A Two-stage Prosodic Structure Generation Strategy for Mandarin Text-to-speech Systems

DONG Yuan[1] ZHOU Tao[1] DONG Cheng-Yu[2] WANG Hai-La[2]

Abstract Prosodic structure generation is the key component in improving the intelligibility and naturalness of synthetic speech for a text-to-speech (TTS) system. This paper investigates the problem of automatic segmentation of prosodic word and prosodic phrase, which are two fundamental layers in the hierarchical prosodic structure of Mandarin, and presents a two-stage prosodic structure generation strategy. Conditional random fields (CRF) models are built for both prosodic word and prosodic phrase prediction at the front end with different feature selections. Besides, a transformation-based error-driven learning (TBL) modification module is introduced in the back end to amend the initial prediction. Experiment results show that the approach combining CRF and TBL achieves an F-score of 94.66 %.

Key words Text-to-speech (TTS), prosodic structure generation, conditional random fields (CRF), transformation-based error-driven learning (TBL)

DOI 10.3724/SP.J.1004.2010.01569

Text-to-speech is an important technique to generate the artificial speech from text dependent application[1]. This technique has been widely applied in many fields, such as telecommunication services, embedded mobile application and entertainment.

When people talk, they rarely speak out a whole sentence without a break. Instead, an utterance is divided into smaller units with perceivable boundaries between them. These phonetic spurts or chunks of speech, which are commonly known as prosodic units, signal the internal structure of the message and serve important functions in helping people to clearly express their ideas as well as their feelings. These units are different in their categories, levels, and boundary strength. Smaller- and lower-level units are contained in larger- and higher-level units to form a prosodic hierarchy. In Mandarin, this hierarchical structure is often simplified to three layers (from bottom to up): prosodic word, prosodic phrase, and intonation phrase[2].

In a current text-to-speech (TTS) system, the input text is firstly processed by a text analysis model, whose output is a string of syntactic words, each with a part-of-speech (POS) tagging. Then, its prosodic structure is expected to be constructed without linguistic information in order to enhance the naturalness and understandability of the synthetic speech[3]. However, as the grammatical structure does not necessarily correspond to its prosodic counterpart, misjudgments often occur in assigning proper prosodic boundary, which has become the major impediment for TTS systems to achieve human-like performance.

That is why the prediction of prosodic structure is an important step for a TTS system. In fact, fluent spoken language is never produced in a smooth and unvarying stream. Furthermore, variation in phrasing can change the meaning by the utterances of a given sentence.

Researches have shown that the relative size and location of prosodic boundaries provide an important cue for resolving syntactic ambiguity. More and more attention has been paid to addressing the problem of automatic prosodic boundary prediction.

Researches have discovered a hierarchical prosodic structure for Chinese prosody, which constitutes the rhythm of Chinese speech. Researchers also found that the main prosodic elements in Chinese speech are prosodic word, prosodic phrase, and intonation phrase.

In the earlier time, rule-based methods were usually adopted[4]. It mainly started from Gee and Grosjean′s work on performance structures, and has had various extensions over the years[5]. For Chinese, similar investigation has also been carried out into the formation of prosodic constitutes, such as those reported by Cao and Wang[6−7]. The central idea of all these works is to find out some explicit rules that are able to recreate the prosodic structure of a sentence from syntax through a large number of experiments and empirical observations. This method is easily explicable and understandable, but also poses strict demand for the system developer to summarize these rules. Moreover, it is hard to update and improve, and the set of rules is usually constrained to one branch of language, which hinders its general application.

With the availability of increasing prosodically annotated corpora and the rapid development of machine learning, stochastic-based approach has been more and more popular in prosodic boundary prediction[8]. As in most cases, it is assumed that syntactic word is the smallest unit (i.e., leaf node) in a prosodic hierarchy tree, the task of building prosodic structure could be reduced to decide the type for each syntactic word boundary, which is actually a classification problem. Thus, many different statistical methods used for classification have been tried, such as classification and regression tree (CART) used by Wang et al.[9], and hidden Markov model proposed by Paul et al.[10]. Researchers have also begun to adopt this approach in Chinese during recent years. Besides those mentioned above, attempt was made to predict prosody phrase break based on the maximum entropy model[11−12]. Generally, these methods relate each boundary site with some features (e.g., length and POS of adjacent words). By extracting and absorbing these features from a large collection of annotated sentences, a statistical model is trained and then applied to unlabeled texts. For each potential boundary site in the text, a probability is estimated for each possible outcome, and the one with the largest likelihood is determined as the correct type.

