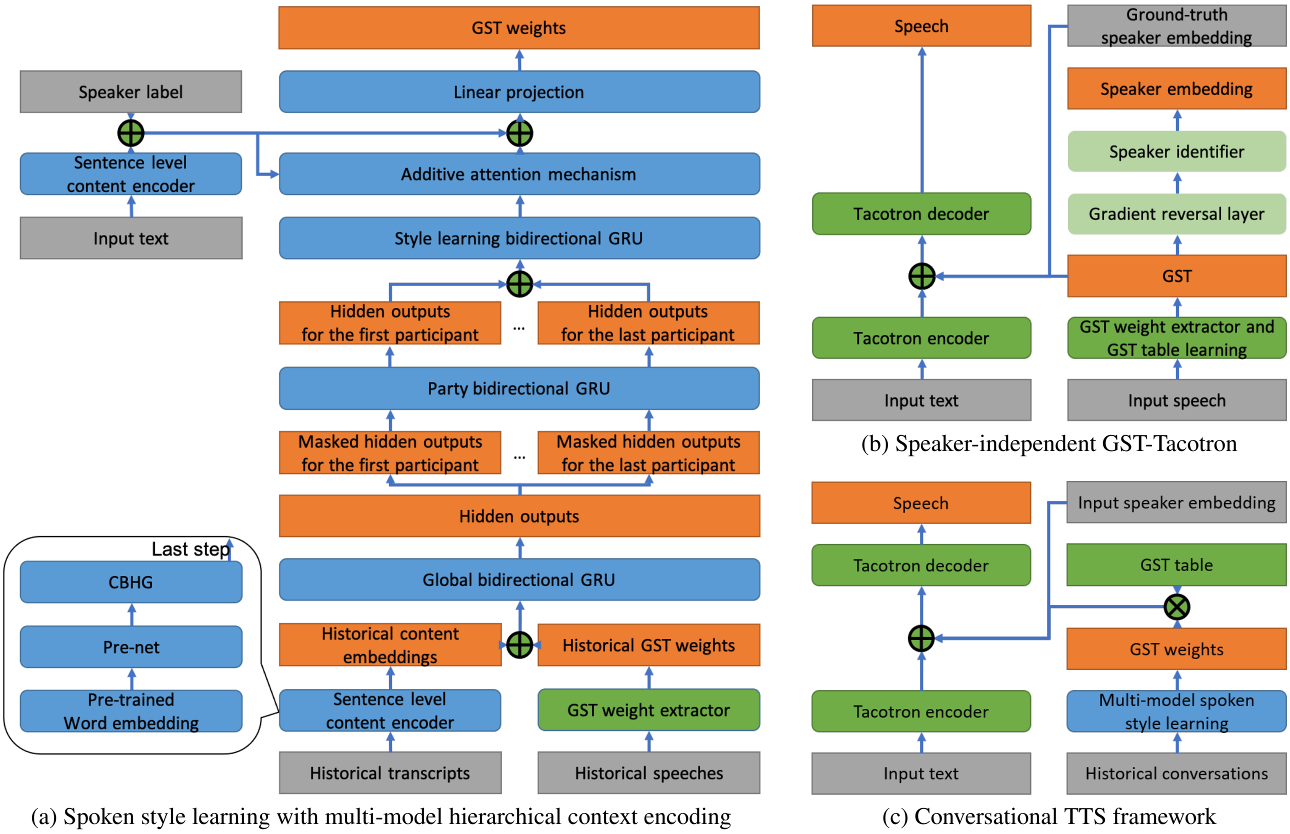
With the development of deep learning, neural network based text-to-speech (TTS) systems have achieve superior performance than conventional TTS systems [1, 2, 3, 4]. However, they usually generate synthesized speech with a fixed spoken style. While for conversational TTS systems with high levels of social ability [5], it is vital that the systems can adjust the spoken styles of synthesized speech according to different contents and spoken styles in historical conversations. In human-human conversations, people are interacting with different social signals such as humor, empathy and compassion [5] through the contents and spoken styles in their speeches. Therefore, for conversational TTS systems with given contents, learning spoken styles from textual information and spoken styles in historical conversations is of great significance for realizing more natural human-computer speech interactions.

