PREDICTING PROSODY FROM TEXT

Keh-Jiann Chen[1], Chiu-yu Tseng[2], Chia-hung Tai[1]

[1] Institute of Information Science, Academia Sinica, Taipei
[2] Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei
kchen@iis.sinica.edu.tw, cytling@sinica.edu.tw, glaxy@iis.sinica.edu.tw

Abstract. In order to improve unlimited TTS, a framework to organize the multiple perceived units into discourse is proposed in [1]. To make an unlimited TTS system, we must transform the original text to the text with corresponding boundary breaks. So we describe how we predicate prosody from Text in this paper. We use the corpora with boundary breaks which follow the prosody framework. Then we use the lexical and syntactic information to predict prosody from text. The result shows that the weighted precision in our model is better than some speakers. We have shown our model can predict a reasonable prosody form text.

1 Introduction

In order to improve the prosody of unlimited TTS, a framework to organize the multiple perceived units into discourse is proposed in [1]. Some preceding study regards fluent speech as a succession of independent sentences. If we only apply succession of discreet and often declination intonations to unlimited Mandarin Chinese TTS (text-to-speech synthesis), the unlimited TTS can not produce satisfactory fluent speech prosody. However in our framework, these units are not equal for perception. Some perceived units are grouped by a higher-level unit. The higher-level unit governs and constrains the lower-level units. Lower-level units in different position presented different acoustic patterns rather than being regarded as the same prosodic unit. In other word, this is a hierarchical framework. As Figure 1 illustrated, these units located inside different levels of boundary breaks across speech flow. The boundaries are annotated using a labeling system that annotated small to large boundaries with a set of five break indices. i.e., B1–B5. The framework can also be viewed as a tree-branching organization of multi-phrase prosody.

From bottom up, the layered nodes are syllables (SYL), prosodic words (PW), prosodic phrases (PPh) or utterances, breath group (BG) and prosodic phrase groups (PG). These constituents are, respectively, associated with break indices B1–B5.

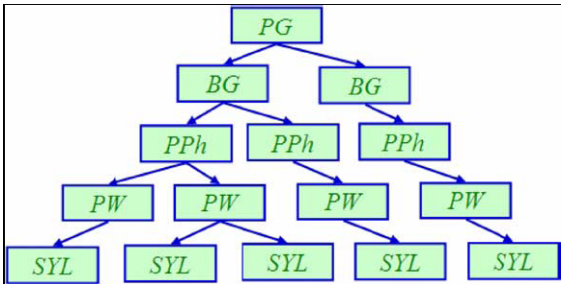

Fig. 1. A schematic representation of the prosody framework

Table 1. Index of Break Hierarchy and Transcription Consistency

	Definition	Characteristics
B1	normal syllabic boundary	Usually with no identifiable pauses, but more of a psycholinguistic unit for native speakers.
B2	prosodic word boundary	Perceived as a boundary where a slight tone of voice change usually follows.
B3	prosodic phrase boundary	A clearly perceived pause.
B4	breath group boundary	Perceived end of exhale cycle followed by inhaling to begin another breathing cycle. It could be where a speech paragraph ends where trailing occurs with final lengthening coupled with weakening of speech sounds. But the speaker may still go on by breathing but not ending the speech paragraph.
B5	prosodic group boundary	A complete speech paragraph ends by final lengthening coupled with weakening of speech sounds. The speaker makes a complete stop, take a new breath, and begin a new speech paragraph.

B1 denotes syllable boundary at the SYL layer where usually no perceived pauses exist. B2 is a perceived minor break at the PW layer. B3 is a perceived major break at the PPhs layer. B4 denotes a boundary break when the speaker is out of breath and takes a full breath and breaks at the BG layer. B5 is when a perceived trailing-to-a-final-end occurs and the longest break follows. Table 1 shows the definition of all breaks and the characteristics of those. When acoustic parameters of unlimited TTS are strung into speech flow, they must adjust and modify to derive satisfactory fluent speech prosody. How acoustic parameters adjust and modify is according to which level of boundary breaks they located inside. To make an unlimited TTS system, we must transform the original text to the text with corresponding boundary breaks. So we describe how we predicate prosody from Text in this paper.

To predict prosody from text we need the corpora with boundary breaks. We describe the corpora we used in more detail in Section 2. The prosody production models are described in the section 3. The section 4 shows experimental results.