This paper proposes to predict prosodic boundaries by using a two-stage strategy: conditional random fields (CRF) models and a transformation-based error-driven le-

Manuscript received March 25, 2009; accepted July 1, 2010
Supported by National Natural Science Foundation of China (90920001), the Key Project of the Ministry of Education of China (108012), and Joint-research Project between France Telecom R&D Beijing and Beijing University of Posts and Telecommunications (SEV01100474)
1. Beijing University of Posts and Telecommunications, Beijing 100876, P. R. China 2. France Telecom R&D (Beijing), Beijing 100190, P. R. China

arning (TBL) modification module[13−14]. The strategy is applied to both prosodic words and prosodic phrase boundary labeling.

The rest of the paper is organized as follows. Section 1 introduces the prosodic structure briefly. The CRF algorithm and corresponding feature selection are presented in Section 2. Section 3 describes the TBL modification module[15]. Experiments and performance results are given in Section 4. Section 5 draws the conclusion.

1 Prosodic structure

Generally, the prosodic hierarchy includes three tiers, which are prosodic word (PW), prosodic phrase (PP), and intonation phrase. Prosodic word is the most elementary rhythmic unit among these three tiers. In real speech, prosodic word should be uttered continuously and closely without breaks. Prosodic phrase and intonation phrase are the combination of several PWs. Thus, the segmentation of prosodic word will affect the segmentation of prosodic phrase and intonation phrase, and will also play an important role in increasing the naturalness of synthesized speech. The segmentation module is able to make boundaries of lexicon words (LW). However, the boundaries of lexicon words are quite different from prosodic structure.

Prosodic word is the basic element of prosodic boundary, which has three types of relationship with lexicon word: 1) A prosodic word is a lexicon word; 2) A prosodic word is combination of several short lexicon words; 3) A prosodic word is part of a long lexicon word. Most of the cases belong to the first two types. The first step of our work is to make a proper prosodic word boundary from the output of segmentation and POS module.

Prosodic phrase can be treated as the combination of several prosodic words. Prosodic phrase usually contains $2\sim3$ prosodic words, even one single prosodic word, with a length of $3\sim9$ characters. Prosodic phrase plays the most important role in prosodic boundary prediction, as it contains most information of human pronunciation attribute. So, the F-score of prosodic phrase takes the majority work in determining the naturalness of voice synthesis in TTS system. In this paper, prosodic phrase boundaries are generated based on the output of the prosodic word prediction.

Intonation phrase can mostly be indicated by the punctuations, which can be easily detected. Punctuations can be regularly marked as a sign of prosodic boundaries. This paper does not build a single CRF model for intonation phrase, but defines it as an intonation phrase by the punctuations.

The framework of the two-stage strategy is shown in Fig. 1.

2 Conditional random fields model

By using a probability map model, CRF expresses the ability of a long-range dependence and overlapping. CRF has the advantage to solve the problem of bias and normalize the overall situation in order to find the optimal solution. Two distinct CRF models are built by incorporating different features for prosodic word and prosodic phrase prediction, respectively.

Given a string of consecutive syntactic words, for each boundary between two of them (say w_i and w_{i+1}), there are three types: LW (w_i and w_{i+1} are within the same prosodic word), PW (within the same prosodic phase but different prosodic words) and PP (within different prosodic phrases). Then, our task comes down to decide the right type for each syntactic word boundary, which could be accomplished with a conditional random fields model.

