RC22917 (C0309-008) September 24, 2003
Computer Science

IBM Research Report

Comparisons among Four Statistics Based Methods of Prosody Structure Prediction

Qin Shi, Wei Zhang, XiJun Ma, WeiBin Zhu, Ling Jin
IBM Research Division
China Research Laboratory
Beijing
P. R. China

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

COMPARISONS AMONG FOUR STATISTICS BASED METHODS OF PROSODY STRUCTURE PREDICTION

Qin Shi, Wei Zhang, XiJun Ma, WeiBin Zhu, Ling Jin
IBM China Research Lab, Beijing, P.R.China
E-mail: shiqin(zhangzw, maxijun, zhuweib, jinling)@cn.ibm.com

ABSTRACT

Prosody structure prediction plays an important role in text-to-speech (TTS) conversion systems. It is the must and prior step to parametric prosody prediction. Dynamic programming (DP) and decision tree (DT) are widely used for prosody structure prediction [1][2][3] but with well-known limitations. In this paper, two other new methods, combination of dynamic programming with decision tree and combination of decision tree with finite state machine (FSM), are proposed. Then, based on a manually labeled corpus, comprehensive comparisons among the four methods are done. It could be concluded from these experiments that combination of dynamic programming with decision tree method is the best choice for prosody word boundary prediction and combination of decision tree with FSM is the best candidate for prosody phrase boundary prediction.

1. INTRODUCTION

In a TTS system, it is well known that naturalness and intelligibility of synthesis voice are strongly influenced by the assigned rhythm. Therefore prosody structure prediction becomes more and more important with the rapid improvement of TTS technology. Although there is a tight relationship between the syntactic information and the prosody structure, the syntactic information is not the only factor that influences the prosody structure. It is also influenced by the pronunciation habit.

Two statistics approaches, dynamic programming and decision tree, have been popularly used for describing the complex relationship between prosody structure and syntactic structure. But, it is still a challenge so far to understand both advantages and disadvantages of these approaches and then adopt them in TTS system effectively.

The methods of prosody structure prediction investigated in the paper include dynamic programming, decision tree, combination of dynamic programming and decision tree, and combination of decision tree and algorithm of finite state machine. A manually labeled corpus is used to train the prediction model. Part-Of-Speech (POS) and syllable numbers in lexical words are employed as statistic features.

The paper is structured as follows: In section 2, A Mandarin concatenative TTS system is introduced briefly. Typical methods used in prosody structure prediction are described in section 3, and the manually labeled corpus for training is also shown here. In section 4, detail of the experiments is presented to find out the relationship of each method's accuracy and different sizes of corpus. The conclusion is given in section 5.

2. INTRODUCTION OF MANDARIN TTS SYSTEM

2.1 The framework of linguistic analysis in Mandarin concatenative TTS system

Highly accurate prosody structure prediction is the target of linguistic analysis component, which is essential to one high-quality TTS system. Figure1 gives out the framework of linguistic analysis component in Mandarin concatenative TTS system [4].

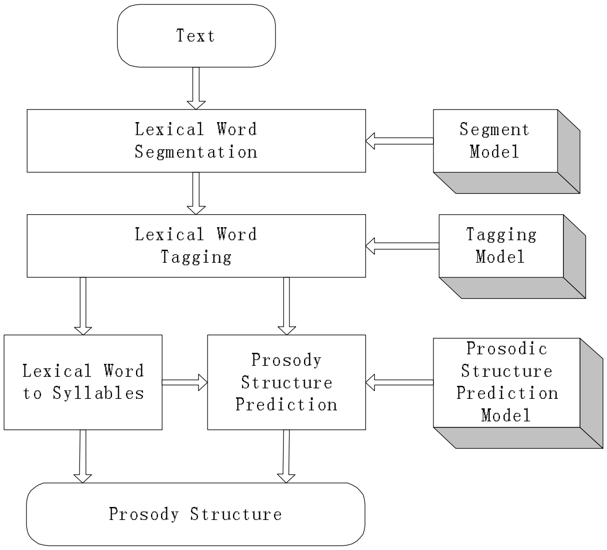
Text
Lexical Word Segmentation
Segment Model
Lexical Word Tagging
Tagging Model
Lexical Word to Syllables
Prosody Structure Prediction
Prosodic Structure Prediction Model
Prosody Structure