2 Materials Used--Text VS. Speech Corpora

COSPRO 01 and 05 speech data from Sinica COSPRO Database [2] were used. COSPRO 01 contains 599 paragraphs (24803 syllables in total) ranging from 2-character simple sentences up to 181-character complex sentences. COSPRO 05 consisted of readings of 26 paragraphs (11592 syllables in total) of text ranging from 85 to 981 characters per paragraph rearranged from the COSPRO 01 for frequency and phonetic controls. The two sets of text overlapped 88%. Four native untrained speakers (2 males M01, M02 and 2 females F01, F02) read the COSPRO 01 at the average speech rate of 304 ms/syllable in COSPRO 01. Another two radio announcers (1 male and 1 female) read the 26 longer paragraphs at the average speaking rate of 200 ms/syllable in COSPRO 05. Segmental identities were first automatically labeled using the HTK toolkit and SAMPA-T notation, then hand tagged by trained transcribers for perceived boundary breaks using the Sinica COSPRO Toolkit [3]. All labeling was also spot-checked by trained transcribers.

The majority of PWs were disyllabic (67%) and tri-syllabic (25%) [4]. Although the length of PPhs are mostly under 10, the variations of PPhs were more complicated than PWs. Figure 2 shows the distribution of the length of PPhs in COSPRO 01.

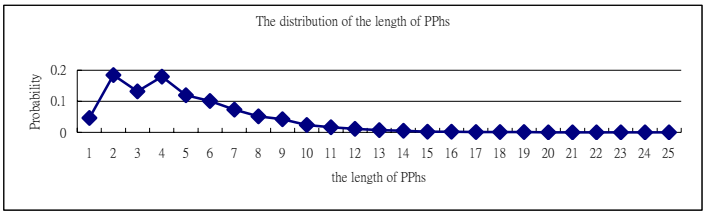

Fig. 2. The distribution of the length of PPhs

The length of PPhs seems not a suitable feature to predict the PPhs. Instead syntactic structures are somewhat related to the structures of PPhs. They do have some common patterns shown in the prosodic structure annotated speech data and syntactic annotated text. For instance, the prosody structure of the sentence "中油公司高級主管昨天表示" is shown in Figure 3 and its syntactic structure is shown in Figure 4. The first PPh is coincident with the NP structure and the second PPh is a partial VP structure. Our predicting models are trained from the prosodic and syntactic structure aligned parallel corpora. We will present our prediction models in more details in Section 3.

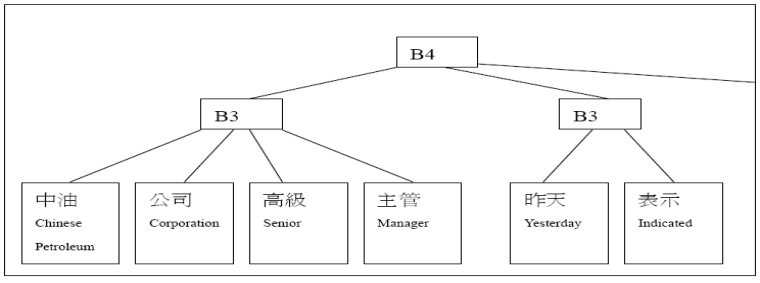

Fig. 3. .Part of prosody structure for "中油公司高級主管昨天表示"

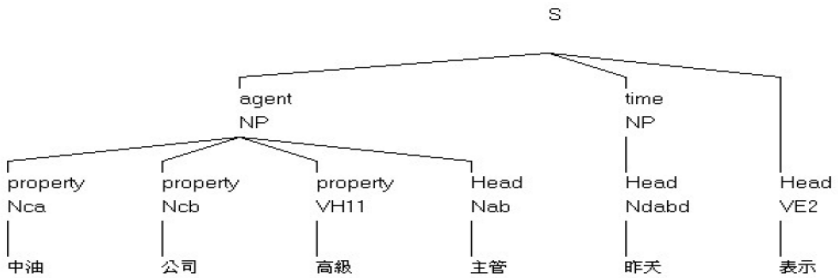

Fig. 4. Syntactic structure of "中油公司高級主管昨天表示"

3 The Model for Predicting Prosody from Text

We propose a series of bottom up models to predict prosody from text. We use word segmentation program (http://rocling.iis.sinica.edu.tw/CKIP/wordsegment.htm), POS tagger and Chinese parser [5] to retrieve the syntactic and lexical information of sentences for training and applying our models. The major features used in our models include lexical words, part-of-speeches (POS), syntactic structures, and lengths. Prediction of B1 is obvious, since character boundaries are natural boundaries of SYL in Chinese. For predicting PWs, length of the word and POS are two essential features. Since there is no gold standard for PW, a consistency checking with human speeches is performed. An average performance of 90% F-score is achieved for PW prediction. Comparing with the average consistency F-score of 92% among human speakers, the model performs quite well. The detail PW model is in [4].

