Chinese Prosody and a Proposed Phonetic Model

CAO Jianfen

1. INTRODUCTION

Prosody of a language usually contains following main aspects: rhythm, stress and intonation. Perceptually, prosody is referred to the perceived impression of so-called "輕重緩急，抑揚頓挫", i.e., the cadence of speech sounds. In natural speech, the three aspects are not completely independent, but integrated with each other, and it is achieved mainly through the common ground of modulations in pitch and duration.

Rhythm is mainly related to the timing behavior of speech, rhythmic organization is a chunking strategy referred to both of speech production and perception. Some studies have noticed that this phenomenon is based on human cognitive mechanism (Laver, 1994). Therefore, it is an important step for TTS to model the rhythmic structure.

Stress is another important factor that influence on speech naturalness. Generally, it is manifested through duration elongation and pitch accent of stressed unit. Processing of stress information is also a complex task.

Intonation is a very thorny subject in the prosodic processing. In natural speech, it is mainly characterized by pitch movement of the whole course of utterance. Unfortunately, however, in tone languages like Chinese, the observable F0 pattern of intonation is a multiple combination of roles of many factors (Shih, 1997) , these factors all strongly affect the last output of intonation. Consequently, in this sense, modeling of intonation is actually to model the global prosody of a language.

The present paper try to discuss this topic based on the analyses to spoken Chinese including read sentences, News'announcements and read style discourses. The main attention will be paid to: (1) description on prosodic information summarized from relevant analyses to spoken Chinese (Cao, 1991, 1992, 1994, 1995a, 1995b, 1998, 1999a, 1999b; Yang, 1997; Lu, et al., 2000a, 2000b; Wang, et al., 2000; Zheng, et al., 2000); (2) proposal on a strategy in processing of prosodic information in TTS.

2. PROSODY OF CHINESE

2.1. Rhythmic organization and temporal structure

2.1.1. Rhythmic hierarchy

According to the data obtained from spoken Chinese, the rhythmic elements are organized as a hierarchy in terms of particular coherent property within a unit and certain boundary maker between units (Cao, 1999b). It consists of three main layers, that is, prosodic word (hereafter PW), prosodic phrase(PP) and intonation phrase(IP).

Generally, PW is a disyllabic or trisyllabic chunk, they are the principal building-block of rhythmic structure. In some cases, it also contains a few monosyllabic words and tetrasyllabic chunks, which is formed by adding function word pre- or post- into disyllabic or trisyllabic chunk. Generally, PW is the right domain for some phonological processes, for example, tone sandhi and lexical durational pattern is taken place in this size.

As the intermediate chunk, PP is the most common and functional rhythm unit used in speech production and perception. Generally, it is larger than word but smaller than syntactically defined phrase or clause. The span of this chunk is usually limited within 9 syllables, i.e., about 7 2 syllables, especially when these syllables occur in relatively unstressed positions.

IP is a rhythmic group contains one or more PPs, it is identified to syntactically defined sentence.

2.1.2. The dynamic variation of syllable duration

In natural Chinese speech, syllable duration contains a wide range of variation due to the effects of multiple factors. These factors are mainly coming from two aspects, namely, intrinsic and extrinsic aspects. The intrinsic ones are phonetically and phonologically motivated, it is related to syllable itself including difference of phonological constituents, tonal distinct, lexical stress contrast, and so on. The extrinsic factors are context-dependent and more related to linguistic constraint, which contains the variables in speech rate, speech mood and speech style, syllable location in context, stress status in phrase or sentence, and so on. Among them, the most powerful effect is stress status and location difference of the syllable in sentence. Usually, the duration of stressed syllable can reach as much as 1.4-4.3 times longer than that of unstressed one's, and the duration of PP-final syllable is 1.3 times of that of PP-initial one's in average, while the situation in IP is in reversed direction. The details will be specified in 2.1.4.

2.1.3. Temporal structure of PW

According to the data measured from polysyllabic words, temporal structure of PW can be described as follows:

(1) For disyllabic PW, it is MID-LONG of normal type in stressed position, LONG-MID of normal type in relatively unstressed position, and LONG-SHORT of neutral type in either cases.

