INTEGRATED APPROACHES TO PROSODIC WORD PREDICTION FOR CHINESE TTS

Guohong Fu and K.K. Luke

Department of Linguistics, The University of Hong Kong,
Pokfulam Road, Hong Kong China
ghfu@hkucc.hku.hk, kkluke@hkusua.hku.hk

ABSTRACT

This paper focuses on integrated prosodic word prediction for Chinese TTS. To avoid the problem of inconsistency between lexical words and prosodic words in Chinese, lexical word segmentation and prosodic word prediction are taken as one process instead of two independent tasks. Furthermore, two word-based approaches are proposed to drive this integrated prosodic word prediction: The first one follows the notion of lexicalized hidden Markov models, and the second one is borrowed from unknown word identification for Chinese. The results of our primary experiment show these integrated approaches are effective.

Keywords: Prosodic word prediction, Text-to-speech synthesis, Lexicalized HMMs

1. INTRODUCTION

It is proved that prosodic word (P-Word) is an important prosodic unit in Mandarin TTS [Chu and Qian, 2001]. In general, Chinese utterance can be structured as a prosodic hierarchy, which contains three main levels of prosodic units, i.e. prosodic word, prosodic phrase and intonation phrase [Li and Lin, 2000]. As the lowest level of prosody, prosodic word not only plays an important role in predicting higher levels of prosodic phrases, but also is an essential factor in generating other prosodic features, such as intonation, stress, duration and pause. However, there is very little explicit information of prosodic words in plain Chinese texts. The objective of prosodic word prediction is therefore to predict the implicit prosodic word boundaries in written texts.

Prosodic word prediction is by no means a trivial task, especially for Chinese. On the one hand, Chinese text is character based. There are no explicit delimiters to indicate word boundary, except for some punctuations. On the other hand, prosodic words are formed dynamically in real utterance. In theory, any combination of Chinese character or lexical words (L-Words) may be a potential prosodic word. In fact, all prosodic words form an open-set. It is impossible to collect exhaustively all possible prosodic words into in a pre-defined lexicon.

Another important challenge in prosodic word prediction for Chinese is the inconsistency between lexical words and prosodic words. It is proved that using prosodic words as the basic prosodic unit, instead of lexical words, will result in more natural synthetic speech [Chu and Qian, 2001]. However, most previous work Chinese TTS take lexical words as the basic unit for prosodic phrasing. In practice, lexical words are not exactly coincident with prosodic words. As mentioned in [Chu and Qian,

0-7803-7902-0/03/$17.00 © 2003 IEEE.

2001], only 70.70% of lexical words are identified with prosodic words in real speech. In particular, a prosodic word may be made up of one or more lexical words and vice versa. For example, the numeral-quantifier phrase 一对 (yi1 dui4, one pair) is often uttered as one prosodic word in Chinese and is syntactically segmented as two lexical words "一" (yi1, one) and "对" (dui4, pair). But for the number 二万七千二百一十七 (er4 wan4 qi1 qian1 er4 bai3 yi1 shi2 qi1, twenty seven thousand two hundred and seventeen), it is often considered as an independent lexical word, but is naturally uttered as a sequence of prosodic words in real speech, i.e. "二万/七千/二百/一十七/".

This paper focuses on integrated prosodic word prediction. To avoid the problem of inconsistency between lexical words and prosodic words, we take lexical word segmentation and prosodic word prediction as one process rather than two independent tasks. Furthermore, two word-based statistical models are also given to assign prosodic word breaks at proper places of the text. The first model follows the notion of the lexicalized hidden Markov models (LHMMs)[Lee, et al., 2000]. In this framework, word sequence, word juncture type sequence and their interaction are combined to perform correct lexical word segmentation and juncture type assignment. The second model is borrowed from unknown word identification in Chinese word segmentation. In this framework, prosodic words are considered as a special group of unknown lexical words and a hybrid model for unknown word segmentation is modified and further extended to score equally all possible lexical word candidates and prosodic word candidates of the text. In this way, different features such as prosodic word-formation patterns, word juncture and contextual information are statistically computed and incorporated for this integrated prosodic word prediction.

The rest of this paper is organized as follows: Section 2 describes in detail the lexicalized HMMs for prosodic word prediction. In section 3, a hybrid model for unknown word identification is modified and extended for locating prosodic word boundaries in texts. In section 4, we report our experiments on a speech corpus, and in the final section we draw some conclusions on this work.

