AUTOMATIC DETECTION OF PROSODY PHRASE BOUNDARIES FOR TEXT-TO-SPEECH SYSTEM

Xin Lv Tie -jun Zhao Zhan -yi Liu Mu -yun Yang
School of Computer Science and Technology
Harbin Institute of Technology
Box 321, No.92, XiDaZhi Jie, NanGang DIST
Harbin, Heilongji ang, 150001, P.R.China
{lxfancy, tjzhao, zhanyi,ymy}@mtlab.hit.edu.cn

Abstract

Automatic acquisition of the prosodic phrase boundary detecting rules from the text and speech corpora has always been a difficulty for TTS systems. We collected over 5,000 sentences as the corpus, introduced a method based on the transform-based error-driven learning to get the rules for detecting prosodic phrase boundaries, and then used trees to organize the rules in the TTS system. For using the transformation -based error-driven learning, we designed a set of templates especially. Using 1,000 sentences to get rules for the TTS system can reach 92% accuracy in close -test and 73% accuracy in open-test.

1 Introduction

Building a Chinese text-to-speech (TTS) system involves three major steps. In the first, text is converted to syllables, the symbols representing in a rough way the categories of Chinese Mandarin speech sounds . A second stage involves questions of prosody, i.e., the intonation and pausing; and the third stage is the backend, the component responsible for the production of the sounds from the specifications provided by the first two components. Today for TTS systems, the research in improving the naturalness focuses on two aspects: first, trying to get the prosody characters from the text input by natural language processing; second, on the prosody rules trying to synthesize good output speech by using some prosody modification algorithms. For the first one, there are always many difficulties in it. Here we focus on the detection of prosodic phrase boundaries which effect directly the naturalness of Chinese Mandarin speech output.

It is normal for a human speaker to pause at various places in his or her speech —to think, to find a word, to emphasize. Human listeners expect pauses when they listen to speech, and a functional TTS system must give its listeners those expected pauses. Without them, the task of listening to extended synthetic speech becomes a burdensome task, and the listener's attention will rebel. In research of the Chinese sentence structure, a sentence is always separated into several chunks. There is the same result in researches of continuous speech analysis: breaks appear not only between the sentences but also inside a sentence. That is to say that spoken Chinese always has a certain rhythm. To describe it, prosody has been introduced. Prosody concerns the supra-segmental aspects of spoken language, and has to be processed with phrasing, loudness, duration and speech intonation [1]. Prosodic phrase was firstly used by Sheon(1995) to name phrases between tentative pauses when he analyzed the acoustic characteristics of the vicinity of the prosodic phrase boundary. Whether the prosodic phrase boundary is properly detected will affect directly the naturalness and correctness of TTS. Here we mainly focus on where we should insert a break and don't care the time duration of each break. This problem will remain for our later research.

To predict prosodic parameters, many researchers have used statistical modeling techniques such as neural networks (Haykin, 1994), hidden Markov model (HMM) (Huang et al., 1990) and CART (classification and regression trees) (Breiman et al., 1984) and achieved limited success in prosodic phrasing (Fujio et al., 1995;

Wang and Hirschberg, 1992), in segmental duration prediction (Riley, 1992), in prosodic label prediction (Ross and Ostendorf, 1996) and in fundamental frequency generation (Ljolje and Fallside, 1986; Traber, 1992) [2]. But none of these methods are created for Chinese TTS systems.

Corpus-based techniques are wildly used in speech processing and they often have good performance while ignoring the true complexities of language, based on the fact that complex linguistic phenomena can often be indirectly observed through simple superficies. Brill (1992) put forward an approach named transformation-based error-driven learning to make progress in corpus-based natural language processing. This algorithm has been applied to solve many natural language problems, including part-of-speech tagging, prepositional phrase attachment disambiguation, syntactic parsing, building pronunciation networks for speech recognition. Here we want to use this algorithm to solve the problem on automatic detecting prosodic phrase boundaries.