However, such study of learning spoken styles from the historical conversations for conversational TTS systems is still in its early stage. Recently, a conversational context encoder [6] was proposed to learn a historical embedding from the historical conversations and adjust the spoken styles for conversational TTS by processing the sentence embeddings of historical conversations through a uni-direction gated recurrent unit (GRU) [7] network. This approach only deals with textual features in historical conversations, neglecting the spoken styles in historical dialogue. Moreover, only the global states [8] which model the interactions between different speakers are considered in this approach. However, the party states [8] which model the self interactions inside each speaker are also proved to be highly related to the state of the current speaker, are not considered.

In this paper, in order to improve the spoken style learning for conversational TTS, we propose a spoken style learning method with multi-modal hierarchical context encoding. Global style token (GST) [9] weights are extracted by a pre-trained multi-speaker GST enhanced Tacotron [1] system (hereafter, GST-Tacotron) in a speaker independent manner and used as the ground-truth spoken styles for the utterances in conversations. Then the sentence level content embeddings in the textual domain and the GST weights in the acoustic domain in historical dialogue are used to learn the spoken style related features through multiple hierarchical bidirectional GRUs consisting of a global bidirectional GRU and a party bidirectional GRU, aiming to respectively model the global and self interactions in the conversation. Attention mechanism [10] is further used to summarize the features that are most relevant to the current sentence from the output of the hierarchical bidirectional GRUs into historical context encodings. Then the historical context encodings are used to predict the GST weights of the current sentence. The encoder, decoder and GST table in the GST-Tacotron and the proposed spoken style learning model form our final conversational TTS framework. By giving the text and historical conversations, the conversational TTS framework predicts the GST of the current sentence from the historical conversations and synthesizes the speech using the pre-trained encoder and decoder.

Experimental results in both objective and subjective evaluations on a spontaneous conversational corpus demonstrate the effectiveness of our proposed method over the baseline approach using context encoding that is learnt from transcripts only in global aspects. The MSE of the predicted GST weights decreasing from 3.17×10^{-4} to 2.86×10^{-4}, MOS score on the naturalness of the synthesized speech increasing from 3.138 to

[∗] Corresponding author.



(a) Spoken style learning with multi-model hierarchical context encoding

(b) Speaker-independent GST-Tacotron

(c) Conversational TTS framework

Figure 1: *(a) Architecture of the proposed spoken style learning method with multi-model hierarchical context encoding. (b) Architecture of the speaker-independent global style token enhanced Tacotron framework. (c) Architecture of the conversational TTS framework.*

3.408 and ABX rate on the naturalness of synthesized speech exceeding the baseline by 36.45%. The effectiveness of the multi-model information and the hierarchical architecture are also explored and demonstrated through ablation studies.

2. Methodology

As shown in Figure 1, our work consists of three parts. In Section 2.1, we introduce our proposed spoken style learning method with multi-model hierarchical context encoding for conversational TTS. The ground-truth GST weights are extracted from a speaker-independent GST-Tacotron model which we will describe in Section 2.2. Then the encoder, decoder and GST table in GST-Tacotron and our proposed spoken style learning model form our final conversational TTS framework which is shown in Figure 1(c). By giving the text of the current sentence and historical conversations, the spoken style learning part predicts the GST weights from the historical conversations. The GST weights are further converted to the GST of the current sentence with the pre-trained GST table. Then the encoder and decoder in the GST-Tacotron synthesizes speech with the predicted GST and the given speaker embedding.

2.1. Spoken style learning with multi-modal hierarchical context encoding in conversational TTS

To model the spoken style related features from the content and spoken styles in historical conversations in both global and party aspects, we propose a spoken style learning approach with multi-modal hierarchical context encoding.