For PPh prediction, A conditional probability $P(B3 \mid Ph, PL, MPhYN, B, X)$ of a location X to be a PPh boundary B3 was proposed to model the production of PPhs. Where the conditional feature *Ph* is the name of the phrase contained the prosodic word at left of *X*. *PL* is the length of *Ph*. *MPhYN* is a value of yes/no which indicates whether the *Ph* is an embedded phrase or not. *B* is the boundary type of *X*. There are four different types. They are "| |", "| (", ") |", and ") (". "| |" means that the PWs in the both sides of *X* are in the same phrase. "| (" means that *X* is the left boundary of an embedded phrase. Similarly, the ") |" means that *X* is the right boundary of an embedded phrase. The ") (" means that *X* is located between two embedded phrases.

Table 2. The occurrence probabilities of B3 at different types of boundaries

Boundary representation	The probability of PPh
| |	0.214669
| (	0.316559
) |	0.380176
) (	0.589354

The probabilities of being a PPh boundary for different boundary types observed from COSPRO corpus are demonstrated in Table 2. The probability of PPh in ") (" is much higher than others which means that having a PPh break between two complete syntactic units is preferred.

$P(B3 \mid Ph, PL, MPhYN, B, X)$ can be derived from annotated training corpora by Maximum-likelihood or Maximum Entropy estimations. The complete PPh production model is shown below.

PPh Production Model:

Input: A sequence of sentences with word, POS, PW and syntactic structure annotated.

Algorithm: For each input sentence,

Step 1. Assign B3 to every place with punctuation markers of comma, period, question mark, exclamation mark, and semicolon.

Step 2. For each PW boundary X, derive the value of $P(B3 \mid Ph, PL, MPhYN, B, X)$.

Step 3. Determine the number of PPhs m in the input sentence by a control parameter n which is an integer value proportional to the intended speech rate.
m=[Length of sentence/n] ,where n are usually set to 5 or 7 for normal speed.

Step 4. Assign m number of B3 at X1,X2,…,Xm which have the highest accumulated probabilities of $\sum_{1}^{m} P(B3 \mid Ph, PL, MPhYN, B, Xi)$, such that no resulting PPh contains only single PW.

The Procedure of prosody prediction
(1) Input Data= text of multiple paragraphs
(2) Text with word boundaries, POS and syntactic structure annotation is produced by a syntactic parser.
(3) Identify PWs for each Sentence in Input Data
(4) If (SentenceLength(Sentence)>10) then
apply PPh model to identify PPhs.
(5) Identify PG.
(6) Identify BG.
(7) Output Data: text of multiple paragraphs with boundary breaks

Fig. 5. The algorithm for predicting prosody from text

Figure 5 shows the algorithm of producing complete prosody. In Step (1) of the algorithm, we read in a text of multiple paragraphs with punctuations. In Step (3) we use a PW model [4] to predict PW boundaries B2. For long sentences, which are longer than 10 characters, the PPh production algorithm will be applied to mark B3. After we decide PPhs, at step (5) we mark B5 before identify breath group BG, since the location of a B4 depends on the length of PG and speech rate. Since PG is a discourse unit and usually is a complete paragraph, naturally we use periods and

question marks to predict PG. On the other hand, BG is caused by physical constrain of human exhale cycles. It is obvious that predicting of breath groups depends on speech rate and length of PGs. Normally, 20~30 syllables are produced in each exhale cycle. Table 3 shows the statistics of the 4 speakers on COSPRO 01 data. Within a long PG, we need to find natural stopping points for inhale and next exhale cycle. For every PG, we use following heuristic rules to mark B4 in the step (6).

(1) Every end of a sentence is a possible candidate of B4 and obviously B5 is mandatory a B4.
(2) For each B4 candidate, if the number of characters to the next B5 is greater than 40 or the followed sentences has more than 30 characters, then we mark it as B4.

