Prosody Prediction with Discriminative Representation Method

Jipeng Zhang †
School of Information Science and Engineering Xinjiang University of China; Xinjiang Key Laboratory of Signal Detection and Processing
zhangjipeng@stu.xju.edu.cn

Hankiz Yilahun*
School of Information Science and Engineering, Xinjiang University of China; Xinjiang Key Laboratory of Multilingual Information Technology, China
hansumuruh@xju.edu.cn

Xiaoqin Feng
Mobvoi AI Lab
Suzhou, China
xiaoqin.feng@mobvoi.com

Yunlin Chen
Mobvoi AI Lab
Suzhou, China
yunlinchen@mobvoi.com

Xipeng Yang
Mobvoi AI Lab
Suzhou, China
xipeng.yang@mobvoi.com

Askar Hamdulla
School of Information Science and Engineering
Xinjiang University
Xinjiang Key Laboratory of Multilingual Information Technology, China
askar@xju.edu.cn

***Abstract*—Rhythm affects the naturalness and intelligibility of Text-To-Speech (TTS). However, rhythm prediction remains a great challenge, usually in two aspects: 1) the united annotation is a relatively difficult task, which depends on expert's experience. 2) traditional methods based on conditional random field (CRF), which heavily rely on feature engineering, such as word segmentation, part of speech(pos) etc. For above problems, we propose a method to reduce the dependency for united annotation data and conduct the joint experiment which use one unified model on independent data. Meanwhile, we also propose an algorithm of Layer Look Up Table (LLUT): use an embedding layer to learn a discriminative representation for different level of prosody data without any feature engineering. By using this method, the classifier can share the parameters and predict for different prosody level separately, which reduces the number of trainable model parameters. In order to better represent the input text, we use the pre-training model, like BERT, to provide the semantic information. Our experiment shows that the method of LLUT, is better able to acquire the discriminative meaning of different prosody levels. And also, our algorithm is proved to be general for sequence annotation tasks thus we can do extra task, like polyphone-prosody prediction.**

Keywords—prosody, CRF, layer look up table, BERT, pre-training model, discriminative representation

I. INTRODUCTION

End-to-end approaches based on deep learning have been very successful in text-to-speech (TTS) synthesis. In particular, TTS systems based on sequence-to-sequence models (e.g., Tacotron [1]) enable models to map character sequences directly to acoustic features, thus eliminating the need for complex text processing front-ends. The front-end of a TTS system depends on the language. For Chinese, it includes various modules, such as text normalization, segmentation, G2P conversion [2], and prosody prediction [3][4]. In this study, we focus on Chinese prosody prediction, which is one of the important tasks to improve naturalness of speech. Previous studies have utilized traditional statistical methods [5][6] such as decision trees, HMM, CRF, etc., and several experiments have shown that CRF works best in the task of prosody prediction.

However, in Chinese speech synthesis, CRF-based rhythm prediction has two main shortcomings. Firstly, it relies on strictly united annotation data, which requires high quality for data annotation. second, it relies on strict feature engineering, which requires high experience of annotators; finally, it cannot fully exploit the semantic information of the text by constructing statistical windows of the context.

In recent years, the research on text representation has entered the stage of state-of-art, among which research led by pre-trained language models enables researchers from different institutions to obtain excellent experimental results. For example, BERT [7], GPT, ELMO based on Transformer bi-directional encoder can be fine-tuned on a variety of text tasks such as QA, DG, sequence tasks, etc. to take full advantage of the pre-trained models [8]. Based on such architectures, the main contributions of this paper are as follows: (1) The conventional way of jointly labeling data requires strict labeling rules. By disentangling the different hierarchical rules of rhyming data, we solve the problem of not easily obtaining the complete joint dataset and improve the efficiency of data labeling. (2) We propose a discriminative feature representation under different prosody levels to facilitate the design and extension of multi-layer rhythm models. (3) Leverage the pre-trained model BERT to obtain rich semantic features without any feature engineering. (4) We propose several types of training methods for the discriminative representation task, which can achieve consistent results compared to training on joint data. We use the CRF model as the benchmark model to compare the proposed method in an objective evaluation based on F1 score, and meanwhile, conduct some ablation experiments to demonstrate the effectiveness of the proposed method and strategy.

†Work done during internship at Mobvoi AI Lab

II. PROPOSED METHOD

The goal of prosody prediction is to predict the correct pauses of phrases from the input text. This work can be viewed as a sequence labeling task [9], that is, after the text processing module, each character is marked with pauses or not. That is, given a source sequence $X = \{x1, x2, .., xt\}$ and the tag sequence $Y = \{y1, y2, ... yt\}$, the goal is to predict the label sequence from the character sequence. The ith element in the label sequence Y can be defined as a four-category label with non-prosodic pauses, prosodic words, prosodic phrases, or intonation phrases.

