



Improving Mandarin Prosodic Structure Prediction with Multi-level Contextual Information

Jie Chen^{1,†,‡}, Changhe Song^{1,†}, Deyi Tuo³, Xixin Wu⁴, Shiyin Kang^{2,*},
Zhiyong Wu^{1,4,*}, Helen Meng^{1,4}

¹ Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

² XVerse Inc., Shenzhen, China ³ Huya Inc., Guangzhou, China

⁴ The Chinese University of Hong Kong, Hong Kong SAR, China

{chenjie20, sch19}@mails.tsinghua.edu.cn, {wuxx, hmmeng}@se.cuhk.edu.hk,
tuodeyi@huya.com, zywu@sz.tsinghua.edu.cn, kangshiyin@xverse.cn

Abstract

For text-to-speech (TTS) synthesis, prosodic structure prediction (PSP) plays an important role in producing natural and intelligible speech. Although inter-utterance linguistic information can influence the speech interpretation of the target utterance, previous works on PSP mainly focus on utilizing intra-utterance linguistic information of the current utterance only. This work proposes to use inter-utterance linguistic information to improve the performance of PSP. Multi-level contextual information, which includes both inter-utterance and intra-utterance linguistic information, is extracted by a hierarchical encoder from character level, utterance level and discourse level of the input text. Then a multi-task learning (MTL) decoder predicts prosodic boundaries from multi-level contextual information. Objective evaluation results on two datasets show that our method achieves better F1 scores in predicting prosodic word (PW), prosodic phrase (PPH) and intonational phrase (IPH). It demonstrates the effectiveness of using multi-level contextual information for PSP. Subjective preference tests also indicate the naturalness of synthesized speeches are improved¹.

Index Terms: prosodic structure prediction, multi-level contextual information, hierarchical encoder

1. Introduction

In human conversation, speakers tend to insert different levels of breaks at appropriate positions according to the semantic meaning of the text and the intentions to be conveyed. Such phenomenon is called prosodic phrase grouping [1, 2]. The inserted breaks chunk the utterance text into different prosody constituents including PW, PPH and IPH, which can be organized in a hierarchical way as the prosodic structure [3]. In TTS synthesis, PSP aims to predict the aforementioned hierarchical prosodic structure from the input text, which is crucial to the naturalness and intelligibility of the synthesized speech [4].

The PSP task is generally regarded as a sequence-to-sequence based classification problem to predict whether there is a prosody break (i.e. PW, PPH or IPH boundary) after each character of the input text. Previous works have carried out relevant studies on not only feature engineering but also model structures for the PSP task. For feature engineering, early studies have investigated several human designed features and trainable word embedding [5], which are later replaced by bidirectional encoder representation from Transformers (BERT) [6, 7]

to provide richer linguistic information. Syntactic features such as phrase structure tree and dependency tree also benefit PSP [8, 9, 10]. With the development of neural network based methods, an end-to-end PSP model BLSTM-CRF [11, 12] was proposed which integrates bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) [13, 14] and conditional random field (CRF) [15, 16]. Maximum entropy model is also employed in PSP [17]. RNN is further replaced with multi-head self-attention to better model long-term dependencies [6, 18]. Moreover, separate classifiers for PW, PPH and IPH are integrated under a framework [12], which utilizes dependencies between three tasks and improves overall prediction performance.

However, current state-of-the-art PSP models still lack of ability to fully model naturalness and expressiveness to resemble human speech, especially in multi-sentential discourse [19]. Several works in literature have proved the existence of supra-sentential prosody patterns in discourse segments, e.g. the increase in numbers of breaks, pause and lengthening before sentential boundary due to the availability of neighboring semantic information [20, 21], the declination of pitch through both intra- and inter-sentential units [22], etc. Thus, introducing inter-utterance linguistic information from adjacent utterances can have the potential to benefit PSP. Some studies in TTS adopt hand designed features or text embeddings of adjacent utterances to provide inter-utterance linguistic information for predicting acoustic features directly [23, 24, 25, 26, 27]. Previous work also suggests that effective incorporation of inter-utterance linguistic information can improve automatic speech recognition [28].