After those steps, we had text of multiple paragraphs marked with different levels of boundary breaks as output file. Then the prosody of the text is established by the boundary breaks.

Table 3. Statistics of the lengthes of BGs of the corpora COSPRO 01 and 05

Corpus	Speaker	Maximum	Minimum	Average	Most
COSPRO01	F01	92	3	25.5	23
	F02	104	8	32.3	23
	M01	148	1	27.5	23
	M02	109	3	22.3	17
COSPRO05	F051	133	6	29.8	25

4 Experimental Results and Evaluations

Cross-validation was applied on the data COSPRO 01. The COSPRO 01 was split into six subparts 100 paragraphs each. Each subpart was tested in turn with other 5 subparts as training data. We also used COSPRO 05 as testing data for open test.

4.1 Evaluation Metrics

To evaluate the performances of prediction models, we propose three different sets of evaluation metrics. Each set of evaluation metrics consists of recall, precision, and balanced F-score, the harmonic mean of precision and recall, but with slightly different senses.

$$\text{Precision} = \frac{\text{number of correctly predicted boundary breaks}}{\text{number of predicted boundary breaks}}$$

$$\text{Recall} = \frac{\text{number of correctly predicted boundary breaks}}{\text{number of real boundary breaks}}$$

$$\textit{Balanced } F-score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The first set of evaluation metrics takes each human performance as standard. As a result it shows the degree of consistency between machine and human performances. We proposed a second evaluation metric called weighted precision to evaluate the quality of our prediction. The idea is that more speakers agree upon the boundary break which gets more weight. If one speaker agrees with the position, we give the weight 0.25. If two speakers agree with the position, we give the weight 0.5. If all four speakers agree with the position, we give the full weight 1. The third set of metrics of evaluation is called "general precision" which a prediction of break type matching any one of speaker is considered correct.

4.2 Evaluation results and analysis

We evaluate the performances of each individual model and compare them with human produced prosody. The first section is the evaluation results of PPh model, and the next section contains evaluations for B4 (breath groups) and B5 (prosodic phrase groups).

4.2.1 The evaluation results of PPh model

PPh model was applied by controlling parameter of speech rates at two different values *n*=5 and 7. The results of cross validation on COSPRO 01 with respect to four different speakers F01, F02, M01, and M02, are showed in Table 4 and Table 5. The F-Score of our model is around 73%. We also calculate the consistency among speakers. The consistency among four speakers of prosodic phrases on COSPRO 01 is shown in Table 6. The F-Score of speaker's consistency on COSPRO 01 are around 75%. The Results show that our model performs almost comparable with the human speaker's consistency.

Table 4. The result of n=5 for PPhs on COSPRO 01

	F01	F02	M01	M02
Recall	0.6990	0.6966	0.7913	0.6628
Precision	0.7782	0.7815	0.6632	0.7918
F-score	0.7365	0.7366	0.7216	0.7216
Weighted Precision	0.80			
General Precision	0.88			

Table 5. The result of n=7 for PPhs on COSPRO 01

	F01	F02	M01	M02
Recall	0.6679	0.6587	0.7715	0.6258
Precision	0.8423	0.8413	0.7342	0.8486
F-score	0.7450	0.7389	0.7524	0.7204

Weighted Precision	0.85
General Precision	0.92

Table 4 and 5 also show the weighted precision in different *n*, and the weighted precisions are over 80%. In Table 7, the lowest weighted precision of four speakers in CORPOS 01 is 80%, and the weighted precision in our model is comparable with human speakers. Regarding the general precision of our model, over 88% of our predictions are marked as PPh by at least one of those four speakers. These evaluation results show that our model performs well and can consistently identify prosodic phrases.

Table 6. The consistency of PPhs among human speakers on COSPRO 01

	F01	F02	M01	M02
Recall	0.789	0.793	0.643	0.815
Precision	0.747	0.726	0.857	0.711
F-Score	0.763	0.754	0.735	0.756

Table 7. The weighted precisions of B3 among human speakers on COSPRO 01

F01	F02	M01	M02
0.83811	0.82513	0.91738	0.80268

We use the COSPRO 05 for our open test. Table 8 shows the evaluation results of PPh model in comparing with two speakers M051 and F051 at different speech rates *n*. The F-Score of our model is around 78%. It is close to the F-Score of human speaker's consistency of 80% shown in the Table 9. Because there are only two speakers in COSPRO 05, we do not evaluate the weighted precision and general precision.