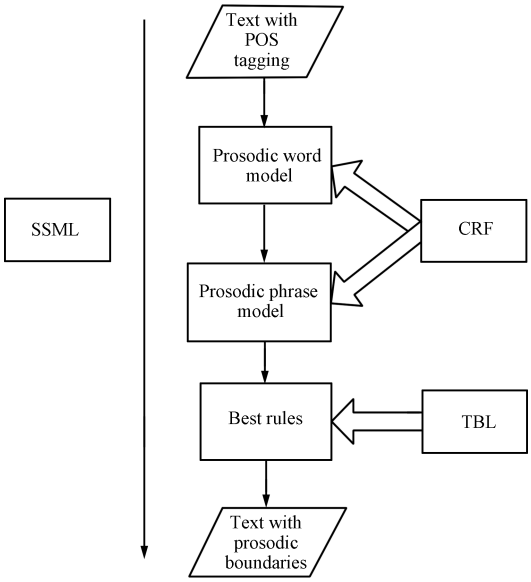


Fig. 1 Flow of prosodic structure prediction

2.1 CRF algorithm modeling

Consider probability distributions over sets of random variables $V = X \cup Y$, where X is a set of input variables that are assumed to be observed, and Y is a set of output variables that we wish to predict. Every variable $v \in V$ takes outcomes from a set ν , which can be either continuous or discrete although we discuss only the discrete case in this section. We denote an assignment to X by x, and we denote an assignment to a set $A \subset X$ by χ_A, and similarly for Y. We use the notation $1_{\{x=x'\}}$ to denote an indicator function of x which takes the value 1 when $x = x'$ and 0 otherwise.

A graphical model is a family of probability distributions that factorize according to an underlying graph. The main idea is to represent a distribution over a large number of random variables by a product of local functions each of which depends on only a small number of variables.

$$Z = \sum_{x,y} \prod_{A} \Psi_A(x_A, y_A)$$

The constant Z is a normalization factor defined as above, which ensures that the distributions sum to 1. The quantity Z, considered as a function of the set F of factors, is called the partition function in the statistical physics and graphical models communities. Computing Z is intractable in general, but much work exists on how to approximate it. Graphically, we represent the first factorization by a factor graph. A factor graph is a bipartite graph $G = (V, F, E)$, in which a variable node $v_s \in V$ is connected to a factor node $\Psi_A \in F$ if v_s is an argument to Ψ_A.

A directed graphical model, also known as a Bayesian network, is based on a directed graph $G = (V, E)$. A directed model is a family of distributions that factorize as

$$p(y, x) = \prod_{v \in V} p(v|\pi(v))$$

where $\pi(v)$ are the parents of v in G.

We use the term generative model to refer to a directed graphical model in which the outputs topologically precede

the inputs, that is, no $x \in X$ can be a parent of an out put $y \in Y$. Essentially, a generative model is one that directly describes how the outputs probabilistically "generate" the inputs.

2.2 Feature selection

Feature selection is an important part in the research of machine learning. Good data can make a proper supplement to the corresponding algorithm. In this paper, the feature "template" is manually designed.

For specific application, most commonly used features include POS (part of speech tagging), WLen (length in syllables), and Word (the word itself) of the words surrounding the boundary, which have also proved to be the most important determinants of prosodic boundary types. On account of this, we add them into both templates, with a window length of 2 for POS, i.e., we consider the POS of 2 words immediately before and after the boundary in question, and a window length of 1 for WLen and Word. A point to note, though, is that "word" has different meanings under the two scenarios. For prosodic word, it indicates syntactic word and is readily available from the input text. For prosodic phrase, which is built upon prosodic words rather than syntactic words, the meaning accordingly changes to prosodic word. Here, "POS" property of a prosodic word is acquired by simply concatenating POS′s of the syntactic words it contains.

Another feature "LastType" is also introduced into the templates, which denotes the last prosodic boundary type. The motive for adding this information came from the observation that current boundary type is influenced by that of the last one, which applies to prosodic word as well as prosodic phrase boundary. For example, a "LastType" of PP could well reduce the possibility that current boundary is still PP.

So, the feature selection in this paper is mainly focusing on three facts: POS, WLen, and Word. Based on the analysis of the result to the segmentation and POS tagging, six types are generated for both prosodic words model and prosodic phrases model. These six types are presented in Table 1.

Table 1 Types of features

Feature	Meaning
LastType	The type of last prosodic boundary
Word	The current word
WLen	The length of the current word
POS	The POS information of the current word
Word&&POS	The POS information and current word
WLen&&POS	The word length and POS information

Besides, two more types, DistPre and DistNxt, are specially designed for the prosodic phrase model. To some extent, insertion of prosodic phrase boundaries in natural spoken language is to balance the length of the constituents in the output. Hence, it is not surprising that most PP breaks occur in the middle part of a long sentence, and a prosodic phrase is usually $5 \sim 7$ syllables long, but rarely shorter than 3 or longer than 9 syllables. For this reason, we took into consideration length measures by including DistPre and DistNxt in our templates for prosodic phrase prediction, which means the distance from current boundary to the last and next nearest PP location.

