CHINESE PROSODY AND PROSODIC LABELING OF SPONTANEOUS SPEECH *

Li Aijun

Institute of Linguistics, Chinese Academy of Social Sciences
liaj@linguistics.cass.net.cn

ABSTRACT

In this paper some prosodic research on read and spontaneous speech is introduced first, then the difference between read and spontaneous speech will be depicted, and finally the prosodic labeling system C-ToBI will be described.

1. INTRODUCTION

From 1950s to 1980s, our research focused on read speech in the exploration of the acoustics, psychology and physiology of speech. In the recent 10 years, the research on spontaneous speech has become increasingly important for the applied speech technology and the development of theories for speech production and perception. How to automatically generate, analyze and recognize the prosody in spontaneous speech is one of the unresolved problems confronting many speech synthesis and recognition systems.

In the first part of this paper, a cursory description is given of the different models on Chinese intonation, mainly based on read speech or lab speech. Unlike read speech, spontaneous speech is unscripted and it is produced without plan and much control. Its prosodic feature is easily affected by the interlocutors and the talking environment. Many differences exist between read and spontaneous speech in acoustic domain such as prosody and segment. So the second part of the paper will give some illustrations and make a comparative study between read and spontaneous speech.

The third part of the paper describes the C-ToBI, a TOBI-like transcription system developed for Chinese prosodic annotation [12,14]. Since ToBI's beginning in 1991 [20,22,31], many ToBI systems for different languages have been proposed such as J-ToBI for Japanese, K-ToBI for Korean, GR-ToBI for Greek, G-ToBI for German, M-TOBI for Mandarin, SP-ToBI for Spanish, Cantonese-ToBI and Pan-Mandarin ToBI[21]. Our first Chinese ToBI-C system was developed in 1996 when we made prosodic labeling for 52 read dialogues. But for the lack of a workable intonational theory for Chinese spontaneous speech, we have to employ different prosodic research results in our labeling system. The ToBI-C reported here is the third version for spontaneous speech labeling.

2. CHINESE INTONATION

In the early 20th century, the tone and intonation research for Chinese entered into a new phase due to two phoneticians: Dr. Liu Fu (Ban-nong) and Dr. Chao Yuan-ren (Y.R.Chao). Chao pointed out that the syllabic tone patterns can be modified by the sentential attitudinal intonation, just like "the small ripples riding on top of large waves"[1]. It makes clear the relation between syllabic (also phrasal) tone patterns and the sentential intonation contours, and gives a key solution to a number of problems in intonation analysis. He claims that register is used to describe the English intonation, while contour is used in studying Chinese intonation [3].

Contours and phonological notations of the four citation tones in Standard Chinese (SC) are as follows:

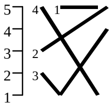
Tone Contour	Tone Type	Tone Letter	Tonal Feature
	Tone 1 Tone 2 Tone 3 Tone 4	5 5 3 5 214 5 1	H-H (H) L-H (R) L-L (L) H-L (F)

Fig.1 Tone-contour patterns, tone types, Chao's tone letters and phonological notations of four SC tonemes. NB: H=High, R=Rising, L=Low, F=Falling.

There is a phonological tone sandhi rule in standard Chinese: if a tone 3 is followed by another tone 3, it will change to tone 2. Additionally, the tonal feature of neutral tone is L following tones 1/2/4 and H following tone 3.

Pitch contour bears tone and intonation information in Chinese. The factors that affect the pitch contour are multiple. For example, different stress patterns and sentence moods can change the pitch contour on the linguistic level; different initials like voiceless fricatives and aspirated affricates can raise pitch contour while the sonorant and lip sound can lower pitch contour on segmental level. In addition, the mechanisms of tone perception and production also contribute to the complexity of tonal contours [36].

Generally speaking, the component of Chinese intonation includes three major parts [8]:

1. The first part is the pitch variation which consists of the variation of pitch range, pitch register and the global trend of the pitch contour. Pitch range or tonal range is the pitch variable range of different tones. The variation of pitch range or tonal range refers to its compression or expansion. Pitch register or tonal register is the "key" of the pitch range used by Wu Zongji. Chao mentioned that there are two kinds of register change: entire pitch raise or entire pitch fall. The variation trend of the pitch contour is characterized by the variation of the top line and bottom line.

The project is funded by CASS foundation and the 10th 5-year National Social Sciences Foundation. This paper was read at Speech Prosody 2002, held in France.

2. The second part is the prosodic structure which is conventionally called rhythmic structure. The rhythmic group is a word, phrase or sentence, while the perceived pause is called physiological pause, grammatical pause or psychological pause.

3. The third part is stress structure. Traditionally Chinese has word stress and sentence stress. The sentence stress consists of logical and attitudinal prominence. Unlike French, Chinese stress is variable. A syllable can be stressed, weak or medium. A weak syllable or neutralized syllable can change the word meaning. For instance, "东西 dong1 xi1" means "east and west", but it means "an object" when the second syllable bears a weak stress as in "dong1 xi0". The remainder of this section will introduce the state of the art of Chinese intonation study based on the aforementioned three components of Chinese intonation.

