Low-Resource Mandarin Prosodic Structure Prediction Using Self-Training

Xingrui Wang, Bowen Zhang, Takahiro Shinozaki
Tokyo Institute of Technology
www.ts.ip.titech.ac.jp

***Abstract*—For Chinese text-to-speech (TTS) system, prosodic structure prediction plays an important role in it. The prosodic information can improve the synthesize result tremendously. Recurrent neural network (RNN), conditional random fields (CRF), and self-attention have been used in this task. These approaches use large amount of data. However, the data marking process can be really time-consuming, and needs people with relevant knowledge to participate, which makes the prosodic information dataset expensive. On the other hand, the unlabeled data, are easy to get. In this paper, we use self-training to make use of unlabeled data and solve the labeled data insufficient problem. With the help of pre-trained language model, we start training with little labeled data and use active learning to generate more data. Experiments show that only starting training with 4k of labeled data, the proposed method can achieve equivalent or better result than baseline, while the baseline used more than 40k labeled data.**

I. Introduction

With the rapid development of deep learning, end-to-end speech synthesis (or text-to-speech, TTS) system has replaced traditional pipeline approaches and become the mainstream, and simplifies the developing process. Tacotron 2 [1] is the representative end-to-end TTS model, which is shown to be capable for generating high quality speeches for English speech synthesis.

Compared with English TTS, Chinese Mandarin TTS is more complicated. In Chinese text, there is no space between words. Therefore, in Chinese speech, prosody is used to represent the boundary between two words or topic turning between two parts. Prosody contains a lot of information, which affects the naturalness of the synthesis result. Wrong prosody boundaries can result in significant performance drops in speech synthesis quality. Therefore, it is necessary to model the prosody information separately so that the prosody information contained in text can be extracted properly. Fig. 1 shows an example of an incorrectly separated prosodic boundary, a pause is added by mistake in the middle of the word “pattern”. Research [7] has shown the improvement brought by introducing prosodic information to Chinese TTS.

At the early time, some statistical methods were used to generate prosodic information, such as hidden Markov model [2], maximum entropy model [3] and conditional random fields [4]. With the development of the deep learning model, Recurrent neural network (RNN) was used in this task successfully [5], Transformer [6] model has achieved outstanding results by enhancing the ability of extracting long-term contextual information, which filled up the shortcoming of Long short-term memory (LSTM) or RNN. [8] And [9] make use of self-attention mechanism to model the prosodic information.

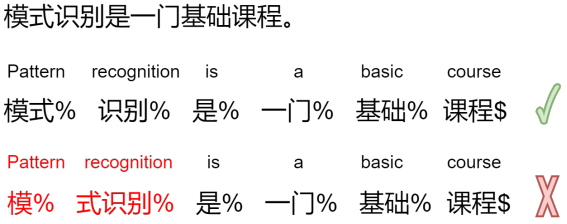

Fig. 1. An example of Chinese prosodic structure. Chinese and English words are one-to-one correspondence. The red part shows an incorrect boundary insertion. “%” is the sign of prosodic word or prosodic phrase(PW/PPH), “$” is the sign of intonational phrase (IPH).

Data marking process is very time-consuming and requires the participation of people with certain expertise, which makes the prosodic information dataset expensive and difficult to get. On the other hand, unlabeled data are usually abundant and easy to obtain. Self-training is used in such situation. Self-training has achieved many successes in classification tasks [14,15] and sequence generation [16,17]. A possible intuitive question on this method is “Will those bad pseudo labels lower the model performance?” [18] has shown that a high level of noise will lower performance but a low level of noise can actually improve performance due to its smoothing effect.

Previous works focus on accuracy improvement more. In this research, we focus on using less labeled data. We proposed a self-training way to solve the data insufficient problem. First use small amount of labeled data for training on small models, then use well-trained small model to mark unlabeled data. After that we use labeled data and data marked by the model together to train a complex model. We make use of unlabeled data and improve the accuracy effectively. By using only 4k labeled training data, we achieve higher accuracy than the baseline which uses about 40k labeled data.

