A NEW PROSODIC STRENGTH CALCULATION METHOD FOR PROSODY REDUCTION MODELING

Honglei Cong[1,2], Zhiyong Wu[1,2], Lianhong Cai[1,2] and Helen M. Meng[1]

[1] Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems, Graduate School at Shenzhen, Tsinghua University, Shenzhen

[2] Department of Computer Science and Technology, Tsinghua University, Beijing

ABSTRACT

To improve the naturalness of synthetic speech, prosody models in text-to-speech (TTS) system should be able to describe different prosody variations in natural speech. In this paper, prosody variation patterns behind the partial reduction phenomena are analyzed. In order to model the prosody reduction effect and incorporate it into the prosody model for speech synthesis, prosodic strength is introduced and a new prosodic strength calculation method is proposed. The method aims to model the sentence planning of prosody reduction and is based on the concept that the objective of prosodic strength should complete the planned target of the speech unit. The approach on how to integrate prosodic strength into speech synthesis system is also introduced. Experiments show that the estimated prosodic strength values by the proposed method have good correlations with both prosody structure and acoustic features.

Index Terms— speech synthesis, prosodic strength, prosody reduction, prosody model

1. INTRODUCTION

One important work in high naturalness speech synthesis is prosody prediction. It is usually performed in two-steps [1]. The first step is prosodic events prediction, which is done by text analysis module; the second step is acoustic prosodic parameters prediction, which is the responsibility of prosody model. Most prosody models in current speech synthesis systems are based on data-driven methods, such as decision tree [2], probability-based statistical model [3], etc. They predict acoustic prosodic parameters based on the context clustering technique [4]. The context information used in clustering mainly includes prosody structure information (i.e. prosodic word, prosodic phrase, sentence, etc), pre-syllable and post-syllable type, the information about the target itself, etc. Because of the limited context information, these methods can only predict some common prosodic variations, and are unable to describe those relatively sophisticated prosody phenomena, such as prosody reductions, in natural speech.

It is widely agreed that, in natural human speech communication, different syllables in one prosodic rhythm are likely to get different degrees of emphasis or weakening. And these emphasisses and weakenings make human speech more colorful and rich. From the stresses and rhythmicity in speech, we can learn more about what the speaker would like to express, such as the speaker's attitude, intention, etc. In fact, this extra information is delivered by speech prosody. And from observations of laboratory speech recordings, we find that there always are some sorts of intonation variations in natural speech, although the speaker had been told to speak in a neutral way and make no special intention about the script. Moreover, there are always some natural emphases or semantic pivots in sentence text (e.g. turn words), they could naturally get stressed in speech. And for those less important elements (e.g. accessorial words), they may get reduced naturally. This means that, in natural speech communication, the semantic inequality in text would inevitable give rise to prosodic inequality effect among different speech units. We believe this phenomenon is one indispensable part of natural speech prosody. Hence, to make synthetic speech more natural, the prosody model of the speech synthesis system should be capable to describe such prosody variations.

The outline of this paper is as follows: Section 2 analyzes the prosody reduction effects in natural speech and explains the concept about prosody strength. Section 3 presents a new prosodic strength calculation method based on the target prosody model. Experiments and their results analysis are then given in Section 4, which illustrates the effectiveness of our method. Section 5 presents how the prosodic strength is integrated into speech synthesis system. Finally, Section 6 summarized the paper.

2. PROSODY REDUCTION ANALYSIS

2.1 Background on prosody reduction and prosodic strength

The prosody reduction phenomena in human speech have been studied from various aspects. [5] studies the phenomena from the difference between laboratory speech and spontaneous speech, and points out that one problem in current prosody models is the missing link between sentence planning and articulation gestures. This missing link also causes the difficulty in modeling the prosody reduction behavior in human speech.

The prosody reduction phenomena are explained with the Stem-ML prosody model in [5]. In Stem-ML model [6], each speech unit is associated with a weight parameter to

represent the speaker's effort which is put in the articulation of the unit. It is generally held that, in natural speech, the speakers always like to make use of least effort to achieve best articulation effect. From the perspective of "weight", it is a balancing action between more accurate production and less weight. During speech communication, there is a weights planning process before articulation, and then each speech unit is articulated according to its planned weight. In [7], a term *prosodic strength* is introduced to represent this unit weight, and a prosodic strength calculation method is proposed. The proposed calculation method *only* makes use of pitch contour information. In the method, the pitch contour of the speech unit (e.g. syllable) is first fitted by Stem-ML model to obtain a series of Stem-ML parameters that fit the pitch contour best. The parameters are then compared with the tone templates. The distance between the best fitted Stem-ML parameters and the tone template is finally used to calculate the prosodic strength values of the syllable. Bigger distance means more seriously the syllable articulation deviates from the template, and means bigger prosodic strength is needed to produce the syllable.