As shown in Figure 1(a), the transcripts of the historical conversations are converted to word embeddings by a pre-trained word embedding model [11]. Then a sentence level content encoder is used to encode the word embeddings into The content encoder consists of a pre-net and a CBHG network [12]. The pre-net consists of two fully connected layers which have 128 and 256 hidden units respectively with dropout probability 0.5. The CBHG network has the same structure with the CBHG network used in Tacotron [1], which consists of a bank of 1-D convolutional filters, a max pooling layer, four highway networks [13] and a bidirectional GRU network. After encoding, the output of the last step in the bidirectional GRU network is used as the content embedding of a given sentence.

We employ GST weights as the spoken style information of historical speeches. The GST weights are extracted by a GST weight extractor trained in a speaker-independent GST-Tacotron model which we will describe in Section 2.2.

Sentence embeddings which are the concatenation of the content embeddings and the GST weights for each sentence are used as the inputs for the following hierarchical networks. Three kinds of bidirectional GRU are hierarchically stacked to model the global and self interactions and learn spoken style related features in conversations, which all have 256 hidden units and dropout rate 0.9. The global bidirectional GRU takes the concatenated sentence embeddings of the historical conversations as inputs to model the interactions between the participants in the conversation. The outputs of the global bidirectional GRU are masked for each participant, resulting in multiple masked output sequences. In detail, for each participant, the outputs of the sentence spoken by the other participants are masked to zero. Party bidirectional GRU is used for each participant to model the self interactions inside each participant from the corresponding masked sequence. Then the outputs of party bidirectional GRU for each participant are concatenated and processed by the style learning bidirectional GRU to learn the spoken style related features in historical conversations. The above procedures can also be formulated as the following equations:

$$O^G = BiGRU^G(S) \tag{1}$$

$$O_i^P = BiGRU^P(O^G \odot Mask_i) \tag{2}$$

$$O^P = [O_1^P, \ldots, O_N^P] \tag{3}$$

$$O^S = BiGRU^S(O^P) \tag{4}$$

where S is the concatenated sentence embeddings of the historical conversations, $BiGRU^{\{G,P,S\}}$ are the global, party and style learning bidirectional GRU respectively, O^G is the outputs of the global bidirectional GRU, i is the i-th participant in the conversation, $Mask_i$ is a binary matrix in which only the elements that belong to the sentences spoken by the i-th participant are set to 1, O_i^P is the outputs of the party bidirectional GRU for the i-th participant, O^P is the concatenation of the outputs of the party bidirectional GRU for all N participants and O^S is the output of the style learning bidirectional GRU.

The attention mechanism [10] is further used to summarize the features that are most relevant to the current sentence. The content embedding of the current sentence and the one-hot label of the current speaker are concatenated as the sentence embedding of the current sentence. The relations between the sententce embedding of the current sentence and the outputs of the style learning bidirectional GRU are calculated by the additive attention mechanism [14] as attention weights. Then the outputs of the style learning bidirectional GRU are weighted summed into a conversational context embedding with the attention weights.

The sentence embedding of the current sentence and the conversational context embedding are then concatenated and projected to GST weights. A fully connected layer with dropout probability 0.5 is used for the projection.

The mean squared error (MSE) between the predicted and ground-truth GST weights is used as the loss function for our proposed spoken style learning approach.

2.2. Speaker-independent global style token learning

GST-Tacotron [9] is proposed to unsupervisely learn the non-textual information of speeches as different spoken styles. But the original GST-Tacotron may also embed speaker information into GSTs for conversations with multiple speakers. To proactively discourage the GSTs from also capturing speaker information in the conversational scenarios, we propose a speaker-independent GST learning method based on the original GST-Tacotron.

