An RNN-based Algorithm to Detect Prosodic Phrase for Chinese TTS

Zhiwei Ying and Xiaohua Shi

Intel China Research Center
6th Floor, North Office Tower, #06-01 Beijing Kerry Centre
1, Guanghua Road, ChaoYang District, Beijing 100020, PRC

Email: victor.ying@intel.com, xiaohua.shi@intel.com
http://www.intel.com

Abstract

The goal of the work presented here is to automatically predict the prosodic phrase boundaries from the text for Chinese TTS (text-to-speech) by using the trigram of the POS (part-of-speech) with the info of the breaks between the prior two word-pairs by using a RNN (recurrent neural network). Prosodic phrase boundaries are very important to a Chinese TTS system because it will influence the prosodic model for speech synthesis. In this paper, the algorithm tried to use RNN to find some mapping relationship between the POS sequence and prosodic phrase boundaries, and hoped to improve the naturalness of synthesized speech.

Keyword

Chinese Text-to-speech, Prosodic Phrase, Part-of-speech

1. Introduction

In general, there are three key modules in the Chinese TTS system: the text analysis, the prosodic model and the speech synthesis. The first step of text analysis for Chinese language processing is word segmentation, since there is no space between the words in Chinese text as English. Because of the difficulty of syntactic parsing for Chinese, most of the previous Chinese TTS systems just segment the words in the text analysis procedure. And limited by the intrinsic properties of the Chinese words, the average length of the words after the segmentation is about 1.6 syllables, which means a small pause will be inserted every 1.6 syllables during the speech synthesis if there is no other higher level linguistic info, such as prosodic word, prosodic phrase and intonational phrase, than the word level, thus we can not get very fluent speech. In continuous utterances, native speakers tend to group words into phrases whose boundaries are marked by duration and intonational cues, and many phonological rules are constrained to operate only within such phrases, usually termed prosodic phrases [6]. Prosodic phrase will help the TTS system produce more fluent speech, while the prosodic structure of the sentence will also help improve the intelligibility and naturalness of the speech. So placing phrase boundaries is very important to ensure a naturally sounding TTS system.

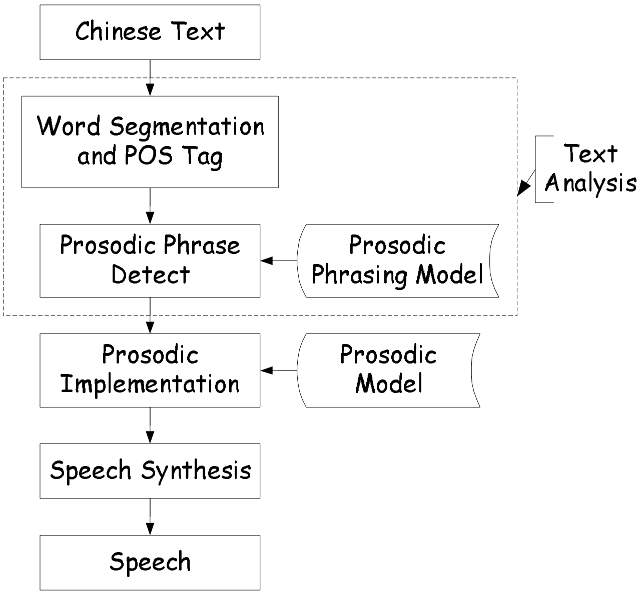

Figure 1 General overview of the Chinese TTS system

If correct prosodic phrases could be detected from text, then we can make out high quality prosodic model, and provide the acoustic parameters, which include pitch, energy, duration and so on, for the speech synthesis.

A lot of methods have been introduced to extract prosodic phrase boundaries from English text, such as statistic model, CART (Classification and Regression Tree), FSA (Finite State Automata), MM (Markov Model), and so on. Some approaches use the language information to parse the text, and then map from the syntactic structure to prosodic structure, some methods make use of POS to extract prosodic phrase from the text. In this paper, a new algorithm based on recurrent neural network, is introduced. The basic idea of this algorithm is to use a model to establish the mapping from Chinese text and its POS tags sequence to prosodic structure, and predict the prosodic phrase boundaries to help build proper prosodic model for speech synthesis. *Figure 2* shows the block diagram of prosodic phrase detect algorithm introduced in this paper.