2.2 Our method for prosodic strength calculation

We hope to improve the prosody model of our text-to-speech (TTS) synthesis system by adding the concept of prosodic strength. The prosody model of the TTS system is based on the target prosody model [8] rather than the Stem-ML model. In order to incorporate prosodic strength into the target prosody model, a new prosodic strength calculation method is required.

Our prosody model agrees on most basic concepts of Stem-ML model. We further suppose that there is a latent target associated with each speech unit. And the objective of prosodic strength is to complete the target. From this point of view, the process of speech articulation is to complete all speech units' targets as good as possible meanwhile maintaining the overall consumption of prosodic strength as less as possible.

Two steps are involved in our method to calculate the prosodic strength of a speech segment (i.e. syllable). First, the target completion degrees of the syllables are estimated. Then the prosodic strengths of the syllables are calculated based on the assumption that speech units will have similar prosodic strength values when they are with the same final and tonal type (see section 3) and have similar target completion degrees.

3. PROSODIC STRENGTH CALCULATION

3.1 Syllable target completion degree

Chinese is a tonal language. Each Chinese syllable can be subdivided into an optional initial (i.e. the consonant that starts the syllable), the final (i.e. the vowels that ends the syllable), and the tone. Tone plays a very important role in Chinese speech communication. Before calculating the prosodic strength, a function is defined to measure the syllable target completion degree for each tone.

The realization of tone is crucial in Chinese syllable articulation. In acoustic parameter layer, tone is mainly manifested by some pitch variation patterns. Therefore, in our method, all the syllable tone completion degrees are estimated from the syllable pitch contours, and are taken as the syllables' target completion degree.

There are four tones in Chinese Mandarin (neutral tone is not considered). Tone1 is high tone, tone3 is low tone, and their pitch contours are mainly in the level state. For these two tones, a good pitch realization should have a level pitch value with small variations (high average pitch value for tone 1, and low for tone3). Tone2 is raising tone, tone4 is falling tone, and they have dynamic targets. Their good realization is supposed to have a large pitch range. Based on these facts, the target completion degree calculation functions are defined as:

批注 [c1]: a ?

$$\text{tone1: } (p/p_m + s_m/s)\cdot t/t_m \,, \quad (1)$$

$$\text{tone2: } (p_{high}/p_{m_{high}} + r/r_m)\cdot t/t_m \,, \quad (2)$$

$$\text{tone3: } (p_m/p + s_m/s)\cdot t/t_m \,, \quad (3)$$

$$\text{tone4: } (p_{m_{low}}/p_{low} + r/r_m)\cdot t/t_m \,, \quad (4)$$

where p is the average syllable pitch value, s is the variance of syllable pitch value, p_{high} is the highest syllable pitch value, p_{low} is the lowest syllable pitch value, r is the pitch range, t is the syllable duration. The Chinese syllables ended with the same type of finals will have similar acoustic prosodic representations. Hence the average mean values are considered, where s_m is mean pitch variance, p_m is mean pitch value, r_m is mean pitch range, and t_m is the mean duration of the syllables with the same syllable final type.

3.2 Prosodic strength calculation

In the calculation of prosodic strength value, we use neural function to represent the mapping from acoustic parameters to prosodic strength. All syllables are grouped into different classes by syllable final and tonal types. For each class, a neural function is trained. There are 40 types of finals and 4 tones (neutral tone is not considered) in Chinese Mandarin. Hence, there are 160 prosodic strength functions in total.

Each syllable is divided into five segments equally. Each segment is represented by 4 acoustic parameters, including the mean pitch value p_i, the pitch variation p_i', the pitch value after being removed the superposition pitch contour [9] p_i^*, and the mean of energy E_i. The prosodic strength of each segment can then be represented with:

$$F_i = f(p_i, p_i', p_i^*, E_i) \quad (5)$$

where f is the neural mapping function. The superpositional pitch contour calculation method we used is similar to the method proposed in [9].

The prosodic strength F of a syllable is the summation of the prosodic strengths F_i of all segments multiplied by the syllable duration:

$$F = \text{sum}(F_i) * duration. \quad (6)$$

The training procedure is as follows.

