Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 6438843, 11 pages
https://doi.org/10.1155/2022/6438843

Research Article

Improving Fluency of Spoken Mandarin for Nonnative Speakers by Prosodic Boundary Prediction Based on Deep Learning

Hongwu Yang,[1,2] **Dong Li**,[2,3] **and Yajing Yan**[3]

[1]*School of Educational Technology, Northwest Normal University, Lanzhou, China*
[2]*National and Local Joint Engineering Laboratory for Learning Analytics Technology of Internet Education Data, Lanzhou, China*
[3]*College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, China*

Correspondence should be addressed to Hongwu Yang; yanghw@nwnu.edu.cn

Received 22 October 2021; Revised 31 March 2022; Accepted 20 April 2022; Published 14 May 2022

Academic Editor: B. B. Gupta

Nonnative Mandarin speakers always have some unnatural pauses when speaking Mandarin due to their native pronunciation habits. Accurately predicting the prosodic structure of Chinese sentences is the key to improving fluency in Mandarin for nonnative speakers. This paper investigated the influence of the Chinese prosodic boundaries on the Mandarin spoken by international students. First, we proposed a new method to predict the prosodic word and prosodic phrase boundaries from Chinese sentences to obtain the prosodic boundaries automatically. Then, we used the predicted results to improve the Mandarin spoken fluency of international students. To train the prosodic boundary prediction model, we firstly constructed a Chinese prosodic boundary corpus that includes 100,000 Chinese sentences with manually labeled prosodic boundaries under the guidance of a linguist. We also proposed an end-to-end Chinese prosodic boundary prediction model based on the sequence-to-sequence model with a new feature named number of syntax hierarchy (NSH). Finally, we assess the fluency score of Mandarin using 1300 utterances recorded by six international students and a native Mandarin speaker. The utterances are recorded without/with the predicted prosodic boundaries. The experimental results show that the $F1$ scores of the prosodic word prediction model and the prosodic phrase prediction model are 98.14% and 85.24%, respectively. The fluency assessment results show that the fluency score labeled with the prosodic boundaries is higher than the fluency score of the international students when they read freely. The improvement of the score is between 7 and 16. Therefore, our method can be applied to the Mandarin education system to improve the spoken Mandarin fluency of nonnative speakers.

1. Introduction

Popularization and learning of Chinese for international students are now one of the priorities of higher education in China, with the gradual increase in the number of international students coming to China. However, because of the particular prosodic structure of Chinese, international students with poor Chinese proficiency are always not fluent in Mandarin. The prosodic structure of Mandarin is mainly reflected in speed and pause. An effectively organized prosodic structure contains the emotions and thoughts of the person when the sentence is expressed [1]. It is known that prosodic boundary can divide continuous speech into several prosodic units of various sizes to produce correct pauses in sentences, which directly affects the understanding of speech. For example, "打/死老虎" (hit the dead tiger) and "打死/老虎" (kill the tiger) have different meanings because of the pauses produced by the different prosodic boundaries [2].

Different languages have different prosody segmentations, and the same language will also produce different prosody hierarchies under different segmentation methods. For example, the English prosodic structure can be divided into phonological utterance, intonation phrase, phonological phrase, prosodic word, foot, syllable, and mora in descending order of syntactic standards [3]. According to the intonation mode, the English prosodic structure can be divided

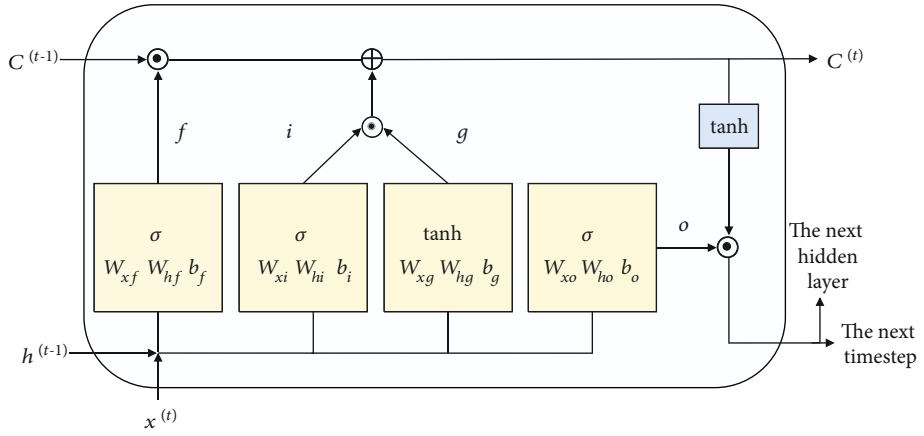

Figure 1: An LSTM memory cell.

into the middle phrase and intonation phrase [4]. For the Chinese prosodic structure, Li [5] followed the prosody structure of English to divide the Chinese prosody level from small to large into the foot, prosodic word, minor prosody phrase, major prosodic phrase, and intonation phrase. Zheng [6] made six division levels in more detail: degenerate syllable, normal syllable, minor phrase, main phrase, breath group, rhythm groups, etc. From the perspective of phonology, the prosodic hierarchy structure in small-to-large includes mora, syllable, foot, intonation segment, and tone group segment [7]. Chinese scholars generally agree that the prosodic hierarchy structure is mainly divided into three levels, including prosodic word, prosodic phrase, and intonation phrase [8]. In general, prosodic words refer to several closely and continuously connected syllables in pronunciation. A prosodic word is usually composed of two or three syllables and no pause within the prosodic word, while the prosodic phrase consists of one or a few prosodic words, with a relatively stable phrase intonation mode and phrase stress configuration mode. There are usually apparent pauses between intonation phrases. In oral communication, the listener's understanding of the meaning of an utterance can be obtained by the speaker setting prosodic boundaries to pause or accent while speaking [9].