For some type, we make an extension in the range from previous two words to the following two words. Take feature POS for example, five features are designed, as shown in Table 2.

What is more, the certain combination of the two types can hold some special information for prosodic boundaries. Some features are generated in this way as shown in Table 3.

Table 2 The extension of features

Feature	Meaning
POS−2	The POS of the word before last word
POS−1	The POS of last word
POS	The POS of current word
POS+1	The POS of next word
POS+2	The POS of the word after next word

Table 3 The combination of features

Feature	Meaning
Word−2POS−2	The word before last word and its POS
Word−1POS−1	The last word and its POS
Word0POS0	The current word and its POS
Word+1POS+1	The next word and its POS

The above four features are separately standing for: the word before the last one and the POS of the word before the last one; the last word and the POS of the last word; the current word and the POS of the current word; the next word and its POS.

Altogether, 34 features are selected for prosodic word model and 38 features are designed for the prosodic phrase model.

3 Transformation-based error-driven learning modification module

In our preliminary experiment using only CRF model for prediction, we find that there are always some obvious mistakes that humans would never commit as they obviously contradict to some accustomed "fixed patterns". It then happened to us that these mistakes might be corrected by rules. The transformation-based error-driven learning modification module is added at the back-end of the system to amend the mistakes left by CRF model.

3.1 The design of TBL modification module

Transformation-based error-driven learning, commonly referred to as "transformation-based learning" or TBL, is an automatic machine learning technique. The output of TBL is an ordered list of rules, whose application to data can result in a reduction of error.

TBL algorithm starts from an initial state, and by use of a series of transformation rules, it modifies the result bit by bit to achieve the best score according to the objective function used.

Fig. 2 illustrates how the transformation-based error-driven learning works. First, unannotated text is passed through an initial-state annotator. The initial-state annotator can range in complexity from assigning random structure to assigning the output of a sophisticated manually created annotator. For syntactic parsing, we have explored initial state annotations ranging from the output of a sophisticated parser to random tree structure with random nonterminal labels.

Once the text passes through the initial-state annotator, it is compared to the truth. A manually annotated corpus is used as our reference for truth. An ordered list of transformations is learned that can be applied to the output of

the initial-state annotator to make it better resemble the truth. There are two components to a transformation: a rewrite rule and a triggering environment.

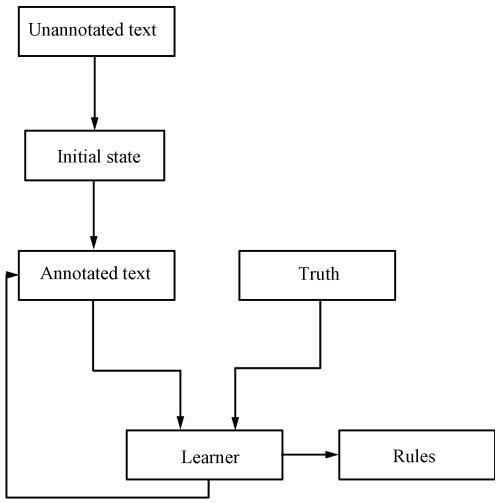


Fig. 2 Framework of TBL

In all of the applications, we have examined to date the following greedy search is applied for deriving a list of transformations: At each iteration of learning, the transformation is found whose application results in the best score according to the objective function being used; that transformation is then added to the ordered transformation list and the training corpus is updated by applying the learned transformation. Learning continues until no transformation can be found whose application results in an improvement to the annotated corpus. Other more sophisticated search techniques could be used, such as simulated annealing or learning with a look-ahead window, but we have not yet explored these alternatives.

3.2 Applying TBL to prosodic structure prediction

This paper applies TBL to the prosodic structure prediction of Mandarin TTS system.

According to the counting result of the performance with CRF model, this paper presents 20 Chinese characters to TBL learning for further disambiguation. These 20 Chinese characters occur with most mistakes frequently in the test corpus, such as "会", "的", and "同样".

For each character, we prepare two sets of corpus, one for training and another for testing. The two corpora are extracted from the annotated People′s Daily Corpus 2000, which was annotated manually by the Institute of Computational Linguistics.

The whole TBL modification module contains two processes: the training process and the testing process.

The training process accepts the annotated corpus. The corpus is first processed by the corpus preprocessor component, which extracts the features used for prosodic structure prediction. The second component is the TBL trainer component, which learns a set of rules by the transformation-based error-driven learning algorithm. This component also needs the prepared rule template as the input. The raw rule template is stored in a file and parsed by the template parser component.