As shown in Figure 1(b), most parts of the model are same to those in the original GST-Tacotron. The GST of speech is represented as a weighted average of the GST bases in the GST table through a reference encoder and attention mechanism. Based on the original GST-Tacotron, we further employ domain adversarial training [15] to encourage the framework to encode GSTs in a speaker-independent manner by introducing a speaker identifier based on the GST and a gradient reversal layer [15]. The speaker identifier consists of a single fully connected layer and tries to predict the ground-truth speaker embeddings which are x-vectors [16] extracted by a pre-trained model from GSTs. The gradient reversal layer is inserted prior to the speaker identifier with weight -0.1. To recover the speech from the speaker-independent GST, the ground-truth speaker embedding are used as an additional input for the decoder. The MSE between the predicted and ground-truth x-vectors is used as an additional adversarial loss function.

After trained, the reference encoder and attention mechanism for GST learning form the GST weight extractor for our proposed spoken style learning framework. And the encoder, decoder and GST table are transferred to the conversaional TTS framework.

3. Experiments

3.1. Baseline approach

We implement a baseline approach based on a state-of-the-art conversational TTS system [6], which is shown in Figure 2.

The sentence level content encoder in our proposed approach is also employed to get the content embedding of the current sentence. The transcripts of the historical conversations are converted to the sentence embeddings by a pre-trained BERT model [17, 18]. A uni-directional GRU with 512 hidden units is used to encode the concatenation of the sentences embeddings and the one-hot speaker labels of the historical conversations into a state vector. Finally the state vector, the content embedding of the current sentence and the one-hot speaker label of the current sentence are concatenated and then projected through a fully connected layer to predict GST weights. The loss function is also the MSE between the predicted and ground-truth GST weights.

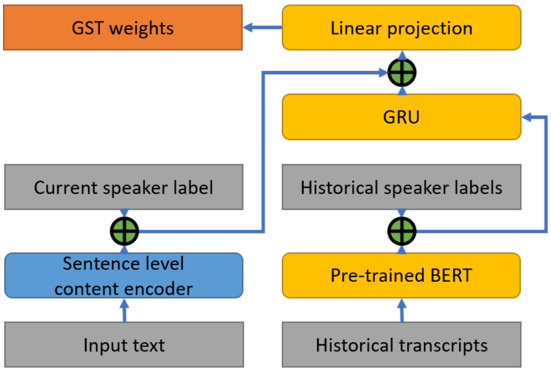


Figure 2: *Architecture of the baseline approach.*

3.2. Training setup

Our training process involves training the speaker-independent GST-Tacotron, followed by training the proposed spoken style learning model with the learnt GST weights as the ground-truths and inputs.

We use a spontaneous conversational corpus to train our models which has 16 hours of human-human conversations (dialogue sessions). These conversations are recorded in a sample rate of 16,000 from 194 native Chinese speakers lived in Beijing, which have rich spoken styles and a wide topic range. After dropping utterances with heavy spontaneous behaviors [19] such as multiple repeated words, the conversations are cut into

8,959 short chats. Each short chat consists of 6 utterances from two speakers, from which 5 utterances are used as the historical context.

We merge the utterances from all short chats to train the speaker-independent GST-Tacotron. The model is developed based on an open sourced TensorFlow implemention of the original GST-Tacotron [20]. Mel spectrograms are computed through a short time Fourier transform (STFT) using a 50 ms frame size, 12.5 ms frame hop, and a Hann window function. We transform the STFT magnitude to the mel scale using an 80 channel mel filterbank spanning 125 Hz to 7.6 kHz, followed by log dynamic range compression. The filterbank output magnitudes are clipped to a minimum value of 0.01. The learning rate is 10^{-3} exponentially decaying to 10^{-4} after 40,000 iterations. The model is trained for 300,000 iterations on a NVIDIA 2080 Ti GPU with a batch size of 32.

8,639 short chats extracted from 92 conversations in the corpus are used as the training set and 406 short chats extracted from another 5 conversations are used as the validation set to train our proposed spoken style learning model. The model is developed based on an open sourced PyTorch implemention of Tacotron [21]. The learning rate is fixed to 10^{-3}. The model is trained for 5,000 iterations on a NVIDIA 2080 Ti GPU with a batch size of 32.

For the baseline approach, the setup is same to the setup of the proposed approach.