The paper unfolds as follows. Section 2 briefly introduces how we construct the text and speech corpora for our study. In Section 3 we separately discuss how we specify the parameters of the error-driven learning and how we use the transformation-based error-driven learning to detect prosodic phrase boundaries automatically. In Section 4 we introduce a method to organize the rules in the TTS system which can speed up the handling of producing prosodic phrases in the system. Then the experiment results and conclusion are given in Section 5.

2 Building The Experiment Corpus

Since the goal of TTS systems is to synthesize speech given an unlimited text, the corpus for training prosody generation models should cover the variability of the language. We build a text and speech corpus which have 5,725 Chinese Mandarin sentences and was read by a female speaker. The collected sentences include simple, complex, declarative, interrogative and exclamatory sentences.

In order to get the right prosodic phrase boundaries, we let some skilled persons annotate breaks in the sentence by watching the speech waves and listening the speech carefully. For example, in Figure 1 the wave form of the sentence “五名死者包括 名妇女和两名儿童，” which means “Among five decedents there are a woman and two children” is shown and the breaks annotated manually are under it.

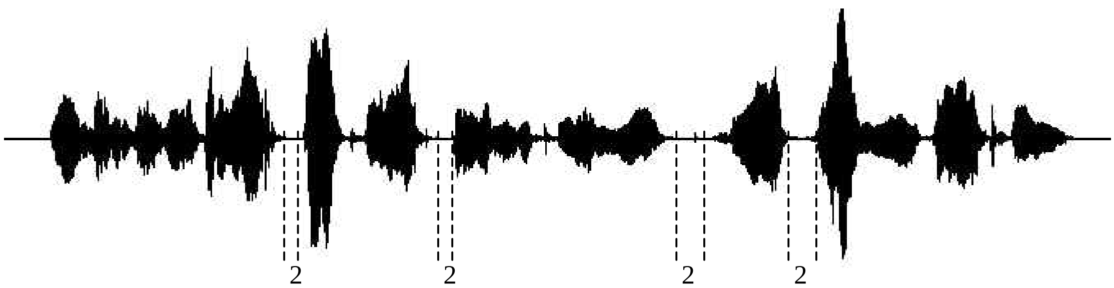


Figure 1: Wave form and break annotation of a sentence

Before detecting the prosodic phrase boundaries, the TTS system has to do morphological analysis firstly. The result of the morphological analysis is very important to the rule learning for detecting the prosodic phrase, so we must use a method that has very high accuracy in the word segmentation and part-of-speech tagging. We use the strategy of multi-step processing [3]: disambiguation of pseudo-ambiguities, full segmentation of sentence, determinate segmentation for some words, processing of numeral string, processing for reduplication of words, statistical identification for unknown words and final correction for segmentation ambiguities with part-of-speech which is integrated in the tagger. There are 52 part-of-speech tags used in it (Table 1). By these processing the segmentation and part-of-speech tagging can get the accuracy above 98%. Here is an example: we input a Chinese sentence “大学生运动会将在北京举行，”, and then we can get the segmentation and part -of-speech result through the process above, that is "大学生/nc 运动会/ng 将/vz 在/p 北京/nd 举行/vg ，/wj". Then what we have to do is to determine which words should be combined into a prosodic phrase, so this segmentation and part-of-speech result is fairly important.

3 Applying The Transformation -based Error-driven Learning On Prosodic Phrase Boundary Detecting

3.1 The Selection Of The Template Parameters

Chinese sentence is made up of words, and a speaker usually inserts breaks in a sentence according to the word and the sentence structure. So firstly the system should do word segmentation and tag the part of speech of each word. Though the word segmentation in TTS is still based on syntax dictionary, the prosodic phrase is usually not the same as the syntax phrase. It may be a noun phrase, a verb phrase or a preposition phrase and can also be made up of a syntax phrase together with its precursor or subsequence or both.

Since the part of speech can represent the sentence structure, surely it can be used to detect the prosodic phrase boundary. On the other hand, on the experiments done by some linguists we get that the average numbers of syllables between two breaks is 3.6 [4]. This shows the prosodic phrase also relates to the syllable number, so we can use the number as another important factor.