In this work, we propose to use multi-level contextual information, which includes not only intra-utterance linguistic information of the current utterance but also inter-utterance linguistic information from adjacent utterances, to improve the performance of PSP. To integrate the multi-level contextual information into PSP, we propose a novel PSP model based on encoder-decoder architecture which has a hierarchical encoder and a MTL decoder. The hierarchical encoder consists of a character encoder, an utterance encoder and a discourse encoder and is responsible for extracting multi-level contextual information from the input text. The character encoder is composed of several Transformer encoder blocks [29]. The utterance encoder and the discourse encoder employ multi-layer convolutional neural network (CNN) [30]. A group of adjacent utterances in the same context are sent to the character encoder to produce character representations, which are further processed by the utterance encoder to get utterance representations. Then,

[‡] Work performed when Jie Chen was interning at Huya Inc.

[†] Equal contribution. * Corresponding author.

¹ Synthesized speech samples at: <https://thuhsi.github.io/mlc-PSP/>

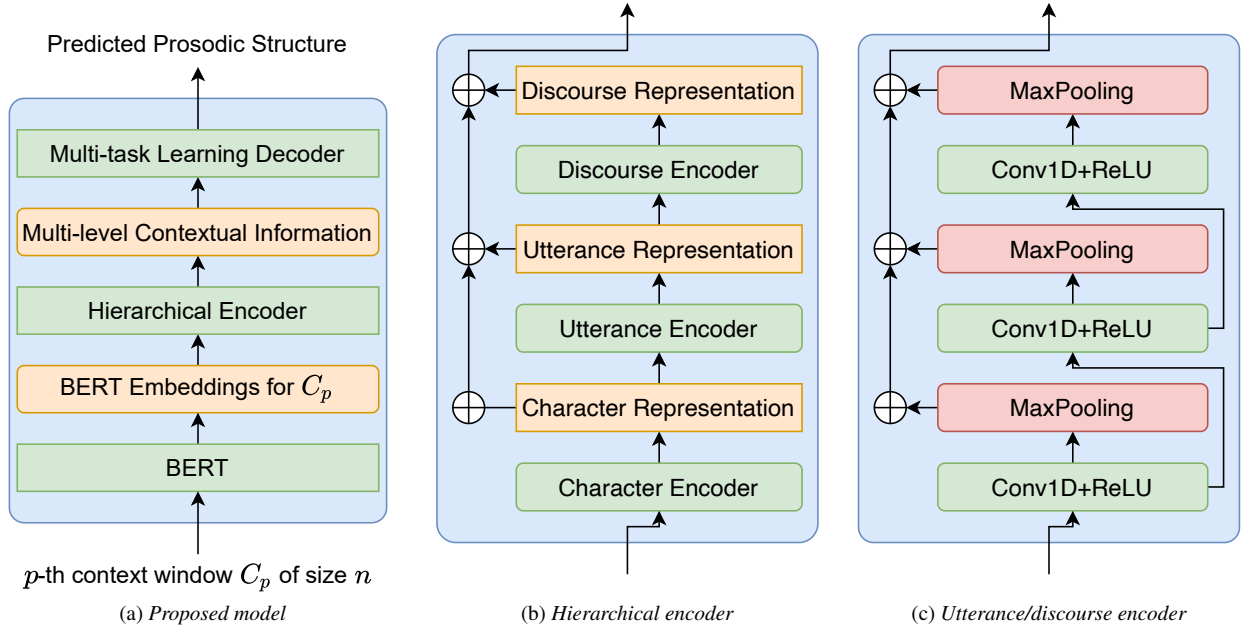


Figure 1: The overall architecture for proposed model.

discourse encoder capturing inter-utterance linguistic information from utterance representations produces a discourse representation. Finally, these three-level representations are concatenated together as the multi-level contextual information which is fed into a MTL decoder to predict PW, PPH and IPH boundaries.