In the past, many researchers used rule-based methods or traditional machine learning-based methods for predicting Mandarin prosodic structure. The C4.5 and transformation-based learning (TBL) [10] are typical rule-based learning algorithms. The maximum entropy model (ME) [11] and conditional random field (CRF) [12] are used to predict the prosodic phrase boundary. The rule templates of these methods are determined manually. Although the system is relatively simple and easy to understand or practice, the manual operation has many limitations. In recent years, deep learning-based methods have been widely used for prosodic boundary prediction, like recurrent neural networks (RNN), bidirectional long-short memory (BLSTM), and the BLSTM-CRF model formed by combining BLSTM and CRF [13–16]. Du et al. [17] applied the self-attention mechanism [18] to the task of prosodic structure prediction. Pan et al. [19] proposed a mandarin prosodic boundary prediction model based on multitask learning (MTL) architecture. Lu et al. [20] model the relationships between prosodic boundaries and lexicon words for prosodic boundary prediction by combining self-attention with MTL and setting word segmentation as an auxiliary task. Because the prosodic structure of Mandarin is related to the syntactic structure [21], the traditional shallow linguistic feature has been augmented or replaced by embedding features and some syntactic features [22–26]. There is still room to improve the prosodic prediction of Chinese and use the predicted result to perfect international students' speaking Mandarin.

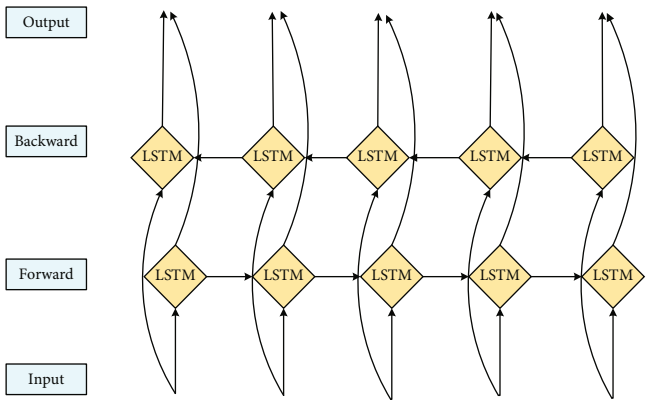

Figure 2: A BLSTM network.

The paper proposed a new Chinese prosodic boundary prediction method to help international students improve fluency in speaking Mandarin. We firstly constructed a corpus for modeling the Chinese prosodic structure. Then, we proposed a new method to predict the boundaries of both prosodic words and prosodic phrases of Chinese sentences. Finally, we use the predicted result to help international students studying Mandarin. The contributions of this paper are as follows:

(1) We constructed a Chinese prosodic boundary corpus with the prosodic words (PW) labeling and prosodic phrase (PPH) labeling of the lexicon under the guidance of a linguistic expert

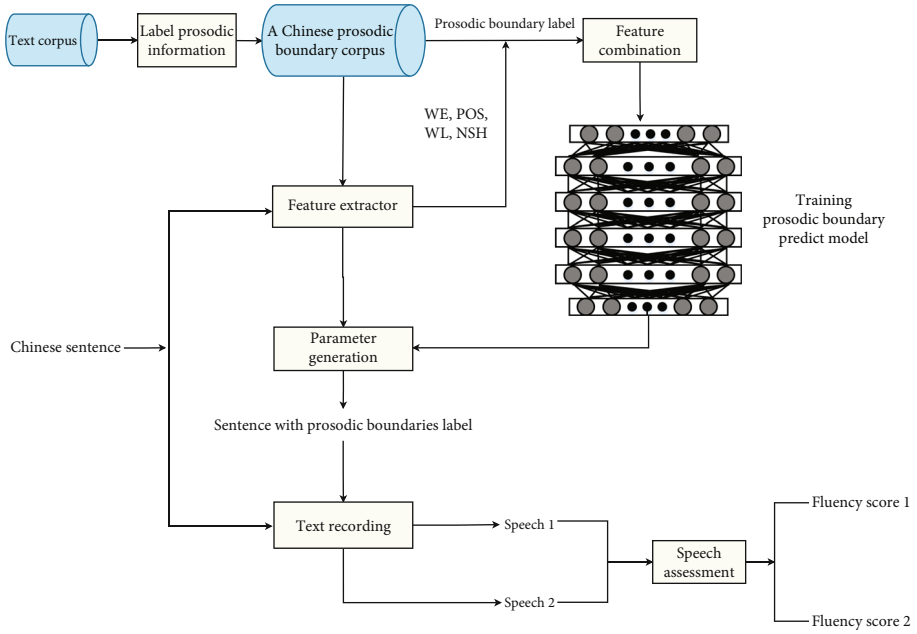

Figure 3: Framework for assessing fluency of international students in speaking Mandarin.

(2) We proposed an end-to-end Chinese prosodic prediction method by adding a new feature named the number of syntax hierarchy (NSH)

(3) We access the results of prosodic boundary prediction for improving the fluency of international students in speaking Mandarin

The rest of the paper is organized as follows. We first introduce the temporal sequences modeling in Section 2. Then, we present our framework for assessing fluency of international students in speaking Mandarin and explain each module in Section 3. The experimental setup and experimental results are presented in Section 4, while the discussion of the results is included. Finally, a brief conclusions and future works are provided in Section 5.

2. Temporal Sequence Modeling

For temporal sequences modeling, RNN always plays an important role. However, the standard RNN structure has limited ability to model long-term dependencies. Therefore, a variant of RNN, named long short-term memory (LSTM), is proposed to solve the problem effectively. This section presents a brief description of the LSTM and BLSTM networks and introduces how attention mechanisms solve the problem of hidden state information loss due to LSTM networks.

2.1. LSTM. The key components of LSTM are memory cells and gates. Figure 1 shows the memory cell of a single LSTM. This structure can retain information at many time steps and can effectively capture long-term time dependence. The calculation details of LSTM are as follows:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) = \sigma\left(W_{fh} h_{t-1} + W_{f_x} x_t + b_f\right), \tag{1}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) = \sigma\left(W_{ih} h_{t-1} + W_{i_x} x_t + b_i\right), \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh\left(W_{ch} h_{t-1} + W_{cx} X_t + b_c\right), \tag{3}$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) = \sigma\left(W_{oh} h_{t-1} + W_{ox} x_t + b_o\right), \tag{4}$$

$$h_t = o_t \odot \tanh\left(c_t\right), \tag{5}$$

where σ is the Sigmoid function and f_t, i_t, c_t, o_t, and h_t are the output of forget gate, input gate, cell memory, output gate at t time. h_t and X_t stand for hidden layer outputs and input vectors at t time. h_{t-1} and X_t form the input vectors for the current moment, w and b are the weight parameter matrix and the bias vector. $\odot$ represents the element-wise product.

