INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Knowledge-based Linguistic Encoding for End-to-End Mandarin Text-to-Speech Synthesis

Jingbei Li[1], *Zhiyong Wu*[1], *Runnan Li*[1], *Pengpeng Zhi*[2], *Song Yang*[2], *Helen Meng*[3]

[1]Department of Computer Science and Technology, Tsinghua University
[2]AI Lab, TAL Education Group
[3]The Chinese University of Hong Kong

jb-li15@mails.tsinghua.edu.cn, zywu@sz.tsinghua.edu.cn, lirn15@mails.tsinghua.edu.cn,
{zhipengpeng, yangsong1}@100tal.com, hmmeng@se.cuhk.edu.hk

Abstract

Recent researches have shown superior performance of applying end-to-end architecture in text-to-speech (TTS) synthesis. However, considering the complex linguistic structure of Chinese, using Chinese characters directly for Mandarin TTS may suffer from the poor linguistic encoding performance, resulting in improper word tokenization and pronunciation errors. To ensure the naturalness and intelligibility of synthetic speech, state-of-the-art Mandarin TTS systems employ a list of components, such as word tokenization, part-of-speech (POS) tagging and grapheme-to-phoneme (G2P) conversion, to produce knowledge-enhanced inputs to alleviate the problems caused by linguistic encoding. These components are based on linguistic expertise and well-designed, but trained individually, leading to errors compounding for the TTS system. In this paper, to reduce the complexity of Mandarin TTS system and bring further improvement, we proposed a knowledge-based linguistic encoder for the character-based end-to-end Mandarin TTS system. Developed with multi-task learning structure, the proposed encoder can learn from linguistic analysis subtasks, providing robust and discriminative linguistic encodings for the following speech generation decoder. Experimental results demonstrate the effectiveness of the proposed framework, with word tokenization error dropped from 12.81% to 1.58%, syllable pronunciation error dropped from 10.89% to 2.81% compared with state-of-the-art baseline approach, providing mean opinion score (MOS) improvement from 3.76 to 3.87.

Index Terms: end-to-end text-to-speech system, knowledge-based learning, linguistic encoding, multi-task learning

1. Introduction

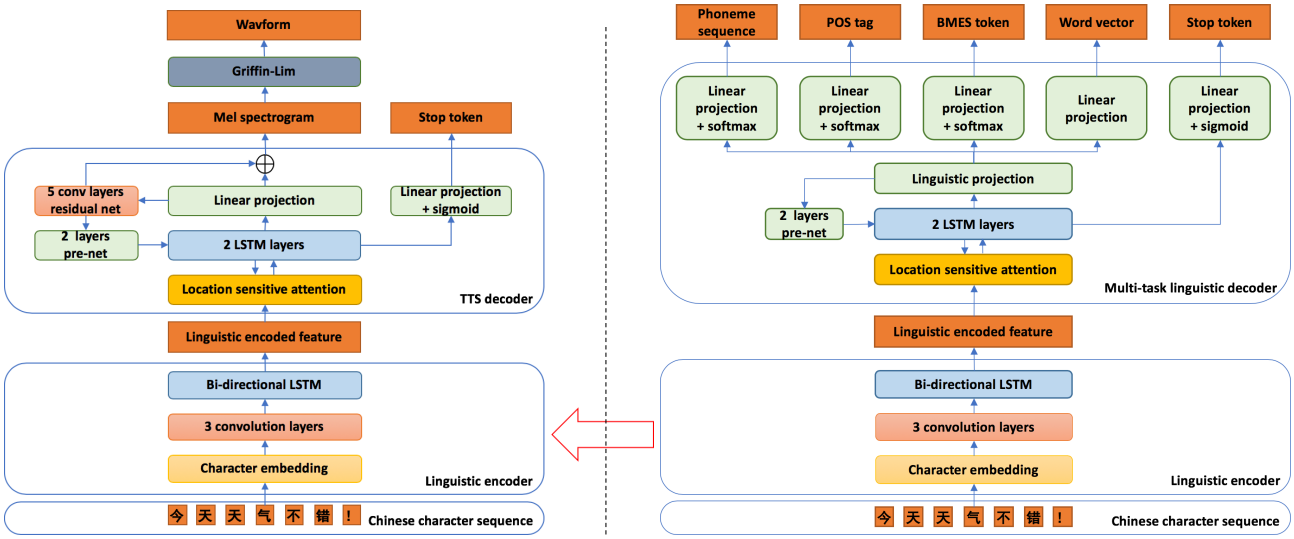
Text-to-speech (TTS) system, as an important component in human computer interaction (HCI) frameworks, aims to generate natural speech from given text [1]. Conventional TTS systems employ complicated pipelines for speech generation [2], separated with linguistic feature generation component, duration prediction component, acoustic feature prediction component and a vocoder for speech synthesis. These components are well-designed and investigated to complete the respective subtasks in the speech generation pipelines, but developed and trained separately, resulting in errors compounding in steps [3, 4].