Results from objective experiments show that our model achieves better prediction performance in PW, PPH and IPH subtasks, which shows the effectiveness of using multi-level contextual information for the PSP task. Furthermore, ABX preference tests also indicate our method can derive better naturalness of synthesized speech.

2. Methodology

Fig.1(a) shows the overall architecture of our proposed model including a pretrained BERT, a hierarchical encoder and a MTL decoder.

2.1. Hierarchical encoder

The hierarchical encoder belongs to PSP models with self-attention [6, 18] and consists of a character encoder, an utterance encoder and a discourse encoder. Fig.1(b) shows its architecture. Assume a Chinese document $D = \{s_1, s_2, s_3, \dots, s_N\}$ has N utterances, where s_i is the i -th utterance in the original document. A sequence of n utterances from s_p to s_{p+n-1} form the p -th context window of size n in D , denoted by $C_p = \{s_p, s_{p+1}, s_{p+2}, \dots, s_{p+n-1}\}$. Let C_{pj} be the j -th utterance in C_p , and C_{pjt} be the t -th Chinese character of C_{pj} . For C_p , all utterances will be embedded with a character-level BERT to get the feature $B_p \in \mathbb{R}^{n \times l \times d}$, where l is the length of the padded utterance, d is the dimension of the BERT feature. Then B_p is passed to the hierarchical encoder to extract multi-level contextual information.

2.1.1. Character encoder

The character encoder utilizes Transformer encoder blocks [29] to extract character representations, in which the self-attention mechanism helps learn long-range dependencies inside each utterance by connecting two arbitrary characters directly regardless of their distance [18]. Positional encodings [29] are added to B_{pj} to assist Transformer encoder blocks distinguish different positions of the input sequence. The character representation $CR_{pj} \in \mathbb{R}^{l \times d}$ of the utterance C_{pj} is produced by the character encoder. That is:

$$CR_{pj} = \text{CharacterEncoder}(B_{pj} + \text{PE}) \quad (1)$$

where PE is the positional encoding.

2.1.2. Utterance encoder

The utterance encoder is responsible for aggregating the information from one utterance into a single vector, which is shown in Fig.1(c). For this encoder, we use multi-layer CNN. The character representation CR_{pj} of utterance C_{pj} is processed by a stack of 1-D convolution layers. Assume we have a total of m convolution layers. The r -th convolution layer CONV_r has k_r convolution kernels. All convolution layers have ReLU activation function. The output Y_r of the r -th convolution layer is:

$$Y_r = \text{ReLU}(\text{CONV}_r(X_r)) \quad (2)$$

where $X_1 = CR_{pj}$, $X_{r+1} = Y_r$, and $Y_r \in \mathbb{R}^{l \times k_r}$ is the output of the r -th convolution layer. The output of each convolution layer is passed to a max-over-time pooling operation [30] to capture the most important feature. Finally, the outputs of the pooling layers are concatenated to produce the utterance representation UR_{pj} of the utterance C_{pj} :

$$UR_{pj} = [\text{MP}(Y_1), \text{MP}(Y_2), \dots, \text{MP}(Y_m)] \quad (3)$$

where $[\cdot]$ is the concatenating operation, MP is the max-over-time pooling operation, $UR_{pj} \in \mathbb{R}^{\sum_{i=1}^m k_i}$.

2.1.3. Discourse encoder

The discourse encoder, which shares the same architecture as the utterance encoder in Section 2.1.2, is responsible for producing discourse representations capturing inter-utterance linguistic information. Total n utterance representations $UR_p \in \mathbb{R}^{n \times \sum_{i=1}^m k_i}$ of C_p are fed into the discourse encoder to derive a single vector as the discourse representation $DR_p \in \mathbb{R}^q$, where q is the sum of numbers of convolution kernels in discourse encoder.

2.2. Multi-task learning decoder

For decoder, MTL framework is adopted, where the prediction of PW, PPH and IPH tags are regarded as three different but related subtasks of PSP. Each subtask has separate gated recurrent unit (GRU) network. To predict the prosodic boundary tag after character C_{pjt} , the aforementioned character, utterance and discourse representations CR_{pjt} , UR_{pj} and DR_p are concatenated as the multi-level contextual information $MLCI_{pjt}$ of C_{pjt} .