Fig. 1 Linguistic Analysis in IBM Mandarin TTS System

The behavior of this component can be illustrated through below example,
The sentence: 虽说应用界面不如中文之星丰富，但正在抓紧完善，一定能后来居上。
After word segmentation and POS tagging, the basic syntactic information - lexical word and its POS – are given.
虽说(cf) 应用(vg) 界面(ng) 不如(vg) 中文(ng) 之(zh)星(ng) 丰富(ag) ,(w2) 但(cbc) 正在(dr) 抓紧(vg)完善(vg) ，(w2) 一定(dr) 能(va) 后来居上(vg) 。(w1)
According to above information, the syllables and prosody structure are generated as the following.
Lexical Word & Syllable Layer:
虽说(sui1 shou1) 应用(ying1 yong4) 界面(jie4 mian4) 不如(bu4 ru2)

中文(zhong1 wen2) 之(zhi1) 星(xing1) 丰富(feng1 fu4) 但(dan4)正在(zheng4 zai4) 抓紧(zhua1 jin3) 完善(wan2 shan4) 一定(yi2 ding4) 能(neng2) 后来居上(hou4 lai2 ju1shang4)

Prosody Structure:

1. Prosody Word Layer:

虽说 应用 界面 不如 中文之星 丰富 但 正在 抓紧 完善 一定 能 后来居上

2. Prosody Phrase Layer:

虽说应用界面 不如中文之星丰富 但正在抓紧完善
一定能后来居上

3. Intonation Phrase Layer:

虽说应用界面不如中文之星丰富 但正在抓紧完善
一定能后来居上

2.2 Prosody generation, unit selection and synthesizer in Mandarin Concatenative TTS system.

In the Mandarin concatenative TTS system, a decision tree based prosody generation model is trained from the corpus. The linguistics analysis result of text to be synthesized is the features of the input of the prosody generation model, which leads to a context dependent expectation of prosody parameters, such as pitch, duration and energy values. Beam search is then used to get the best candidate sequence by the expectation of prosody parameters, and then synthesis voice is generated through a synthesizer with prosody and spectrum smoothing.

3. THE METHODS USED IN THE PROSODY STRUCTURE PREDICTION

In this section, the training corpus and the new statistic methods for prosody structure prediction are introduced.

3.1 The training corpus used in the prosody structure prediction

The Mandarin TTS corpus [5] includes about 22,000 sentences. The script is carefully designed to have wide coverage for various speech phenomena. In order to describe the prosody event, the prosody structure is then annotated manually after recording. The principle of annotation is "to label what you hear". Considering the speech character of Mandarin, the symbols of annotation are defined as:

1. BP2: the boundary of intonation phrase
2. BP1: the boundary of prosody phrase
3. BP0: the boundary of prosody word with the break in a supernormal level.
4. Blank Space: the boundary of prosody word.

The relationship among the different levels of prosody structure is illustrated in Figure2.

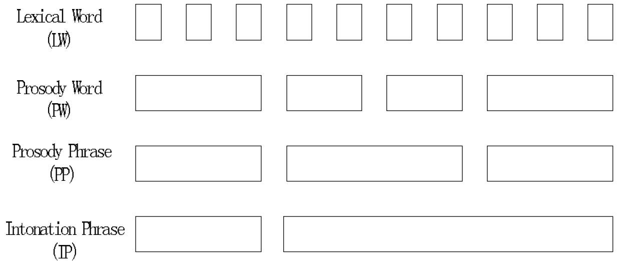

Figure2. The levels of prosody structure

The perceptive cues of the judgment for annotation are:

- The discontinuity of the pitch
- The change of the rhythm
- The length of the pause
- The duration lengthening

The punctuation information can be used for intonation phrase prediction. However, so far our work focus on prediction of prosody word layer and prosody phrase layer only.

3.2 Typical methods for prosody structure prediction

There are many statistic methods which can be deployed for prosody structure prediction: SVM (Support Vector Machine), DP (Dynamic Programming), DT (Decision Tree)[4], FSM (Finite State Machine) and so on. But each method has its inherent advantages and disadvantages. How to use these methods efficiently and avoiding their weak points is always a research topic. In the paper, we try to combine the different methods to leverage each method's advantages for the best prediction accuracy.