1. Calculate all syllables' target completion degrees using the predefined target completion functions in Equation (1) – (4).
2. Initialize all prosodic strength functions.
3. Calculate all syllables' prosodic strength using their correspondent prosodic strength function.
4. Take statistics of prosodic strength values.
   We assume that syllables of the same final and tonal type with similar target completion degrees should have similar prosodic strength values. Hence, for each syllable *k* in the training set, all the syllables which are in the same type and have similar target completion degrees with syllable *k* are grouped together. The prosodic strength values in this group are modeled by a Gaussian distribute, and are then used to estimate the prosodic strength distribution $N(m_k, \sigma_k)$ of syllable *k*, where m_k and σ_k are the mean and variance of the Gaussian distribution respectively.
5. Recalculate all syllables' prosodic strength values.
   Suppose that the prosodic strength value of the entire prosodic phrases is 10, all syllables' prosodic strength values can then be re-estimated by making use of the syllable prosodic strength distributions calculated in step 4. The re-estimation formula is:
   $$s^* = \arg\max_{s=(s_1,\ldots,s_n)} \prod_{i=1}^{n} N(S_i; m_i, \sigma_i), s.t. \sum s_i = 10 \quad (7)$$
   where *n* is the syllable number in phrase. Actually we can relax the assumption of fixed value of prosodic strength (i.e. 10) for the phrase, and improve the re-estimation formula as:
   $$s^* = \arg\max_{s=(s_1,\ldots,s_n)} \left[\sum_{i=1}^{n} \left(\frac{s_i - m_i}{\sigma_i} \right)^2 - w \left(\sum_{i=1}^{n} s_i - 10 \right)^2 \right] \quad (8)$$
   where *w* is a fixed weight parameter; and it is set to be 0.25 in our work.
6. Replace all the syllables' prosodic strength values with the new estimated value $\mathbf{s}^*$.
7. Train all the prosodic strength functions using the re-estimated prosodic strength values.
8. Check for convergence. If it is not, go back to step 3. Otherwise, end.

4. EXPERIMENT AND DISCUSSION

The speech corpus used in the experiment contains 5000 sentences. The prosodic structures are annotated manually. 7500 prosodic phrases, each with more than 4 syllables, are chosen from the corpus as the training data.

To evaluate the calculation method proposed in the previous section, statistical analysis of the target completion degrees and prosodic strength values is performed. Table 1 shows the result.

It should be noted that a direct comparison of target completion degree between different tonal syllables is not useful, because their calculation methods are different. From Table 1, we can find that target completion degrees of tone1 syllables are relative high, and the target completion degrees of tone3 syllables are low. This is coincident with the fact that the articulations of tone3 syllables in our speech corpus are mostly partial reduced, while most tone shapes of tone1 syllables are good. But the mean prosodic strength value of tone3 syllable is still larger than that of tone1 syllable, although the mean target completion degree of tone3 syllables is just a little more than 0.5. This agrees with our empirical knowledge that the articulation of tone3 syllable is the hardest among the four tones, and the production of tone1 is easy.

Table 1. Statistics of the target completion degrees (Trg) and the prosodic strength (PS) values. (var: variance).

	Trg mean	Trg var	PS mean	PS var
Tone1	1.943	1.957	1.509	0.914
Tone2	1.462	1.921	1.909	0.601
Tone3	0.518	0.795	1.842	0.833
Tone4	1.253	1.062	2.211	0.877

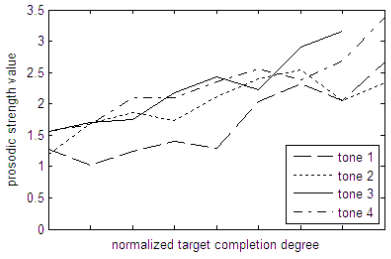

Figure 1. Correlation between prosodic strength and target completion degree.

Figure 1 shows the relations between prosodic strength and target completion degree for different tones. It can be seen clearly that prosodic strength value of each class increases with its target completion degree. Because in calculation, we only make the assumption for the syllables with similar target completion degrees; and no assumption or constrains is made when syllable target completion degrees are different. The correlations prove the function of our calculation method.

To evaluate the relationship between prosodic strength and upper layer prosody events (e.g. prosody structure), we computed the correlation between syllable prosodic strength and syllable position in prosodic words. Syllable position in prosodic word takes 3 values: 1 for the initial position in the word, 3 for the final position and all others are 2. The result is shown in Table 2.

In Table 2, most prosodic strength values decrease as the syllable position increase in the word; especially for tone 4 syllables, their average prosodic strength values in the final position are much lower than the values at other positions. This prosodic strength variation trend is agreed with the metrical patterns reported in [7].

Table 2. Correlation between prosodic strength and syllable position in prosodic word.

	1	2	3
Tone1	1.546	1.467	1.459
Tone2	1.889	1.814	1.993
Tone3	1.896	1.883	1.743
Tone4	2.384	2.247	1.993

To illustrate the relation between prosodic strength and acoustic parameters, the correlation between prosodic strength value and syllable duration is calculated. Figure 2 shows the result.