2.1 Intonational theories

Professor Wu Zongji put forward a set of tone and intonation changing rules which are called *Natural Tone sandhi rule underlining sentence* and *Optional Tone-change rule up sentence*. [34] On the word level, there are rules of bisyllabic tone sandhis, tri-syllabic tone sandhis, quadro-syllabic tone sandhi and D*omino rule* of tone-sandhi, which means that the tone sandhi starts from the most deeply embedded immediate constituent (IC), and then it applies from the underlying form to surface forms successively. The order of its application depends on the grammatical constraints such as tightness of the two domains. On the phrase level, he uses the *transposition rule*. His idea is that the phrasal tone sandhi patterns with different basic registers are similar in musical calibration. Moving the chunks of phrasal contours higher or lower to a certain degree can meet the requirement of logical and attitudinal prominence through the transposition of their basic registers. From the experimental result given by Wu, we find that the phrasal register is categorized into 3 keys.

Basing his research on intonation theory, Wu also designs a labeling system called APLT to label phonetic feature for TTS system [35], e.g. "[TR12]" stands for tonal range is 12 semitones and "[BG-5]" stands for the key of beginning syllable, which is 5 semitones lower than the basic register.

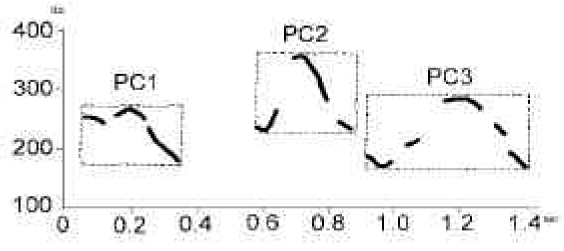


Fig2、Wu's intonation model: Phrase PC1,PC2 and PC3 are transposited by different change keys decided by attitude while the range of each phrase are same in musical calibration for Statement sentence.

Shen Jiong's view about intonation is quite different from Wu [25-28]. He suggested that intonation contour is produced by regulating a set of pitch ranges and the tonal contour is the result of pitch changing in the tonal range. Intonation has a modification rule for the tonal range whose variation exhibits the quantitative change of the tonal contour. The variation of the top line and the bottom line of the pitch range is the cue to the intonation pattern. Generally speaking, top line variation is the result of semantic related prominence variation, while the variation of the bottom line is related to the rhythmic structure. The raising and falling tunes of the interrogative and declarative sentence are produced by the variation of pitch range which is quite different from that of toneless language. When the range expands, the tonal contour enlarges; when the range raises, the contour raises, and vice versa. Shen proposed a systematic intonation structure for Standard Chinese such as top line and bottom line, rhythmic stress and sentence stress. He also proposed a marking system of intonation features based on his research of read speech as shown in table 2.

Table 2. Shen's intonation marking system: T for Top-line, B for Bottom-line, U for unmarked, M for marked.

Intonation of 4 function sentences		**attitudinal intonation**	
AP: declarative intonation	TU+BU	AL: narrowband low tune	TU+BU+NB
BP interrogative intonation	TM+BM	BM: narrowband mid tune	TM+BM+NB
CP imperative intonation	TU+BM	CH: narrowband high tune	TU+BM+NB
DP exclamative intonation	TM+BU	DW: wideband tune	TM+BU+WB

Professor Lin Maocan's research focuses on statement utterances. [17] The pitch contour is decomposed into prosodic boundaries and top and bottom lines. Fig 3 shows the normalized f0 top and bottom line patterns of three phrases. There are three patterns described by Lin: SW is stress-weak pattern (upper panel), WS is weak-stress pattern (mid-panel) and WSW is weak-stress-weak pattern. The top and bottom lines are drawn by connecting the highest and lowest points of each prosodic trunks, so they are hierarchically organized into top and bottom lines of prosodic word and prosodic phrase.

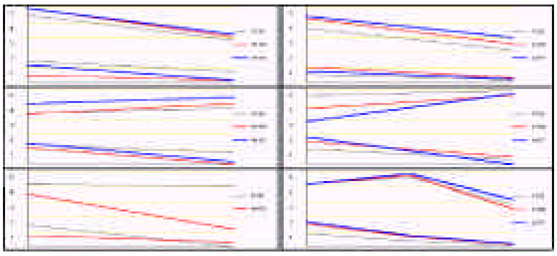


Fig 3. F0 top and bottom lines for three speakers in Y.R.Chao's five degrees of "Tone Letter"(after Lin).