2.2. BLSTM. The disadvantage of LSTM is that it can only access previous inputs. BLSTM uses a bidirectional structure to access the presiding and succeeding inputs. We unfold the time step of the BLSTM forward and backward, as shown in Figure 2. Every box represents an LSTM memory cell. Forward LSTM read input sequence $(x_1, x_2, \cdots, x_{Tx})$ and calculate hidden state $(\overrightarrow{h}_1, \overrightarrow{h}_2, \cdots, \overrightarrow{h}_{Tx})$, backward LSTM read sequence $(x_{Tx}, \cdots, x_2, x_1)$ and calculate hidden state $(\overleftarrow{h}_1, \overleftarrow{h}_2, \cdots, \overleftarrow{h}_{Tx})$.

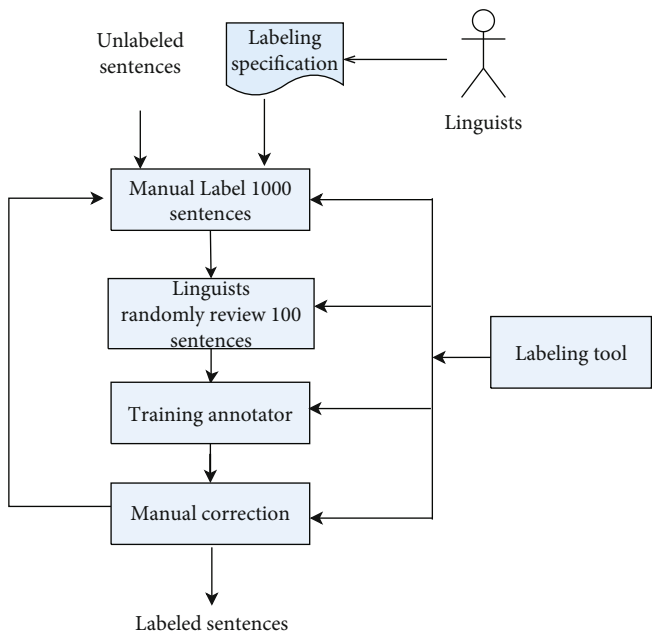

Figure 4: The diagram of the prosodic labeling process.

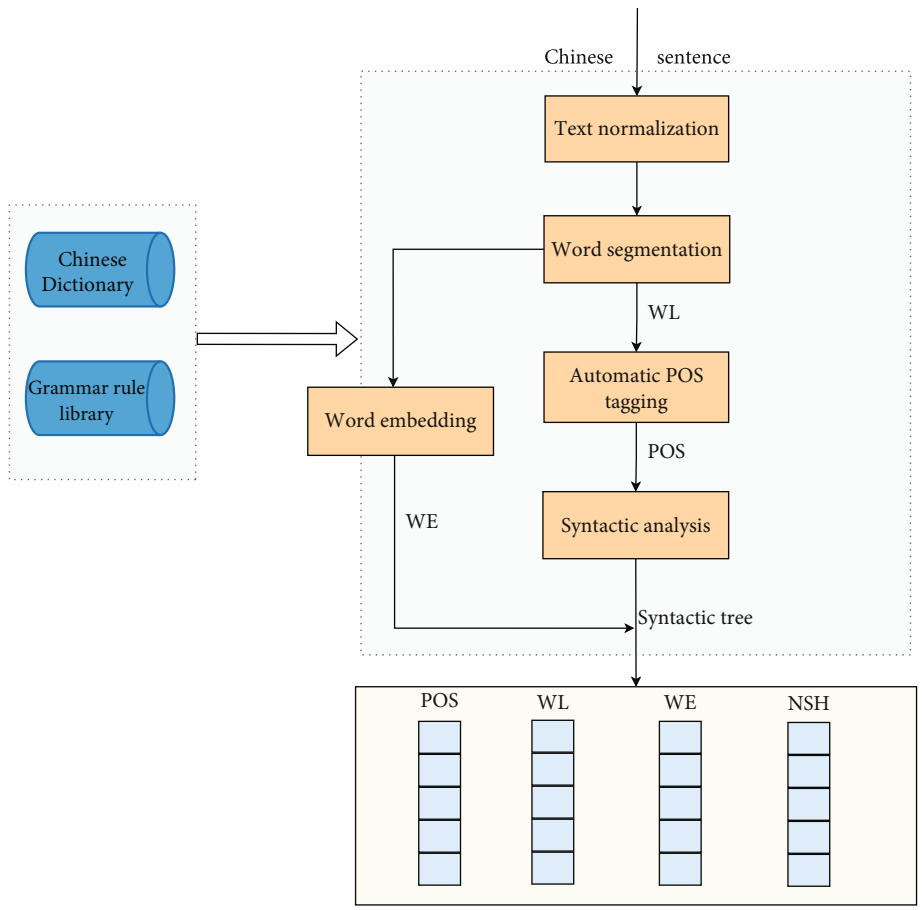

Figure 5: The framework of feature extractor.

```
(ROOT
  (IP
    (NP
      (DNP
        (NP (NN 人类)      1
          (NN 文明 ))  2
        (DEG 的))      2
      (NP (NN 发展 )))      3
    (PU ，  )      1
    (VP
      (VP
        (ADVP (AD 即将))      2
        (VP (VV 进入)      1
          (NP
            (QP (CD 一)      1
              (CLP (M 个 )))      3
            (ADJP (JJ 新))      2
            (NP (NN 世纪 )))))      5
      (PU ，  )      1
      (VP (VV 开启)      1
        (NP
          (QP (CD 一)      1
            (CLP (M 个 )))      3
          (ADJP (JJ 新))      2
          (NP (NN 千年 )))))      5
(PU 。 )))      3
```

Figure 6: An example of the syntax analysis with the labeled number of syntactic hierarchies, where the number is the NSH of lexicon words.

