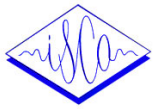

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

PREDICTING PROSODIC WORDS FROM LEXICAL WORDS--A FIRST STEP TOWARDS PREDICTING PROSODY FROM TEXT

Hua-jui Peng[1] *and Chi-ching Chen*[2], *Chiu-yu Tseng*[2], *Keh-jiann Chen*[1]

[1] Institute of Information Science, Academia Sinica, Taipei
[2] Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei
E-mail: phr@iis.sinica.edu.tw

ABSTRACT

Much remains unsolved in how to predict prosody from text for unlimited Mandarin Chinese TTS. The interactions and the governments between syntactic structure and prosodic structure were still unresolved challenges. By using Part-of-Speech tagging (hence POS), lexical information of text was required, we aimed to find significant patterns of word grouping from analyzing real speech data and such lexical information. This paper reported discrepancies found between lexical words (hence LW) parsed from text and prosodic words (hence PW) annotated from speech data, and proposed a statistical model to predict PWs from LWs. In statistical model, both length of the word and the tagging from POS are two essential features to predict PWs, and the results showed approximately 90% of prediction for PWs, however, it did leave more room for extension. We believe that evidence from PW predictions is a first step towards building prosody models from text.

1. INTRODUCTION

Much remains unsolved in how to predict prosody from text for unlimited Mandarin Chinese TTS. Linguistic analyses of text have been insufficient to provide specifications required for speech prosody, both in terms of prosodic units and boundaries, and in intonation contours for connected fluent speech. Though syntactic analyses provide possible boundaries and intonation specification for phrases, location of boundaries and breaks in connected speech require more specification, and prosody of fluent speech goes beyond concatenating simple-sentence intonations into strings. Aiming to build a prosody model for connected fluent speech from the bottom upward, our first step was to set up models that could sufficiently predict PW from LW, and to serve as a base for building speech prosody.

In hierarchical rhythmic structures [1], PW is fundamental prosodic unit, while LW is basic syntactic unit in syntactic structure. However gaps and discrepancies were in each layer of syntactic and prosodic structures. Only 67.5% of PWs and LWs were coincident in our prosodic structure tagged corpora (in section 2.3). In this paper we proposed a statistical model for predicting PWs by grouping lexical words. The issues of grouping words to form PWs have been studied in [2, 3], a good word grouping strategy helped construct the temporal organization of speech and rendered spoken utterances natural and fluent. In the following sections, we focused on finding an optimal word grouping strategy by combining lexical information i.e., POS tagging and analyses of real speech data, and studying how LWs form PWs.

2. MATERIALS USED--TEXT VS. SPEECH CORPORA

Two modalities of the same corpus were used, namely, text prepared for read speech and speech data collected subsequently. Two sets of text were used. One set was 599 paragraphs (24803 syllables in total) ranging from 2-character simple sentences up to 181-character complex sentences. These paragraphs were controlled for word frequency using the CKIP database (http://godel.iis.sinica.edu.tw/CKIP/) and phonetic balance for segments and tones. Another set was 26 longer paragraphs (11592 syllables in total) of text ranging from 85 to 981-character paragraphs rearranged from the 599 paragraphs for frequency and phonetic controls. The two sets of text overlapped 88%. These texts served as materials for linguistic analysis via a lexical analysis algorithm (Section 2.2.) to derive LWs. Two sets of speech corpora were collected. Four native untrained speakers (2 males M01, M02 and 2 females F01, F02) read the 599 paragraphs at the average speech rate of 304 ms/syllable. Another two radio announcers (1 male and 1 female) read the 26 longer paragraphs at the average speaking rate of 200 ms/syllable. The two sets of speech data were referred as slower speech vs faster speech.

2.1. PW segmentation consistency among different speakers and speech rates

Both the slower speech data and the faster speech data were labeled manually by trained transcribers for perceived boundaries and breaks (pauses) using a self-designed labeling system [4]. The labeled data were further checked for transcriber consistency [5, 6, 7], the results were consistent with the high ratio of agreement at over 98% on the locations of PW boundaries reported in [8]. Then PW overlaps across speakers as well as speech rates were checked. Average PW overlaps across speakers for the slower speech and the fast speech were at 91.57% and 92.45%, respectively. We then compared cross-speaker overlaps of PW segmentation across the two sets of speech data to see if speech rate has any effect on PW. An overlap at 90.35% was found. The overlap indicated that PW was a reliable prosodic unit in speech production across speakers and speech rate. Hence predicting PWs is a feasible task for prosody generation.

