SYNTACTIC REPRESENTATION LEARNING FOR NEURAL NETWORK BASED TTS WITH SYNTACTIC PARSE TREE TRAVERSAL

Changhe Song[1], *Jingbei Li*[1], *Yixuan Zhou*[1], *Zhiyong Wu*[1,2,*] , *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
{sch19, lijb19, zhouyx20}@mails.tsinghua.edu.cn, {zywu, hmmeng}@se.cuhk.edu.hk

ABSTRACT

Syntactic structure of a sentence text is correlated with the prosodic structure of the speech that is crucial for improving the prosody and naturalness of a text-to-speech (TTS) system. Nowadays TTS systems usually try to incorporate syntactic structure information with manually designed features based on expert knowledge. In this paper, we propose a syntactic representation learning method based on syntactic parse tree traversal to automatically utilize the syntactic structure information. Two constituent label sequences are linearized through left-first and right-first traversals from constituent parse tree. Syntactic representations are then extracted at word level from each constituent label sequence by a corresponding uni-directional gated recurrent unit (GRU) network. Meanwhile, nuclear-norm maximization loss is introduced to enhance the discriminability and diversity of the embeddings of constituent labels. Upsampled syntactic representations and phoneme embeddings are concatenated to serve as the encoder input of Tacotron2. Experimental results demonstrate the effectiveness of our proposed approach, with mean opinion score (MOS) increasing from 3.70 to 3.82 and ABX preference exceeding by 17% compared with the baseline. In addition, for sentences with multiple syntactic parse trees, prosodic differences can be clearly perceived from the synthesized speeches.

Index Terms— Syntactic representation learning, Neural network based text-to-speech, Syntactic parse tree traversal, Prosody control

1. INTRODUCTION

Recently neural network based text-to-speech (TTS) systems have achieved certain success in prosody and naturalness of synthesized speech over conventional methods [1, 2, 3, 4]. By applying encoder-decoder framework with attention [5], these systems can directly predict speech parameters from graphemes or phonemes by learning acoustic and prosodic patterns via a flexible mapping from linguistic to acoustic space. However, the learnt prosodic patterns only contain part of prosodic structural information [4], resulting in poor prosody and naturalness performance, even improper prosody.

To further improve prosody and naturalness of synthesized speech, adding prosodic structure annotations such as tones and break indices (ToBI) labels [6] or other prosodic structure labels [7] to the input sequence of neural network based TTS models has been proposed. Prosodic structure annotations need to be subjectively labeled from speech, which is time-consuming. Although these annotations can be automatically annotated by training another prosodic structure prediction model [8], the accuracy of predicted prosodic structure labels is still limited by using subjectively labeled annotations as the ground-truths.

The high correlation between syntactic structure and prosodic information has been proved by successful syntactic-to-prosodic mapping [9, 10]. A set of rule-based syntactic features such as part-of-speech (POS) and positions of the current word in parent phrases are proposed and used in hidden Markov model (HMM) based acoustic model [11]. To utilize more syntactic structure information, phrase structure based feature (PSF) and word relation based feature (WRF) are proposed in neural network based TTS [12]. PSF and WRF expand the set of syntactic features used in HMM model. More features such as highest-level phrase beginning with current word (HBCW) and lowest common ancestor (LCA) are further introduced to model syntactic structure [12].

However, the expanded features are still manually designed features rather than automatically learned high-level representations. PSF only contains features from limited layers of the whole syntactic tree structure. WRF only exposes the information of partial nodes and edges from the whole syntactic parse tree.

To make better use of the syntactic information, motivated by the syntactic parse tree traversal approach in neural machine translation [13], we propose a syntactic representation learning method to further improve the prosody and naturalness of synthesized speech in neural network based TTS. Syntactic parse tree is linearized into two constituent label sequences through left-first and right-first traversal. Then syntactic representations are extracted from the constituent label sequences using different uni-directional GRU network for each sequence. After which, the syntactic representations are up-sampled from word level to phoneme level and concatenated with phoneme embeddings. Tacotron 2 is employed to generate spectrogram from the concatenated syntactic representations and phoneme embeddings, with Griffin-Lim [14] to reconstruct the waveform. Nuclear-norm maximization loss (NML) is introduced to the constituent label embedding layer to enhance discriminability and diversity. Compared to only hiring left-first traversal [13], right-first traversal is proposed to alleviate the ambiguity.