II. Method

In this paper, we propose a convenient and efficient method to shorten the time of data preparation. The prosodic structure of a sentence is highly related to lexical and sentence structure. To extract and make use of this information, the pre-trained

language model is an ideal choice. In this research, first, we use BERT to preprocess the sentences and embedding them into latent space, then prosodic predict model generate prosodic annotation using embedded sentences and attention mask. To improve the accuracy, we also leverage those unlabeled data by using self-training.

A. Bidirectional Encoder Representations from Transformers (BERT)

BERT is a pre-trained language model that generates representations with high-level linguistic knowledge from input texts. It can be used in various NLP tasks without complex structural adjustments. Excellent performance and simple deployment make it widely used in various NLP tasks.

BERT uses a masked language model for training [12], that is, inputting a partially masked sentence and use models to fill these masked parts, so that the model can learn linguistic patterns from it. For Chinese, it is segmented and masked at the granularity of characters, without considering the Chinese word segmentation in traditional NLP. In this research, we use BERT with whole word mask [13] to preprocess and embed the text. As shown in Fig. 2, when masking a sentence, the BERT with whole word mask masks an entire word (shown in red) instead of only one character (shown in blue). The model has to predict the whole masked word during training. In Chinese, words are the smallest language unit that can be used to express meanings independently. Most words are composed of two or more characters. Masking whole words forces BERT to model the linguistic patterns at the language unit level, which is more in line with the natural form of Chinese. Meanwhile, the prosodic structure is highly related to word-level linguistic features, a good encoding can accurately reflect the meaning of words, which is necessary for further process.

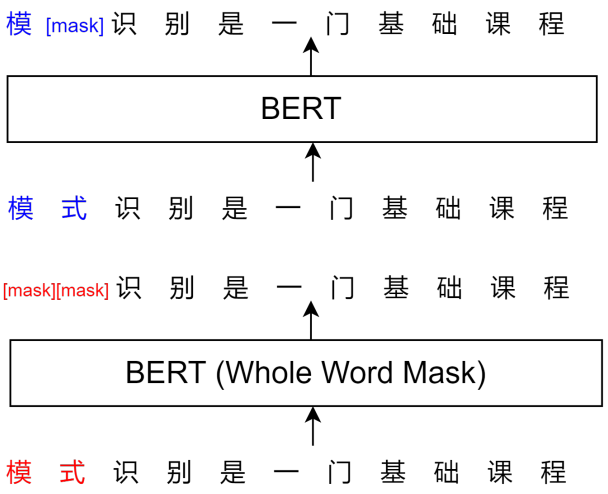

Fig. 2. An example of Chinese whole word mask. The red part showed that the word "pattern" is masked as a whole.

B. Self-training

Self-learning is an effective semi-supervised method. The process is as follows: first, a base model is trained on labeled data, acts as a "teacher". Then it is used for labeling the unlabeled data. These data labeled by the model are called pseudo labeled data. Then use both pseudo labeled data and labeled data to train a "student" model. The process is described in Algorithm 1.

Comparing with labeled data, unlabeled data are sufficient and easy to collect. Self-learning is suitable for this scenario. First, we train a teacher model, then use it to label the unlabeled data. We then train a student model on the extended dataset merged by pseudo-labeled data and labeled data. While training the student model with extended dataset, we divide the whole process into two steps: firstly, pre-training the model only on pseudo-labeled data; secondly, fine-tuning the model with labeled data. In the pre-training stage, the model captures common linguistic features, then in the fine-tuning stage, the model adapt to downstream tasks and make adjustments.