2.2. Linguistic analysis of text into LWs

A lexical segmentation algorithm was used to segment the text into LW [9]. In 599 paragraphs set, comparisons between LW and PW were made as shown in Table 1.

Table 1. Comparison of syllable numbers in LW and PW (%)

# of syl / %	1	2	3	4	5	6	7	8	9
LW	38	54	5	2	0.1	0.07	0.19	.01	.01
PW	5	67	25	2	0.2	0.02	0	0	0

A clear discrepancy between LW and PW was found. Results from lexical analysis showed that monosyllabic (38%) and disyllabic (54%) LWs constituted the majority of LWs (92%) while the amount of syllabic>2 LWs was insignificant. However, results from labeled read speech data of the same text showed that the disyllabic (67%) and the tri-syllabic (25%) PWs were the majority of PWs (92%) whereas the monosyllabic (5%) and the syllabic>3 PWs were insignificant. Further analyses showed that 67.5% of PWs equaled LWs, the rest 32.5% of PWs consisted of multi-LWs in which 89% of PWs were consisted of 2 LWs and 11% of PWs were consisted of 3 or at most 4 LWs. Table 2 and 3 demonstrated the distribution of PWs respectively in situation (1.) PWs=LWs and (2.) PWs consisted of multi- LWs.

Table 2. Distribution of PWs while PWs =LWs

# of syllable	1	2	3	4	5~
%	7.1	84.1	8.6	0.2	0

Table 3. Distribution of PWs while PWs consisted of multi-LWs

# of syllable	1-1	1-2	2-1	1-1-1	1-2-1	3-1	Others
%	30.7	23.2	31.9	6.4	2.4	1.3	4.2

The above analyses suggests that (1.) lexical words were not sufficient to cover prosodic words, (2.) monosyllabic LWs were combined with their neighboring LWs to form disyllabic and tri-syllabic PWs, and (3.) possible solution might dwell in capturing how monosyllabic LWs behaved in PW formation.

A rule-based model was experimented to see the behavior of monosyllabic LW, and the results are showed in Table 4. In rule-based model, we extracted the prosodic parameters by observing the labeled speech data of 599 paragraphs, and generalized the rules. The experimental results showed that the recall is lower than the precision, meaning that most part of the boundaries was predicted, but the real speech had more breaks and pauses. By observing the errors we concluded the following two disadvantages of rule-based model; 1) rules were set up by human and could not enumerate all possible combinations, 2) rule-based model did not handle the cases of combining three lexical words to form PWs. Therefore, further development of statistical model was made to resolve these defects.

Table 4. Results of rule-based model

	F01	F02	M01	M02
Recall	85.62%	86.91%	84.36%	83.96%
Precision	92.45%	91.31%	90.05%	91.17%
F-score	88.9%	89.05%	87.11%	87.41%

3. STATISTICAL MODEL FOR PROSODIC-WORD SEGMENTATION IN TEXT

Statistical model was adopted and tested to experiment how PWs could be better predicted, in particular how monosyllabic LWs form PWs. It is noted that the POS identities of majority of monosyllabic LWs were adverbs, prepositions, aspects, quantifiers, personal pronouns, particles, and conjunctions etc. These monosyllabic LWs tended to combine with preceding or following word to become a possible PW. Our statistical model will use both features of POS and length of LWs to predict PWs. Since speech rate did not affect PWs (Section 2), we used the larger speech corpus, i.e. the slower speech data from 4 speakers, for subsequent experiments.

3.1. Statistical model

The aim of a statistical model was to find the most optimal combination for PWs from LWs. In this paper we modeled PW generation as a tagging problem. There are only two different tags L and M for LWs in our model. If

a LW has an L tag which means this LW should be combined with its left LW to form a PW no matter the left LW has tagged L or M. If a LW has an M tag and its right neighbor LW has also an M tag, it means that this LW stands along as a PW. We can derive PWs of a text from its {L,M} tagged LW structure and vise versa. For instance a PW "以四個月" has its lexical {L,M} tagged structure as "以(M) 四個(L) 月(L)". The problem of deriving PW structures of texts from LW structures becomes a tagging problem. We will design a statistical model for tagging {L,M} of LWs. The probabilities used in our model can be estimated from PW segmented corpora.