Experimental results show that our proposed model outperforms the baseline in terms of prosody and naturalness. Mean opinion score (MOS) increases from 3.70 to 3.82 compared with the baseline approach (t-test, p=0.0079). ABX preference rate exceeds the baseline approach by 17%. For sentences with multiple different

* Corresponding author.

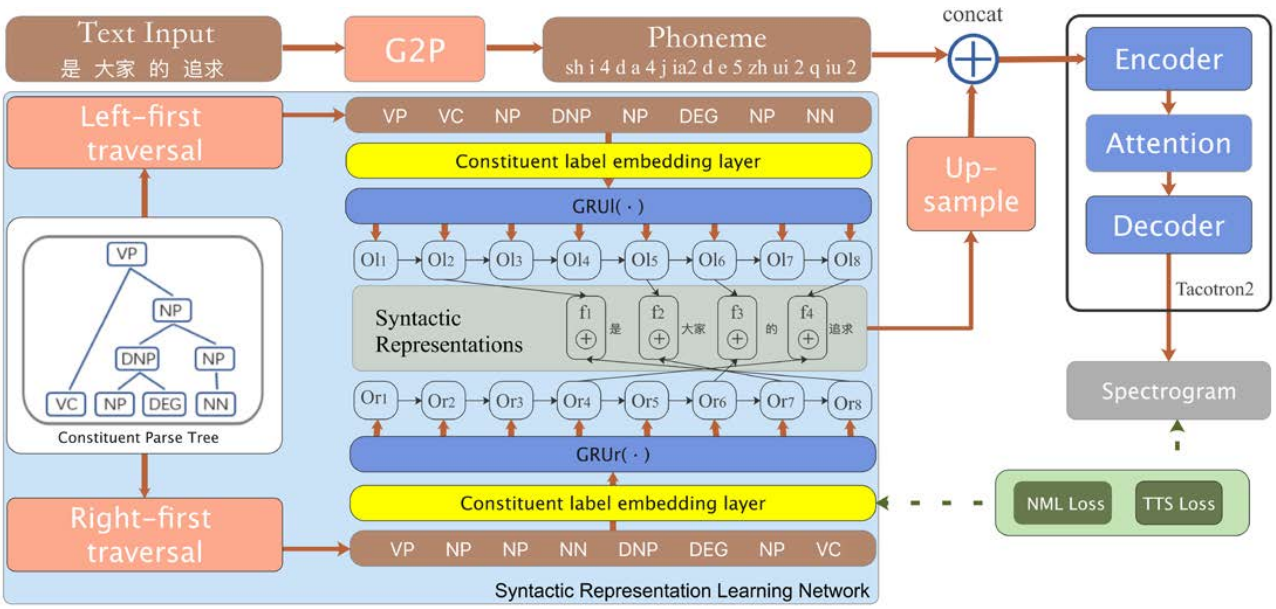


Fig. 1. The structure of proposed syntactic representation learning network.

syntactic parse trees, prosodic differences can be clearly perceived from corresponding synthesized speeches.

2. METHODOLOGY

Fig.1 shows the framework of our proposed method. Our work mainly focuses on introducing a trainable syntactic structure information extractor as part of neural network based TTS system to improve the prosody and naturalness of the synthesized speech.

2.1. Syntactic representation learning

To provide high-level syntactic representations with rich syntactic information to neural network based TTS system, we propose a syntactic representation learning network based on syntactic parse tree traversal. Constituent parse trees are extracted including labels and tree structure of constituents.

To represent the tree structure for neural network based TTS, depth-first traversals are possible to use to linearize the syntactic parse tree to constituent sequences. Since any single tree traversal algorithm will map multiple syntactic parse trees to a same sequence, both left-first and right-first are proposed to use to alleviate the ambiguity. The sequence of constituent labels generated by the two traversals can be formulated as the following equations:

$$\begin{aligned} C_l &= [c_l^1, ..., c_l^j, ..., c_l^m] \\ C_r &= [c_r^1, ..., c_r^j, ..., c_r^m] \end{aligned}, j \in \{1, 2, ..., m\} \tag{1}$$

where C_l and C_r are the constituent label sequences generated from the left-first and right-first traversals respectively, c_l^i and c_r^i are constituent labels, m is the length of generated constituent label sequences. The constituent labels are then embedded by a shared embedding layer and modeled by two different uni-directional GRU networks, one GRU network for each sequence. The process can be represented as:

$$\begin{aligned} \hat{c}_l^j &= Embedding(c_l^j) \\ \hat{c}_r^j &= Embedding(c_r^j) \\ \hat{C}_l &= [\hat{c}_l^1, ..., \hat{c}_l^j, ..., \hat{c}_l^m] \\ \hat{C}_r &= [\hat{c}_r^1, ..., \hat{c}_r^j, ..., \hat{c}_r^m] \end{aligned} \tag{2}$$