Algorithm 1: Self-training process

Input: labeled dataset $\mathcal{X}$, unlabeled dataset $\mathcal{U}$;
Initialize a model θ;
Train model θ on $\mathcal{X}$;
for *each training epoch* **do**
  Apply θ on unlabeled data $\mathcal{U}$;
  Select a subset $S \subset \{(u, \theta(u)) \mid u \in \mathcal{U}\}$;
  Train model θ on $\mathcal{X} \cup S$;
end
Return: well trained model θ

III. Experiment

A. Model

As shown in Fig. 3. We use the BERT-like model to preprocess the text. The open-source pre-trained models are provided by [13]. RoBERTa-wwm-ext and RBT3 are used in this experiment. RBT3 is a tiny version of RoBERTa. The parameter amount of RBT3 is 38% of RoBERTa. Further details of the model depend on the different experiments, which are shown in section III.D.

B. Dataset

In this experiment, we use two open-source datasets: AISHELL-3 [19] and THUCNews [20].

AISHELL-3 is a large-scale and high-fidelity multi-speaker Mandarin speech corpus. The corpus contains roughly 85 hours of emotion-neutral recordings spoken by 218 native Chinese mandarin speakers and a total of 88035 utterances. For this dataset, we only use its prosody annotations. The prosody annotations were divided into two types: "%" means the boundary after the prosodic word and the minor prosodic phrase, and "$" means the major prosodic phrase respectively, representing the boundary for topic turning and the intonation boundary. As a result, in this experiment, there are three types

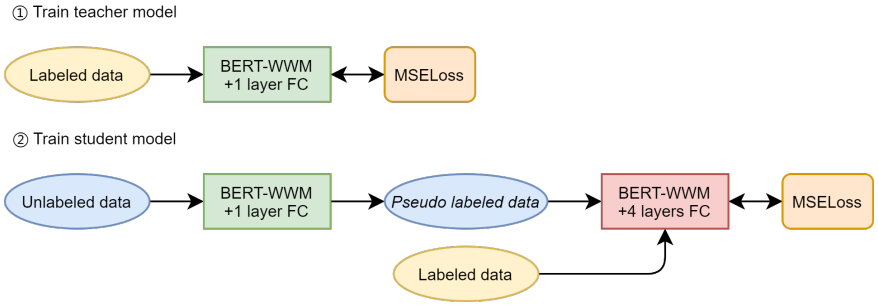

Fig. 3. Model illustration. The square brackets represent the size of tensor.

of prosody annotations: No Boundary (NB), Prosodic Word or Prosodic Phrase (PW/PPH), which is marked as "%" in the dataset, and Intonational Phrase (IPH), which is marked as "$".

THUCNews is generated based on the historical data of the Sina News RSS subscription channel from 2005 to 2011 and contains 740,000 news documents. It is divided into 14 categories. For THUCNews, we use part of text under the "social"category as unlabeled data, these data are used for generating pseudo labels. We split the paragraph into sentences for further process.

C. Evaluation metrics

We use total accuracy (T-ACC) to evaluate the prediction, it is calculated as:

$$T - ACC = \frac{N_{Correctly\_predicted}}{N_{Total\_number\_of\_labels}}, \quad (1)$$

where $N_{Correctly\_predicted}$ is the number of correct predictions, $N_{Total\_number\_of\_labels}$ is the number of prosodic label in the test set. We also use accuracy to evaluate the prediction results of each type labels.

There is only one IPH in a sentence in most situations but may have multiple NB or PW/PPH, which results in unbalanced data distribution. We also compute the F1 scores to to make the result more convincing.