The prediction of PW tag for character C_{pjt} doesn't depend on other subtasks, so $MLCI_{pjt}$ is directly sent to PW GRU to produce its hidden state H_{pjt}^{PW} :

$$H_{pjt}^{PW} = \text{GRU}(MLCI_{pjt}) \quad (4)$$

For PPH GRU, in addition to $MLCI_{pjt}$, it also accepts H_{pjt}^{PW} as part of its input to produce its hidden state H_{pjt}^{PPH} :

$$H_{pjt}^{PPH} = \text{GRU}([MLCI_{pjt}, H_{pjt}^{PW}]) \quad (5)$$

For IPH GRU, it accepts not only $MLCI_{pjt}$ but also hidden states from PW and IPH GRUs, and generates its hidden state H_{pjt}^{IPH} as:

$$H_{pjt}^{IPH} = \text{GRU}([MLCI_{pjt}, H_{pjt}^{PW}, H_{pjt}^{PPH}]) \quad (6)$$

To predict whether there is a PW, PPH or IPH boundary break after character C_{pjt} , three feedforward neural networks (FNNs) with softmax activation function are adopted for the three subtasks respectively, each of which accepts the hidden state of corresponding GRU as input. Summation of the losses from three subtasks is used as the total loss for optimization.

By conditioning the IPH tag prediction on PPH and PW subtasks, and conditioning the PPH tag prediction on PW subtask, we provide richer information from other subtasks and model the hierarchical dependencies between PW, PPH and IPH, which can improve the overall performance.

3. Experiments

3.1. Datasets

As there is no public dataset for PSP, we prepared two datasets to validate the performance of our proposed model in different scenarios. The first dataset is transcribed from a massive open online courses (MOOC) and prosodic structures are manually labeled according to the corresponding audios. The second dataset is manually labeled from the recordings of an audiobook, and its text contains both dialogue and narration. Both datasets are divided into training and test sets with ratio 9:1. The statistics of the datasets is shown in Table 1.

3.2. System setup

Two representative models in recent years are selected for comparison. All models use pre-trained Chinese character-level

Table 1: Statistics of the datasets.

Dataset	Type	Count
MOOC	utterance	4,772
	character	160,719
	PW	35,096
	PPH	13,239
	IPH	12,362
audiobook	utterance	11,078
	character	441,819
	PW	73,333
	PPH	46,445
	IPH	43,777

BERT² embeddings as input. The parameters of BERT are frozen.

- **BLSTM-CRF**: A baseline model uses BLSTM as encoder and CRF as decoder and has similar architecture to [11]. MTL framework is adopted, where the prediction of IPH tag is conditioned on the predicted binary PPH tags and binary PW tags, and the prediction of PPH tags is conditioned on binary PW tags. The value one of the binary tag indicates there's a prosodic boundary after C_{pjt} and zero otherwise.
- **Transformer**: A baseline model has similar architecture to [18]. Different from Proposed, it doesn't have utterance encoder or discourse encoder.
- **Proposed**: Our proposed model. The character encoder has 2 Transformer encoder blocks, and each block has 4 attention heads. The dimension of input to the character encoder is 768 and the dimension of feedforward network is 2048. The utterance encoder has 3 convolution layers whose kernel sizes are 3, and numbers of kernels for each layer are 128, 64, 64 respectively. The discourse encoder has the same structure as the utterance encoder, but numbers of kernels for each layer are halved. The size of the context window (i.e. n) is 8.

3.3. Objective evaluation results and analysis

We choose F1 score as the evaluation metric. Experimental results are shown in Table 2.