3.2. Methodology

In our statistical model, both length of the word and the tagging from POS are two essential features to predict PWs and we try to maximize $Arg_{PW}\ P(PW|LW)$.
$Max_{\ PW}\ P(PW|LW) = Max_{\ T}\ P(T|LW) = MAX_{\ T}\ P(t_1|LW)*P(t_2|t_1,LW)*\ldots*P(t_n|t_1,t_2,\ldots t_{n-1},LW)$, where T={t1, t2, …,tn} is the tag sequence of a PW sequence. For PW generation, an input sentence would be segmented into LWs with POS tagging first. Equation (1) is our statistical model. $P(t_m \mid C_{m-1}, L_{m-1} + f(t_{m-1})L_{m-2}, C_m, L_m)$ is the probability of the m-th LW that it should be combined or isolated. t_m indicated the method of the combination of the m-th LW. C_m indicates the POS-category of the m-th LW , and the length of LW denoted as L_m. A Boolean function $f(t_{m-1})$ was used to decide whether the length of the m-2th LW would be considered, if t_{m-1}=L, return 1, else return 0.

$$P(t_m|t_1,t_2,\ldots t_{m-1},LW) \sim= P(t_m \mid C_{m-1}, L_{m-1} + f(t_{m-1})L_{m-2}, C_m, L_m) \quad (1)$$

$$f(t_{m-1}) = \begin{cases} 1, t_{m-1} = L \\ 0, t_{m-1} = M \end{cases}$$

$$P(t_m \mid C_{m-1}, L_{m-1} + f(t_{m-1})L_{m-2}, C_m, L_m) = \frac{Count(t_m, C_{m-1}, L_{m-1} + f(t_{m-1})L_{m-2}, C_m, L_m)}{Count(C_{m-1}, L_{m-1} + f(t_{m-1})L_{m-2}, C_m, L_m)} \quad (2)$$

Backoff strategy was used as follows: equation (1) is the main equation with features of categories and the length of the word, if the probability is 0, then goes to equation (3), here we ignore the category of the m-1th word. The probability of equation (3) may still get 0, then equation (4) would be used, but the length of the m-2th word remained important to be reserved. The equation (5) is the last probability that we can get. In equation (4) and (5), the length of LW should be constrained. If the length of LW was 4 or more than 4 syllables, the length of LW would be 4. A sequence of each LW may have two alternatives, combine or not to combine. In a sequence of N LWs, we may have 2^N paths, the optimal combination was therefore found by dynamic programming.

$$P(t_m \mid L_{m-1} + f(t_{m-1})L_{m-2}, C_m, L_m) \quad (3)$$

$$f(t_{m-1}) = \begin{cases} 1, t_{m-1} = L \\ 0, t_{m-1} = M \end{cases}$$

$$P(t_m \mid L_{m-1} + f(t_{m-1})L_{m-2}, L_m) \quad if\ L_m \geq 4, L_m = 4 \quad (4)$$

$$f(t_{m-1}) = \begin{cases} 1, t_{m-1} = L \\ 0, t_{m-1} = M \end{cases}$$

$$P(t_m \mid L_{m-1}, L_m) \quad \begin{matrix} if\ L_m \geq 4, L_m = 4 \\ if\ L_{m-1} \geq 4, L_{m-1} = 4 \end{matrix} \quad (5)$$

3.3. Results and analysis

Cross-validation was used in our experiments. The 599 paragraphs were split into six subparts. 6 data sets were tested in turn. Each data set had 5 subparts about 499 paragraphs as training data, and one subpart about 100 paragraphs as testing data. Performance evaluation is based on precision, recall and F-score.

$$\text{Precision} = \frac{\text{numbers of correctly predicted PW boundary}}{\text{numbers of predicted PW boundary}} \quad (6)$$

$$\text{Recall} = \frac{\text{numbers of correctly predicted PW boundary}}{\text{numbers of real PW boundary}} \quad (7)$$

$$F - score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

The average results from the statistical model are showed in Table 5.

Table 5. Results of statistical model

	F01	F02	M01	M02
Recall	86.99%	88.55%	85.92%	85.24%
Precision	92.46%	91.28%	89.70%	90.72%
F-score	89.64%	89.89%	87.76%	87.88%

Compare to Table 4, the results by the statistical model have slight edge over the rule-based model. To further improve the model, we looked further into the speech data for prosodic patterns and found that quadri-syllabic LWs such as idiomatic phrases were usually broken into two disyllabic PWs. For example, the idiom "一脈相傳" should be segmented into two PWs "一脈" and "相傳". Table 6 showed the results of the statistical model with adjustment for multi-syllabic PWs where precision, recall and F-score were improved though by a slight margin.

Table 6. Results in statistical model with long LW segmentation

	F01	F02	M01	M02
Recall	88.46%	89.99%	87.48%	86.67%
Precision	91.83%	90.59%	89.20%	90.10%
F-score	90.11%	90.28%	88.31%	88.34%

In order to test how our model performs in comparison with other models, we replicated the models by Qian et al

[8]. Their two statistical rule-based methods, Simple POS set and Word length indicated POS set, used bi-gram statistic of POS and length of word to determine whether two consecutive LWs form a PW. To decide whether two LWs should be merged, a threshold θ, was used. In our replication, best results were obtained when θ=0.5. Two features, POS tagging and the length of the word were taken into account for both statistical models. The greatest difference existed in the constraint of the length of the word, our statistical model could combine more than 3 LWs.