Testing process applies the rules generated by TBL trainer for prosodic structure prediction. The features used for prosodic structure prediction are word, part-of-speech, character, prosodic word boundary, and prosodic phrase boundary. In real application, the information can be gotten from word segmenter, part-of-speech tagger, and pinyin-lookup module. In our package, the testing corpus is extracted from the annotated corpus and first processed by the corpus preprocessor component, which extracts the features used for prosodic structure prediction from the annotated corpus. Then, we apply all the rules learned from the TBL training system to prosodic structure prediction.

3.3 Feature selection and template design

The previous research shows that the following features are useful for prosodic structure prediction: character and word feature: LC (lexicon character), LW (lexicon word); syntax feature: POS (part-of-speech); other features: POSITION (the position in the syntax word), LEN (the length of each word), BEGIN (if in the begin of the sentence). Besides, we design other features such as PW (prosodic word boundary) and PP (prosodic phrase boundary) for prosodic structure prediction.

The current system makes use of the features of character (LC), word (LW), prosodic word boundary (PW) and prosodic phrase boundary (PP), and part-of-speech (POS).

In order to investigate the disambiguation ability for a different feature set, 18 different kinds of templates are designed with performance in Table 4.

Table 4 Templates for TBL learning

Rule template (P)	Type	Rule template (P)	Type
R0 (0.921 775)	T0	R9 (0.910 721)	T4+T0
R1 (0.869 216)	T1	R10 (0.910 432)	T5+T0
R2 (0.865 277)	T2	R11 (0.930 861)	T1,T0
R3 (0.868 538)	T3	R12 (0.932 088)	T2,T0
R4 (0.862 322)	T4	R13 (0.927 600)	T3,T0
R5 (0.859 158)	T5	R14 (0.931 081)	T4,T0
R6 (0.908 081)	T1+T0	R15 (0.919 302)	T5,T0
R7 (0.910 238)	T2+T0	R16 (0.929 745)	T1+T0,T0
R8 (0.907 211)	T3+T0	R17 (0.931 252)	T2+T0,T0

Each rule template consists of the atom templates: T0, T1, T2, T3, T4, and T5. The atom templates are described as Table 5.

Table 5 Atom templates

Atom	Type	Contents		
T0	POS itself		POS_0	
T1	Char	Char_n ($n = -3, -2, -1, 1, 2, 3$)	Char_n&Char_$n-1$	Char_n&Char_$n-1$&Char_$n-2$
T2	Word	Word_n	Word_n&Word_$n-1$	Word_n&Word_$n-1$&Word_$n-2$
T3	Prosodic word	PW_n	PW_n&PW_$n-1$	PW_n&PW_$n-1$&PW_$n-2$
T4	Prosodic phrase	PP_n	PP_n&PP_$n-1$	PP_n&PW_$n-1$&PW_$n-2$
T5	Context POS	POS_n	POS_n&POS_$n-1$	POS_n&POS_$n-1$&POS_$n-2$

3.4 Rule design and generation

Every item of the template is made up of the combination of several features referred. And every rule is formatted, for example:

$$\mathrm{POS}(Y,0)\&\mathrm{LW}(Y,-1) : A \rightarrow B$$

where "Y" indicates the feature value. "-1" indicates the relative position to the character (we constrain the offset of each feature within the range of $(-3, 3)$), "A" and "B" indicate the initial prosodic boundary and the transformed prosodic boundary, respectively, ":" is used to separate the condition and prosodic boundary tag, "$\rightarrow$" is used to separate the current tag (left) and correct tag (right), "&" is used to separate each sub-condition. Each sub-condition comprises two parts: function name and its parameter list. All the items are composed of five features: "POS", "LC", "LW", "PW", and "PP". Through TBL learning, we could generate the corresponding rule from the template whose items just match the context of the prosodic structure boundary.

4 Experiments

Experiments are designed to test the performance of the two-stage prosodic structure generation strategy. Experiments are designed and implemented in BaiLing TTS system, a real TTS system authorized by France Telecom R&D, Beijing. Experiments here can not only reflect the performance of the two-stage strategy but also be able to show the great practicality in actual speech synthesis.