With the development of end-to-end architectures in deep learning field, end-to-end TTS techniques are also proposed to integrate the components in conventional TTS systems [4, 5]. By merging different components into one trainable framework, end-to-end based systems can benefit from the joint training and alleviate the problems caused by errors compounding, providing more robust prediction performance while requiring less laborious feature engineering. These end-to-end TTS techniques, such as Tacotron [4] and Deep Voice 3 [6], are mainly developed on encoder-decoder structure [7], using encoder to produce linguistic encodings and decoder to directly generate raw spectrogram. Specially, attention mechanism [8] is applied with decoder, such as location-sensitive attention in [9], learning alignment between the given text and output spectrogram. Trainable vocoder is also proposed to use in end-to-end TTS framework [5], further improves the overall performance by replaying the complex signal-processing-based vocoder with learning-based generation model. To avoid the information bias after multiple recurrent processing in the recurrent neural networks (RNNs) and its memory enhanced variants LSTM [10] and GRU [7] in training the encoder-decoder structure based TTS models, Transformer [11] based framework is further proposed to use [12] to make better use of wider contextual information, which achieves better generation performance.

End-to-end structure based TTS models have been successfully implemented in various languages [4, 5, 13, 14, 15], providing superior performance in generating natural speech with high quality. However, researchers still employ linguistic feature generation pipelines rather than the genuine end-to-end structure in constructing Mandarin TTS systems[12, 16, 17, 18]. Related to the complex linguistic structure of Chinese, using Chinese characters directly can lead to poor linguistic encodings, resulting in improper word tokenization and pronunciation errors. To ensure the naturalness and intelligibility of synthetic speech, state-of-the-art Mandarin TTS systems employ a list of components, such as word tokenization, part-of-speech (POS) tagging and grapheme-to-phoneme (G2P) conversion, to produce knowledge-enhanced inputs to alleviate the problems caused by linguistic encoding. These components are based on linguistic expertise and well-designed, but trained individually, leading to errors compounding for the TTS system.

In this paper, to address the challenges in developing character-based Mandarin end-to-end TTS system, we proposed a knowledge-based linguistic encoder. Compared with conventional encoders, the proposed linguistic encoder is trained in a multi-task learning framework, using a list of linguistic analysis subtasks to enhance the performance of linguistic encodings generation. These linguistic analysis subtasks, including word tokenization, POS tagging and G2P conversion, wording embedding and stop token prediction, are well-designed and closely related to the Chinese linguistic syntax. Trained with the multi-task framework, the linguistic encoder can produce robust and discriminative linguistic encodings to the following TTS decoder, dramatically reduce the mispronunciations in synthetic

http://dx.doi.org/10.21437/Interspeech.2019-1118



(a) Architecture of the Chinese character-based end-to-end TTS system

(c)*

(b) Architecture for training the knowledge-based linguistic encoder

Figure 1: *The architecture of the proposed Chinese character-based end-to-end TTS system and the architecture for training the knowledge-based linguistic encoder.* * *After trained, the knowledge-based linguistic encoder is transferred to the Chinese character-based end-to-end TTS system.*

speech and further improve the overall performance of Mandarin TTS systems. Experimental results demonstrate the effectiveness of the proposed framework, with word tokenization error dropped from 12.81% to 1.58%, syllable pronunciation error dropped from 10.89% to 2.81% compared with state-of-the-art baseline approach, providing mean opinion score (MOS) improvement from 3.76 to 3.87.