In conclusion, we decide to use the part of speech and the syllable number as the template parameters in learning.

3.2 Specifying The Start State Of Learning

During the process of the transformation -based error-driven learning, unannotated text is presented and pre-specified initial state knowledge is used to annotate the text. This initial state can be at any level of sophistication, ranging from an annotator that assigns random structure to a mature hand -crafted annotator [5]. Here the initial state is not difficult to obtain but contains information derived automatically from a corpus. In order to get a good start state to shorten the time of learning, we count the syllable number and part of speech of each word in the prosodic phrase from the hand-annotated corpus and get the most probable transform rule to produce the annotated text as the start state.

3.3 Designing Of The Training Template

A set of transformation templates specifying the types of transformations which can be applied to the corpus must be pre-specified. Unlike other learning approaches, the transformation templates are very simple, do not contain any deep linguistic knowledge and the number of transformation templates is also small [5]. However, that is not to say we can design the templates casually. If the template set was not designed rationally, the accuracy of detecting the prosodic phrases by using the rules we got would be very low or no right rules we could get from the training. So we generate templates after serious consideration.

Since we have decided to use the part of speech tags and the number of syllables as the template parameters, what we should do next is to decide the type of the templates and how to sort these templates.

Though the prosodic phrase boundary is not completely detected according to the syntax information, the first rule bearing in a speaker's mind is the logic of a sentence while speaking, and usually this logic is mostly shown in the syntax structure especially part-of-speech tag. So we specify the first template is:

if 0:POS=X->PAUSE=*

"0" indicates the current word, "POS=X" indicates the word's part of speech tag is X and "PAUSE=*" indicates whether the word is the end boundary of a prosodic phrase. If "*" is "2", the word is the end boundary of a prosodic phrase; If "*" is "1", the word is still as a word; If "*" is "0", the word is attached in a prosodic phrase and not as the boundary of a phrase. For the number of syllables is the second parameter of the templates, we specify the second template with adding it into the first one:

if 0:POS=X&0:LENGTH=Y->PAUSE=*

“LENGTH=Y” indicates the number of syllables of the word is Y. Certainly we can ’t just consider the current word and ignore the contexts, so the information of the previous word and following word have to be added. This is similar with a tri-gram model. Then all the templates are:

(Class 1) if 0:POS=X->PAUSE=*

(Class 2) if 0:POS=X&0:LENGTH=Y->PAUSE=*

(Class 3) if -1:POS=X&0:POS=Y->PAUSE=*, if 0:POS=X&1:POS=Y->PAUSE=*

(Class 4) if 0:POS=X&-1:POS=Y&-1:LENGTH=Z->PAUSE=*
if 0:POS=X&1:POS=Y&1:LENGTH=Z->PAUSE=*
if 0:POS=X&0:LENGTH=Y&-1:POS=Z->PAUSE=*
if 0:POS=X&0:LENGTH=Y&1:POS=Z->PAUSE=*

(Class 5) if 0:LENGTH=X&0:POS=Y&-1:POS=Z&-1:LENGTH=U->PAUSE=*
if 0:LENGTH=X&0:POS=Y&1:POS=Z&1:LENGTH=U->PAUSE=*

(Class 6) if -1:POS=X&1:POS=Y&0:POS=Z->PAUSE=*

(Class 7) if -1:POS=X&1:POS=Y&0:LENGTH=Z&0:POS=U->PAUSE=*
if -1:POS=X&1:POS=Y&-1:LENGTH=Z&0:POS=U->PAUSE=*
if -1:POS=X&1:POS=Y&1:LENGTH=Z&0:POS=U->PAUSE=*

(Class 8) if -1:POS=X&-1:LENGTH=Y&1:POS=Z&0:LENGTH=U&0:POS=V->PAUSE=*
if -1:POS=X&1:POS=Y&0:LENGTH=Z&0:POS=U&1:LENGTH=V->PAUSE=*
if -1:POS=X&1:POS=Y&-1:LENGTH=Z&0:POS=U&1:LENGTH=V->PAUSE=*