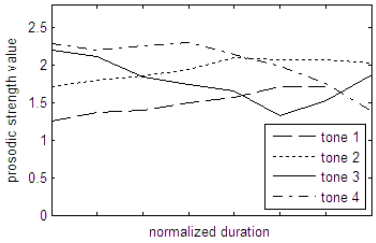

Figure 2. Correlation between prosodic strength and duration

In Figure 2, the prosodic strength values of tone1 and tone2 syllables show good correlation with their syllable durations. But tone4 syllables show an opposite behavior, its prosodic strength values decrease after exceeding a particular duration. This may be caused by the fact that tone4 is a fall tone and may have different characteristic on how prosodic strength completes its dynamic target. Statistical analysis also shows that tone4 syllables with high target completion degrees generally have smaller duration than those syllables whose target completion degrees are low. As for the curve of tone3 in Figure 2, it is the result of partial completions of tone3. In [7], similar correlationship between prosodic strength and duration is also found, and it also reported that this correlationship does not always exist for those phrase final syllables which usually have longer duration but low strength values.

These experiment results show that the calculated prosodic strength values not only have close relationships with acoustic parameter and target completion degree, but also correlate with prosody structure very well. Hence it can serve as the missing link between sentence planning and articulation gestures.

5. USE OF PROSODIC STRENGTH IN PROSODY MODEL FOR SPEECH SYNTHESIS

We have built a model for prosodic strength variation patterns in prosodic word with clustering technique. In the clustering, each prosody word is present with a vector whose elements are the prosody strengths of the syllables it contains, and is labeled by a vector whose elements include position in phrase, positions in sentence, syllable number, and syllable types. Next we plan to study the interactions between neighbor syllables under different prosodic strength context through the correlationship between prosodic strength and target completion degree.

During synthesis, we select units which minimize the overall cost function combining prosody cost and prosodic strength distance. The form of the cost function is:

$$C = w_1 C_t(\mathbf{t},\mathbf{u}) + w_2 C_c(\mathbf{u}) + w_3 C_s(\mathbf{s}_t,\mathbf{s}_u) \quad (9)$$

where $\mathbf{t}$ is the predicted prosody target, $\mathbf{u}$ is the candidate units. $\mathbf{s}_t$ is the predicted target prosodic strength patterns, $\mathbf{s}_u$ is the strength values of $\mathbf{u}$. C is the entire cost, C_t is the target prosody cost, C_c is the concatenative prosody cost, and C_s is the matching cost between the predicted prosodic strength patterns and the prosodic strength of the candidates. C_s ensures the unit selection result comply with the prosody reduction patterns predicted by the prosody model.

6. CONCLUSIONS

In this paper, we first make an analysis of the partial reduction phenomenon in natural speech, and conclude that it is necessary to introduce prosodic strength into current prosody model to improve prosody modeling for more sophisticated prosody variations. Then we introduced a prosodic strength calculation method based on the target prosody model. The experiment results show that the estimated prosodic strength values have good correlations with prosody hierarchy and some acoustic observables such as durations.

7. ACKNOWLEDGEMENTS

This work is jointly supported by the research funds from the Innovation and Technology Fund – Guangdong-Hong Kong Technology Cooperation Funding Scheme (GHP024/06) of Hong Kong SAR, the National Natural Science Foundation of China (60433030) and the National Basic Research Program of China (2006CB303101).

8. REFERENCES

[1] Min Chu, "The Uncertainty in Prosody of Natural Speech and Its Application in Speech Synthesis", Journal of Chinese Information Processing, Vol,18 No.4, 2004.

[2] Antoine Raux, Alan W Black, "A Unit Selection Approach to F0 Modeling and its Application to Emphasis", Proceedings of ASRU 2003, 700-703, 2003.

[3] Xijun Ma, Wei Zhang, Weibin Zhu, Qin Shi, Ling Jin, "Probability Based Prosody Model for Unit Selection", Proceedings of ICASSP 2004, 649-652, 2004.

[4] Alan W. Black and Paul Taylor, "Automatically clustering similar units for unit selection in speech synthesis", in Eurospeech 97, Rhodes, Greece, 1997, pp. 601-604.

[5] Chilin Shih, Greg Kochanski and Su-Youn Yoon, "The Missing Link Between Articulatory Gestures and Sentence Planning", Proceedings of ICPhS 2007, 35-38, 2007.

[6] Greg Kochanski, Chilin Shih, "Prosody modeling with soft templates", Speech Communication 39(3-4), 311–352. 2003.

[7] Greg Kochanski, Chilin Shih, "Quantitative measurement of prosodic strength in Mandarin", Speech Communication 41(4), 625–645. 2003.

[8] Yi Xu, "Speech melody as articulatorily implemented communicative functions", Speech Communication 46, 220-251, 2005.

[9] Shinsuke Sakai, Han Shu, "A Probabilistic Approach to Unit Selection for Corpus-based Speech Synthesis", Proceedings of InterSpeech 2005, 81-84, 2005.