2.3. Attention. The attention mechanism is mainly proposed to deal with the problem of information loss in the hidden state [27]. Given the input sentence, the target sentence is generated after encoding-decoding operations, that is, we calculate the conditional probability of each possible word to search for the most likely word. The formula is as follows:

$$p(y_i|y_1, \cdots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i), \tag{6}$$

where s_i represents the hidden layer state of the decoder at t time:

$$s_i = f(s_{i-1}, y_{i-1}, c_i), \tag{7}$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j, \tag{8}$$

where c_i is the vector obtained by adding the hidden vector sequence $(\overrightarrow{h}_1, \overrightarrow{h}_2, \cdots, \overrightarrow{h}_{Tx})$ in the encoding process according to the weight.

Since the encoder uses a BLSTM network, h_i not only refers to the i^{th} word in the input information but also the information before and after the word. The vector sequences of these hidden layers are added according to their weights, and they have different proportions of attention distribution when generating the j^{th} output, as follows:

$$\alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{k=I}^{T_x} \exp(e_{ik})}, \tag{9}$$

$$e_{ij} = a\left(s_{i-1}, h_j\right), \tag{10}$$

where e_{ij} mainly evaluates the relationship between the input of position i and j; it mainly relies on the hidden state s in the LSTM and the i^{th} label h_j in the input sequence. The larger of α_{ij}, the more attention the i^{th} output allocates to the j^{th} input, and the greater the influence of the j^{th} input when generating the i^{th} output. Taking a as the parameter of the forward neural network enables the gradient of the loss function to propagate in different directions and be trained with the model.

3. The Framework for Assessing Fluency of International Students in Speaking Mandarin

To assess the effect of prosodic boundaries on the fluency of international students in speaking Mandarin, we further our work of [25] to predict the prosodic boundaries of Chinese sentences, as shown in Figure 3. First, in the model training stage, we performed text feature extraction on the Chinese sentences to extract features for the prosodic boundary prediction model training. Then, in the prediction stage, the extracted features of input sentences are fed into the model to predict the prosodic boundary labels. Finally, in the speech assessment stage, we used the 1300 recorded sentences to obtain the spoken Mandarin fluency scores of six international students by the speech evaluation system and analyzed the results.

3.1. Text Corpus with Prosodic Boundary Labeling. We selected 100,000 Chinese sentences from the "People's Daily" in 1998 and 2000 as the original text corpus. The corpus mainly consists of news and information, including social, financial, military, history, culture, science and technology, automotive, real estate, sports, entertainment, and health.

Through the statistical analysis, each sentence contains an average of 51.46 syllables and 25.73 grammatical words. The original sentences are first segmented into Chinese words and automatically labeled the POS with a lexical analysis tool. Then, we manually labeled prosodic boundaries (prosodic word boundaries and prosodic phrase boundaries) guided by a linguistic expert according to a labeling specification drawn up by the linguists. Figure 4 shows the labeling process of the corpus. In labeling, linguists randomly checked some sentences for review and correction. As a result, we have achieved a high degree of labeling precision on prosodic boundaries of consistency with linguists through our continuous corrections.

3.2. Feature Extraction. We use the shallow semantic features, including the part-of-speech (POS) of lexicon word, lexicon word length (WL), lexicon word embedding (WE), and a new deep syntactic feature named the number of syntax hierarchy (NSH) as the features to train the prosodic boundary prediction model. We designed a feature extractor to obtain all features, as shown in Figure 5. The Chinese input sentence

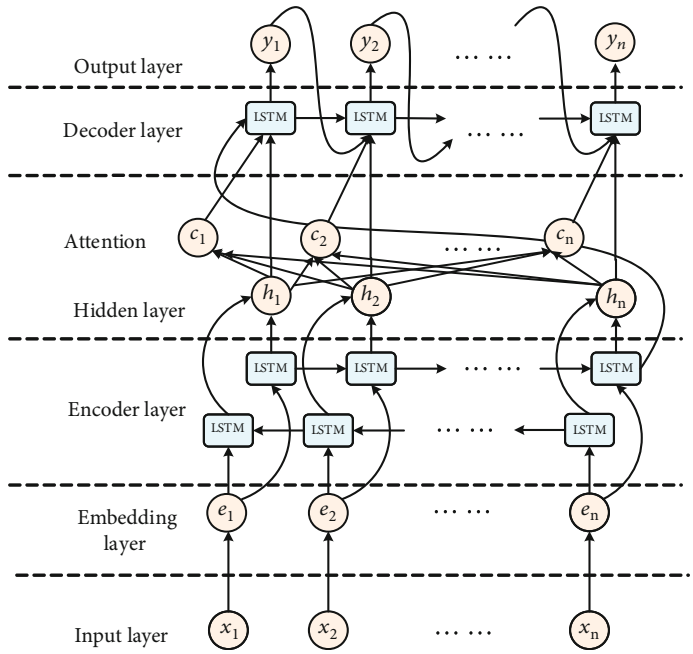

Figure 7: The framework of the seq2seq with attention model.

is first normalized with the text normalization algorithm. The sentence is then automatically segmented into lexicon words to obtain the WL and WE. Finally, a grammar rule library and a Chinese dictionary are employed to obtain each lexicon word's POS.

Because there is a strong relationship between grammatical words and prosodic boundaries, we conduct a syntactic analysis [28] of the normalized sentence again to obtain the syntactic tree of the sentence for calculating the NSH. The NSH is the height of the lexicon word on the syntactic tree that determines the sentence's syntactic hierarchies. The larger value of NSH takes the higher syntactic hierarchy and indicates the greater probability of the appearance of prosody boundaries. An example of calculating the NSH is shown in Figure 6.

3.3. Architecture of the Prosodic Boundary Prediction Model. We train the prosodic boundary prediction model by adopting the sequence to sequence (seq2seq) model based on the encoder-decoder structure [29], as shown in Figure 7. The encoder changes the input information of the model into a fixed-length semantic vector to compress the information. However, the encoder will lose the initial information of the sequence in information compression. As a result, the last generated meaning vector contains incomplete information so that the prediction sequence generated during decoding is inaccurate. Therefore, we added an attention mechanism to improve the prediction accuracy.