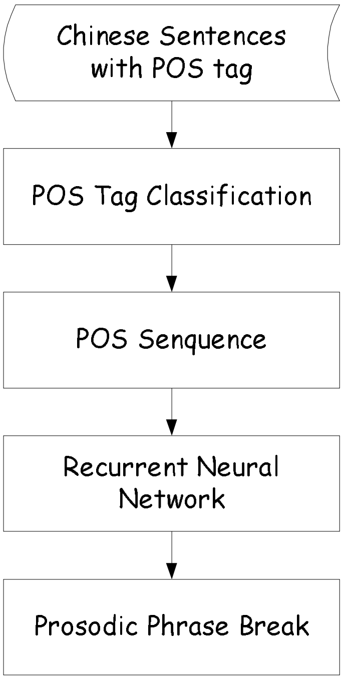

Figure 2 Block diagram of prosodic phrase detect algorithm

2. Preparation of Corpus

There are total 2000 sentences (17897 words) in the corpus. The corpus is designed for general purpose, not limited in any specific domains. 90% of the corpus is used to train the RNN, and the other 10% used to test.

The corpus is labeled well with words, POS tags and prosodic phrase boundaries.

2.1 Word segmentation: The words are segmented by the in-house word segment system automatically. The word segment system includes more than 130k Chinese words, uses Maximal Matching method and linguistic rules to segment words, and the word segment accuracy is about 97%.

2.2 POS tagging: There is a lexical analysis procedure in the text analysis part. Markov Model is implemented into this procedure to tag the POS of the words automatically. In fact, this part of work is combined with the word segment part closely, word segment is the premises of the lexical analysis, but lexical analysis can also help remove some ambiguity for word segmentation. There are total 26 different tags in the POS tag set, it is neither necessary nor practical to use all of them to train the model. If all of them are put into use, there will be 26*26*26 = 17576 possible trigrams, at the same time there are only 17897 words in the corpus. In fact if the corpus were big enough, we could just use the words themselves to train the model. So we cluster them into 11 classes, which include *adjective, adverb, noun, verb, number, quantifier, preposition, conjunction, idiom, punctuation,* and *others*. After clustering there are total 11*11*11 = 1331 kind of trigrams, and the experiments proved that the classification is correct and effective.

2.3 Prosodic phrase tagging: The prosodic phrase boundaries are labeled half automatically. Although it is generally agreed that prosodic phrases have some relationship with syntactic phrases, the two are not isomorphic. If all of the prosodic phrase boundaries are labeled manually, there would be too much syntactic phrases in the training corpus. So we use some simple rules, which include silence (pause), phrase lengthening and pitch declination, to extract prosodic phrases from speech for reference, and then check them manually. *Figure 3* shows the block diagram of this procedure.

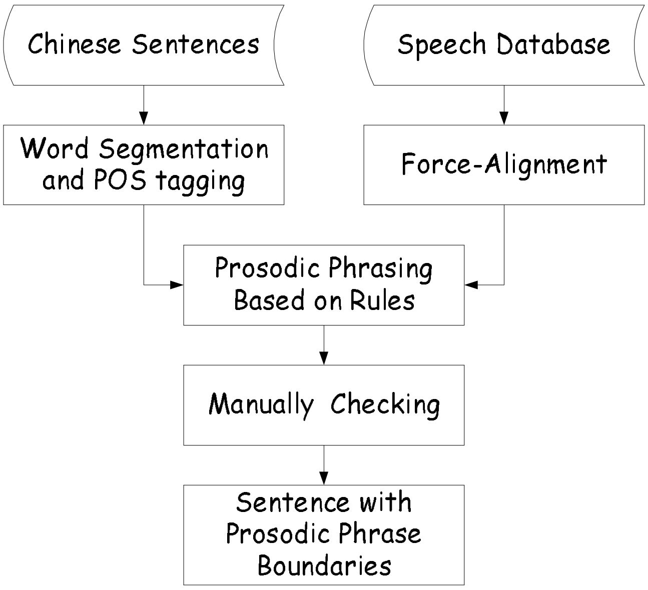
Chinese Sentences
Speech Database
Word Segmentation and POS tagging
Force-Alignment
Prosodic Phrasing Based on Rules
Manually Checking
Sentence with Prosodic Phrase Boundaries

Figure 3 Block diagram of prosodic phrasing half automatically

So the sentences in the corpus look like follows:

W_1/T_1 W_2/T_2 | W_3/T_3 W_4/T_4 W_5/T_5 ....