(Class 9)if-1:POS=X&1:POS=Y&0:LENGTH=Z&0:POS=U&1:LENGTH=V&-1:LENGTH=W->PAUSE=*

“-1” indicates the previous word and “1” indicates the following word. We can see that the cover range of these templates is descending with adding restriction gradually. The produced r ules should appropriately follow this principle. Since these templates have different cover ranges, we must classify them into several classes according to the cover range so that using each class of templates we can get the least redundant new rules to correct the errors produced by the original rule set.

3.4 Process Of The Transformation-based Error-driven Learning

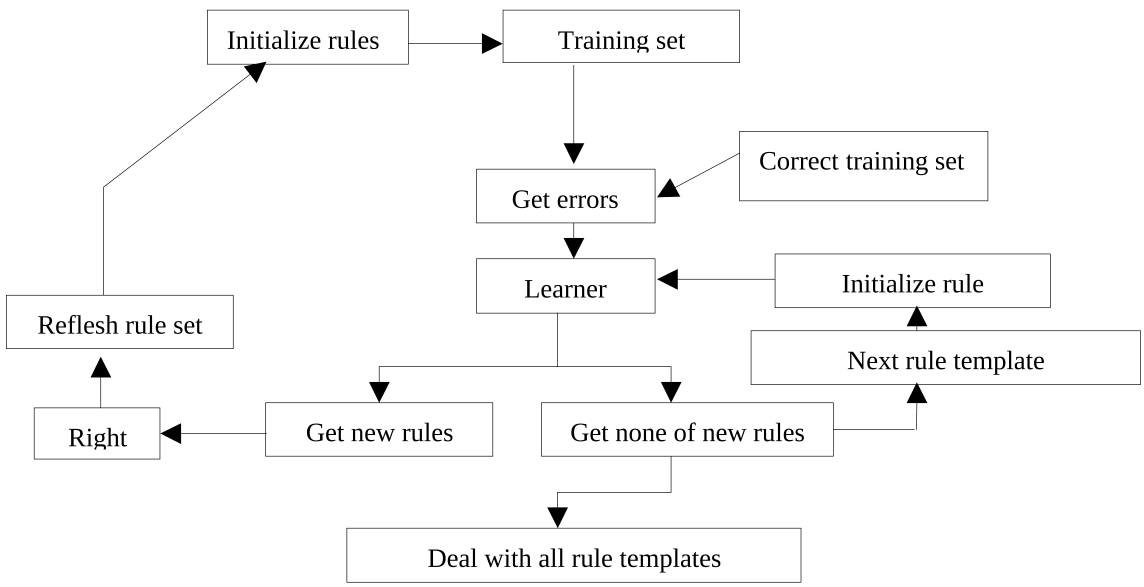


Figure 2: The process of automatic acquisition of the prosodic phrase boundary detecting rules

After we get the resource of the t ransformation-based error -driven learning, we can build a system to do automatic acquisition of the prosodic phrase boundary detecting rules (Figure 2).

We use the rules in the set to produce the prosodic phrases and then compare the boundaries of each phr ase with the manual annotated boundaries. If a boundary is not the same as the corresponding one in the annotated corpus, we regard there is an error. Using one class of templates we can use the learning algorithm to produce new rules and then put the new rules into the rule set. These new rules must be able to modify the errors with the number above a certain threshold. Certainly if the new rules conflict the rules in the set, the new rules would not be put into the set. And if we can ’t get any new rules t hat fit the requirements, the learner will use the next class of templates to produce new rules. Do this process until the system can ’t get any rules. The threshold is directly related to the number of the rules we would get and the accuracy of the prosodic phrase boundary detecting, so we have done several experiments to choose a proper threshold. We choose 1,000 sentences which have been manual annotated as the training corpus and the result of these experiments with different thresholds is listed in Table 1.