On the MOOC dataset, Proposed outperforms BLSTM-CRF. Compared with BLSTM-CRF, F1 scores of Proposed on PW, PPH and IPH tasks are improved by 0.27%, 0.91% and 0.19% respectively. On the audiobook dataset, compared with BLSTM-CRF, Proposed has an absolute improvement of 0.70%, 1.14% and 1.05% in terms of PW, PPH and IPH F1 scores. When the MTL framework is removed, Proposed* still achieves superior F1 scores than BLSTM-CRF* on the two datasets. Compared with Proposed, F1 scores of PW, PPH and IPH prediction tasks of Transformer decrease 2.16%, 2.18% and 1.39% on the MOOC dataset, and decrease 1.18%, 1.02% and 1.09% on the audiobook dataset. Moreover, when the MTL framework is removed, Transformer* still gets inferior performance than Proposed*. The superior performances from both Proposed and Proposed* show that multi-level contextual information does help in all three subtasks of PSP, especially for

²<https://huggingface.co/bert-base-chinese>

Table 2: Performance of the compared models on MOOC and audiobook datasets. When the utterance encoder and the discourse encoder of Proposed are removed, Proposed becomes Transformer. * means the MTL framework is removed from the corresponding model.

Model	F1 score			Dataset
	PW	PPH	IPH	
Transformer	93.19%	72.97%	76.50%	MOOC
BLSTM-CRF	95.08%	74.24%	77.70%	
Proposed	95.35%	75.15%	77.89%	
Transformer*	93.29%	73.20%	76.73%	audiobook
BLSTM-CRF*	94.97%	74.11%	76.52%	
Proposed*	95.09%	75.22%	77.48%	
Transformer	93.93%	86.87%	87.67%	audiobook
BLSTM-CRF	94.41%	86.75%	87.71%	
Proposed	95.11%	87.89%	88.76%	
Transformer*	93.92%	86.79%	87.32%	audiobook
BLSTM-CRF*	94.41%	86.74%	87.15%	
Proposed*	95.11%	88.33%	88.92%	

PPH and IPH. These experimental results also indicate the importance of the utterance encoder and the discourse encoder in our model design.

We also found that F1 scores of PPH and IPH from all models are consistently lower than PW in our experiments, which is also a common phenomenon in previous works [12, 11, 18, 7]. We inspected precision and recall of all models on test set for PPH and IPH, and found that recall is lower than precision. According to Table 1, this can be explained by the imbalanced dataset where positive labels are significantly less than negative ones for PPH and IPH. We will explore solutions for this issue in future work.

3.4. Experiments on context window size

We also conduct experiments to determine the best context window size. Results are shown in Table 3. Generally speaking, a larger window brings better prediction performance, but with diminishing returns. Note that when the size of the context window is set to 1, Proposed becomes a PSP model which has more parameters but doesn't employ inter-utterance linguistic information, and it has the worst prediction performance. On two datasets, the prediction performance begins to saturate when context window size is larger than 8.

We should also notice that bigger context window will introduce extra computation overhead during training. Assume n is the size of the context window, N is the number of utterances in the training corpus. Each training sample in the training set is C_p containing n utterances, where $p \in [1, 2, 3, \dots, N - n + 1]$. Thus, computation for each training sample grows linearly with the increase of context window size. To better balance prediction performance between computation overhead of training, the size of the context window for Proposed is set to 8 in our experiments.

3.5. Subjective evaluation results and analysis

To further evaluate the PSP performance of different methods and their impact on TTS synthesis, we further conduct ABX preference tests on the naturalness of the synthesized speech.

Table 3: Experiments on context window size for Proposed on MOOC and audiobook datasets.

Window Size	F1 score			Dataset
	PW	PPH	IPH	
1	94.47%	73.20%	76.57%	MOOC
2	94.83%	73.99%	76.99%	
4	94.95%	74.63%	77.18%	
8	95.35%	75.15%	77.89%	
12	95.31%	75.01%	77.77%	
16	95.39%	74.94%	77.95%	
1	94.45%	87.30%	88.19%	audiobook
2	94.83%	87.63%	88.79%	
4	94.99%	87.82%	88.94%	
8	95.11%	87.89%	88.76%	
12	95.20%	87.81%	88.66%	
16	95.17%	87.85%	88.76%	