(2) For trisyllabic PW, it is LONG-MID-LONG of normal type in stressed position, LONG-MID-MID of normal type in unstressed position, LONG-LONG-SHORT or LONG-SHORT-SHORT of neutral type in either cases.

Roughly speaking, the durational ratio of LONG vs. MID vs. SHORT is about 100: 80: 60.

2.1.4. Temporal structure of PP and IP

From Table I, we can see that in PP, the duration of the first syllable is systematically shorter than that of the last syllable ; however, in IP, the duration of the first syllable is systematically longer than that of the last syllable. This specification indicates that, in Chinese, both of the beginning and ending of utterances exhibit regular adjustment in speech tempo, though the direction of such adjustment in PP and IP is reversed.

<Table I> Distribution of syllable duration in phrase-initial and phrase-final: (1)in PP; (2)in IP

Speaker	Average In general Mean Sd.	Average in Phrase-Initial syllables Mean Sd.	Average in phrase-Final syllables Mean Sd.
Female	179 60	(1) 168 47 (2) 190 56	(1) 298 61 (2) 177 33
Male	155 59	(1) 154 53 (2) 219 51	(1) 250 106 (2) 142 26

In addition, as the basic building-block of PP or IP, the duration manifestation of PW is considerably sensitive to context, especially to the change of sentence stress. Table II lists a set of data measured from read sentences which consists of the same syllable string but different in stress allocation. Specifically, in S1 and S2, the stress is located at the PP of "bu jie shi", and further concentrated on the PW "jie shi" in S1. Whereas, in S3 and S4, it is located at the PP of "zhe shuang xie", and further on the PW "zhe shuang" in S3. From the figures listed in the table, we can observed that, the durational ratio of stressed PWs are standing out clearly.

<Table II>durational ratio(%) of PWs in different stress status in sentences

PWs Sentences	Zhe shuang xie	Bu jie shi	Zhe shuang	Jie shi
S1	48.9	**51.1**	31.3	**42.7**
S2	47.8	**52.4**	30.3	33.8
S3	**55.7**	44.3	**39.7**	36.6
S4	**58.0**	42.0	32.7	36.7

2.1.5. Distribution of silent pause and pre-boundary lengthening at rhythmic boundary

The size of pause interval and pre-boundary lengthening is varied depending on the boundary strength.

<Table III> Temporal distribution at rhythmic boundary:
A(%):Lengthening/shortening in terms of durational ratio of the rhymes in pre-boundary syllables;
B(ms):Duration of silent pauses between PPs and IPs:
a. Sentence end, b. Paragraph end

	A(%)		B(ms)	
Phrase Type / Speech Style	PP-final	IP-final	B/w PPs	B/w IPs
Female Speaker	1.68	0.93	154	a. 538 b. 1112
Male Speaker	1.55	0.86	397	a. 719 b. 2000
News Speech	1.62	0.90	276	a. 629 b. 1020
Declaim Speech	1.33	1.05	59	a. 548 b. 2000

From the figures listed in column B of Table III, we can see that, pause duration is in following order: pause between paragraphs> between IPs > between PPs. This order is identified with that of found in a psycho-acoustical study (Yang, 1997). Usually, there is no silent pause occurred between syllables within a PW, and seldom occurred between PWs within a PP. However, note that, sometimes there may exist a short silent interval between syllables within a PW or between PWs within a PP, but in most of the case, especially in the case of syllable initial is a stop or affricative consonant, this silent interval is phonetically belonging to syllable itself, in stead of a real pause. Pre-boundary lengthening is another important prosodic feature, it is usually

complemented with silent pause. Generally, the maximum lengthening occurs at PP-final, but usually no lengthening taken place at the end of IPs, where a slight shortening occurred in most of the case.

2.2. STRESS

2.2.1. The contrast of lexical stress

In some literatures, lexical stress in Chinese to be classified into three degrees, i.e., stressed, moderate and neutralized. However, experimental investigations have revealed that stress contrast in word level is exited only between normal type and neutral type (Lin, 1984; Cao, 1995). Syllables that received normal type stress have a moderate pitch and duration pattern, while that of the neutral type ones is significantly neutralized and reduced.