2. P-WORD PREDICTION USING HMM

2.1 The problem

In practice, it is very difficult to indicate the exact differences between lexical words and prosodic words. For convenience, lexical words refer to the words that are included in the lexicon used, and the prosodic words, on the contrary, refer to the words that are out of the lexicon.

Thus, we can define the problem of integrated prosodic word prediction as follows: an input text consists of a sequence of characters $C = c_1 c_2 \cdots c_n$. There are usually a number of candidate lexical word sequences. Let $W = w_1 w_2 \cdots w_m$ denote a certain sequence of lexical word candidates. Between each pair of lexical words is a word juncture. In particular, there are two types of junctures in prosodic word prediction: *prosodic word boundary* and *non-prosodic word boundary*, denoted by t_B and t_N respectively. Obviously, each lexical word sequence contains one specific sequence of junctures, denoted by $J = j_1 j_2 \cdots j_m$. But there may be more than one possible sequence of word juncture types $T = t_1 t_2 \cdots t_m$ for one juncture sequence. The goal of integrated prosodic word prediction is therefore to find the most appropriate lexical word sequence $\hat{W}$ and its related proper sequence of juncture types $\hat{T}$, with which the lexical word sequence can be further segmented into a meaningful prosodic word sequence.

From the point of view of probability theory, this process is equivalent to find a best sequence $\hat{W}$ of lexical words and a proper sequence $\hat{T}$ of word juncture types that maximizes the conditional probability $P(T,W|C)$, i.e.

$$\Psi(W,T) = \arg\max_{W,T} P(W,T|C) \qquad (2.1.1)$$

2.2 The general model

Equation (2.1.1) gives a general description about prosodic word prediction. Using Bayes' theorem, it can be be rewritten as follows:

$$\Psi(W,T) = \arg\max_{W,T} P(C \mid W,T)P(W,T)/P(C) \qquad (2.2.1)$$

For an input character string C, the probability $P(C)$ is fixed. Therefore, this term can be dropped from above equation. For simplification, the term $P_r(C \mid W,T)$ can also be ignored in that W and T involves all information of C. Thus, we obtain a general statistical model for prosodic word prediction:

$$\begin{aligned}\Psi(W,T) &= \arg\max_{W,T} P(W,T) = \arg\max_{J,T} P(J,T) \\ &= \arg\max_{J,T} \prod_{i=1}^{m} P(j_i \mid t_{1,i}, j_{1,i-1}) P(t_i \mid t_{1,i-1}, j_{1,i-1})\end{aligned} \qquad (2.2.2)$$

Where, $j_i = w_{i-1}w_i$. Note that a sequence of word juncture is equivalent to the relevant sequence of lexical words.

Actually, Equation (2.2.2) is the general statistical model for prosodic word prediction. However, it is non-computable in practice because it has too many. To make it tractable and avoid the problem of data sparseness, two kinds of assumptions are employed to simplify this model.

2.3 Standard HMMs

The first kind of assumptions comes from the independent hypothesis in standard HMMs: The appearance of current juncture j_i depends only on current juncture type t_i, and the assignment of current juncture type t_i depends only on its previous juncture type t_{i-1}. Thus,

$$\Psi(W,T) = \arg\max_{J,T} \prod_{i=1}^{m} P(j_i \mid t_i) P(t_i \mid t_{i-1}) \qquad (2.3.1)$$

Equation (2.3.1) gives the standard HMMs for prosodic word prediction. Where, $P(j_i \mid t_i)$ refers to the model of word sequence, and $P(t_i \mid t_{i-1})$ denotes the model of juncture type sequence.

As mentioned earlier, [Taylor and Black. 1998] also proposed HMMs for prosodic phrasing. However, their model is par-of-speech based while the model in Equation (2.3.1) is word-based.

In maximum likelihood estimation (MLE), the relevant probabilities in Equation (2.3.1) can be approximated by their relative frequencies, viz.

$$\begin{cases} P(j_i \mid t_i) = \dfrac{Count(w_{i-1}t_i w_i)}{Count(t_i)} \\ P(t_i \mid t_{i-1}) = \dfrac{Count(t_{i-1}t_i)}{Count(t_{i-1})} \end{cases} \qquad (2.3.2)$$

2.4 Lexicalized HMMs

The second type of assumption follows the notion of lexicalized HMMs. In this assumption, the appearance of current juncture j_i or word pair $w_{i-1}w_i$ depends not only on current juncture type t_i but also its previous juncture j_{i-1}; and the assignment of current juncture type t_i depends both its previous juncture j_{i-1} and juncture type t_{i-1}. Thus, Equation (2.2.2) can be simplified as:

$$\Psi(W,T) = \arg\max_{J,T} \prod_{i=1}^{m} P(j_i \mid t_i j_{i-1}) P(t_i \mid t_{i-1} j_{i-1}) \qquad (2.4.1)$$

Actually, Equation (2.4.1) gives lexicalized HMMs for predicting prosodic word boundaries in text. In this case, both contextual words and juncture types, and their interaction are combined for prosodic word prediction.