The results are showed in Tables 7 and 8. Comparison between our statistical model and Qian et al's models (Tables 6 vs. Tables 7 and 8) indicated that our precision was much better than both Simple POS set and Word length indicated POS set. Our F-score was better as well. Overall, our model yielded better performance in general.

Table 7. Replication results of Simple POS set (Qian et al 2001) with the threshold at 0.5

	F01	F02	M01	M02
Recall	90.23%	90.93%	87.75%	88.61%
Precision	82.01%	80.13%	78.32%	80.62%
F-score	85.88%	85.15%	82.72%	84.38%

Table 8. Replication results of Word length indicated POS set (Qian et al 2001) with the threshold at 0.5

	F01	F02	M01	M02
Recall	90.29%	91.15%	87.99%	88.76%
Precision	83.13%	81.39%	79.57%	81.81%
F-score	86.53%	85.97%	83.54%	85.12%

4. CONCLUSION AND FUTURE WORKS

We believe that successful prediction of a lower level prosodic unit PW is a first step towards building prosody models from text. Nevertheless, we conceive the production of speech rhythm as a multidimensional task, phrase grouping of the upper layers in the rhythmic structure and the temporal organization of the speech seem to be the main impediments. Rhythm is defined as an organization of meaning across the alternation of accents, sound effects, and prosodic organization [10, 11]. In other words, rhythm should be seen as a key prosodic tool for signaling an overarching semantic organization. The higher layers of the rhythmic structure reveal the more intensive correlation with the semantics. To clarify the exact nature of relations between semantics and prosody become an essential work in the future.

In this paper, by means of statistical approach, results from our statistical model rather than rule-based model perform quite good prediction in word grouping. Also, the data-driven statistical model allowed more room for extensions than the linguistic rule-based model. In next stage, we will focus on how phrase grouping could be piled up from word grouping and disambiguate semantic relation in phrase grouping.

6. REFERENCES

[1] Tseng, C, S. Pin and Y Lee, "Speech Prosody: Issues, Approaches and Implications", in Fant, G., H. Fujisaki, J. Cao and Y. Xu Eds. *From Traditional Phonology to Mandarin Speech Processing*, Foreign Language Teaching and Research Process, Beijing, China, 2004, pp. 417-438

[2] Zellner Keller B. "Revisiting the Status of Speech Rhythm." in Bernard Bel & Isabelle Marlien (eds.), 2002. *Proceedings of the Speech Prosody 2002 conference*, Aix-en-Provence: Laboratoire Parole et Langage. pp. 727-730, 11-13 April 2002

[3] E., Keller, B. Zellner, "A timing model for fast French", *York Papers in Linguistics*, 17, University of Yorks. Pp. 53-57, 1996

[4] F. Chou, C. Tseng and L. Lee, "Automatic Segmental and Prosodic Labeling of Mandarin Speech Database", *Proc. Of ICSLP*1998, Sydney, Australia, 1998

[5] C. Tseng, "Investigating Mandarin Chinese Prosody through Speech Database", *Proc. Of 1999 O-COCOSDA*, Taipei, Taiwan, R.O.C. pp. 65-68, 1999

[6] F Chou, C. Tseng, L. Lee, "Automatic Corpus Processing for Mandarin Speech Synthesis", *Proc. Of 1999 O-COCOSDA*, Taipei, Taiwan, R.O.C. pp.141-144

[7] C. Tseng and F. Chou, "A Prosodic Labeling System for Mandarin Speech Database", *Proc. Of XIVth International Congress of Phonetic Science*, San Francisco, California. Pp. 2379-2382

[8] Y. Qian, M. Chu, and H. Peng, "Segmenting Unrestricted Chinese Text into Prosodic Words Instead of Lexical Words", *Proc. of ICASSP2001*, Salt Lake City, 2001

[9] W., Ma, K., Chen, "A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction," *Proceedings of ACL workshop on Chinese Language Processing 2003*, pp. 31-38, 2003

[10] Dessons G., Meschonnic, H., "*Traité du rythme*", Dunod, Paris, 1998

[11] B., Zellner (Forthcoming). "Prediction of Temporal Structures for Various Speech Rates", In, N. Campbell. (Ed.) *Volume on Speech Synthesis*. Springer-Verlag