2.2.2. Sentence stress category and its strength

It is a controversial issue on the category of sentence stress in Chinese. Practically, it can be summarized into grammatical stress and logical stress. In running speech, sentence stress always fall onto certain syllable of a unit that baring semantic focus, and always to be manifested via certain type of lexical stress patterns. Thus, the stress status in sentence may be roughly divided into three degrees, i.e., strong , moderate and weak.

2.2.3. Acoustic-phonetic correlates of stress

(1) Pitch accent: In Chinese, pitch accent of stressed syllable is generally manifested by the elevating of pitch register and / or the expansion of pitch range, and it is mainly satisfied by the movement of the high point of pitch range (Shen, 1994; Wang, 2000). Moreover, the specific direction of pitch register movement also depending on the intrinsic characteristics of tonal distinction. For example, the 3rd tone in Chinese is characterized by low register, therefore, when it is accented, its register usually is not elevated, but further lowered.

(2) Duration elongation: Elongation of syllable duration is usually co-occurred as a paragenesis phenomenon with pitch accent. In addition, a recent psychological-acoustical investigation (Zheng, 2000) reports that this elongation can serve as a factor to distinguish the grammatical stress from the logical stress.

2.3. INTONATION SKELETON

2.3.1. Intonation contour universals and special characteristics

As an universal characteristics, there does exist a global declination tendency of pitch movement in Chinese. It is physiologically motivated. However, the way of specific manifestation is quite different from that in intonation languages. Because Chinese is a typical tone language, the contour of rising or falling and so on in pitch movement has been locally employed as tone shapes, these shapes are lexically given and can not be changed arbitrarily. Consequently, the intonation contour in Chinese does not directly take the shape as those occurred in the languages like English, but appear as a complex combination of local tones and underlying intonation skeleton in a special way. That is a way of so-called "algebraic sum of big wave plus small wave" (Chao, 1933).

2.3.2. Characteristics of so-called "algebraic sum of big wave plus small wave"

According to the data measured from both of News' speech and prose declaim, a tendency of declining in global pitch register of sentences can be observed clearly. And it is rhythmically reseted at either the phrase and sentence boundaries. Thus, this declination tendency is obviously undulated and forms as "big waves" like the solid lines shown in the bottom of Fig. 1 (see the last page). It builds up the basic skeleton of intonation. However, such declination contour is not represented directly by the actual pitch trajectory of local syllables, but embodied indirectly through the alignment of pitch register of local syllables (as schematized by the solid lines shown in the top of Fig.1). Because the pitch movement of a syllable in running speech, i.e., so-called "small wave", basically consists of two aspects. The one is the time-varying pitch contour of the syllable (as those shown in the mid line of Fig.1), it is constrained by its tonal distiction including lexical tone sandhi rules. The other is the movement of its relative pitch register. It is similar to musical melody and can be shifted up-/or down-ward (i. e., change key, see Wu, 1994). When a syllable is stressed or located nearby the peak of "big wave", then its pitch register will be elevated clearly, and vice versa. Consequently, in natural speech, each syllable carries the information either of tone and intonation in a way that, on one hand, the pitch contour is relatively kept in constant, so that to remain tonal contrast; on the other hand, its pitch register is moved up- or down-ward, so that to carry the information of underlying intonation.

2.3.3. The last shape of intonation contour

The last shape of intonation contour is a result of multiple regulations. Besides the

physiological mechanism and tonal effect described above, it is also modified by certain speech mood, different location of sentence stress and different speech rate. For example, the general situation is that, the faster the speech rate, the steeper the intonation slope, and vice versa. Fig. 2 (see last page) shows an example of the influences from speech mood and stress allocation. Roughly speaking: (1) the basic intonation contour of a declared sentence is falling, while that of simple interrogative sentence is rising; (2) the closer to sentence beginning the stress location, the steeper the intonation slope, and vice versa; (3) the whole pitch register of interrogative sentence is systematically higher than that of declared sentence.