Similarly, if we have a corpus that has been annotated with prosodic word boundaries, we can easily estimate the relevant probabilities using the following formula:

$$\begin{cases} P(j_i \mid t_i j_{i-1}) = \dfrac{Count(w_{i-2}w_{i-1}t_i w_i)}{Count(t_i)} \\ P(t_i \mid t_{i-1} j_{i-1}) = \dfrac{Count(w_{i-2}t_{i-1}w_{i-1}t_i)}{Count(w_{i-2}t_{i-1}w_{i-1})} \end{cases} \qquad (2.4.2)$$

To avoid the problem of sparse data in above estimation, a simplified back-of smoothing technique [Lee, et al., 2000] is also employed in our work.

3. P-WORD PREDICTION AS UNKNOWN WORD IDENTIFICATION

In this section, a hybrid model for unknown word identification is revised to predict prosodic word breaks in text.

3.1 Prosodic words vs. unknown lexical words

In practice, prosodic words and unknown lexical words have a number of similar characteristics. First, both prosodic words and unknown lexical words are not listed in the lexicon used. Second, both of them are made up of known lexical words in the lexicon. Furthermore, it is observed that the rule of prosodic word formation is similar to that of unknown lexical word formation. For example, some function words such as 的 (de5, of) never present itself at the initial position of a prosodic word, while some prefixal lexical words, such as 阿 (a1), hardly occur at the final position of an unknown lexical word. Due to these similarities, prosodic words can be viewed as a special group of unknown lexical words to some extent. Thus, prosodic prediction becomes a process of identifying special unknown words in text to some extent. Based on this point, some previous techniques for unknown lexical word identification can be applied for prosodic word prediction.

3.2 P-word prediction as unknown word identification

We have developed a hybrid model for unknown word identification. Here, we revise it for predicting prosodic word breaks in text.

Given a sequence of Chinese character string $C = c_1 c_2 \cdots c_n$, there is usually more than one possible sequence of words $W = w_1 w_2 \cdots w_m$, which consists of unknown prosodic words and known lexical words. The prosodic word prediction aims to find the most appropriate word sequence $\hat{W} = w_1 w_2 \cdots w_m$ that maximizes

$$\begin{aligned} P_H(W) &= P_{plm}(W) P_{WJM-I}(W) P_{WJM-O}(W) P_{bigram}(W) \\ &= \prod_i P_{plm}(w_i) P_{WJM-I}(w_i) \prod_i P_{WJM-O}(w_i) P_{bigram}(w_i \mid w_{i-1}) \end{aligned} \quad (3.2.1)$$

Equation (3.2.1) indicates a hybrid model for integrated prosodic word prediction. Where, $P_H(W)$ denote the overall probability of a possible word sequence for the text; $P_{WJM-I}(w_i)$ denotes the probability of internal word-junctures inside the word w_i; $P_{WJM-O}(w_{i-1}, w_i)$ denotes the probability of the external word juncture between two successive words $w_{i-1} w_i$; $P_{bigram}(w_i \mid w_{i-1})$ is the word bigram probability.

Let $t(xy)$ denote certain type of a word juncture xy, and $P(t(xy)) \stackrel{def}{=} \frac{Count(xty)}{Count(xy)}$ denote the relevant conditional probability. Given a word $w_i = e_1 e_2 \cdots e_h$ (where e_j is a component word of w_i, $1 \le j \le h$), then its internal juncture probability $P_{WJM-I}(w_i)$ can be calculated by equation (3.2.2).

$$P_{WJM-I}(w) = \begin{cases} \prod_{j=1}^{j-1} P_r(t_N(e_j e_{j+1})), & \text{if } w \text{ is a P - word} \\ 1, & \text{if } w \text{ is a L - word} \end{cases} \quad (3.2.2)$$