In order to get the prosody structure, we start from the lexical word layer, which is generated after segmentation and tagging. The higher prosody layers are generated from the lower prosody layers. The prediction problem could be described as: There is a sequence of LW units: $U = \{u_{1,...,}u_n\}$, and every unit u_i has a feature vector: x_i. For every unit, a prosody label a_i, which presents the prosody boundary, should be assigned according to the unit's context features. The probability of assigning the prosody labels is defined as:

$p(x^n, a^n) = p(x_1 ... x_n, a_1 ... a_n)$

Then prosody structure prediction can be expressed as maximizing $p(x^n, a^n)$.

The statistic methods mentioned above can be used separately to solve the problem, but they are limited by their own property. In the following, two new methods are tried.

3.2.1 Combining Dynamic Programming with Decision Tree

From the above viewpoint, using DP methods to predict PW (Prosody Word) layer from LW (Lexical Word), $a_i = 0$ or 1. 0 means: the LW boundary is not a PW boundary, 1 means: it is a PW boundary.

When predicting PP (Prosody Phrase) layer from PW layer, $a_i = 0$ or 1 also. 0 means: the PW boundary is not a PP boundary, 1 means: it is a PP boundary.

In the training data, the units, which are grouped, are presented as $(u_j, u_{j+1}, ..., u_{j+m-1})$. The prosody labels could be presented as $(00...01)$. Every unit is mapped into the feature: $u_i \rightarrow x_i$. Then the frequency of the grouped features could be described as: $freq(x_j ... x_{j+m-1}, 00...01)$. We could use $p_{DP}(x_j{}^m, a_j{}^m)$ to present it.

$$p_{DP}(x_j{}^m, a_j{}^m) = freq(x_j, ..., x_{j+m-1}, 00...01) \qquad (1)$$

In the following prediction step, all grouping paths' probabilities from the statistic result can be gained. According to them, the DP method could give out the optimum-grouping path that makes the probability $p(x^n, a^n)$ maximum.

The following graph demos the paths generated from the training data:

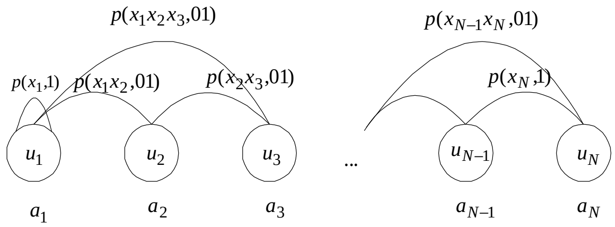

Figure3. The grouped path for prediction

Although DP could give out the optimum path from the global viewpoint, it does not fully utilize the context information of the feature. DT just has this advantage. We use the DT probability to adjust the path score of DP. The function $p_{DP}(x_j{}^m, a_j{}^m)$ is revised as $p_{DP+DT}(x_j{}^m, a_j{}^m)$.

$$\begin{aligned} & p_{DP+DT}(x_j{}^m, a_j{}^m) \\ & = p_{DP}(x_j{}^m, a_j{}^m) \cdot p_{DT}(x_j{}^m, a_j{}^m) \\ & = p_{DP}(x_j{}^m, a_j^m) \cdot [\prod_{i=j}^{j+m-2} p_{DT}(x_i, 0)] \cdot p_{DT}(x_{i+m-1}, 1) \end{aligned} \qquad (2)$$

3.2.2 Combining Decision Tree with FSM

If staring from LW layer to predict the PW and PP layer at the same time, DT method could be used. Under this circumstance, $a_i = 0$, 1 or 2. 0 means: The LW is only the LW boundary; 1 means: it is a PW boundary but not a PP boundary; 2 mean: it is a PP boundary. The decision tree could give out the label's probability according to the context information, but the relationship among the labels is ignored. In order to overcome the disadvantage, the combination of DT with FSM to predict prosody structure [6][7] is given below.