We use the BLSTM as the model's encoder while the LSTM as the decoder. After a word embedding layer, we feed the feature vectors obtained from the feature extractor into the encoder. The encoder generates a hidden state for every time step and initializes the initial hidden state of the decoder. Next, the attention mechanism calculates the correlation between the hidden state of the decoder and all the hidden states of the encoder to obtain different attention weights. Finally, the decoder reads the hidden state of the sequence forward and generates a prosodic boundary label sequence represented by 0 or 1.

3.4. Fluency Assessment. Since this work wants to verify whether the correct prosodic boundary can help improve international students' fluency in speaking Mandarin, we also ask six international students to be the subjects for recording Chinese sentences. The subjects all have an undergraduate degree and are between 19 and 23 years old and studied Chinese for one to two years, so they have an introductory level of Chinese. In order to compare with the level of native speakers, we also invited a native Mandarin female graduate student with the Mandarin Proficiency Test secondary-level A certificate as the speaker to record the same sentences. We selected 100 sentences based on syllable phoneme coverage to record. All sentences are automatically labeled prosodic boundaries with the trained prosodic prediction model. All speakers are first asked to read the sentence without labeling prosodic boundaries according to their understanding and then read the same sentence with prosodic boundary labeling according to the prosodic boundaries.

Because international students are unfamiliar with Chinese characters, the reading text is presented to them in both Chinese and Pinyin to avoid not being fluent in speaking caused by understanding the sentence. Before recording, each participant had 10 minutes to become familiar with the text, ensuring there were no text recognition and understanding barriers. In the recording, each subject is asked to keep reading at a constant speed as much as possible to reduce the influence of the difference in speaking speed on the experimental results. As a result, we finally recorded 1300 utterances. All recordings were first saved in the Microsoft Windows WAV format as sound files (monochannel,

Table 1: Architectures of BLSTM, seq2seq, and seq2seq+attention.

Type	Type of layer(s)	Number of layer(s)	Number of units
BLSTM	BLSTM	2	256
seq2seq	LSTM	1	256
	BLSTM	2	256
seq2seq+attention	LSTM	1	256
	BLSTM	2	256

Table 2: Experimental results of PW and PPH boundary prediction.

Model	Precision (%)		Recall (%)		$F1$ score (%)	
	PW	PPH	PW	PPH	PW	PPH
BLSTM	94.63	87.84	97.38	74.76	95.98	80.77
seq2seq	95.54	83.40	97.77	81.76	96.64	82.57
seq2seq+attention	97.50	84.21	98.78	81.59	**98.14**	**82.88**

Table 3: Experimental results of PPH boundary prediction with NSH feature.

Model	Precision (%)	Recall (%)	$F1$ score (%)
seq2seq+attention	84.21	81.59	82.88
seq2seq+attention+NSH	89.10	81.71	**85.24**

Table 4: Experimental results of different methods for PPH boundary prediction.

Model	Precision (%)	Recall (%)	$F1$ score (%)
ME	71.77	77.24	74.40
CRF	78.61	81.55	80.05
BLSTM	87.84	74.76	80.77
BLSTM-CRF	—	—	82.95
seq2seq	83.40	81.76	82.57
seq2seq+attention+NSH	89.10	81.71	**85.24**

signed 16 bits, sampled at 16 kHz) and then fed into the iFlytek Speech Assessment System (https://www.xfyun.cn/services/ise) to obtain the required fluency score. The iFlytek Speech Assessment Technology, which can score the speech on a scale of 0 to 100 based on the percentage of incorrect pauses, has been approved by the State Language Commission and has reached the practical level.

4. Experiments

We conducted several experiments to evaluate the proficiency of the prosodic boundary prediction model and assess the effects of the prosodic boundaries on the fluency of Mandarin read by international students.

Table 5: Preliminary fluency score of international student (S1 ~ S6) and fluency score of the native Mandarin speaker (standard).

	S1	S2	S3	S4	S5	S6
Initial	48.45	51.34	53.39	58.71	60.33	62.61
Standard			81.52			
Fluency level		Bad			Good	

4.1. Experimental Setup. We combine original word embedding (WE) and prosody features for the prosodic boundary prediction model to get a new embedding. Original WE are trained by the result of automatic word segmentation with 74,497 words. The context window size is five during training, and the word embedding dimension is 128. Finally, we concatenate the POS, WL, NSH, and original WE by the last dimension as the eventual input of the model.

Three kinds of models, including BLSTM, seq2seq, and seq2seq+attention, were compared in the experiments. The architectures of these frameworks are shown in Table 1. The batch size of the three models is 64, the learning rate is 0.003, and the decay rate is 0.2.

4.2. Experiment Results on Prosodic Boundary Prediction. We used precision (P), recall (R), and $F1$ score ($F1$) to evaluate the experimental results. Precision refers to the ratio of correctly identified prosodic boundaries to the total identified prosodic boundaries. The high precision takes low misrecognition. Recall refers to the ratio of correctly identified prosodic boundaries to the total prosodic boundaries in the test set. The higher recall takes lower missed identifications. We used $F1$ score to reconcile P and R, as defined in

$$F1 = \frac{2PR}{P+R}. \tag{11}$$

Predicting PW and PPH by the proposed the seq2seq +attention prosodic boundary prediction model is compared with the BLSTM and seq2seq model, as shown in Table 2. The $F1$ scores predicted by the seq2seq+attention model for PW and PPH reached 98.14% and 82.88%, respectively. We can see from Table 2 that the proposed prosodic boundary prediction model achieves the highest $F1$ score on both PW and PPH.

To evaluate the effect of the NSH on PPH prediction, we compare the results of the proposed seq2seq+attention model by adding/removing the NSH, as in Table 3. We can see from Table 3 that the $F1$ score of the seq2seq+attention+NSH model is improved by 2.36% compared to the seq2seq+attention model. It proves the effectiveness of the proposed NSH feature for PPH boundary prediction.

We also compared the PPH prediction performance of our model with others' models after adding NSH features using the same data set, as shown in Table 4. We can see from Table 4 that the proposed seq2seq+attention model with NSH reaches the highest $F1$ score.