where $\hat{C}_l$ and $\hat{C}_r$ are the embedding sequences of the constituent labels for left-first and right-first traversals respectively. Then the outputs of the GRU network are:

$$\begin{aligned} O_l &= GRU_l(\hat{C}_l) \\ O_r &= GRU_r(\hat{C}_r) \end{aligned} \tag{3}$$

where $GRU_l(\cdot)$ and $GRU_r(\cdot)$ are two different uni-directional GRUs, O_l and O_r are the outputs of $GRU_l(\cdot)$ and $GRU_r(\cdot)$ respectively.

Syntactic features are the concatenations of the outputs of GRUs for each word. Given the i-th word of an input text with w words, the positions of its corresponding constituent labels in the generated sequences C_l and C_r are recorded as p_l^i and p_r^i respectively, then the learnt syntactic representation f_i for the i-th word is formulated as:

$$f_i = [O_l^{p_l^i}, O_r^{p_r^i}], i \in \{1, 2, ..., w\} \tag{4}$$

2.2. Nuclear-norm Maximization Loss

To improve the discriminability and diversity of the embeddings of the syntactic labels, global nuclear-norm maximization loss (NML) [15] is proposed to increase the rank of the embeddings of all possible constituent labels. The NML is defined as:

$$\begin{aligned} \hat{C} &= Embedding(C) \\ \mathcal{L}_{NML} &= -\frac{1}{N}\|\hat{C}\|_*. \end{aligned} \tag{5}$$

where C is the set of all possible constituent labels, $\hat{C}$ and N are the embedding and length of C respectively. $\|\hat{C}\|_*$ is computed as:

$$\|\hat{C}\|_* = \sum_i \sigma_i \tag{6}$$

where σ_i is the i-th singular value of $\hat{C}$.

2.3. TTS with syntactic representations

The learnt syntactic representations are word related, which are upsampled to phoneme level and concatenated with phoneme embeddings. Syntactic representation is copied to match the phoneme sequence length of current word. Tacotron 2 [4] is employed to generate spectrogram from the concatenated syntactic representations and phoneme embeddings, and Griffin-Lim [14] is further utilized to reconstruct the waveform. The whole model is trained with a loss function which can be formulated as:

$$\mathcal{L} = \mathcal{L}_{TTS} + \lambda \mathcal{L}_{NML} \tag{7}$$

where $\mathcal{L}_{TTS}$ is the loss function defined in Tacotron 2 and λ is the loss weight for NML.

3. EXPERIMENT

3.1. Training setup

We train models on public Chinese female corpus [16], which includes 10-hour professional speech and 10000 sentences. 500 sentences are used for validation and other sentences are used for training. We down-sample the speech to 16k Hz sampling frequency, The tacotron 2 part in our model is trained with vanilla setups except setting frequency to match our speech. The learning rate is fixed to 10^{-3} and the loss weight of NML is 0.05.

We train the WRF based TTS [12] as baseline approach. The parser used in WRF [17] is replaced by state-of-the-art syntactic parsing model Benepar [18].

We program all the models based on an open sourced Tensorflow implemention of Tacotron 2 [19]. We train all the models for 50k iterations with a batch size of 16 on a NVIDIA 2080 Ti GPU.

3.2. Subjective evaluation

We randomly select 30 sentences from Internet as test set, 5 of which are sentences with multiple different syntactic parse trees. Synthesized speeches are shifted in random order and rated by 20 native speakers on a scale from 1 to 5, from which a subjective mean opinion score (MOS) is calculated.

As show in Table.1, the proposed system receives a MOS of 3.82 while the baseline approach receives a MOS of 3.70, with a comparable variance. T-test reveals that our proposed approach significantly outperforms the baseline with a p of 0.0079.

We also conduct a ABX preference test between pairs of systems on the synthesized speech. The listeners are presented with the speeches synthesized by the baseline and proposed approaches in random order, and decide which one has the better prosody and naturalness. As show in Table.2, the proposed approach receives 45.8% preference rate exceeding the baseline approach by 17%.

Table 1. Comparision between baseline and the proposed method. MOS variances are given in brackets.

	WRF	Proposed
MOS	$3.70(0.05)$	$3.82(0.06)$

Table 2. ABX preference comparision between baseline and the proposed method. ABX-PR means preference rate of ABX test.