3. STRATEGY AND TACTICS ON THE PROCESSING OF PROSODIC INFORMATION

According to the situation summarized above, all the prosodic elements, including rhythm, stress and intonation, are integrated with each other based on two main variables, the one is stretching or contraction of duration; the other is the variation of pitch register and pitch range. Consequently, prosodic information processing for TTS can be achieved by control these variables, and a generation framework of prosody for Chinese TTS may be built up through following steps.

3.1. Build up an optimal database

(1) Considering of Chinese TTS systems are mostly using PSOLA method, and the main synthetic unit is syllable. Therefore, as the first step, have to build up a large scale natural speech corpus, it should be cover necessary prosodic environments referencing to stress, rhythm and intonation. Then, build up an applied database based on the speech corpus. It can be made by selecting proper syllable tokens from speech corpus, and try to cover all possible prosodic phenomena. To achieve this goal, the syllable templates should be selected from designed context, so that can be served as the relatively stable prosodic mode for certain context. Different sets of templates represent different hierarchical characteristics. Thus, if the database can cover enough such sort of templates, then, the main prosodic information will be available for TTS system.

(2) To enrich the knowledge for the database, and improve its usage in TTS, both segmental and prosodic labeling are preferred. It should be conducted based on the phonetic and linguistic approaches.

3.2. Make a basic rhythmic structure

(1) Make basic rhythmic chunks by dividing the designed syllable string into PWs, and adjust their pitch and duration in terms of tone sandhi rules and lexical duration patterns. In tactics, the pitch adjustment can be conducted by using 14 main tonal variants which summarized from context, and the duration adjustment can be achieved by using an applied model of: D=Di fw fb fs (Lu, et al., 2000b), which is established according to the situation described above in 2.1.

(2) Define a prosodic hierarchy by inserting different size of silent pause and pre-boundary lengthening at proper PWs' boundary, so that to provide the chunking information of PP and IP, and form an outline of discourse structure in temporal dimension. The size scale of pause and lengthening can take reference to the figures summarized in 2.1.5.

3.3. Establish a basic intonation skeleton

(1) Modify the pitch register and pitch range of each PW, and align them in a gradually downward and contracted way, so that to achieve a declined tendency.

(2) Reset declination tendency by adjusting the pitch register and pitch range at PP and IP boundaries, so that to form an intonation skeleton.

3.4. Achieve a last output of prosody

(1) Make prominence for stressed syllable by adjusting its pitch register, pitch range and duration.

(2) Adjust pitch register, pitch range and duration for phrase-initial and phrase-final syllables respectively, so that to further highlight the prosodic hierarchy.

In tactics, the tasks both of (1) and (2) can be conducted by modifying the top line of pitch movement according to stress status of the PW in PP or IP, and modifying the base line according to the position of PW in PP or IP (Lu, et al., 2000b).

(3) At last, regulate the slope of intonation contour in terms of certain speech mood, speech rate and the location of sentence stress, so that to achieve a last output of the whole prosody. This task can be achieved by applying the rule described in 2.3.3.

REFERENCES

Cao, Jianfen. Basic temporal structure of a sentence in Standard Chinese. *Journal of Chinese Linguistics, Vol. 7,* 1995(a).

_____, Tone sandhi and stress contrast. Zhongguo Yuwen, No.4, 1995(b).

_____, Some aspects on Chinese intonation, *Proc. of the PCPLC, Hong Kong, May 28-30,* 1998.

_____, Acoustic-phonetic characteristics on the rhythm of Standard Chinese, *Proc. of 4th National Conference on Modern Phonetics,* Beijing, August 25-27, 1999(b).

Chao, Yuanren, Tone and intonation in Chinese, *145th Meeting of the American Oriental Society,* 1933.Lu, Shinan et al. A comparison between synthetic speech and natural speech of Chinese. *Proc. of ISCSLP 2000, Beijing, Oct. 14-15,* 2000 a.

_____, Prosodic control in Chinese TTS system. *Proc. of ICSLP2000, Beijing, Oct. 17-20,* 2000b.

Wang, Bei, et al., The pitch movement of word stress in Chinese. *Proc. of ICSLP2000, Beijing, Oct. 17-20,* 2000.