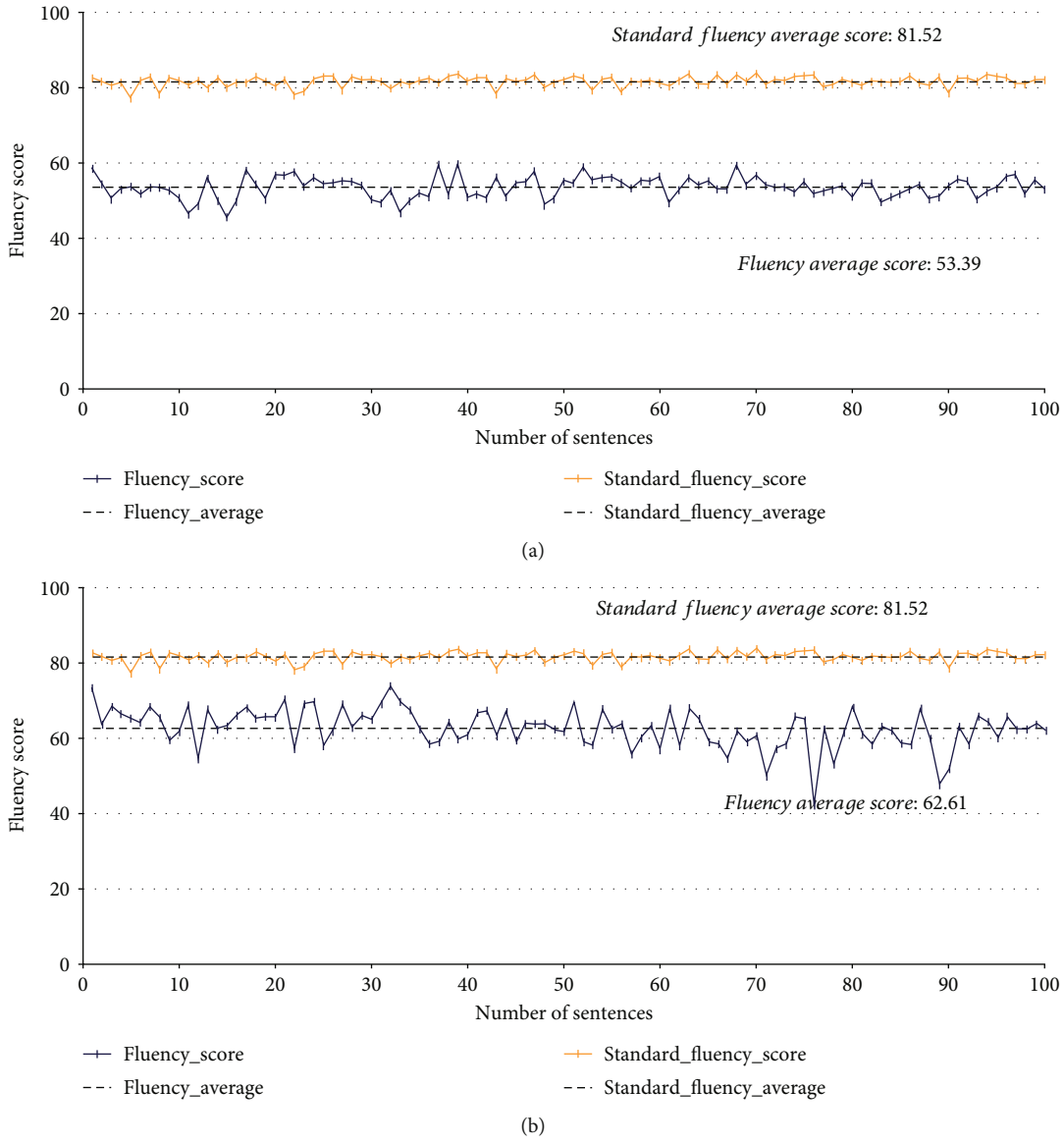

Figure 8: Comparing the preliminary Mandarin fluency scores of students S3 (a) and student S6 (b) with the standard fluency score.

4.3. Fluency Assessment. In order to analyze the influence of prosodic boundaries on fluency for international students speaking Mandarin, we conducted two experiments with the iFlytek Speech Assessment System. The first experiment used the recordings without prosodic labeling, while the second used the recordings with prosodic labeling.

In the first experiment, we assess each international student's preliminary fluency in Mandarin (S1 to S6), as shown in Table 5. We can divide the international students into two groups: poor reading fluency (S1 to S3) and good reading fluency (S4 to S6), based on the statistical results of Table 5.

Figure 8 shows the fluency score of S3 and S6. We can see from Figure 8 that the fluency score for native speakers is relatively stable, with an average score of 81.52. International students' fluency scores are generally lower, fluctuating more around the average score. After the 70th text sentence, the fluency score went down significantly. The possible reason is that the predominance of long texts after the 70th sentence is more difficult for international students. In addition, another possible reason is incorrect stopping positions for words and phrases during reading.

Table 6: International student (S1 ~ S6) fluency score labeled with prosodic boundaries.

	S1	S2	S3	S4	S5	S6
Average score	62.89	65.11	68.69	70.80	70.07	69.92

In the second experiment, we assess the effect of the prosodic boundaries on the fluency score for international students, as shown in Table 6. Similarly, we selected S3 and S6 for visual analysis of fluency scores, as shown in Figure 9. The statistics of