4.1 Corpus preparation

Our raw corpus comprises 20 000 sentences randomly selected from the corpus of People′s Daily 2000. Each sentence had been manually segmented into syntactic words with POS tags according to "Specification for Corpus Processing at Peking University: Word Segmentation, POS Tagging and Phonetic Notation".

On average, one sentence contains 45.24 syllables and 25.15 syntactic words. Then, this corpus was prosodically annotated by two trained people; more than 90 % of their annotating results were consistent with each other.

Based on the annotation of syllables and syntactic words, both prosodic word and prosodic phrase boundaries were marked out. The whole process was guided and supervised by Cao Jian-Fen of Chinese Academy of Social Sciences.

19 000 sentences were randomly selected for CRF model training and rest 1 000 sentences were used for testing of all the following experiments. The two sets do not overlap with each other.

In this paper, CRF++ 0.51 tool was used for the model training, which can be downloaded from: http://crfpp.sourceforge.net/.

For TBL module training, 30 words were selected which are mostly happening with errors in prosodic boundaries generation in the above test experiments with CRF model.

4.2 Evaluation criteria

Evaluation criteria are precision (P), recall (R), and F-score (F). P shows the ratio between the number of correctly tagged prosodic boundaries and the number of all tagged prosodic boundaries, while R indicates the proportion of correctly tagged prosodic boundaries number and all manually tagged prosodic boundaries number. Besides, F-score obeys the following equation:

$$F = \frac{2 \times P \times R}{P + R}$$

4.3 Experiments results

The baseline uses the maximum entropy (ME)-based method for Mandarin prosodic structure prediction. In Table 6, the ME-based method achieves the F-score of 87.05%.

Table 6 Experiment performances

	Precision (%)	Recall (%)	F-score (%)
ME (PW)	87.95	86.16	87.05
CRF (PW)	90.21	92.94	90.67
CRF&TBL (PW)	93.77	95.56	94.66
ME (PP)	71.77	77.24	74.40
CRF (PP)	78.61	81.55	80.05
CRF&TBL (PP)	83.91	85.33	84.61

The ME-based method and CRF-based method are using the same training and testing corpus.

As shown in Table 6, compared with the ME-based method, the two-stage prosodic structure prediction strategy of CRF & TBL, respectively, improves the F-score by 7.61%.

It shows that the two-stage strategy is much more effective than the ME-based method for prosodic boundaries prediction. Moreover, compared with the simple CRF based method, the method that combines CRF and TBL works well.

Still, some inflexible errors occur in our test of prosody boundary prediction. Such as the following sentence "到聊天室去聊天", the result of the two-stage strategy is: "到/vq/PW 聊天室/n/PP 去/vi/LW 聊天/vi/LW", and the standard answer is: "到/vq/PW 聊天室/n/PW 去/vi/LW 聊天/vi/LW".

The strategy gives the "PP" prediction after the word "聊天室", as well as the real answer should be "PW" boundary. Two reasons can be brought to explain such case. For one reason, the length of the word "聊天室" is three and "聊天" is a verb. The strategy may probably treat it as a V-O construction, in order to make the PP boundary prediction. For another reason, the word "聊天室" is quite new to the corpus of People′s Daily 2000. This word rarely or even not appear in the corpus, so that the strategy is not able to get enough information neither in CRF model training nor in TBL rule learning. Finally, the strategy get the wrong prosody boundary prediction.

According to the two reasons above, the next work will be making certain rule to fix the error boundary and updating the corpus timely.

Due to difference in the corpora and evaluation metric, these results may not be comparable in all respects. Yet from the statistics above, the approach presented in this paper is a successful attempt towards prosodic boundary prediction.

5 Conclusion

This paper makes an extensive investigation on Mandarin prosodic structure prediction of TTS system. Based on the analyzing of a large-scale corpus, a two-stage prosodic structure prediction strategy is proposed: conditional random fields model for the first prediction and transformation-based error-driven learning modification module for the back-end of the system to amend the errors from CRF prediction.

Experiments results indicate that this approach is able to achieve a good performance with high precision, recall, and F-score. Moreover, the approach of the two-stage predic-

tion strategy for Mandarin prosodic boundaries works well in real TTS system. It also shows great general application in further processing of TTS system.