Similarly, the external juncture probability $P_{WJM-O}(w_i)$ of the juncture between w_i and its previous word $w_{i-1} = e'_1 e'_2 \cdots e'_l$, can be formulated as

$$P_{WJM-O}(w_i) = \begin{cases} P(t_B(w_{i-1} w_i)), \text{if both } w_{i-1} \text{ and } w_i \text{ are } L-words \\ P(t_B(e'_l e_1)), \text{ if both } w_{i-1} \text{ and } w_i \text{ are } P-word \\ P(t_B(e'_l w_i)), \text{ if } w_{i-1} \text{ is } P-word \text{ and } w_i \text{ is } L-word \\ P(t_B(w_i e_1)), \text{ if } w_{i-1} \text{ is } L-word \text{ and } w_i \text{ is } P-word \end{cases} \quad (3.2.3)$$

As for the word bigram probability, it can be computed by equation (3.2.4).

$$P_{bigram}(w_i \mid w_{i-1}) = \begin{cases} P_r(w_i \mid w_{i-1}), \text{if both } w_{i-1} \text{ and } w_i \text{ are known} \\ P_r(e_i \mid e_{i-1})), \text{ if both } w_{i-1} \text{ and } w_i \text{ are unknown} \\ P_r(e_i \mid w_{i-1})), \text{ if } w_{i-1} \text{ is known and } w_i \text{ is } un\text{known} \\ P_r(w_i \mid e_{i-1}), \text{ if } w_{i-1} \text{ is } un\text{known and } w_i \text{ is known} \end{cases} \quad (3.2.4)$$

Where, e_{i-1} and e_i denote the final component word of w_{i-1} and the initial component word of w_i respectively.

If a prosody-labelled corpus is available, the probabilities in equation (3.2.2)-(3.2.4) can be easily estimated using the maximum likelihood estimation. The details can be seen in [Fu and Luke, 2003].

4. EXPERIMENTS

This section reports the relevant experiments on above approaches.

4.1 Experimental Data and Evaluation

In evaluating our system, we conduct an experiment on a large speech corpus. This corpus contains 17,830 sentences and is manually annotated with lexical and prosodic word boundary. As shown in Table 1, 90% of this corpus, namely about 16,047 sentences are used as training data or close-test data, and the rest 10% are used for the open-test.

	#sentences	#words	#L-words	# P-words
Training data	16,047	78,234	35,399	42,835
Test data	1,783	8,784	3,954	4,830
Total	17,830	87,018	39,353	46,665

Table 1: Experimental corpora

In our experiments, three measures, i.e. recall, precision and F-score are used to evaluate the performance of our system. Recall (denoted by R) is defined to be the number of correctly predicted (prosodic) words divided by the total number of standard prosodic words in test data, and the precision (denoted by P) is defined to be the number of correctly predicted prosodic words divided by the total number of automatically identified prosodic words. As for F-score (denoted by F), it is the weighted harmonic mean of precision and recall, i.e.

$$F = \frac{(\beta^2 + 1)RP}{\beta^2 R + P} \qquad (4.1.1)$$

Here, we use the balanced F-score (viz. $\beta^2 = 1$) to evaluate the overall performance of our system in prosodic word prediction in that it is not clear that which one, recall or precision, is more important for other modules in text-to-speech synthesis.

4.2 Results and discussions

In addition to the lexicalized HMMs based approach (denoted by M1), the standard HMMs based approach (denoted by M2) in section 2 and the integrated unknown-word identification technique in section 3 (denoted by M3), other methods for unknown word identification are also introduced into our experiment for comparison, including the two-stage segmentation incorporating word-based word-formation patterns, word juncture models and word bigram and (denoted by M4, shown in [Fu and Luke, 2003]) and the two-stage segmentation incorporating character-based word-formation patterns, character juncture models and word bigram (denoted by M5, [Wang, et al., 2000]). Furthermore, we compute following measures in our experiments, i.e. the overall F-measure (F), the overall recall (R), the overall precision (P), the F-measure on lexical words (F_{LW}), the recall on lexical words (R_{LW}), the precision on lexical words (P_{LW}), the F-measure on prosodic words (F_{PW}), the recall on prosodic words (R_{PW}) and the precision on prosodic words (P_{PW}). We hope these measures can give a complete and objective evaluation on these approaches. What is more, we also hope our experiments can answer how much contribution different strategies and models make to achieve correct prosodic word prediction and which method for unknown word identification is still effective for prosodic word prediction.

Our experiment consists of two tests, i.e. a close test on the training data and an open test on the test data. The results of these two tests are summarized in Table 2 and Table 3 respectively.

