Improving Mandarin Prosody Boundary Detection by Using Phonetic Information and Deep LSTM Model

Ju Lin[1,2]
[1] Beijing Language and Culture University, Beijing, China.
[2] Clemson University, Atlanta, United States of America.
e-mail: jul@clemson.edu

Zhuanzhuan Ji, Wenwei Dong
Beijing Language and Culture University, Beijing Advanced Innovation Center for Language Resources, Beijing, China.
e-mail: Jizhuanzhuan2017@126.com, dongwenwei_blcu@163.com

Yanlu Xie
Beijing Language and Culture University, Beijing Advanced Innovation Center for Language Resources, Beijing, China.
e-mail: xieyanlu@blcu.edu.cn

Jinsong Zhang*
Beijing Language and Culture University, Beijing Advanced Innovation Center for Language Resources, Beijing, China.
e-mail: jinsong.zhang@blcu.edu.cn

***Abstract*—Automatic prosodic boundary detection is useful for automatic speech processing, such as automatic speech recognition (ASR) and speech synthesis. In this paper, we propose two techniques to improve the boundary detection performance. First, in addition to prosody features (e.g, pitch, duration and energy), phonetic information (word/articulatory information) is integrated into the framework of prosodic boundary detection. We compared two forms of phonetic information: word form and articulatory form. Moreover, boundary detection can be regarded as a sequence labeling task. A deep Long Short-Term Memory (LSTM) is adopted for this task, which replaces the traditional Deep Neural Networks (DNN) model. The experimental results showed that the boundary detection performance can be improved by the additional phonetic information, with relative 5.9% (word form) and 9.8% (articulatory form) improvements respectively in contrast with the system that only used prosody features modeled. The articulatory information and prosody features with deep LSTM achieved the best result, with further performance enhancement from 76.35% to 77.85% (relative 6.3%) compared with that modeled by DNN.**

Keywords-Prosodic boundary detection, articulatory information, sequence labeling, LSTM

I. INTRODUCTION *(HEADING 1)*

Prosody plays an important role in speech production and comprehension. An effectively-organized prosody structure enables the speakers to more clearly convey their intended meanings. Meanwhile, it also helps the listeners understand the intention of the speakers more accurately. It is known that prosodic boundary indicates the degree of disjuncture between adjacent syllables in Mandarin, which divides continuous speech into several prosodic units of various sizes, such as prosodic word, intermediate phrase and intonational phrase. Such prosodic effects are highly informative for listeners, with similar results found across languages. As one of the key issues concerning prosody, prosodic phrasing affects the comprehension of the sentence in speech. For example, “打/死老虎” (hit the dead tiger) and “打死/老虎” (kill the tiger) have different meanings since they have different prosodic boundaries. It is thus greatly beneficial to take prosodic boundaries into consideration in automatic speech processing, such as automatic speech recognition (ASR) and speech synthesis, etc.

Previous studies have presented various features and approaches on the automatic prosodic boundary detection. Ostendorf et al. [1] proposed to apply decision trees and a Markov sequence model to predict the prosody boundary of the text in English and achieved 77% accuracy. An automatic phrase boundary labeling method for speech synthesis database annotation was presented by Chen et al. [2], which used context-dependent hidden Markov models (CD-HMM) and n-gram prior distributions, and improved the F-score of phrase boundary labeling from 72.2% to 79.6% on Boston University Radio Speech Corpus (BURNC). Ni et al. [3] proposed a hierarchical prosodic break classification method, which utilized the acoustic, lexical and syntactical features, and achieved 78.25% correct rate for the testing set. Yang et al. [4] used an unsupervised method based on CD-HMMs for labeling the phrase boundary positions of a Mandarin speech synthesis database and obtained an F-score of 77.64%. Lin et al. [5] proposed to utilize tone nucleus based prosodic features and DNN model to improve the detection performance.

As mentioned above, the performance of boundary detection still needs to improve. Prosody in speech is manifested by variations of loudness, exaggeration of the pitch so that low pitches are lower and high pitches are higher, and exaggeration of consonant and vowel properties, such as vowel height and aspiration [6]. Among different prosodic boundaries, this degree of exaggeration would be the difference. Thus, the speech attribute, also known as articulatory information, can be described this degree of exaggeration. It would be useful for improving the performance of boundary detection.