References

1 Chou F C, Tseng C Y, Lee L S. Automatic generation of prosodic structure for high quality Mandarin speech synthesis. In: Proceedings of the 4th International Conference on Spoken Language. Philadelphia, USA: IEEE, 1996. 1624−1627

2 Cao Jian-Fen. Prediction of prosodic organization based on grammatical information. *Journal of Chinese Information Processing*, 2003, **17**(3): 41−46 (in Chinese)

3 Chu M, Lu S N. A text-to-speech system with high intelligibility and high naturalness for Chinese. *Chinese Journal of Acoustics*, 1996, **15**(1): 81−90

4 Niu Z Y, Chai P Q. Segmentation of prosodic phrases for improving the naturalness of synthesized Mandarin Chinese speech. In: Proceedings of the International Conference on Spoken Language Processing. Beijing, China: ISCA, 2000. 350−353

5 Gee J P, Grosjean F. Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 1983, **15**(4): 411−458

6 Cao J F, Zhu W B. Syntactic and lexical constraint in prosodic segmentation and grouping. In: Proceedings of the Conference on Speech Prosody. Aix-en-Provence, France: ISCA, 2002. 203−206

7 Wang Hong-Jun. Prosodic words and prosodic phrases in Chinese. *Zhongguo Yuwen*, 2000, **23**(6): 525−536 (in Chinese)

8 Mao X N, Dong Y, Han J Y, Huang D Z, Wang H L. Inequality maximum entropy classifier with character features for polyphone disambiguation in Mandarin TTS systems. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Honolulu, USA: IEEE, 2007. 705−708

9 Wang M Q, Hirschberg J. Predicting intonational boundaries automatically from text. In: Proceedings of the Workshop on Speech and Natural Language. Pacific Grove, USA: Association for Computational Linguistics, 1991. 378−383

10 Taylor P, Black A W. Assigning phrase breaks from part-of speech sequences. *Computer Speech and Language*, 1998, **12**(4): 99−117

11 Zhou T, Dong Y, Huang D Z, Liu W, Wang H L. A three-stage text normalization strategy for Mandarin text-to-speech systems. In: Proceedings of the 6th International Symposium on Chinese Spoken Language Processing. Kunming, China: IEEE, 2008. 1−4

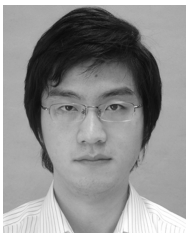
12 Mao X N, Dong Y, Han J Y, Wang H L. A comparative study of diverse knowledge sources and smoothing techniques via maximum entropy for polyphone disambiguation in Mandarin TTS systems. In: Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China: IEEE, 2007. 162−169

13 Lafferty J D, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann, 2001. 282−289

14 Mao X N, Dong Y, Fang W B, He S K, Wang H L. A two-step approach for Chinese named entity recognition based on conditional random fields and maximum entropy. In: Proceedings of the 7th International Conference on Chinese Computing. Wuhan, China: Publishing House of Electronics Industry, 2007. 434−442

15 Brill E. Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics*, 1995, **21**(4): 543−565

DONG Yuan Received his Ph. D. degree in telecommunication from Shanghai Jiao Tong University in 1999. From 1999 to 2001, he was with Nokia Research Center as a research and development scientist, working on voice recognition on Nokia mobile phone. From 2001 to 2003, he worked as a post doctoral research staff at the Engineering Department, Cambridge University, UK, working on European speech recognition project — CORETEX. Since 2003, he has worked as an associate professor at Beijing University of Posts and Telecommunications. He is also a senior research consultant at the France Telecom R&D Beijing. His research interest covers speaker recognition, speech synthesis, speech recognition, and multimedia content indexing.
E-mail: yuandong@bupt.edu.cn



ZHOU Tao Received his master degree in signals and information processing from Beijing University of Posts and Telecommunications in 2010. His research interest covers text-to-speech synthesis and natural language processing. Corresponding author of this paper.
E-mail: zhou.tao.bupt2009@gmail.com

DONG Cheng-Yu Received his master degree from Beijing University of Posts and Telecommunications in 2007. Now, he is a researcher at France Telecom R&D Beijing. His research interest covers speech recognition, speech synthesis, and speaker recognition.
E-mail: chengyu.dong@orange-ftgroup.com

WANG Hai-La Received his Ph. D. degree from University of Paris, France in 1986. Then, he worked at France Telecom R&D for six years. From 1998 to 2004, he was with several France telecom departments. Since 2004, he has been with France Telecom R&D Beijing as a chief technology officer. His research interest covers value-added multimedia service.
E-mail: haila.wang@orange-ftgroup.com