Table 8. The evaluation results of PPhs model at different speech rates on COSPRO 05

M051	Recall	Precision	F-score
n=5	0.7791	0.6398	0.7026
n=10	0.7431	0.7558	0.7494
n=15	0.6972	0.8444	0.7638
F051	Recall	Precision	F-score
n=5	0.7945	0.6444	0.7116
n=10	0.7644	0.7678	0.7661
n=15	0.7280	0.8707	0.7930

Table 9. The human speaker's consistency on PPhs production at COSPRO 05

	Recall	Precision	F-Score
M051-Based	0.801	0.811	0.806

4.2.2 The evaluation results of predicting BGs and PGs

Because COSPRO 01 is not composed by complete text units, we use only COSPRO 05 to evaluate. Table 10 shows the results of BGs on COSPRO 05, and the F-scores are around 55%-60%. The human speaker's consistency of BGs in COSPRO 05 shown in Table 12 is about 0.59. The inconsistency of BGs may be due to the physical difference between the human speakers and the broader scope of BGs. The variation of BGs makes the difficulty of prediction. Our predictions of BGs are close to the human speaker's consistency of BGs.

Table 10. The results of BGs prediction on COSPRO 05

	Recall	Precision	F-score
B4-M051	0.5723	0.5360	0.5535
B4-F051	0.6064	0.5994	0.6028

Table 11 shows the result of PGs on COSPRO 05. The human speaker's consistency of PGs on COSPRO 05 is 63%. Compare to Table 12, the F-score of our prediction is much lower than the F-Score of consistency. The main reason is we do not have paragraph mark in the text. So we mark every period punctuation as prosodic phrase group. It results in the low precision in PG prediction. Another reason may be that the trained transcribers used not only text information but also acoustic information. We only use text information, so the precisions of our prediction are much lower.

Table 11. The result of PG predictions on COSPRO 05

	Recall	Precision	F-score
B5-M051	0.7822	0.3222	0.4564
B5-F051	0.75	0.3388	0.4668

Table 12. The human speaker's consistencies of BGs and PGs at COSPRO 05

	Recall	Precision	F-Score
B4	0.609	0.577	0.592
B5	0.669	0.610	0.638

5 Conclusions and Future Works

This is the first attempt to build a model to predict prosody from text. We used the syntactic structure of text to predict prosodic phrase and used heuristic rules and punctuations to predict breath group and prosodic phrase group. Because the low consistency means the variety of possibility, it makes the difficulty to predict the boundary breaks. Our weighted precision for PPhs on COSPRO 01 is better than some speakers. We have shown our model can predict a reasonable prosody form text. Although we have predicted the prosody model from text, how to use semantic information to group prosodic phrase group is another way to improve our predictions. Using semantic information to predict end of paragraph may help to predict prosodic phrase group. Because we only use punctuation information to determine the end of paragraph, how to use semantic information to detect the change of topic will be our future research.

References

1. Chiu-yu Tseng, Shao-huang Pin, Yeh-lin Lee, Hsin-min Wang, and Yong-cheng Chen (2005). “Fluent speech prosody: framework and modeling,” Speech Communication, Vol.46, issues 3-4,(July 2005), Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation, 284-309.
2. Chiu-yu Tseng, Yun-Ching Cheng and Chun-Hsiang Chang (2005). “Sinica COSPRO and Toolkit—Corpora and Platform of Mandarin Chinese Fluent Speech” Proceedings of Oriental COCOSDA 2005,(Dec. 6-8, 2005), Jakarata, Indonesia, 23-28
3. Sinica COSPRO and Toolkit: http://www.myet.com/COSPRO/
4. HuaJui Peng, Chiching Chen, Chiuyu Tseng, Kehjiann Chen ,”PREDICTING PROSODIC WORDS FROM LEXICAL WORDS--A FIRST STEP TOWARDS PREDICTING PROSODY FROM TEXT”, International Symposium on Chinese Spoken Language Processing, ISCSLP, 2004.
5. Yu-Ming Hsieh, Duen-Chi Yang, and Keh-Jiann Chen. 2005. Linguistically-motivated grammar extraction, generalization and adaptation. In Proceedings of the Second International Join Conference on Natural Language Processing, pages 177-187, Jeju Island, Republic of Korea.