For a FSM model,

$$p(a_1, ..., a_n) = p(a_1) \prod_{i=2}^{n} p(a_i \mid a_{i-1}) \qquad (3)$$

Combining decision tree with FSM model, the overall probability is given:

$$\begin{aligned} & p(x_1 ... x_n, a_1 ... a_n) \\ & = p(x_1 \mid a_1) p(a_1) \prod_{i=2}^{n} p(x_i \mid a_i) p(a_i \mid a_{i-1}) \\ & = [\prod_{j=1}^{n} p(x_j)] L(a_1 \mid x_1) p(a_1) \prod_{i=2}^{n} L(a_i \mid x_i) P(a_i \mid a_{i-1}) \end{aligned} \qquad (4)$$

Where

$$L(a \mid x) = \frac{p(a \mid x)}{p(a)} = \frac{p(x \mid a)}{p(x)} \qquad (5)$$

$p(a \mid x)$ is given by the decision tree. $p(a)$ is the marginal probability of a , and $p(a_i \mid a_{i-1})$ is given by FSM. The goal is to choose the best label sequence:

$$\begin{aligned} \hat{a}^n & = \arg \max_{a^n} \; p(x^n, a^n) \\ & = \arg \max_{a^n} L(a_1 \mid x_1) p(a_1) \cdot \prod_{i=2}^{n} L(a_i \mid x_i) p(a_i \mid a_{i-1}) \end{aligned} \qquad (6)$$

The maximization problem could be solved by the dynamic programming algorithm.

4. EXPERIMENT

Several experiments are conducted to compare the performance of the different methods under the different training corpus.

Four methods investigated are:

1. DP: Only use DP method
2. DP+DT: Combine DP with DT
3. DT: Only use DT method
4. DT+FSM: Combine DT with FSM

In order to compare the prediction result with the reference, the evaluation criterion is defined:

1. Evaluate prosody word layer:

Firstly we split the words to characters. If the character is a PW boundary, it is labeled as 1. Else it is labeled as 0.

For example: A prosody word layer is presented as:

这 座 鼎 || 既 || 具有 || 深远 的 || 历史 || 意义 || ， || 又 不乏 || 观赏 || 价值 || 。 ||

After labeling, the string became:

001 1 01 001 01 01 1 001 01 01 1

Comparing the reference and prediction result, there could be four kinds of condition.

Reference labeling	0	0	1	1
Prediction labeling	0	1	0	1
Count Number	A	B	C	D

Table 1. Different Labeling in Prediction

Three kinds of evaluation rate are defined,

Recalling Rate (RR): D/(C+D)

Precise Rate (PR): D/(B+D)

Matching Rate (PR): D/(B+C+D)

2. Evaluate prosody phrase layer

The same criteria are used to evaluate the accuracy of PP layer prediction.

There are five training corpus: 5K, 10K, 15K, 20K, and 22K. The testing data is 2K, which is out of the 5K~20K training data, but is in the 22K training data. The 22K training data is for testing which method is sensitive to the data style.

Using different methods and different training corpus, the accuracy of PW prediction is showed in Table2.

Training data	DP	DP+DT	DT	DT+FSM
5K (RR/PR)	94.8/91.3	96.4/91.5	93.3/94.0	94.3/93.8
(MR)	87.0	88.5	88.1	88.8
10K(RR/PR)	95.5/94.3	96.4/94.0	93.3/94.9	94.3/94.5
(MR)	90.4	90.8	88.9	89.4
15K(RR/PR)	95.2/94.9	96.2/94.6	93.6/95.1	94.3/94.9
(MR)	90.5	91.2	89.3	89.7
20K(RR/PR)	95.2/95.0	96.3/94.8	93.8/95.1	94.3/94.8

(MR)	90.6	91.5	89.4	89.7
22K(RR/PR)	95.7/96.3	96.6/96.1	96.1/94.0	94.4/94.9
(MR)	92.3	92.8	89.7	89.9

Table 2. The Accuracy of Prosody Word Prediction

In Table3, The accuracy of the PP prediction is given out.