2. Methodology

Our work consists of two components, shown in Figure 1: (a) a Chinese character-based end-to-end Mandarin TTS system which directly converts Chinese character sequence to mel spectrograms with the knowledge-based linguistic encoder, and then synthesizes the speech by the Griffin-Lim algorithm; (b) a multi-task learning framework for training the knowledge-based linguistic encoder to learn the rich-knowledge linguistic embeddings from raw Chinese characters with the help of different natural language processing (NLP) tasks.

2.1. Chinese character-based end-to-end Mandarin TTS

The backbone of the proposed Chinese character-based end-to-end Mandarin TTS system is an encoder-decoder model with attention [8], consisting of a knowledge-based linguistic encoder and a TTS decoder. The knowledge-based linguistic encoder takes Chinese character sequence as input and produces knowledge-rich linguistic encodings, then the TTS decoder takes the linguistic encodings as input and produces mel spectrograms that are then converted to waveforms by the Griffin-Lim algorithm.

The knowledge-based linguistic encoder is a neural network which has the same architecture as the Tacotron 2 encoder [5]. Input Chinese characters are represented by the learned 512-dimensional character embeddings, and then are passed through a stack of 3 convolutional layers each containing 512 filters with shape 5×1 and ReLU activations. The output of the final convolutional layer is passed into a single bi-directional LSTM layer containing 256 units in each direction to generate the linguistic encodings.

We employed the Tacotron 2 decoder [5] as the TTS decoder in the proposed system to converted the linguistic encodings to mel spectrograms along with the stop token. Then the Griffin-Lim algorithm is used to synthesize speech from the mel spectrograms.

Although both the encoder and decoder used in the proposed system have the same architecture as those used in Tacotron 2, the knowledge-based linguistic encoder is trained in another network and then transferred into the proposed system, providing richer linguistic information than the encoder in Tacotron 2.

2.2. Knowledge-based linguistic encoding

To provide rich linguistic information in the encodings for Chinese character-based end-to-end Mandarin TTS, we proposed a multi-task learning framework to train the knowledge-based linguistic encoder. Tasks closely related to the Chinese linguistic analysis and popularly researched in the NLP field are used as the sub-tasks in the multi-task learning framework to alleviate the syllable pronunciation error and word tokenization error and improve the naturalness and intelligibility in the synthesized speech.

To alleviate the syllable pronunciation error in the synthesized speech, we include G2P as the first sub-task for training the knowledge-based linguistic encoder, which converts the input sequence of Chinese characters to the corresponding sequence of phonemes. And to alleviate the word tokenization error, we include word tokenization as the second sub-task, which predicts the BMES (Begin, Middle, End, Single) tokenization tags [19] for each character. To improve the performance of the first two sub-tasks and enrich the semantic knowledge in the lin-

guistic encodings, POS tagging and word embedding are also included in the multi-task learning framework. The POS tagging sub-task predicts the POS tags for input characters which are duplicated from the corresponding POS tags of words. And the word embedding sub-task predicts the word vectors for input characters which are also duplicated from the corresponding word vectors of words. Here the target word vectors are embedded from a well-trained Word2Vec model trained on the Chinese Wikimedia corpus [20]. The stop token predicting used in Tacotron 2 is also included as the last sub-task in the multi-task decoder.

The backbone of the proposed multi-task learning framework is also an encoder-decoder model with attention, consisting of a knowledge-based linguistic encoder and a multi-task decoder. The knowledge-based linguistic encoder also takes Chinese character sequences as input and produces linguistic encodings, then the multi-task decoder takes the linguistic encodings as input to produce predictions for the sub-tasks.