	WRF	Proposed	Neutral
ABX-PR	28.8%	45.8%	25.4%

3.3. Ablation study

To visualize the contribution of NML, we train our model without NML with the same settings. Another ABX preference test is conducted on same test set and listeners. The listeners are presented with the speeches synthesized by our models with and without NML, and decide which one has the better prosody and naturalness. As show in Table.3, the approach with NML receives 58.3% preference rate exceeding the approach without NML by 27%.

Table 3. ABX preference result with or without NML.

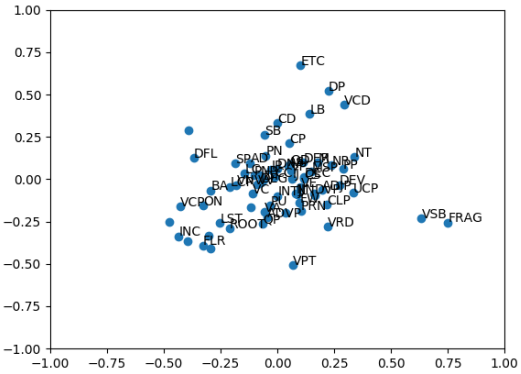
	Proposed w/o NML	Proposed w/ NML	Neutral
ABX-PR	31.3%	58.3%	10.4%

We visualize the learnt embeddings of constituent labels with and without NML by principal components analysis (PCA). As show in Fig.2, embeddings with NML are more scattered than the embeddings without NML, which demostrates the effectiveness of NML in improving discriminability and diversity.

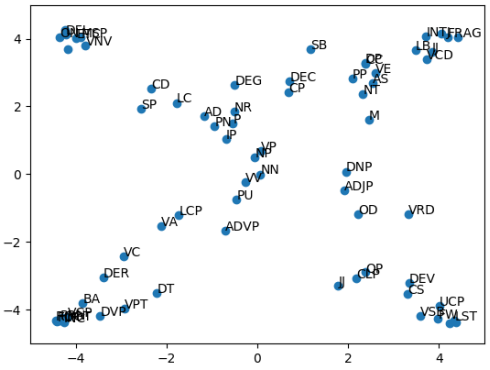
3.4. Analysis and discussion

We further conduct case studies by comparing the spectrogram and pitch contours of the speeches generated from different methods, as shown in Fig.3. The speech from baseline approach has an unexpected long pause between the 5-th character and 6-th character in such long sentence. Besides, the word from the 16-th to 17-th character, "ji4 yi4" that are both of fourth tone in Chinese, is synthesized with an unnatural up-and-down tone in WRF. Instead, the spectrogram and pitch contour of the speech synthesized by our proposed method are more stable, indicating the stability of the proposed syntactic representations from bidirectional traversals. For the last three characters, there is an unexpected stress (high pitch value) on 23-th character "xia4" in WRF result, while proposed method shows a gradually decreasing pitch contour at the end of the sentence leading to higher naturalness. This is caused by the WRF features in baseline approach which consider a uni diridectional information in the syntactic parse tree.

With the proposed syntactic representation learning method, it is possible that a sentence with the same text but different syntactic parse trees might lead to synthesized speeches with different prosody expressions, which provides a possibility for prosody control of speech synthesis. To validate this, we conduct further experi-



(a) Without NML



(b) With NML

Fig. 2. PCA results of constituent label embeddings on NML ablation study.

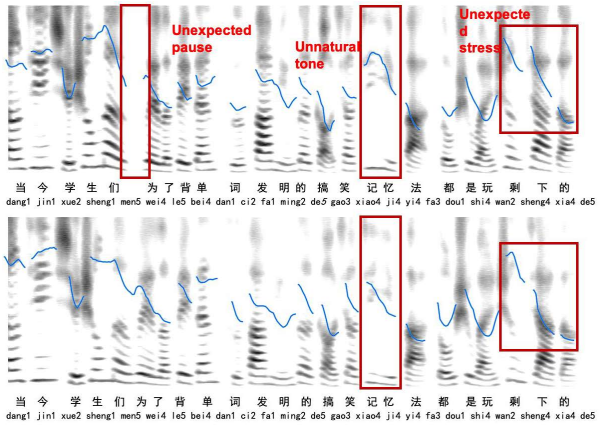
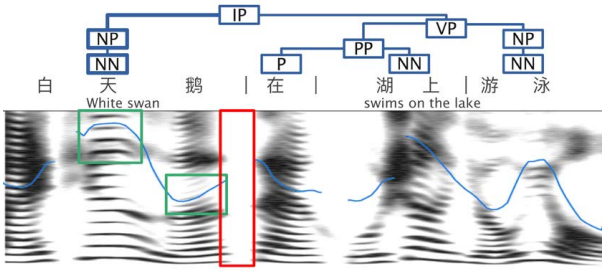
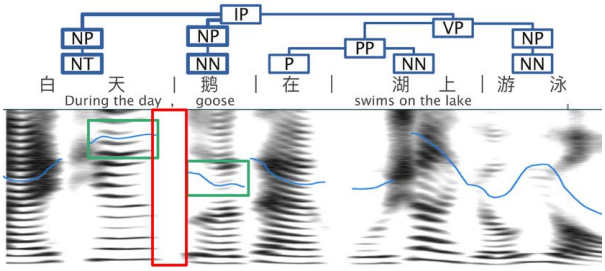