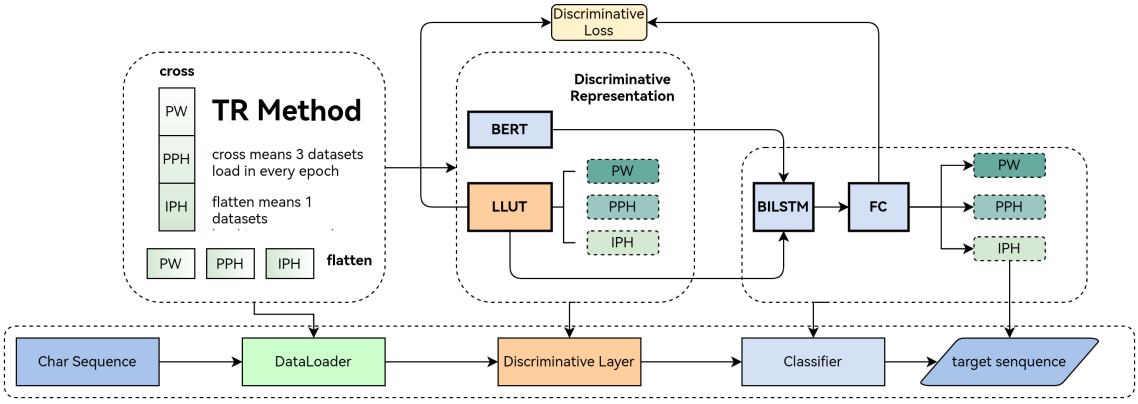

Fig. 1. Proposed model architecture: prosody prediction with discriminative representation structure

We use transformer-based BERT to model the sentence representation. The following is our proposed network structure, which mainly includes three parts: Data Loader, Discriminative Layer and Classifier. We will introduce the main structures in the following sessions, in which section A introduces the framework of BERT, section B introduces the Discriminative representation method proposed in this article, and section C introduces different training methods proposed in this article.

A. Bert Architecture

The pre-trained language model has the following advantages. Using massive unlabeled corpus data, it can learn common language representation and improve the effect of downstream tasks; it can better initialize other models, speed up model training and improve the effect; pre-training can reduce downstream tasks. Risk of overfitting on small data, equivalent to a regularization method.

BERT is a recently popular language model [10] that consists of a bidirectional Transformer encoder. The model can be used as an encoder, taking a series of characters as input and generating a word embedding for each token. The multi-headed self-attentive mechanism in the Transformer blocks enables the model to capture word dependencies in left and right-side contexts without any restrictions on word position in the sentence. After two unsupervised pre-training tasks, namely Masked Language Model (MLM) and Next Sentence Prediction (NSP), BERT is used to fine-tune the downstream tasks. Thanks to this pre-training, the model is believed to be able to capture rich Chinese contextual and semantic information, thus facilitating subsequent NLP tasks. In this paper, namely, we use BERT as sentence representation for prosody prediction, and introduce lstm to further extract the pre-trained features by adding sequence loss crf or softmax to construct the loss between target and prediction.

B. Discriminative Representation Structure

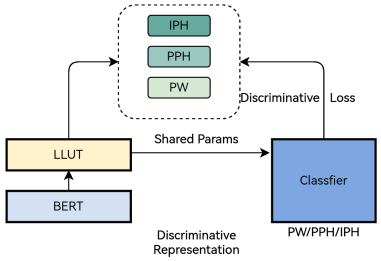

Fig. 2. Discriminative representation structure

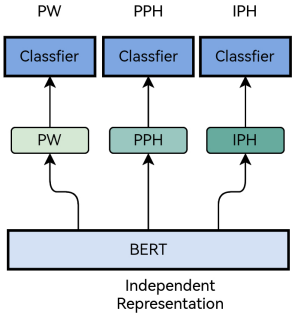

Fig. 3. Independent representation structure

After general modeling of sentences based on Bert, we constructed a discriminative representation called LLUT (Layer Look-Up Table). The motivation of this approach is to model the different level of prosody, achieving the effect of both differentiating the representation [11] classes and sharing the parameters of the intermediate layer as well as the classification layer for joint learning. We import an embedding layer to construct this module. During training, the corresponding layer is passed in and the embedding of the layer is contact to the output layer of Bert, where this embedding can be trained and updated to continuously adjust to obtain the best differentiated representation. In addition, our losses are also calculated for the current layer type and are not calculated jointly. We believe that this is a discriminative loss scheme, and learns the connection between layers in the intermediate shared parameters. We later performed a series of ablation experiments to demonstrate the effectiveness of our proposed method. At the same time, our proposed method is a general methodology, which will play a certain guiding role for similar tasks.