Previous studies have shown that several models are used for prosodic boundary detection, such as decision trees, SVM, HMM, and DNN, etc. DNN has a strong ability in feature learning, which can map the input features into a better feature representation via non-linear transformation of several hidden layers. Deep Recurrent Neural Network (RNN) not only has the abilities of DNN’s feature learning

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

but also can model long context and sequence information. Moreover, deep RNN with LSTM architecture [7] can address the vanishing problem that exists in standard RNN. Therefore, we propose to a prosodic boundary detection approach based on Deep LSTM, which combine the prosodic and articulatory information.

The paper is organized as follows: In Section 2, a description of the LSTM model is presented. Section 3 presents the features we used in this paper. It is followed by the experiments and results in Section 4. The paper is concluded with our directions for future work in Section 5.

II. LSTM MODEL

Prosodic boundary detection is a sequence to sequence labeling task, which is mean that each syllable of the sentence needs to design a boundary index. As we knew, LSTM can model a longer context than DNN and standard RNN. This is the reason why LSTM is suitable for this task. We first give an example of the same boundary index with same tone pattern but at different positions in the sentence. As shown in Figure 1, the upper one is at the beginning of the sentence and the nether one is at the ending of the sentence. The pitch at the beginning of the syllable in the upper figure is higher than that in the nether figure. This is due to the tone level will adjust itself to be consistent with the sentential intonation structure. For instance, the pitch contour of a declarative utterance generally declines gradually [8]. We expect that the deep LTSM architecture could model this useful information for prosodic boundary detection.

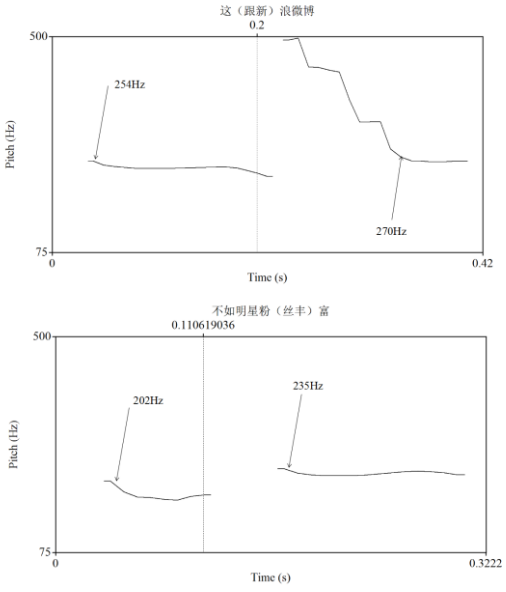

Figure 1. An example of the same boundary index at different positions in the sentence, both of them has the same tone pattern (tone1 and tone 1).

The structure of LSTM we used in this paper is as shown in Figure 2. Here is an example of the unfolded architecture of unidirectional LSTM with three consecutive steps. To deal with a variable length of input sequences, zero-padding was performed to pad a shorter sentence based on the maximum length for training LSTM. If the padding values are set as zero, or some other pre-defined values, the training and testing results will be highly biased. Thus, a masking mechanism was adopted to overcome the potential padding values problem. The Time Distributed function adds an independent layer for each time step in the recurrent model. This is used for gaining the output at each time step. The LSTM transition equations are defined as follows:

$$i_t = relu(W_i x_t + U_i h_{t-1} + b_i) \quad (1)$$

$$f_t = relu(W_f x_t + U_f h_{t-1} + b_f) \quad (2)$$

$$o_t = relu(W_o + U_o h_{t-1} + b_o) \quad (3)$$

$$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

Where the operation $\odot$ denotes the element-wise vector product. At time t, the input gate, the forget gate, and the output gate, denoted as i_t,f_t, and o_t respectively. c_t is the memory cell and h_t is the hidden layer representation. W_* and b_* denote the weight matrix and bias vector of corresponding gate functions.

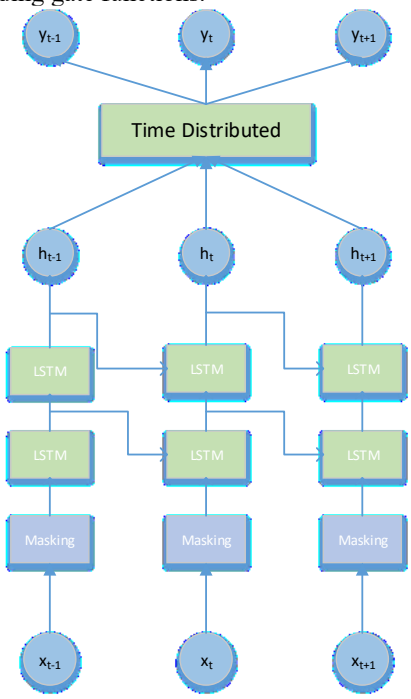

Figure 2. LSTM structure for prosodic boundary detection.