W_n, T_n and | represent Chinese word, POS tag of the word and the prosodic phrase break respectively.

3. Implementation of RNN

This algorithm makes use of the POS trigram probabilities. The trigram is defined as below:
If there are three words in a sentence, looks like
…W1 W2 W3…
Accordingly its POS tag sequence is
… T1 T2 T3…
What we want to know is whether there is a break between W2 and W3. Although we can just calculate the probability of this trigram and place a break when the probability is over a certain threshold T, it is not good since you really need to take into account the fact whether there is a prosodic phrase break in the neighbor word-pairs. For example if there has been a break between W1 and W2, the probability of a break between W2 and W3 is very small. So we expand the trigram T1T2T3 into a quingram B1T1B2T2T3, B1 represents the status of the break before W1, B2 the status of the break between W2 and W3. Both B1 and B2 have two states: 1 or 0, that means there is a break or not.
For convenient, we add a punctuation tag at the front of sequence of POS tags, and set boundary both before and after this tag. So we don't need to design special algorithm for the possible break between the first and second words in a sentence since we use the trigram to predict whether there is a break between second and third tags.
The algorithm is realized by implementing a recurrent neural network. *Figure 4* shows the block diagram of the RNN. After well training, the RNN acts as a parser to generate proper prosodic phrase boundaries from Chinese text tagged with POS.
At first, initializing the RNN system, set the B1 = B2 = 1, set the T1 and T2 the first two POS tags in the sequences. The input layer will remain only one parameter, the third POS tag in the trigram.
The inputted POS tags should be orthogonalized into 11-dimension vectors, since there is no direct linear relationship among the tags themselves. And the break symbols, which represent the state of the break, will be orthogonalized into 2-dimension vectors too.
The output is the state of the break between T2 and T3. After every step, the T3 (Input), Output, T2 and B2 will be feedback to T2, B2, T1 and B1.

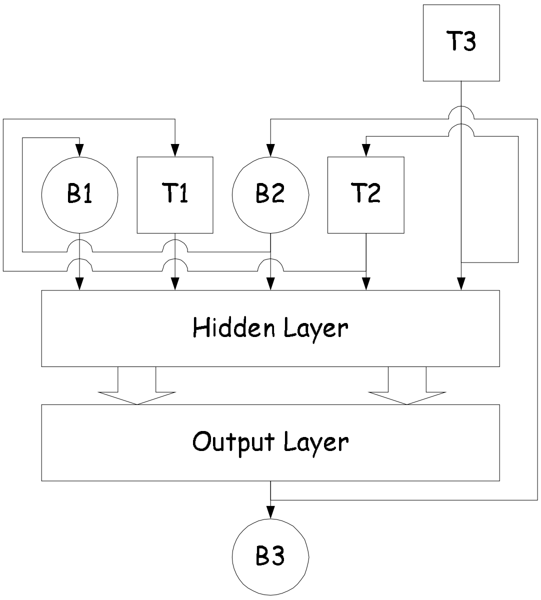

Figure 4 Block diagram of the RNN prosodic phrase detect algorithm

4. Results

According to Sanders [2], there will be two scores to evaluate the results. The first score is the number of correct word pairs divided by the total number of word pairs. This is a simple percentage score of the accuracy the algorithm got. This score in the algorithm is 84.7%. And the second score is: if the first score is S and the proportion of non-breaks to the number of word-pairs is B then the adjust score is $R = \frac{S-B}{1-B}$. In the corpus, the B = 0.55, S = 0.847. We got 0.66 by this method. We also did the experiments using the methods presented in [2], the scores are 81% and 0.57 respectively.