In this paper, softmax based classification loss as well as CRF based sequence loss are chosen for the experiments, where the corresponding objective functions are as follows:

a) Softmax based classification loss

Using cross-entropy as the training loss, the loss of rhythm labels is calculated and then averaged over the training sentences as follows:

$$L = -\frac{1}{|W^x|}\sum_{\omega \in W^x}\sum_{c} 1\{c=k_\omega\} \times \log y_c \quad (1)$$

where w^x is the index set of rhyming words in the training data, $\mathbb{1}$ is the indicator function, and k_ω is the true label of character ω.

b) CRF based sequence loss

During training, the model is optimized by maximizing the fraction of correctly labeled sequences = $(y1, y2, .., yT)$ while minimizing the fraction of all other sequences.

$$E = -s(y) + \log \sum_{\bar{y} \in \bar{Y}} e^{s(\bar{y})} \quad (2)$$

where s(y) is the CRF score of the sequence, and $s(\bar{y})$ denotes all possible sequences of labels.

C. A Special Training Method

Comparing to the normal rhythm task, we also inspired different training methods because of the inconsistency in the way the data are constructed. For the disentangling prosody dataset, we propose two training methods, which are mainly distinguished by the input method of the data in the training phase and the setting of the corresponding BERT parameters. Among them, method 1-Flatten, as shown in the left of Figure 2, uses different Data Loaders for PW, PPH and IPH, respectively. In each epoch, only one type of data will be sent to the model, that is, the training will be performed in the order of prosody level. The point to note in this method is that the parameter update strategy of the pre-training layer will be adjusted accordingly with the design of the model to suit different model structures. Method 2-Cross, use one Data Loader for PW, PPH, IPH data. At each epoch the three types of data are fed into the model according to the crossover, that is, the crossover of rhyme data is trained. For this reason, different model structures will be able to fine-tune the parameters of the pre-trained parameter layers. We believe that the cross method should be better than flatten, because it can optimize the parameters of the model by cross, rather than the way of sequential coverage. We argue that it is the Discriminative representation that does the trick, completely distinguishing the parameter space of each layer, as we will verify and explain in the experimental section.

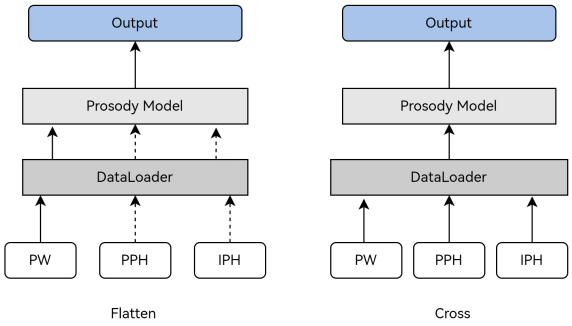

Fig. 4. Different training method design

III. EXPERIMENTS

A. Experiment Setting

Since there is no public dataset for the Chinese prosodic boundary prediction task, the experimental data used in this paper comes from within the company and is annotated by experienced language experts. The annotation results were reviewed to ensure consistency and accuracy. The annotation results are reviewed to ensure consistency and accuracy. Each rhyme in the dataset contains approximately 130,000 sentences, divided into a training set, a validation set, and a test set in an 8:1:1 ratio. For all experiments, we use a 2-layer BiLSTM (with 512 dimensions). The model is trained for 20 rounds using Adam as the optimizer, with batch size and learning rate set to 64, e-5, respectively. The loss functions we use to train the model are cross-entropy loss and CRF loss, while the model is evaluated with F1 scores. If the F1 value does not increase after 20 rounds, we stop the training early.

B. Contrast Experiment I

This set of experiments is mainly to verify the validity of our proposed discriminative representation structure and the experimental comparison with CRF as the benchmark model. Bert-Independent means that three independent classification layers are used for each of the three rhythm levels, and the parameters are updated independently between the layers; Bert-Cascade means that on the basis of Bert-Independent, the input of different levels takes into account the output of the previous level, which belongs to the normal cascade data training method. Bert-LLUT is the proposed model in this paper, in which different levels are represented in the representation layer and share the classification layer, which can be described in Section II. We choose the training method of cross, and the experimental results are shown in Table II.

TABLE I EXPERIMENTAL RESULTS OF BENCHMARK MODELS AND PROPOSED MODEL

Model	Label	ACC	REC	F1
CRF	PW	0.948	0.937	0.927
	PPH	0.908	0.927	0.745
	IPH	0.960	0.646	0.733
BERT-LLUT	PW	0.965	0.963	**0.964**
	PPH	0.922	0.924	**0.923**
	IPH	0.839	0.801	**0.822**

The experimental results in Table I show that our model shows a great improvement in the metrics of ACC, REC and F1 values compared with the baseline experiment.