For the multi-task linguistic decoder, the linguistic encodings are consumed by the location-sensitive attention to summarize the full encodings as a fixed-length context vector for each decoder output step with location features computed using 32 1-dimensional convolution filters of length 31. The prediction from the previous time step is passed through a stack of 2 fully connected layers of 256 hidden ReLU units which is also known and employed as the pre-net in Tacotron 2. The output of the pre-net and attention context vector are concatenated and passed through a stack of 2 uni-directional LSTM layers with 1024 units and zoneout probability 0.1. Then the LSTM output and the attention context vector are concatenated and projected through a linear transform to a 512-dimensional linguistic vector. For the tasks of G2P, word tokenization and POS tagging, the linguistic vector is projected through 3 different linear transforms followed by softmax activations, to predict the probabilities for phoneme tags, BMES tags and POS tags respectively. For the task of word vector prediction, the linguistic vector is projected through a linear transforms followed by linear activation to predict the word vectors. While for the stop token prediction task, the stop token is predicted by the same architecture used in Tacotron 2 [5], by projecting down the concatenation of the LSTM output and attention context vector to a scalar and passing through a sigmoid activation.

In common with the conventional multi-task learning framework, the linguistic encoder can be trained by minimizing the global loss expressed as a weighted sum of the losses from all the sub-tasks. In practice, the weights of sub-tasks are equal. Cross entropy is used as the loss function for the G2P, word tokenization, POS tagging and stop token predicting subtasks, and mean squared error (MSE) is used as the loss function for word embedding sub-task.

The convolutional layers and the pre-net in the network are applied with dropout probability 0.5. The LSTM layers are applied with zoneout probability 0.1. We also apply L_2 regularization with weight 10^{-6}.

2.3. Transfer the knowledge-based linguistic encoder to the Chinese character-based end-to-end Mandarin TTS system

After the knowledge-based linguistic encoder is trained, it is frozen and transferred to the proposed Chinese character-based end-to-end Mandarin TTS system. With the rich linguistic information from the knowledge-based linguistic encoding, the proposed end-to-end TTS system is able to directly synthesize speeches from Chinese characters with lower syllable pronunciation error and word tokenization error.

3. Experiments

3.1. Training setup

Our training process involves first training the knowledge-based linguistic encoder on large-scale text data, followed by training the proposed end-to-end Mandarin TTS system independently with the transferred pre-trained knowledge-based linguistic encoder.

We use the multi-level annotated China Daily dataset [21] to train the linguistic encoder. The ground-truths for word tokenization and POS tagging have been manually annotated in the dataset. The ground-truth for G2P is generated by pypinyin [22] and split_pinyin_sp [23]. The ground-truth for word embedding is generated by a Word2Vec model trained by gensim [24] on the Chinese Wikimedia corpus [20]. After pre-processing, the corpus contains $19,358$ sentences. $18,910$ sentences are used for training and 448 sentences are used for validation. The learning rate is 10^{-4} exponentially decaying to 10^{-5} after $1,000$ iterations.

We then train the proposed end-to-end Mandarin TTS system on a public Chinese female corpus [25] which contains 10-hour professional speech. After pre-processing, the corpus contains $9,998$ utterances, from which $9,550$ utterances are used for training and 448 utterances are used for validation. 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.

We train a vanilla Tacotron 2 [5] also using Chinese character sequences as input and the Griffin-Lim algorithm as the vocoder as the baseline approach.

All the neural networks used in the experiments are programmed based on an open source TensorFlow implemention of Tacotron 2 [26]. We train all the models for $100,000$ iterations on 2 NVIDIA TITAN X (Pascal) with a batch size of 64 on a single GPU.

3.2. Evaluations

Mel-cepstral distortion (MCD) is calculated on the validation set of the Chinese female corpus. MCD is defined as the Euclidean distance between the predicted mel spectrogram and the that of target speech, which can be formalized as:

$$MCD[dB] = \frac{10}{ln10}\sqrt{2\sum_{i=1}^{N}\left(c_i - c_i^{synthesized}\right)^2} \quad (1)$$

where c_i and $c_i^{synthesized}$ refer to the i-th coefficient of the ground-truth and the predicted mel spectrogram frame and N is the dimension of mel spectrogram.

We randomly select 92 sentences from a high school lecture as the test set. Speech files generated on this set are rated by 5 listeners on a scale from 1 to 5 with 0.2 point increments, from which a subjective mean opinion score (MOS) is calculated. Syllable pronunciation error and word tokenization error are also flagged by listeners and calculated from the speech files generated on this set.