D. Experiment design

TABLE I
EXPERIMENT DESIGN

Exp. name	Language model	Prediction model	Labeled data
Teacher-4k	RBT3	1-layer FC	4k
Ablation-4k	RoBERTa	4-layer FC	4k
Student-4k	RoBERTa	4-layer FC	4k
Teacher-8k	RoBERTa	1-layer FC	8k
Ablation-8k	RoBERTa	4-layer FC	8k
Student-8k	RoBERTa	4-layer FC	8k
Ablation-40k	RoBERTa	4-layer FC	40k
Attention [8]	Self-attention		40k

We design two groups of experiments. One is trained with 4k labeled data and the other is trained with 8k labeled data. All experiments use 1k validation data and 1k test data. Detailed information is listed in Table I. A student model and a teacher model are trained for each group. For better comparison, we also add a student model trained with only labeled data, which is called ablation. The teacher model and the ablation model are only trained on labeled data, whereas the student model is trained on both labeled and pseudo labeled data. According to the difference in training data amount, we use a relatively small model as the teacher, and a larger model as the student.

We consider Attention [8] as our baseline, which uses 40k labeled data. To help better understand the experiment results, we also conduct an ablation-40k experiment that uses the same amount of labeled data as the baseline system, as shown in Table III and V.

For all models except for Attention, the text length input to the BERT is 32, learning rate is 5e-6, dropout rate is 0.5, AdamW is used as the optimizer. For teacher model and ablation model, batch size is 32; for the student model, we use a batch size of 128 in the pre-training stage, and 16 in the fine-tuning stage. The parameters for Attention model follow exactly as proposed in [8]. N is set to 6.

E. Result and analysis

The experiment result is shown in Table II, III, IV and V. Despite using much less labeled data, our proposed method Student-4k and Student-8k both achieve better performance than the ablation system Attention [8] which uses 40k labeled data. By comparing the accuracy between student model and ablation model in same groups, we can find that the introduction of pseudo labeled data increases the total accuracy by about 2%.

On the other hand, we can consider the ablation-40k as the topline. With the help of pseudo label, the student model achieves similar results with ablation-40k while using less labeled data. These results prove the effectiveness of our approach.

TABLE II
ACCURACY EVALUATION RESULTS FOR 4K AND 8K EXPERIMENTS

Labeled data amount	4K				8K			
Metrics	NB-ACC	PW-ACC	IPH-ACC	T-ACC	NB-ACC	PW-ACC	IPH-ACC	T-ACC
Teacher	0.8800	0.9112	**0.8794**	0.8890	0.9355	0.9198	0.8697	0.9251
Ablation	0.9119	0.9136	0.8706	0.9088	0.9370	0.9289	**0.8856**	0.9301
Student	**0.9425**	**0.9222**	0.8732	**0.9305**	**0.9668**	**0.9410**	0.8812	**0.9518**

TABLE III
ACCURACY EVALUATION RESULTS FOR 40K EXPERIMENTS

Labeled data amount	40K			
Metrics	NB-ACC	PW-ACC	IPH-ACC	T-ACC
Attention	0.9263	0.9386	0.8864	0.9264
Ablation-40k	0.9694	0.9478	0.9023	0.9572

TABLE IV
F1 SCORE FOR 4K AND 8K EXPERIMENTS

Labeled data	4K			8K		
Metrics	NB-F1	PW-F1	IPH-F1	NB-F1	PW-F1	IPH-F1
Teacher	0.9174	0.8314	0.9041	0.9502	0.8804	0.9035
Ablation	0.9377	0.8585	0.8858	0.9543	0.8886	0.9043
Student	**0.9528**	**0.8883**	**0.9181**	**0.9714**	**0.9210**	**0.9158**

TABLE V
F1 SCORE FOR 40K EXPERIMENTS

Labeled data	40K		
Metrics	NB-F1	PW-F1	IPH-F1
Attention	0.9490	0.8857	0.9117
Ablation-40k	0.9742	0.9290	0.9318

The F1 score of the different experiments is shown in Table IV and Table V. Even the data are unevenly distributed, our method still improves the F1 score for all types of the prosodic label.

We also notice that among all types of prosodic boundaries, IPH has the worst accuracy in each experiment. This is probably caused by the data we use. In the AISHELL-3 dataset, all sentences are single sentences, which means no punctuation is involved. The sentences used for generating pseudo-labeled data also have no punctuation. The punctuation is a sign of IPH, which may provide hints for the model. We force the model to make decisions based on semantic because those punctuation-related IPH can be easily labeled by adding extra judgments in implementation.