TABLE II EXPERIMENTAL RESULTS OF THREE DIFFERENT MODELS

Model	Training Method	Loss	PW	PPH	IPH
Bert-Independent	cross	softmax	0.961	0.921	**0.831**
Bert-Cascade	cross	softmax	0.961	0.920	0.820
Bert-LLUT	cross	softmax	0.962	0.920	0.818
Bert-Independent	cross	crf	0.962	0.921	**0.823**
Bert-Cascade	cross	crf	0.962	0.923	0.827
Bert-LLUT	cross	crf	0.962	0.921	0.827

The experimental results in Table II show that our proposed representation of Discriminative is able to achieve parity with the normal training method for the rhythm classification task, demonstrating the effectiveness of the method.

C. Contrast Experiment II

This set of experiments is mainly designed to verify the effectiveness of our proposed training method. Two types of training methods, Flatten and Cross, are included, and the specific method theory is introduced in Section II. We selected three different types of model structures, and it is known from Table II that crf is effective, so we selected crf as the loss function and used F1 values of different rhythm levels for evaluation, and the experimental results are shown in Table III.

TABLE III EXPERIMENTAL RESULTS OF DIFFERENT TRAINING METHODS

Training Method	Model	PW	PPH	IPH
cross	Bert-Independent	0.962	0.921	**0.831**
flatten	Bert-Independent	0.956	0.904	0.793
cross	Bert-Cascade	0.961	0.920	**0.827**
flatten	Bert-Cascade	0.956	0.906	0.794
cross	Bert-LLUT	0.962	0.920	**0.828**
flatten	Bert-LLUT	0.948	0.879	0.782

With the above result analysis, the cross-based training approach is superior to the flatten approach, which is consistent with our expected results. There is roughly a 1-3 percentage point improvement in the results of each layer. We believe that it is the discriminative representation that does the trick, fully differentiating the parameter space of each layer for learning. Further, we also see that the cascade and flatten model architectures cause little impact on the experimental results.

IV. CONCLUSION

In this paper, inspired by the great success of BERT in many NLP tasks, we propose a prosody prediction model based on the LLUT approach. Through the exploration of the model structure, the experiment proves the effectiveness of our proposed method. 1)By disentangling the different hierarchical rules of prosodic data, we solve the problem of obtaining united annotation data 2)Under the non-joint prosody dataset, a discriminative representation method is proposed towards different prosody levels, which effectively learns the discriminative representation 3)Use the pre-trained model BERT to obtain rich semantic features without any feature engineering 4)Several training methods for the discriminative representation task are proposed, which achieve a better result.

However, we still have many shortcomings. For example, we can propose better methods on data annotation for better rhythm consistency. And also, we will explore the following aspects in the future 1) choosing more pre-trained language models for experiment 2) exploring better data representation methods 3) following lightweight deployment of models.

ACKNOWLEDGEMENT

We would like to thank Mobvoi TTS labeling team in Wuhan for supporting data processing and labeling. Thanks also to our colleagues Yunlin Chen, Xiaoqin Feng and Xipeng Yang et al. for helpful discussions and advice. This work was supported by the Strengthening Plan of National Defense Science and Technology Foundation of China (2021-JCJQ-JJ-0059) and Natural Science Foundation of China (U2003207).

REFERENCES

[1] Wang Y, Skerry-Ryan R J, Stanton D, et al. Tacotron: Towards end-to-end speech synthesis[J]. arXiv preprint arXiv:1703.10135, 2017.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Xie K, Pan W. Mandarin prosody prediction based on attention mechanism and multi-model ensemble[C]//International Conference on Intelligent Computing. Springer, Cham, 2018: 491-502.

[4] Futamata K, Park B, Yamamoto R, et al. Phrase break prediction with bidirectional encoder representations in Japanese text-to-speech synthesis[J]. arXiv preprint arXiv:2104.12395, 2021.

[5] Qian Y, Wu Z, Ma X, et al. Automatic prosody prediction and detection with Conditional Random Field (CRF) models[C]//2010 7th International Symposium on Chinese Spoken Language Processing. IEEE, 2010: 135-138.

[6] Zheng Y, Tao J, Wen Z, et al. BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in a Text-to-Speech Front-End[C]//Interspeech. 2018: 47-51.

[7] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[8] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.

[9] Zheng Y, Tao J, Wen Z, et al. BLSTM-CRF Based End-to-End Prosodic Boundary Prediction with Context Sensitive Embeddings in a Text-to-Speech Front-End[C]//Interspeech. 2018: 47-51.

[10] Qiu, Xipeng, et al. "Pre-trained models for natural language processing: A survey." *Science China Technological Sciences* 63.10 (2020): 1872-1897.

[11] Che H, Li Y, Tao J, et al. Investigating effect of rich syntactic features on mandarin prosodic boundaries prediction[J]. Journal of Signal Processing Systems, 2016, 82(2): 263-271.