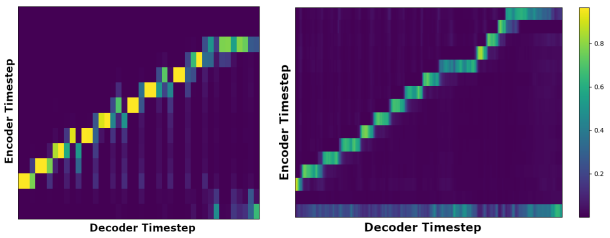
Table 1 shows a comparison of the proposed system against the baseline. In the objective evaluations, the proposed end-to-end Mandarin TTS system outperforms the baseline approach, greatly decreasing the syllable pronunciation error and word tokenization error in the synthesized speech, with MCD dropped from 3.85 to 3.53 on validation set, word tokenization error dropped from 12.81% to 1.58%, and syllable pronunciation error dropped from 10.89% to 2.81% on the test set. In the subjective evaluation, the proposed system receives a MOS of 3.87 while the baseline approach receives a MOS of 3.76.

Table 1: *Performance comparison between the baseline and proposed approaches.*

	Baseline	**Proposed**
MCD (on validation set)	3.85	**3.53**
word tokenization error	12.81%	**1.58%**
syllable pronunciation error	10.89%	**2.81%**
MOS	3.76 ±0.05	**3.87 ±0.05**

3.3. Analysis and discussion

The attention alignment learned when training the knowledge-based linguistic encoder is shown in Figure 2 (a). Comparing with the attention alignments in conventional end-to-end TTS systems, more non-zero weights are shown in the lower triangular of the attention matrix, revealing the context information used to generate the linguistic encoding. Also the attention alignment learned in the proposed TTS system is shown in Figure 2 (b). We can reveal that the decoder has successfully aligned to the linguistic encoding, providing the interpretability of our proposed model.



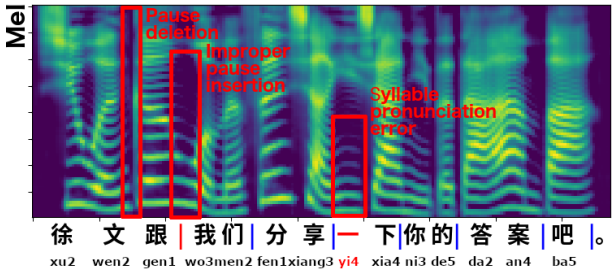
(a) (b)

Figure 2: *Attention alignments (a) learned when training the knowledge-based linguistic encoder (b) learned in the proposed TTS system.*

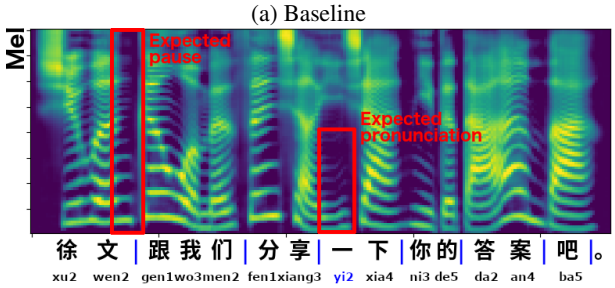
Also a mel spectrogram comparison on a test case is shown in Figure 3. The baseline approach has word tokenization error around the 3-rd character, resulting in pause deletion between the 2-nd and the 3-rd characters and improper pause insertion between the 3-rd and 4-th characters. The baseline approach also has syllable pronunciation error at the 8-th character, incorrectly predicting the pronunciation of the character from $yi2$ to $yi4$. Meanwhile the proposed system has synthesized the speech with expected pronunciations and pauses, demonstrating the effectiveness of the proposed framework.

4. Conclusion and future research

To reduce the complexity of Mandarin TTS system and improve performance by including linguistic information, we proposed



(a) Baseline



(b) Proposed

Figure 3: *Mel spectrogram comparison on the test case "Xu Wen please share your answer with us". The baseline approach has word tokenization error around the name "Xu Wen" and has syllable pronunciation error at the 8-th character.*

a knowledge-based linguistic encoder for the character-based end-to-end Mandarin TTS system. With the same architecture as the encoder used in the state-of-the-art end-to-end TTS systems, the proposed encoder is multi-task learned from linguistic analysis sub-tasks including G2P, word tokenization, POS tagging and word embedding, providing robust and discriminative linguistic encodings for the following speech generation decoder. Comparing with the state-of-the-art end-to-end Mandarin TTS approach, the proposed system can greatly alleviate the syllable pronunciation error and word tokenization error and improve the naturalness and intelligibility in the synthesized speech. Experimental results demonstrate the effectiveness of the proposed framework in both objective and subjective evaluations.