Training data	DP	DP+DT	DT	DT+FSM
5K (RR/PR)	90.2/71.8	91.8/73.5	88.3/84.6	87.5/87.1
(MR)	66.6	69.0	76.1	77.5
10K(RR/PR)	90.5/77.8	92.0/78.7	88.6/85.9	87.7/87.9
(MR)	71.9	74.9	77.4	78.3
15K(RR/PR)	90.2/79.3	91.6/80.4	88.6/86.9	87.7/89.1
(MR)	73.0	75.8	78.2	79.3
20K(RR/PR)	89.7/80.3	91.3/81.6	88.2/87.4	87.1/89.9
(MR)	73.5	79.1	78.3	79.3
22K(RR/PR)	92.0/88.8	93.1/89.8	88.1/88.0	87.3/89.9
(MR)	82.5	84.2	78.6	79.5

Table3 The Accuracy of Prosody Phrase Prediction

Based on MR, Figure4 shows the capacity of predicting prosody word layer by using the four methods.

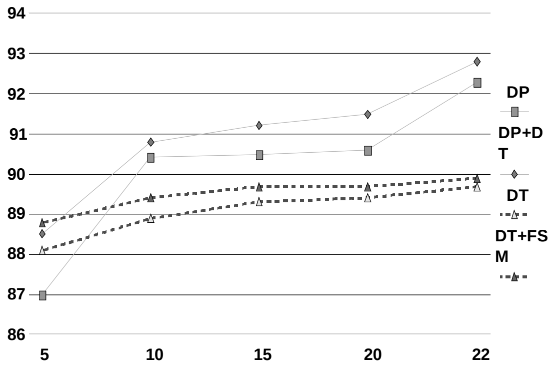

Figure 4. The PW Layer Prediction Results

Based on MR, Figure5 shows the capacity of predicting prosody phrase layer by using the four methods.

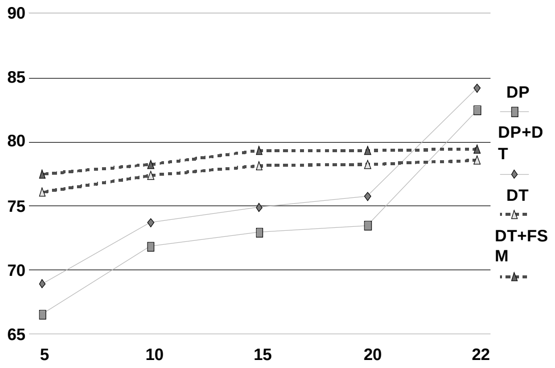

Figure 5. The PP Layer Prediction Results

5. DISCUSSION

From figure4 and figure5 based on MR, several conclusions can be given here,

1.Combination of dynamic programming and decision tree is better than dynamic programming only;

2.Combination of decision tree and finite state machine is better than decision tree only;

3.Dynamic programming and combination of dynamic programming with decision tree are more suitable for local prediction e.g. prosody word, and combination of decision tree and finite state machine is more robust for global prediction e.g. prosody phrase;

4.Decision tree or decision tree dominated combinations are less sensitive, so there is no significant improvement when corpus increases. But dynamic programming or dynamic programming dominated combinations are more sensitive to corpus size. Its accuracy increases quickly when the training set enlarged.

6. REFERENCE

[1] *Fu-Chiang Chou; Chiu-Yu Tseng; Keh-Jiann Chen; Lin-Shan Lee* "A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and non-uniform units" Acoustics, Speech, and Signal Processing, 1997. ICASSP-97,IEEE International Conference on , Volume: 2 , 1997

[2] Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee "Automatic Generation of Prosodic Structure for High Quality Mandarin Speech Synthesis", ICSLP, 1996, pp. 1624-1627.

[3]Ren-hua Wang, Qinfeng Liu and Difei Tang, " A New Chinese Text-to-Speech System with High Naturalness", ICSLP, 1996, pp. 1441-1444

[4] Qin Shi, XiJun Ma, "Statistic Prosody Structure Prediction" IEEE 2002 TTS Workshop

[5]WeiBin Zhu, "Corpus labling for data driven TTS System" IEEE 2002 TTS Workshop.

[6]C.W.Wightman and M. Ostendorf, " Automatic Labeling of Prosodic Patterns," Proc, ICASSP, October, 1994

[7]XiJun Ma,"Automatic Prosody labeling using both Text and Acoustic Information" IEEE 2003 ICASSP, HongKong