Because audios from the MOOC dataset contain too much noise to train a TTS model, we only conduct ABX preference tests on the audiobook dataset. 30 samples that are longer than 30 characters are randomly selected from the test set with different prosodic structures predicted by aforementioned different models and the corresponding speeches are generated using Huya TTS toolkit, which is based on DurIAN [31] and HiFi-GAN [32]. The synthesized speeches corresponding to different models are presented to subjects in random order. 14 subjects were asked to choose a preferred speech in terms of naturalness of the speeches for each speech pair. Results are shown in Fig 2, which demonstrate that compared with BLSTM-CRF and Transformer, our proposed method leads to higher naturalness of the synthesized speech.

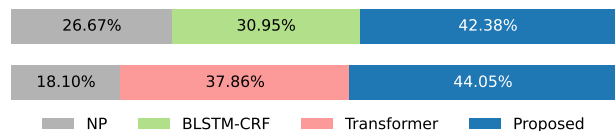


Figure 2: Results of ABX tests for naturalness of synthesized speech on audiobook dataset. NP means no preference.

4. Conclusions

In this paper, we propose a model utilizing multi-level contextual information, which includes both intra-utterance and inter-utterance linguistic information, to improve the performance of PSP. This model consists of a hierarchical encoder and a MTL decoder. Objective experimental results on two datasets show that compared with baseline models, our proposed method achieves better F1 scores on PW, PPH and IPH prediction tasks. We also performed ABX preference tests to evaluate the improvement brought by our method on synthesized speeches, where our method again outperforms baseline models.

Acknowledgement: This work was supported by National Key R&D Program of China (2020AAA0104500), National Natural Science Foundation of China (NSFC) (62076144) and National Social Science Foundation of China (NSSF) (13&ZD189).