Methods	F	R	P	F_{LW}	R_{LW}	P_{LW}	F_{PW}	R_{PW}	P_{PW}
M1	96.3	96.9	95.3	96.6	99.0	94.4	96.0	95.2	96.8
M2	95.7	95.2	96.3	97.1	97.7	96.6	94.5	93.1	96.0
M3	95.7	95.0	96.4	97.7	97.4	98.0	94.1	93.1	95.1
M4	94.5	94.2	94.8	96.3	97.8	94.8	93.0	91.3	94.8
M5	62.5	71.1	55.7	62.8	97.9	46.3	61.9	48.9	84.3

Table 2: Results of the close test for different methods

Methods	F	R	P	F_{LW}	R_{LW}	P_{LW}	F_{PW}	R_{PW}	P_{PW}
M1	66.3	66.9	65.7	72.8	76.2	69.7	60.6	59.4	62.0
M2	53.2	49.7	57.2	64.6	64.5	64.7	42.7	37.6	49.3
M3	56.5	58.4	54.7	70.8	85.1	60.6	40.8	36.5	46.1
M4	54.1	53.1	55.2	66.7	76.6	59.1	40.0	33.8	49.0
M5	50.2	54.8	46.3	60.0	90.0	45.0	34.3	25.9	50.6

Table 3: Results of the open test for different methods

From these results, we can draw some conclusions. Firstly, integrating prosodic word prediction leads to improvement of accuracy in prosodic word prediction. As can be seen in Table 2 and Table 3, the integrated method M3 outperforms the separated method M4 as a whole, though they adopt the same models. Secondly, lexicalized HMMs are helpful to enhance the performance in prosodic word prediction. In comparison with the typical standard HMMs (viz. M2), the lexicalized HMMs improve the overall F-measure on prosodic word prediction (viz. F_{PW}) by 1.5% in the close test and about 18% in the open test. Moreover, lexicalized HMMs achieve the best results among all methods under discussion. Thirdly, some techniques for unknown word identification are still effective for prosodic word prediction, in particular the word-based approaches. In our experiment, M2, M3 and M5 are borrowed from unknown word identification. As shown in Table 2, 94.1%, 93.0% and 61.9% of F-score on prosodic word prediction can be achieved by these three methods respectively in close-test. Finally, the proposed approaches yield satisfactory results in the close test. However, the training data is too small for training word-based models and the serious data sparseness results in degradation of performance in the open test. Therefore, further efforts are still needed to address the problem of data sparseness in open applications.

5. CONCLUSIONS

In this paper, we have discussed the problem of prosodic word prediction for Chinese text-to-speech synthesis. To address the problem of inconsistency between lexical words and prosodic words in Chinese, we take lexical word segmentation and prosodic word prediction as one process instead of two independent tasks. Furthermore, we propose two word-based statistical models for predicting prosodic word breaks in text. The results of our primary experiment show that lexical word segmentation and prosodic word prediction can be resolved effectively by the proposed approaches. In future, we plan to resolve the problem of data sparseness in current system.

References

[1] Chu, M., and Y. Qian. 2001. Locating boundaries for prosodic constituents in unrestricted Mandarin texts. Computational Linguistics and Chinese Language Processing, 6(1): 1-22.

[2] Fu, Guohong, and K.K. Luke. 2003. A two-stage statistical word segmentation system. Proceedings of The 2nd SIGHAN Workshop on Chinese Language Processing, Sapporo, Japan, 156-159.

[3] Fu, Guohong, and K.K. Luke. 2003. An integrated approach for Chinese word segmentation. Proceedings of The 17th Pacific Asia Conference on Language, Information and Computation (PACLIC 17), Singapore. (forthcoming)

[4] Lee, Sang-Zoo, Jun-ichi Tsujii and Hae-Chang Rim. 2000. Lexicalized hidden Markov models for part-of-speech tagging. Proceeding of COLING 2000, Saarbruken, Germany, 481-487.

[5] Li, A., and M.Lin. 2000. Speech corpus of Chinese discourse and the phonetic research. Proceedings of ICSLP2000.

[6] Taylor, P., and A.W. Black. 1998. Assigning phrase breaks from part-of-speech sequences. Computer Speech and Language, 12 (1): 99-117.

[7] Wang, Xiaolong, Guohong Fu, D. S.Yeung, J. N.K.Liu, and R. Luk. 2000. Models and algorithms of Chinese word segmentation. Proceedings of the International Conference on Artificial Intelligence (IC-AI'2000), Las Vegas, USA, 1279-1284.