3.3. Evaluations

28 short chats extracted from 2 conversations which differ to the conversations for the training and validation sets are used for the evaluations. Some of the synthesized samples are available at https://thuhcsi.github.io/interspeech2021-conversational-tts/.

We employ the MSE between the predicted and ground-truth GST weights as the objective evaluation metric. Subjective evaluations are conducted on the naturalness of synthesized speech between the baseline and proposed approaches. Speech files generated on the test set are rated by 25 listeners on a scale from 1 to 5 with 1 point increments, from which a subjective mean opinion score (MOS) is calculated. Meanwhile, the listeners are asked to choose a preferred speech from the speeches synthesized by the baseline and proposed approaches, from which ABX preference rates are calculated.

As shown in Table 1 and 2, the experimental results in both objective and subjective evaluations demonstrate the effectiveness of our proposed method over the baseline approach. The MSE of the predicted GST weights decreases from 3.17×10^{-4} to 2.86×10^{-4}, MOS score on the naturalness of the synthesized speech increases from 3.138 to 3.408 and ABX preference rate on the naturalness of the synthesized speech exceeds the baseline by 36.45%. It is also reported by the listeners that the speeches synthesized by the proposed approach have richer spoken styles such as variable speaking rate, emphasis and prosody.

Table 1: *Objective evaluations for different approaches*

Approach	**MSE**
Baseline	3.17×10^{-4}
+ with hierarchical context encoding	3.01×10^{-4}
Proposed	$\mathbf{2.86} \times 10^{-4}$
- without historical GST weights	2.96×10^{-4}

Table 2: *Subjective evaluations between the baseline and the proposed approaches. * NP stands for no preference.*

	Baseline	**NP***	**Proposed**
MOS	3.138 ± 0.064	-	$\mathbf{3.408 \pm 0.055}$
Preference	19.50%	24.55%	**55.95%**

3.4. Ablation studies

3.4.1. Effectiveness of the hierarchical context encoding

Based on the baseline approach, we implement another variant approach which migrates the hierarchical GRUs in the proposed approach for context encoding to explore the effectiveness of the proposed hierarchical context encoding method. The inputs for the hierarchical GRUs are the sentence embeddings of historical transcripts extracted by the pre-trained BERT in the baseline approach. Objective evaluation of this approach is conducted and shown in the second row of Table 1. Comparing with the baseline approach, the MSE decreases from 3.17×10^{-4} to 3.01×10^{-4} as expected, which demonstrates the effectiveness of the proposed hierarchical context encoding method.

3.4.2. Effectiveness of the historical spoken style information

Based on the proposed approach, we implement a variant approach which drops the historical GSTs from the inputs to explore the effectiveness of the historical spoken style information. Objective evaluation of this approach is conducted and shown in the last row of Table 1. Comparing with this variant approach, the MSE of the proposed approach decreases from 2.96×10^{-4} to 2.86×10^{-4}, showing the effectiveness of using multi-model information in conversational TTS.

4. Conclusions

To improve the prosody and naturalness of synthesized speech for conversaional TTS, in this paper, we proposed a spoken style learning method with multi-modal hierarchical context encoding to learn the spoken style from both the textual and acoustic information in historical speeches. The textual and acoustic information in the historical conversations are processed through multiple hierarchical bidirectional GRU networks to learn the spoken style related features in global and party aspects. Experimental results in both objective and subjective evaluations demonstrate that our proposed approach outperforms a baseline approach using context encoding learnt from only transcripts in global aspect. In ablation studies, the effectiveness of the multi-model information and the hierarchical architecture are also explored and demonstrated.

5. Acknowledgements

This work is supported by National Natural Science Foundation of China (NSFC) (62076144, 61433018), the joint research fund of NSFC-RGC (Research Grant Council of Hong Kong) (61531166002, N_CUHK404/15), the Major Project of National Social Science Foundation of China (NSSF) (13&ZD189). We would also like to thank 2020 Tencent Rhino-Bird Elite Training Program, Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202041) and Tsinghua University - Tencent Joint Laboratory for the support.