III. FEATURES

A. Phonetic information

As mentioned-above, phonetic information can be used for prosodic boundary detection. We compare two ways of feature representations to add phonetic information. One is 1 of V form, which is mean that uses a fixed-size vocabulary of symbols with V members in total, each input symbol can be coded as a vector of size V with all zeros except for the element corresponding to the symbol's order in the vocabulary, which gets a 1. The V in this paper indicates Chinese Pinyin with no tone symbol. Another way is that uses a series of detailed categories based on articulatory movements such as manner of articulation, place of articulation, etc. [8-10]. We adopted articulatory categories consistent with [8]. There are 19 categories in total, including 4 for Initials and 15 for Finals. First, the Pinyin

transcripts are mapped to the Initial and Final sequence. Then, the Initial and Final sequence is encoded in a 19-dimension vector, which has two values are 1 and others are 0.

B. Duration features

Previous studies demonstrated that duration-related features are related to boundary index [3, 11]. It is shown that pauses are more likely to appear at prosody boundaries and the duration of syllables preceding prosody boundaries is longer than that of other positions. The boundary information of Initials, Finals, and silence of the data was generated from forced alignment with a recognizer based on DNN. The durational features we used in this paper is as follows:

- The duration of silence after the current syllable
- The duration of the current syllable.

C. Pitch Features

Pitch is the most commonly used feature for prosodic boundary detection [11]. Prosody boundaries usually give rise to a variation of pitch reset, and the degree of pitch reset is greatly dependent on the level of prosody boundaries [11]. The higher boundary index is, the larger pitch reset there is. In this paper, the pitch reset is not computed explicitly but uses the context information. The pitch information of preceding, current and following syllable are spliced together. We expect the neural network could learn pitch reset information and other information via syllable based context features. Straight toolkit [12] is used for extracting pitch value. For each syllable, the following pitch-related features were calculated:

- Fitting the pitch contour of a current syllable with $f(t) = a + bt$ and {a, b} was used to represent the pitch contour feature.
- The maximum F0 value of the current syllable.
- The minimum F0 value of the current syllable.
- The F0 range of the current syllable.
- The mean F0 value of the current tone nucleus.
- The F0 value of the first point in the current syllable.
- The F0 value of the last point in the current syllable

D. Energy features

Energy-related features are also incorporated for prosodic boundary detection. We use PRAAT toolkit [13] to extract energy features and the features we used are as follows:

- The maximum energy value in the current syllable.
- The minimum energy value in the current syllable.
- The energy range of the current syllable.
- The mean energy value of the current syllable.

IV. Experiments and Results

A. Data set

A large mandarin speech corpus was used in this study, which was designed for TTS and labeled with prosodic ties. All of the text was read by one speaker. The speech signals were recorded in one channel, sampled at 16K Hz and at 16-bit precision. Prosodic boundaries defined in this corpus are similar to C-ToBI [14, 15]. The prosodic boundary was labeled by 1, 2, 3, 4, which represent prosody word boundary, minor prosody phrase boundary, major prosody phrase boundary and intonation group boundary respectively. Syllable boundary not in pre-defined prosody boundaries is set as 0. In the stage of manual annotation of prosodic boundaries, annotators were asked to annotate prosodic boundary according to perceptual listening and acoustic manifestation. Each sentence was annotated by three annotators, the final label of the prosodic boundary was determined by the voting mechanism. Figure 3 is an example of a prosodic boundary annotation. Table 1 shows the break distribution in our corpus in details. In this paper, we selected the 10% of total corpus as a testing set, 10% of total corpus as a validation set, the rest as the training set.