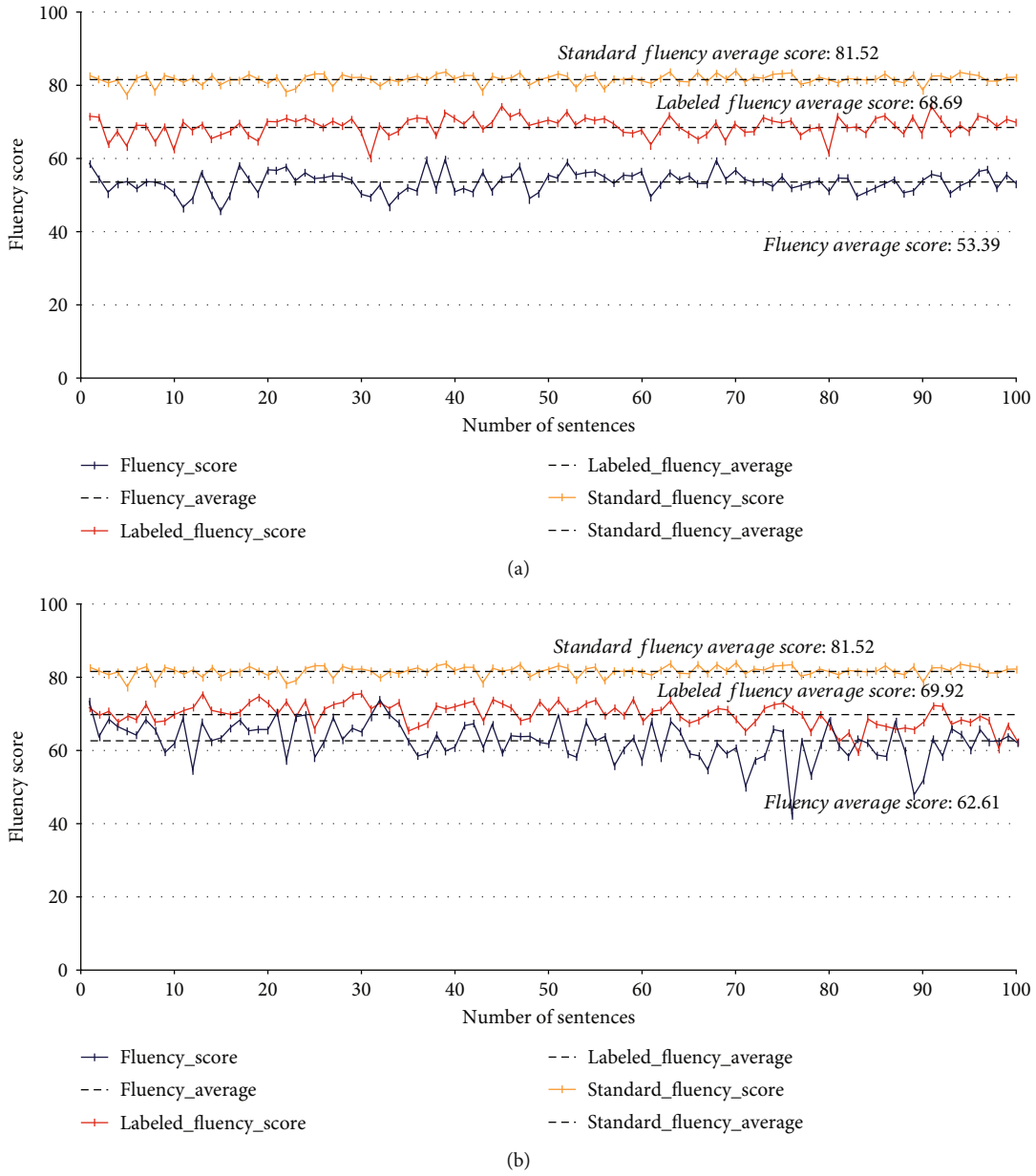

Figure 9: Comparison of the initial Mandarin fluency scores of S3 (a) and S6 (b) with the fluency scores of labeled prosodic boundaries.

the six international students' final assessment results are shown in Figure 10. For international students, labeling the prosodic boundaries can improve their fluency scores notable.

However, there is still a gap between them and native speakers. We believe the reasons are as follows:

(i) The international students did not study Chinese for a long time and were at the elementary Mandarin level. They have not yet been able to systematically grasp Chinese phonetics, grammar knowledge, prosodic characteristics, and the relationship between Mandarin pronunciation, pause, and semantic expression

(ii) The international students lacked Mandarin prosodic awareness. Because the international students lack the understanding of Chinese prosody, they cannot handle the relationship between Mandarin tone and intonation well

After conducting simple prosody training related to "prosody pause" for international students, their oral fluency was greatly improved through the speech evaluation experiment. The improvement score is between 7 and approximately 16. The speech assessment results are consistent with the expected results. It proves that our work can help nonnative Mandarin

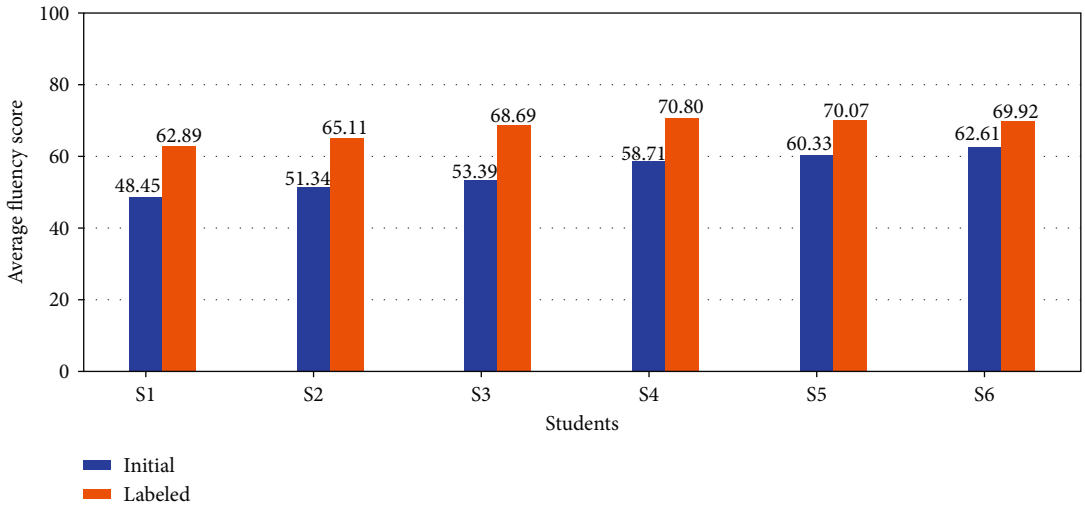

Figure 10: Statistics of the average Mandarin fluency scores results of 6 international students.

learners to grasp the prosody structure better and improve their fluency in speaking Mandarin.