6. References

[1] Y. Wang, R. J. 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," *arXiv:1703.10135 [cs]*, Mar. 2017, arXiv: 1703.10135. [Online]. Available: http://arxiv.org/abs/1703.10135

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *arXiv:1712.05884 [cs]*, Dec. 2017, arXiv: 1712.05884. [Online]. Available: http://arxiv.org/abs/1712.05884

[3] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep Voice 2: Multi-Speaker Neural Text-to-Speech," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 2962–2970. [Online]. Available: http://papers.nips.cc/paper/6889-deep-voice-2-multi-speaker-neural-text-to-speech.pdf

[4] 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)*, May 2020, pp. 7209–7213, iSSN: 2379-190X.

[5] A. Vinciarelli, A. Esposito, E. André, F. Bonin, M. Chetouani, J. F. Cohn, M. Cristani, F. Fuhrmann, E. Gilmartin, Z. Hammal, D. Heylen, R. Kaiser, M. Koutsombogera, A. Potamianos, S. Renals, G. Riccardi, and A. A. Salah, "Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions," *Cognitive Computation*, vol. 7, no. 4, pp. 397–413, Aug. 2015. [Online]. Available: https://doi.org/10.1007/s12559-015-9326-z

[6] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational End-to-End TTS for Voice Agent," *arXiv:2005.10438 [cs, eess]*, May 2020, arXiv: 2005.10438. [Online]. Available: http://arxiv.org/abs/2005.10438

[7] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv:1406.1078 [cs, stat]*, Jun. 2014, arXiv: 1406.1078. [Online]. Available: http://arxiv.org/abs/1406.1078

[8] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," *arXiv:1811.00405 [cs]*, May 2019, arXiv: 1811.00405. [Online]. Available: http://arxiv.org/abs/1811.00405

[9] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," *arXiv:1803.09017 [cs, eess]*, Mar. 2018, arXiv: 1803.09017. [Online]. Available: http://arxiv.org/abs/1803.09017

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[11] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 175–180. [Online]. Available: https://www.aclweb.org/anthology/N18-2028

[12] J. Lee, K. Cho, and T. Hofmann, "Fully Character-Level Neural Machine Translation without Explicit Segmentation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 365–378, 2017. [Online]. Available: https://www.aclweb.org/anthology/Q17-1026

[13] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway Networks," *arXiv:1505.00387 [cs]*, Nov. 2015, arXiv: 1505.00387. [Online]. Available: http://arxiv.org/abs/1505.00387

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv:1409.0473 [cs, stat]*, Sep. 2014, arXiv: 1409.0473. [Online]. Available: http://arxiv.org/abs/1409.0473

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-Adversarial Training of Neural Networks," *arXiv:1505.07818 [cs, stat]*, May 2016, arXiv: 1505.07818. [Online]. Available: http://arxiv.org/abs/1505.07818

[16] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 5329–5333, iSSN: 2379-190X.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, Oct. 2018, arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[18] Y. Cui, W. Che, T. Liu, B. Qin, S. Wang, and G. Hu, "Revisiting Pre-Trained Models for Chinese Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. Online: Association for Computational Linguistics, 2020, pp. 657–668. [Online]. Available: https://www.aclweb.org/anthology/2020.findings-emnlp.58

[19] R. Qader, G. Lecorvé, D. Lolive, and P. Sébillot, "Disfluency Insertion for Spontaneous TTS: Formalization and Proof of Concept," in *Statistical Language and Speech Processing*, ser. Lecture Notes in Computer Science, T. Dutoit, C. Martín-Vide, and G. Pironkov, Eds. Cham: Springer International Publishing, 2018, pp. 32–44.

[20] S. Yang, "A tensorflow implementation of the "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis"," Mar. 2021, original-date: 2018-03-31T08:14:33Z. [Online]. Available: https://github.com/syang1993/gst-tacotron

[21] Y. Ryuichi, "PyTorch implementation of Tacotron speech synthesis model." [Online]. Available: https://github.com/r9y9/tacotron_pytorch