Fig. 3. Mel spectrogram and pitch contour comparison on sentence "At present, the funny memorization methods invented by students to memorize words are leftovers" between proposed model(below) and WRF(upper).

ments by inputting sentences with multiple different syntactic parse



(a) White swan swims on the lake.



(b) During the day, goose swims on the lake.

Fig. 4. Prosodic differences of the sentence with multiple syntactic parse trees.

trees to the proposed model. One example is shown in Fig.4, from which the prosodic differences of the synthesized speeches can be clearly observed. Upper structure regards the first three characters as a word and the sentence means "White swan swims on the lake", while below treats these characters as two words and the sentence means "During the day, goose swims on the lake". Although the graphemes and phonemes are same, the meanings of the sentences are different with each tree. And the prosodic differences match the meanings respectively. Also the prosody of last five characters in either synthesized speech are similar since their corresponding syntactic structure information is similar.

4. CONCLUSION

In this study, we investigate a syntactic representation learning method to automatically utilize the syntactic structure information for neural network based TTS. Nuclear-norm maximization loss is introduced to enhance the discriminability and diversity of synthsized speech prosody. Experimental results demonstrate the effectiveness of our proposed approach. For sentences with multiple syntactic parse trees, prosodic difference can be clearly observed from the synthesized speeches.

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] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio, "Char2wav: End-to-end speech synthesis," in *2017 International Conference on Learning Representations (ICLR)*, 2017.

[2] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[4] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj 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.

[5] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[6] Kim Silverman, Mary Beckman, John Pitrelli, Mori Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg, "Tobi: A standard for labeling english prosody," in *Second international conference on spoken language processing*, 1992.

[7] Massimo Poesio, Florence Bruneseaux, and Laurent Romary, "The mate meta-scheme for coreference in dialogues in multiple languages," in *Towards Standards and Tools for Discourse Tagging*, 1999.

[8] Andrew Rosenberg, "Autobi-a tool for automatic tobi annotation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[9] Xiaotian Zhang, Yao Qian, Hai Zhao, and Frank K Soong, "Break index labeling of mandarin text via syntactic-to-prosodic tree mapping," in *2012 8th International Symposium on Chinese Spoken Language Processing*. IEEE, 2012, pp. 256–260.

[10] Hao Che, Jianhua Tao, and Ya Li, "Improving mandarin prosodic boundary prediction with rich syntactic features," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[11] Rasmus Dall, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, "Redefining the linguistic context feature set for hmm and dnn tts through position and parsing.," in *INTERSPEECH*, 2016, pp. 2851–2855.

[12] Haohan Guo, Frank K Soong, Lei He, and Lei Xie, "Exploiting syntactic features in a parsed tree to improve end-to-end tts," *Proc. Interspeech 2019*, pp. 4460–4464, 2019.

[13] Junhui Li, Deyi Xiong, Zhaopeng Tu, Muhua Zhu, Min Zhang, and Guodong Zhou, "Modeling source syntax for neural machine translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 688–697.

[14] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[15] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian, "Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3941–3950.

[16] Databaker technology Inc. (Beijing), "Open source Chinese female voice database," 2019.

[17] Roger Levy and Christopher D Manning, "Is it harder to parse chinese, or the chinese treebank?," in *proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003, pp. 439–446.

[18] Nikita Kitaev, Steven Cao, and Dan Klein, "Multilingual constituency parsing with self-attention and pre-training," *arXiv preprint arXiv:1812.11760*, 2018.

[19] Rayhane Mama, "DeepMind's Tacotron-2 Tensorflow implementation.," Mar. 2019, original-date: 2017-12-20T16:08:13Z.