Threshold (fraction of the error numbers)	Accuracy	Number of the gotten rules
1/3	91%	8858
9/24	91.5%	8357
5/12	92%	7274
11/24	91.3%	6524
1/2	91%	2091
2/3	87%	2091

Table 1: The accuracy in close-test and the numbers of the gotten rules on different thresholds

4 Organizing Of The Rules In TTS

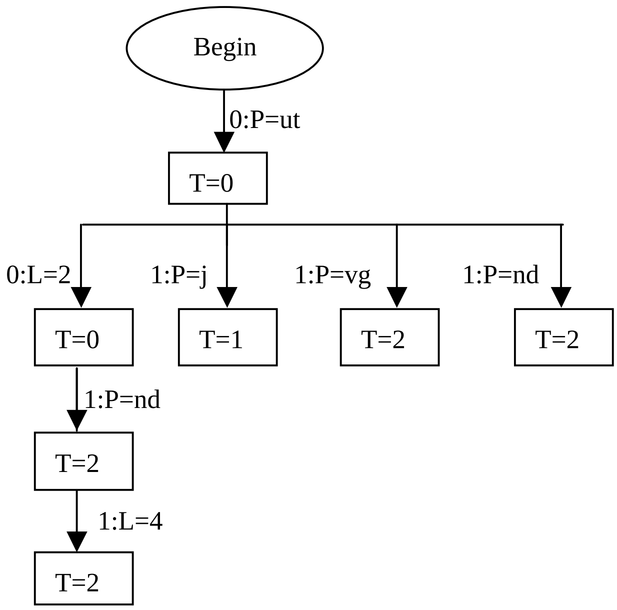


Figure 3: Tree of the prosodic phrase boundary detecting rules

Note: In the figure above, “P” is abbreviation of “part of speech”, “L” is “number of syllable”(LENGTH) for short and “T” is “label of boundary”(PAUSE) for short in a rule.

From the last section we can see that the number of gotten rules is very large, so if we use transformation rule list and let the TTS system searched the proper rule orderly in the list to detect the prosodic phrase boundary, the running time would be very long and it would become a heavy burden to the system. In order to lighten the burden, we decide to introduce the tree to organize the rules. This type of tree is similar to the decision tree. The nodes in th e same level of a tree are having the same part of speech tags of the word or the same syllable numbers. In the rule set the order of the rule indicated by a node is always behind the rule indicated by its left brother node. For example, there are some rules:

if 0:POS=ut->PAUSE=0

if 0:POS=ut&0:LENGTH=2->PAUSE=0

if 0:POS=ut&0:LENGTH=2&1:POS=nd->PAUSE=2

if 0:LENGTH=2&0:POS=ut&1:POS=nd&1:LENGTH=4->PAUSE=2

if 0:POS=ut&1:POS=j->PAUSE=1

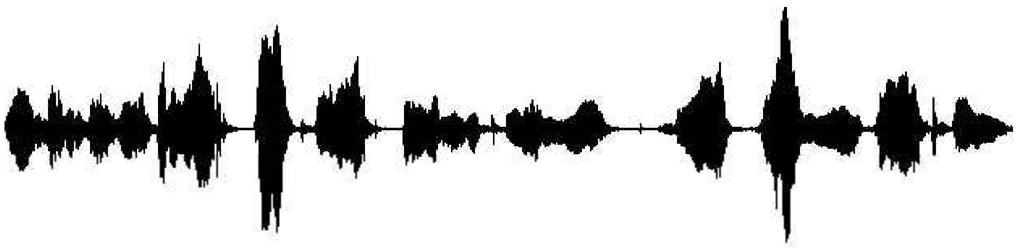
if 0:POS=ut&1:POS=vg->PAUSE=2

if 0:POS=ut&1:POS=nd->PAUSE=2

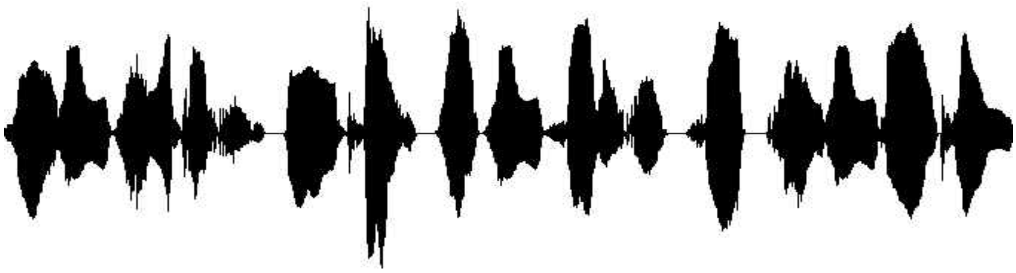
The tree is used to organize these rules is shown in Figure 3. All these rules we got are organized like this tree, we store them and then use depth -first searching to get the proper rule for producing the prosodic phrases. It has been proved that the speed of producing prosodic phrases in TTS is improved.

5 Experiment Results And Discussion

We compared the synthesized speech with the original speech. The typical synthesized speech signal and its corresponding original uttered by the female announcer are shown in Figure 4. The given sentence was "五名死者包括 名妇女和两名儿童，", which means "Among five decedents there are a woman and two children". The utterance consisted of five prosodic phrases, "五名死者", "包括", " 名妇女", "和",and "两名儿童".



(a) Original speech signal uttered by the female announcer



(b) Synthesized speec h signal produced by the TTS system

Figure 4: Original and synthesized speech signals

Number of test sentences	Accuracy in close-test	Accuracy in open-test	Number of gotten rules
1000	92%	73%	7274
5000	87.5%	77.1%	20400

Table 2: Experiment results in different scales of corpus

Based on the best threshold of the learner we have chosen in the transformation -based error -driven learning, the experiment results in several scales of corpus are shown in Table 2 (Note: the threshold we chosen is 5/12 of the err or numbers.) 5000 sentences in the corpora are used for learning and the rest (725 sentences) is used as open-test corpus.

In this paper, we have described a method to do automatic acquisition of the prosodic phrase boundary detecting rules based on the transformation-based error-driven learning. To produce prosodic phrases properly, we constructed a text corpus from various genres, and built a speech corpus of a female speaker. With the help of automatic tagging and manual verifying, we annotated the text and speech corpora (5,725 sentences) including prosodic phrase boundary locations, segmental boundaries and part of speech. Based on the annotated text and speech corpora, parameters proposed for training templates, the form of the templates, we do some experiments to get the proper threshold in the learner and introduced a method to organize the rules into a tree. Finally we do an experiment in large corpus to prove the performance of the transformation-based error-driven learning in detecting prosodic phrase boundaries.

Acknowledgements

We would like to thank all the corpus builders for their hard works in annotating the text and speech corpora. In addition we would like to thank Dr.Chu Min in MSRCN for her valuable suggestions.

References

[1] Terken and Collier. *The generation of prosodic structure and intonation in speech synthesis*. Speech Coding And Synthesis. Elsevier Science, Amsterdam, 1995.

[2] Sangho Lee and Yung-Hwan Oh. *Tree-based Modeling of Prosodic Phrasing and Segmental Duration for Korean TTS Systems*. Speech Communication, Switzerland, 1999.

[3] Tie-jun Zhao, Ya-juan Lv, Hao Yu, Mu-yun Yang and Fang Liu. *Increasing Accuracy of Chinese Segmentation with Strategy of Multi-step Processing*. Journal of Chinese Information Processing, Beijing, P.R.China, 2001.

[4] Hong Ying and Lian-hong Cai. *Research on the Segmentation of the Prosodic Phrase Based on Driven by the Structural Auxiliary Word*. Journal of Chinese Information Processing, Beijing, P.R.China, 1999.

[5] Eric Brill. *A Corpus-Based Approach to Language Learning*. Elsevier Science, 1993.

[6] Eric Brill. *Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging*. Elsevier Science, 1995.