There is still much to be investigated in our framework. Involving more kinds of linguistic knowledges may further improve the performance. And the architectures for the proposed system and the knowledge-based linguistic encoder contain many early design decisions and are ripe for improvement. For example, Transformer is also of possible to use in the proposed Mandarin TTS system and the knowledge-based linguistic encoder. We are currently working on a fully attention based end-to-end Mandarin TTS system with more kinds of linguistic knowledges involved.

5. Acknowledgements

This work is supported by joint research fund of National Natural Science Foundation of China - Research Grant Council of Hong Kong (NSFC-RGC) (61531166002, N_CUHK404/15), National Natural Science Foundation of China (61433018, 61375027) and National Social Science Foundation of China (13&ZD189). The authors also thank Mr. Yan Huang from TAL Education Group for his helpful discussions and advices.

6. References

[1] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[4] 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

[5] 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

[6] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[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] 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

[9] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 577–585. [Online]. Available: http://papers.nips.cc/paper/5847-attention-based-models-for-speech-recognition.pdf

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, . 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

[12] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Neural Speech Synthesis with Transformer Network," *arXiv:1809.08895 [cs]*, Sep. 2018, arXiv: 1809.08895. [Online]. Available: http://arxiv.org/abs/1809.08895

[13] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," *arXiv:1810.11960 [cs, eess, stat]*, Oct. 2018, arXiv: 1810.11960. [Online]. Available: http://arxiv.org/abs/1810.11960

[14] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," *arXiv:1806.04558 [cs, eess]*, Jun. 2018, arXiv: 1806.04558. [Online]. Available: http://arxiv.org/abs/1806.04558

[15] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning," *arXiv:1710.07654 [cs, eess]*, Oct. 2017, arXiv: 1710.07654. [Online]. Available: http://arxiv.org/abs/1710.07654

[16] R. Li, Z. Wu, X. Liu, H. Meng, and L. Cai, "Multi-task learning of structured output layer bidirectional LSTMs for speech synthesis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5510–5514.

[17] S. . Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep Voice: Real-time Neural Text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, pp. 195–204, event-place: Sydney, NSW, Australia. [Online]. Available: http://dl.acm.org/citation.cfm?id=3305381.3305402

[18] 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

[19] M. Zhang, N. Yu, and G. Fu, "A Simple and Effective Neural Model for Joint Word Segmentation and POS Tagging," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 9, pp. 1528–1538, Sep. 2018. [Online]. Available: https://doi.org/10.1109/TASLP.2018.2830117

[20] "Wikimedia." [Online]. Available: https://www.wikimedia.org/

[21] Yu, Shiwen (Peking University), Duan, Huiming (Peking University), and Wu, Yunfang (Peking University), "Corpus of Multi-level Processing for Modern Chinese," 2018. [Online]. Available: http://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/SEYRX5

[22] H. Huang, I. MSK, X. Kong, bors-homu, W. Chen, Y. Yang, D. Wang, T. G. Badger, and M. Wang, *mozillazg/python-pinyin: v0.35.1*, 2019. [Online]. Available: https://doi.org/10.5281/zenodo.2581871

[23] J. Li, "Split pinyin for speech processing." Jan. 2019, original-date: 2018-09-13T14:02:05Z. [Online]. Available: https://github.com/petronny/split_pinyin_sp

[24] R. ehek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, 2010, pp. 45–50.

[25] Databaker technology Inc. (Beijing), "Open source Chinese female voice database," 2019. [Online]. Available: https://www.data-baker.com/open_source

[26] R. Mama, "DeepMind's Tacotron-2 Tensorflow implementation." Mar. 2019, original-date: 2017-12-20T16:08:13Z. [Online]. Available: https://github.com/Rayhane-mamah/Tacotron-2