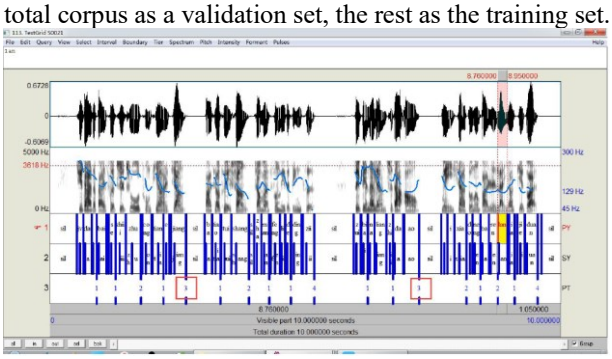

Figure 3. An example of prosodic boundary annotation

TABLE I. The detail information of boundaries in the corpus

Total	0	1	2	3	4
96660	52320	19482	12048	4837	7973
100%	54.12%	20.16%	12.46%	5%	8.24%

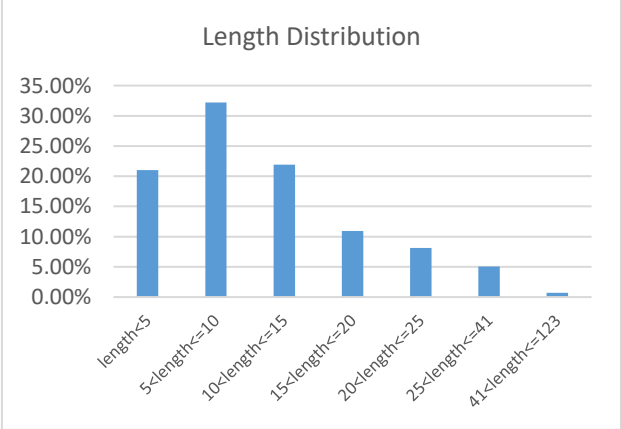

Figure 4. The distribution of sentence length in the corpus.

B. Experimental Setup

The baseline system only used prosodic features (duration, pitch, and energy features) modeled by DNN. We first compared two kinds of phonetic information under DNN framework. DNN_PY indicates that additional phonetic information is 1 of V form and DNN_SA indicates that additional phonetic information is speech attribute or articulatory form, similarly hereinafter. Before training LSTM, parameters, data pre-processing step has been executed to deal with a variable length of input sequences. Zero-padding was performed to pad short sentence based on

the maximum length (123 syllables) in the corpus. As mentioned above, we used a masking layer to address the potential padding values problem in training stage. Figure 4 is the length distribution in the corpus. From Figure 4, we can see that the length of less than 41 account for more than 99%. Thus, the time steps we selected are 41 for training LSTM. The neural networks (both DNN and LSTM) were trained for 100 epochs using SGD with a mini-batch size of 32, 20% dropout [16] in the hidden layers, a cross-entropy objective. We compared different network topologies for the best classification accuracy. Finally, the DNN has three hidden layers each with 128 nodes and the LSTM has three hidden layers each with 512 memory cells. We used KERAS [17] to realize these work.

C. Experimental Results

TABLE II. THE COMPARISON OF DIFFERENT SYSTEM'S PERFORMANCE OF BREAK DETECTION.

System	0	1	2	3	4	Accuracy
DNN	86.7	39.29	56.77	86	95.76	73.8
DNN_PY	85.67	50.02	53.12	88	98.46	75.32
DNN_SA	89.7	41.43	58.14	88	98.17	76.35
LSTM_SA	89.17	51.47	55.63	88	97.69	77.85

From Table 2 we can see that DNN_PY system, which uses conventional prosodic features and additional 1 of V feature vector, improves accuracy from 73.8% to 75.32, compared with DNN baseline, where only uses conventional prosodic features. DNN_SA system improves the boundary detection performance significantly with an absolute increase of 2.55% in contrast with the baseline system. This observation confirms that the additional phonetic information characterized by DNN is helpful for prosodic boundary detection. Moreover, DNN_SA system outperforms DNN_PY, which demonstrates that the form of speech attribute is more effective for boundary detection than the form of 1 of V. The form of 1 of V may result in data sparseness since the appended phonetic vector is a very high-dimensional vector with one-hot encoding. This also could increase model complexity. The form of speech attribute not only reduces the dimension of the appended feature vector but also includes the relationship among words through articulatory movements such as manner of articulation, place of articulation, which is mean that these words with similar articulatory movements are easier to cluster together and have closer boundary index. These may be the reason for the increase by appending speech attribute information.

LSTM with speech attribute information system further improves the performance from 76.35% to 77.85%, compared with the DNN_SA system. This indicates that the context information modeled by LSTM is helpful for prosodic boundary detection. Table 2 also gives the detailed detection performance for each boundary index. We can see that the main improvement by LSTM_SA is for prosody word boundary (boundary index is 1), which is an almost absolute enhancement of 10% compared with DNN_SA. As the input of LSTM is the whole sentence, the tiny difference among closer boundaries with a sentence can be captured by LSTM.