Zheng, Bo, et al., The regular accent of Chinese sentences, *Proc. of ICSLP2000, Beijing, Oct. 17-20,* 2000.

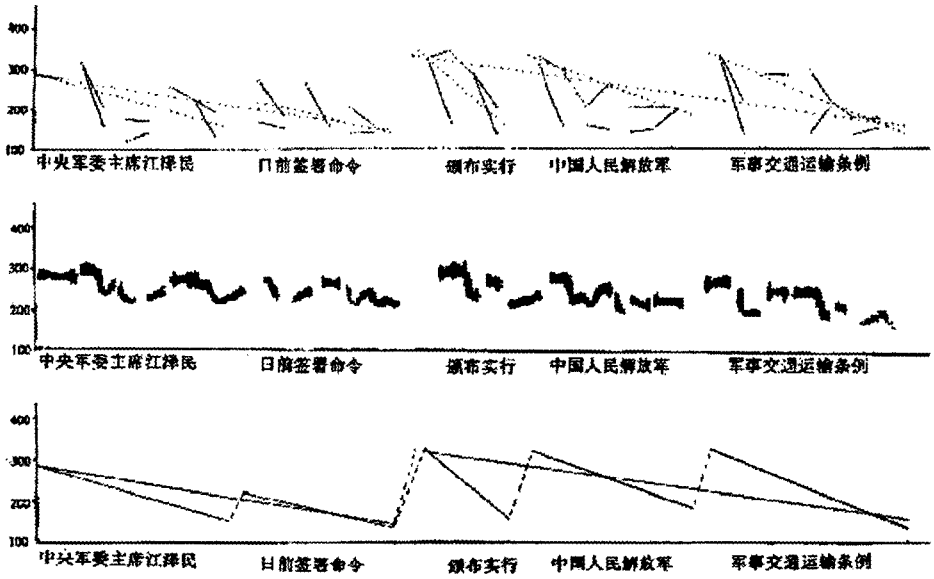

Fig. 1 A schematic diagram for decomposition of surface F0 contour of the discourse into lexical tone (mid: so-called small wave) and basic intonation skeleton (bottom: so-called big wave) in terms of pitch movement.
The top ones represents the relationship of "algebraic sum of small waves plus big waves".

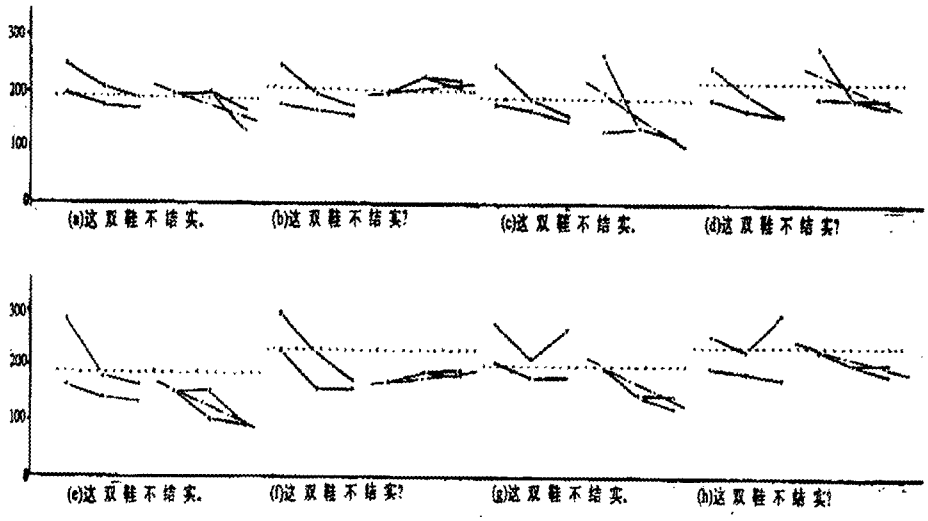

Fig. 2 (1) Dashed lines show the influence of speech mood and stress allocation upon the slope of intonation contour;
(2) doted lines show the influence of speech mood upon the whole pitch register of the sentence.