5. Conclusions

This paper assesses the influence of the Chinese prosodic boundaries on speaking Mandarin for international students. We proposed a novel Mandarin prosodic boundary prediction model based on the seq2seq with an attention mechanism for helping international students learn Mandarin. The model uses a new feature named the NSH to predict PPH. The experimental results show that the proposed model can improve the accuracy of the prosodic boundary prediction. The proposed new feature, NSH, also can further improve the PPH prediction accuracy. We also conducted a fluency assessment experiment with the utterance recorded by the international students read without/with the prosodic boundaries predicted by the proposed model, proving that our work could help international students better grasp the prosodic structure and improve their spoken Mandarin fluency. Because international students also find it challenging to master the Chinese accent, we need to study further the Chinese accent prediction method and the effect of accent on the improvement of international students' Mandarin level in future work. At the same time, because the prosodic boundaries of Chinese have a hierarchical structure, we will further study the method of using hierarchical networks to predict the prosodic structure to improve the accuracy of prosodic boundary prediction.

Data Availability

The data used to support the findings of this study were supplied by the corresponding author under license and so cannot be made freely available. Requests for access to these data should be made to the corresponding author.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62067008 and No. 31860285). Additionally, part of this work was performed in Promoting High-quality Education Development in Ethnic Minority Regions through Collaborative Innovation in Intelligent Education (key scientific research project for Double World-Class Initiative in Gansu Province) (Project No. GSSYLXM-01) and the Science and Technology Program of Gansu Province (Grant No. 21JR7RA117).

References

[1] Z. Xiong, "The prosodic features of natural utterance boundaries and their communicative functions," *Applied Linguistics*, vol. 14, no. 2, pp. 144–144, 2005.

[2] J. Lin, Z. Ji, W. Dong, Y. Xie, and J. Zhang, "Improving mandarin prosody boundary detection by using phonetic information and deep LSTM model," in *International Conference on Asian Language Processing (IALP)*, Shanghai, China, 2019.

[3] E. Selkirk, "On derived domains in sentence phonology," *Phonology Yearbook*, vol. 3, pp. 371–405, 1986.

[4] D. R. Ladd, M. E. Beckman, and J. B. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology*, vol. 3, pp. 255–309, 1986.

[5] A. Li, "Prosodic analysis on conversations in standard Chinese," *Studies of the Chinese Language*, vol. 6, 2002.

[6] Q. Zheng, "The main credibility of prosodic structure in language flow," in *Proceedings of the 6th National Conference on Human-Machine Voice Communication*, Shenzhen, China, 2001.

[7] H. Wang, *Non-Linear Phonology of Chinese*, Peking University Press, Beijing, 2008.

[8] J. Li, G. Hu, and R. Wang, "Prosody phrase break prediction based on maximum entropy model," *Journal of Chinese information processing*, vol. 18, no. 5, pp. 56–63, 2004.

[9] J. Holzgrefe, C. Wellmann, C. Petrone, H. Truckenbrodt, B. Höhle, and I. Wartenburger, "Brain response to prosodic boundary cues depends on boundary position," *Frontiers in Psychology*, vol. 4, p. 421, 2013.

[10] A. Mikheev, *Document Centered Approach to Text Normalization*, ACM, 2001.

[11] J. Cao, *In Research and Exploration Modern Phonetics*, The Commercial Press, Beijing, 2007.

[12] Y. Dong, T. Zhou, and C. Dong, "Prosodic structure prediction based on conditional random field model," *Journal of Beijing University of Posts and Telecommunications*, vol. 32, no. 5, pp. 36–40, 2009.

[13] S. Zhao, "Rule-learning based prosodic structure prediction," *Journal of Chinese Information Processing*, vol. 16, no. 5, pp. 30–37, 2002.

[14] C. Ding, L. Xie, J. Yan, W. Zhang, and Y. Liu, "Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features, Automatic Speech Recognition and Understanding," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 98–102, Scottsdale, AZ, USA, 2016.

[15] 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 *The 19th Annual Conference of the Speech Communication Association (Interspeech)*, pp. 47–51, Hyderabad, India, 2018.

[16] Y. Wang and L. Cai, "Syntactic information and analysis and prediction of prosody structure," *Journal of Chinese Information Processing*, vol. 24, no. 1, pp. 65–70, 2010.

[17] Y. Du, Z. Wu, S. Kang, D. Su, and H. Meng, "Prosodic structure prediction using deep self-attention neural network," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE*, Lanzhou, China, 2019.

[18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171–4186, 2019.

[19] H. Pan, X. Li, and Z. Huang, "A mandarin prosodic boundary prediction model based on multi-task learning," in *The 19th Annual Conference of the Speech Communication Association (Interspeech)*, Graz, Austria, 2019.

[20] C. Lu, P. Zhang, and Y. Yan, "Self-attention based prosodic boundary prediction for Chinese speech synthesis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*, Brighton, United Kingdom, 2019.

[21] H. Che, Y. Li, J. Tao, and Z. Wen, "Investigating effect of rich syntactic features on Mandarin prosodic boundaries prediction," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 263–271, 2016.

[22] A. Rendel, R. Fernandez, R. Hoory, and B. Ramabhadran, "Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5655–5659, Shanghai, China, 2016.

[23] 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," *17th Annual Conference of the International Speech Communication Association (ITERSPEECH)*, pp. 3201–3205, 2016.

[24] 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 *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Scottsdale, AZ, USA, 2015.

[25] Y. Yan, J. Jiang, and H. Yang, "Mandarin prosody boundary prediction based on sequence-to-sequence model," in *IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chongqing, China, 2020.

[26] N. T. T. Trang, N. Ky, A. Rilliard, and C. d'Alessandro, "Prosodic boundary prediction model for Vietnamese text-to-speech," *INTERSPEECH*, vol. 2021, pp. 3885–3889, 2021.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations 2015(ICLR)*, pp. 1–15, California, USA, 2015.

[28] R. Levy and C. D. Manning, "Is it harder to parse Chinese, or the Chinese Treebank?," in *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 439–446, Sapporo, Japan, 2003.

[29] V. O. Sutskever and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 27, pp. 3104–3112, 2014.