V. CONCLUSIONS

In this paper, we compare two ways to incorporate phonetic information for prosodic boundary detection under DNN framework. Experimental results show that both of them can improve the detection performance compared with only utilizing conventional prosodic features (Pitch, energy and duration features). Using the form of speech attribute to represent phonetic information achieves a better result than using the form of 1 of V. The boundary detection performance can be further enhanced by using LSTM, which demonstrates long context information is helpful for this task. For future work, previous studies indicate that spectrum features are related to F0 contour [18-20]. Thus, spectrum features would incorporate into the LSTM based system.

ACKNOWLEDGMENT

This study was supported by Advanced Innovation Center for Language Resource and Intelligence (KYR17005), the Special Program for Key Basic Research fund of Beijing Language and Culture University (the Fundamental Research Funds for the Central Universities) (16ZDJ03), the Fundamental Research Funds for the Central Universities (18YJ030006).

REFERENCES

[1] C. W. Wightman, M. Ostendorf, "Automatic labeling of prosodic patterns" [J]. Speech and Audio Processing, 1994, vol. 2, no. 4: 469-481.

[2] Q. Chen, Z. H. Ling, C. Y. Yang, L. R. Dai, "Automatic phrase boundary labeling of speech synthesis database using context-dependent HMMs and N-Gram Prior Distributions" [C]. INTERSPEECH 2015.

[3] C. J. Ni, A. Y. Zhang, W. J. Liu, and B. Xu, "Classification of mandarin prosodic break based on hierarchical structure of prosodic break" [J]. Application Research of Computers, 2011, vol. 28, no.7.

[4] C. Y. Yang, L. X. Zhu, Z. H. Ling, and L. R. Dai, "Automatic phrase boundary labeling for a Mandarin TTS corpus using the Viterbi decoding algorithm" [J]. Tsinghua Science and Technology, 2011, vol. 51, no. 9, pp: 1276-1281.

[5] J. Lin, Y. Xie, W. Zhang, et al. "Automatic Mandarin prosody boundary detecting based on tone nucleus features and DNN model"[C]. International Symposium on Chinese Spoken Language Processing. IEEE, 2017.

[6] P. A, Ladefoged, "*Course in Phonetics-Second Edition*" [M]. Heinle, 2013.

[7] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory" [J]. Neural Computation, 1997, 9(8):1735-1780.

[8] J. Lin, W. wang, Y. Gao, et al. "Improving Mandarin Tone Recognition Based on DNN by Combining Acoustic and Articulatory Features Using Extended Recognition Networks" [J]. Journal of Signal Processing Systems, 2018(7):1-11.

[9] Z. J. Wu, M. C. Lin, "Experimental phonetics summary"[M], Beijing: Higher Education Press, pp. 153-191, 1989.

[10] J. L, Zhang, "Fundamentals of Chinese Man-Machine communication". Shanghai: Shanghai Scientific & Technical Publishers, 2010.

[11] Z. Y. Xiong, and M. C. Lin, "prosody expression in the position of speech break" [C]. National Conference on Man-machine Speech Communication (NCMMSC 2006).

[12] H. Kawahara, M. Morise, T. Takahashi, et al. "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and

aperiodicity estimation"[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008:3933-3936.

[13] Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program]. http://www.praat.org/

[14] X. X. Chen, A. J. Li, S. G. Hua, "An application of SAMPA-C for standard Chinese" [C]. Sixth International Conference on Spoken Language Processing, 2000.

[15] A. J. Li, "Chinese prosody and prosodic labeling of spontaneous speech" [C]. Speech Prosody, 2002.

[16] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv: 1207. 0580, 2012

[17] Chollet, F., (2015). Keras. GitHub repository, https://github.com/fchollet/keras.

[18] J. Zhang, "The intrinsic fundamental frequency of vowels and the effect of speech modes on formants" [J]. Acta Acustica, 1987, pp.390-393.

[19] I. Lehiste, G. E. Peterson, "Some basic considerations in the analysis of intonation" [J]. The Journal of the Acoustical Society of America, 1961,33(4), 419–425.

[20] N. Ryant, M.Slaney, M. Liberman, E. Shriberg, J. Yuan, "Highly accurate mandarin tone classification in the absence of pitch information" [C]. In SPEECHPROSODY 7 – 7th International Conference on Speech Prosody, May 20-23, Dublin, Ireland, Proceedings, 2014, pp. 673-677.