5. References

- [1] C.-Y. Tseng, “The prosodic status of breaks in running speech: Examination and evaluation,” in *In Proceedings of Speech Prosody 2002*, 2002, pp. 667–670.
- [2] H. Peng, C. Chen, C. Tseng, and K. Chen, “Predicting prosodic words from lexical words - a first step towards predicting prosody from text,” in *2004 International Symposium on Chinese Spoken Language Processing, ISCSLP 2004*, 2004, pp. 173–176.
- [3] M. Chu and Y. Qian, “Locating boundaries for prosodic constituents in unrestricted mandarin texts,” in *International Journal of Computational Linguistics & Chinese Language Processing: Special Issue on Natural Language Processing Researches in MSRA*, vol. 6, 2001, pp. 61–82.
- [4] X. Tan, T. Qin, F. K. Soong, and T. Liu, “A survey on neural speech synthesis,” *arXiv preprint arXiv:2106.15561*, 2021.
- [5] Y. Zheng, Y. Li, Z. Wen, X. Ding, and J. Tao, “Improving prosodic boundaries prediction for mandarin speech synthesis by using enhanced embedding feature and model fusion approach,” in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*. ISCA, 2016, pp. 3201–3205.
- [6] Y. Du, Z. Wu, S. Kang, D. Su, D. Yu, and H. Meng, “Prosodic structure prediction using deep self-attention neural network,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 320–324.
- [7] Y. Zhang, L. Deng, and Y. Wang, “Unified mandarin tts front-end based on distilled bert model,” *arXiv preprint arXiv:2012.15404*, 2020.
- [8] Z. Chen, G. Hu, and W. Jiang, “Improving prosodic phrase prediction by unsupervised adaptation and syntactic features extraction,” in *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, 2010*, T. Kobayashi, K. Hirose, and S. Nakamura, Eds. ISCA, 2010, pp. 1421–1424.
- [9] H. Che, J. Tao, and Y. Li, “Improving mandarin prosodic boundary prediction with rich syntactic features,” in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association*. ISCA, 2014, pp. 46–50.
- [10] Z. Zhang, F. Wu, C. Yang, M. Dong, and F. Zhou, “Mandarin prosodic phrase prediction based on syntactic trees,” in *The 9th ISCA Speech Synthesis Workshop, SSW 2016*, 2016, pp. 160–165.
- [11] Y. Zheng, J. Tao, Z. Wen, and Y. Li, “BLSTM-CRF based end-to-end prosodic boundary prediction with context sensitive embeddings in a text-to-speech front-end,” in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*. ISCA, 2018, pp. 47–51.
- [12] H. Pan, X. Li, and Z. Huang, “A mandarin prosodic boundary prediction model based on multi-task learning,” in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 4485–4488.
- [13] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, “Automatic prosody prediction for chinese speech synthesis using BLSTM-RNN and embedding features,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015*. IEEE, 2015, pp. 98–102.
- [14] Y. Huang, Z. Wu, R. Li, H. Meng, and L. Cai, “Multi-task learning for prosodic structure generation using BLSTM RNN with structured output layer,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*. ISCA, 2017, pp. 779–783.
- [15] G. Levow, “Automatic prosodic labeling with conditional random fields and rich acoustic features,” in *Third International Joint Conference on Natural Language Processing, IJCNLP 2008*. The Association for Computer Linguistics, 2008, pp. 217–224.
- [16] Y. Qian, Z. Wu, X. Ma, and F. K. Soong, “Automatic prosody prediction and detection with conditional random field (CRF) models,” in *7th International Symposium on Chinese Spoken Language Processing, ISCSLP 2010*. IEEE, 2010, pp. 135–138.
- [17] J. Li, G. Hu, and R. Wang, “Chinese prosody phrase break prediction based on maximum entropy model,” in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing*. ISCA, 2004.
- [18] C. Lu, P. Zhang, and Y. Yan, “Self-attention based prosodic boundary prediction for chinese speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7035–7039.
- [19] Å. Peiró Lilja and M. Farrús, “Paragraph prosodic patterns to enhance text-to-speech naturalness,” in *Klessa K, Bachan J, Wagner A, Karpiński M, Śledziński D. Proceedings of the 9th International Conference on Speech Prosody; 2018 June 13-16; Poznań, Poland.[Lous Tourils]: ISCA; 2018. p. 512-6*. International Speech Communication Association (ISCA), 2018.
- [20] J. Kreiman, “Perception of sentence and paragraph boundaries in natural conversation,” *Journal of Phonetics*, vol. 10, no. 2, pp. 163–175, 1982.
- [21] C. Gussenhoven, “A perceptual study of intonation. an experimental-phonetic approach to speech melody,” 1992.
- [22] C. De Looze, I. Yanushevskaya, A. Murphy, E. O’Connor, and C. Gobl, “Pitch declination and reset as a function of utterance duration in conversational speech data,” in *INTERSPEECH*, 2015, pp. 3071–3075.
- [23] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, “Phrase break prediction for long-form reading TTS: exploiting text structure information,” in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*,. ISCA, 2017, pp. 1064–1068.
- [24] S. Le Maguer and I. Steiner, “The marytts entry for the blizzard challenge 2016,” *The Blizzard Challenge*, 2016.
- [25] K. Sawada, C. Asai, K. Hashimoto, K. Oura, and K. Tokuda, “The nitech text-to-speech system for the blizzard challenge 2016,” *The Blizzard Challenge*, 2016.
- [26] Y. Liao, Y. Chai, and C. Tsai, “The nutt’s text-to-speech system for blizzard challenge 2018,” *The Blizzard Challenge*, 2018.
- [27] G. Xu, W. Song, Z. Zhang, C. Zhang, X. He, and B. Zhou, “Improving prosody modelling with cross-utterance bert embeddings for end-to-end speech synthesis,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6079–6083.
- [28] G. Sun, C. Zhang, and P. C. Woodland, “Transformer language models with lstm-based cross-utterance information representation,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7363–7367.
- [29] 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, NIPS 2017*, 2017, pp. 5998–6008.
- [30] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, 2014, pp. 1746–1751.
- [31] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, “Durian: Duration informed attention network for multimodal synthesis,” *arXiv preprint arXiv:1909.01700*, 2019.
- [32] 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, NIPS 2020*, 2020.