IV. CONCLUSION

In this paper, we apply self-training on Chinese mandarin prosodic structure prediction to solve the data insufficient problem. By making use of a pre-trained language model and pseudo labeled data, we have achieved similar results at a very low cost. For future work, we plan to search for more Chinese prosodic datasets to examine the influence of self-learning on generalization ability, we will also try this method on other mandarin-like Asian languages.

REFERENCES

[1] Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R. & Others Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 4779-4783 (2018)

[2] Black, A. & Taylor, P. Assigning phrase breaks from part-of-speech sequences.. (International Speech Communication Association,1997)

[3] Li, J., Hu, G. & Wang, R. Chinese prosody phrase break prediction based on maximum entropy model. *Eighth International Conference On Spoken Language Processing*. (2004)

[4] Qian, Y., Wu, Z., Ma, X. & Soong, F. Automatic prosody prediction and detection with Conditional Random Field (CRF) models. *2010 7th International Symposium On Chinese Spoken Language Processing*. pp. 135-138 (2010)

[5] Vadapalli, A. & Gangashetty, S. An Investigation of Recurrent Neural Network Architectures Using Word Embeddings for Phrase Break Prediction.. *Interspeech*. pp. 2308-2312 (2016)

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances In Neural Information Processing Systems*. pp. 5998-6008 (2017)

[7] Lu, Y., Dong, M. & Chen, Y. Implementing prosodic phrasing in chinese end-to-end speech synthesis. *ICASSP 2019-2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 7050-7054 (2019)

[8] Du, Y., Wu, Z., Kang, S., Su, D., Yu, D. & Meng, H. Prosodic structure prediction using deep self-attention neural network. *2019 Asia-Pacific Signal And Information Processing Association Annual Summit And Conference (APSIPA ASC)*. pp. 320-324 (2019)

[9] Lu, C., Zhang, P. & Yan, Y. Self-attention based prosodic boundary prediction for chinese speech synthesis. *ICASSP 2019-2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 7035-7039 (2019)

[10] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint ArXiv:1810.04805*. (2018)

[11] Scudder, H. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions On Information Theory*. **11**, 363-371 (1965)

[12] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. Deep contextualized word representations. *ArXiv Preprint ArXiv:1802.05365*. (2018)

[13] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S. & Hu, G. Revisiting pre-trained models for chinese natural language processing. *ArXiv Preprint ArXiv:2004.13922*. (2020)

[14] Xie, Q., Dai, Z., Hovy, E., Luong, M. & Le, Q. Unsupervised data augmentation for consistency training. *ArXiv Preprint ArXiv:1904.12848*. (2019)

[15] Miyato, T., Maeda, S., Koyama, M. & Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **41**, 1979-1993 (2018)

[16] Kahn, J., Lee, A. & Hannun, A. Self-training for end-to-end speech recognition. *ICASSP 2020-2020 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 7084-7088 (2020)

[17] Wang, Y., Mukherjee, S., Chu, H., Tu, Y., Wu, M., Gao, J. & Awadallah, A. Adaptive self-training for few-shot neural sequence labeling. *ArXiv Preprint ArXiv:2010.03680*. (2020)

[18] He, J., Gu, J., Shen, J. & Ranzato, M. Revisiting self-training for neural sequence generation. *ArXiv Preprint ArXiv:1909.13788*. (2019)

[19] Shi, Y., Bu, H., Xu, X., Zhang, S. & Li, M. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *ArXiv Preprint ArXiv:2010.11567*. (2020)

[20] Yao, J., Wang, K., Xu, Z. & Yan, J. ClassVector: A Parameterized Prototype-Based Model for Text Classification. *Proceedings Of The 2019 11th International Conference On Machine Learning And Computing*. pp. 322-326 (2019)