Method	Break	Non-Break	Total	R
1	80.3%	81.3%	81%	0.57
2	85.9%	83.2%	84.7%	0.66

Figure 5 Results of two methods

In *Figure 5* we compared the results from different methods. The method 1 used the algorithm introduced in [2]; Method 2 is the algorithm introduced in this paper. The column 'Break' in the *Figure 5* show the results of the accuracy of correct break detect, column 'Non-break' shows the accuracy of non-break detect. It is obvious that after applying our algorithm, the accuracy of prosodic phrase detect is improved.

5. Discussion

We have described the algorithm for detecting prosodic phrase boundaries from the text for Chinese TTS. This algorithm uses RNN method to take advantages of POS. Moreover, the use of the info of the prior two breaks makes the algorithm to get better performance. This algorithm combines the trigram model and the neighbor breaks information. The results seem very interesting. According to [2], while using the distance information, the scores are slight lower than those scores without using the distance information. But In our experiments, the fact is just contrary. Our results, which are generated with the distance information, are higher than those without considering the distance information. I think the reason is about the distance probability used in [2]. As the distance probability is defined as the probability that there is a break after exactly n words given that there hasn't been a break before this word, and is calculated by summing all the phrase-length probabilities from 1 to n. And this kind of distance probability doesn't reflect the distribution of the prosodic phrase boundaries in the sentence. For example if the distance is long enough, there must be a break, whereas next word-pair is more likely to have break among it. And if the first break is marked wrong, it will induce a series of errors.

6. Conclusion

The main advantages of the approach described in this paper are:

This algorithm is really simple and gets high accuracy in finding phrase breaks. While comparing the result with those in English, the results are very similar. That means, I think, there is some relationship between the POS tags and the prosodic phrase break regardless the languages' intrinsic property. So this method can be easily adapted to other languages. Further more this method is easy to train; though the experiments done here are for general purpose, it is still fit to specific domain application.

But in fact, the limitation of this method is also obvious:

1. This algorithm's performance is largely depended on the POS tagging accuracy, which is based on word segmentation and lexical analysis.
2. There is an ultimate limit for such a kind of method to predict the prosodic phrase break by using POS only [10].
3. For a given text, there may be many possible phrase break placings, which are all deemed acceptable, but such a kind of method cannot give the choices.

This algorithm's results will be used to help establish the prosodic model for the Chinese TTS. Those features, such as pause after the phrase boundaries and the final lengthening, will be considered while building the prosodic model. Definitely the prosodic model will help improve the fluency of the synthesized speech and get higher degree of naturalness.

Reference:

1. Sin-Horng Chen, Shaw-Hwa Hwang, Yih-Ru Wang. *An RNN-based Prosodic Information Synthesizer for Mandarin Text-to-speech,* IEEE Trans. Speech and Audio Processing, Vol. 6, No. 3, May 1998

2. Eric Sanders and Paul Taylor. *Using Statistic Models to Predict Phrase Boundaries for Speech Synthesis*. Proc. Eurospeech '95, Madrid

3. Chilin Shih and Richard Sproat, *Issues in Text-to-speech Conversion for Mandarin*, www.bell-labs.com

4. Simon Arnfield, *Word Class Driven Synthesis of Prosodic Annotations*, International Conference of Spoken Language, 1996

5. Hiroshi Shimodaira and Mitsuru Nakai, *Prosodic Phrase segmentation by Pitch Pattern Clustering*. International Conference of Acoustics, Speech, and Signal Processing, 1994

6. C.W. Wightman and M. Ostendorf, *Automatic Recognition of Prosodic Phrases*, International Conference of Acoustics, Speech, and Signal Processing, 1991

7. Joan Bachenko and Eileen Fitzpatrick, *Prosodic Phrasing for Speech Synthesis of Written Telecommunications by the Deaf*, Global Telecommunications Conference, 1991

8. Colin W. Wightman and Mari Ostendorf, *Automatic Labeling of Prosodic Patterns*, IEEE Transactions on Speech and Audio Processing, Vol 2, No. 4. October 1994

9. Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee, *A Chinese Text-to-speech System Based on Part-of-speech Analysis, Prosodic Modeling and Non-uniform Units,* International Conference of Acoustics, Speech, and Signal Processing, 1997

10. Alan W. Black and Paul A. Taylor, *Assigning Phrase Breaks from Part-of-Speech Sequences*, Eurospeech97, pp. 995-998