As for the declination of the F0 contour, Xu Yi interprets it from the perspective of the physiological and non-physiological: "In a sentence consisting of only the high tone, no apparent declination can be observed if there is no particular focus in the sentence. It is not the global trends, but more local shapes, and it is not some obscure declination line, but certain linguistically intonation melodies that are deliberately produced, which are ultimately responsible for the surface F0 contours that is often described as a global declination." He also denotes that there are three distinct focus-related pitch registers, each about 1/2 octave apart, as listed below. [36]

pitch register	Domain of Assignment
high	no-final on focus word

low	post focus words
neutral	all other words

On the contrary, Shih Chilin found that the surface F0 contour is a close approximation of the declination slope, or the declarative sentence intonation. Declination interacts with sentence length and focus. [30]

In our own research, we found the relation between F0 range(=max F0 – min F0), F0 register (=0.5*(max F0+min F0)) and the sentence types by analyzing 52 read dialogues shown in Fig.4[14], where F0 is the normalized f0 in logarithm scale. For F0 range: **S**tatements> **Q**uestions> **E**xclamatory > **I**mperatives; for F0 register, **Q>E>I>S** for speaker A and **E>I>Q>S** for speaker B. So the variation of range and register are not proportionate.

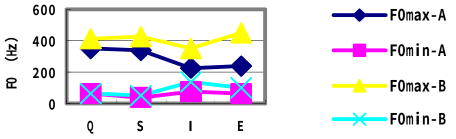


Fig.4a、The maximal and minimal F0 value in four kinds of sentences for speaker A and B

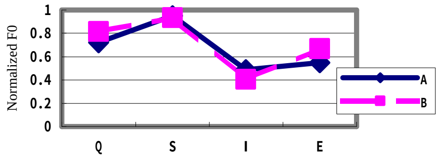


Fig. 4b Normalized F0 range for speaker A and B in 4 kinds of sentence

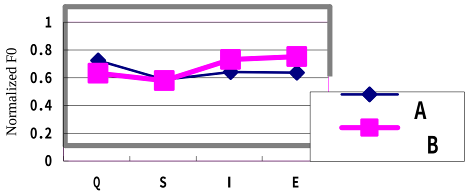


Fig. 4c Normalized F0 register for speaker A and B in 4 kinds of sentence

2.2 Prosodic structure and its acoustic characteristics

Chinese prosodic structure is studied from two perspectives. One is from the phonological and syntactic point of view to define the prosodic units in Chinese. The Chinese prosodic structure is organized hierarchically; it consists of syllable, foot (prosodic word) and prosodic phrase from bottom up. Bisyllabic foot is the standard foot, and it is most frequently used and least constrained. Trisyllabic foot is called supra-foot and monosyllabic foot is called degenerate foot. If a trisyllabic foot has the structure of 2+1 and its structure is N+N by grammatical constituent analysis, then it is a word. If it is V+N and 1+2, then it is a phrase. If syllables are not clear IC, they are grouped into two syllable feet from left to right and the final syllable is grouped into the last foot [5,6,7,29,33]. Only two syllables with IC of V(erb) + C(omplement) is qualified as a word; other VC forms formed with three or more syllables are all phrases. [6]

Prosodic word is a tone sandhi group with fixed grammatical relation. It can be a single foot or a compound two foot word with fixed rhythmic pattern. Prosodic phrases consist of feet with different tonal ranges and rhythmic patterns and it can be divided into word-like phrase and free phrase according to phrasal rhythmic and tonal range variation. [33]

Another perspective to study the prosodic structure is to combine phonology and phonetics. [0,12,14,16,40] Based on our perception experiments, we find that the prosodic structure of Chinese is hierarchically organized from small to large constituent into syllable, prosodic word (PW), minor phrase (MIP), major phrase (MAP) and intonation group (IG) . Prosodic word is a tone sandhi group bearing one word stress. Minor prosodic phrase contains one or more prosodic words bearing one phrasal stress and the perceived break between MIPs is bigger than that of PWs. MAP contains one or more PWs bearing one phrasal stress and the perceived break between MAPs is bigger than that of MIPs. Usually, there is pitch reset between prosodic phrases. [16,40]

Obviously, it is different to define the prosodic unit from syntax and perception. Some lexical words or syntactic words or phrases are prosodic words according to speech perception. For example, “**雨伞厂**(umbrella factory)”,“**小雨伞**(small umbrella)”, and“**卖雨伞** (sell umbrella)”are prosodic word, word-like phrase and free phrase respectively in Wang’s paper. However they can all be classified as prosodic word by the phonetic criterion. In our prosodic labeling system, we employ the phonetic definitions for the prosodic units rather than the syntactic definitions.

In addition, there are many works concerned with prosodic boundary, timing and rhythm of the prosodic features. The utterance is divided into three prosodic tiers: basic rhythmic unit, rhythmic group and sentence [10]. Break perceived in utterances can be a filled pause or a silent pause. The acoustic features signaled the break magnitude is the F0 range preceding or following the boundary for filled pause, while the F0 reset and syllable lengthening is the major acoustic feature signaling the break magnitude for silent pause. [16].

Longer silent pause occurring frequently at major boundary and syllable lengthening may occur at left or right of minor boundary. But a large amount of lengthened syllables do not occur at boundary. In a perception test for the stress of 50 utterances, lengthened syllables are closely related with stress or prominence of a sentence. In many cases, when the syllable group is stressed, all syllables in the group are lengthened. [42]

The durational characteristics of major and minor phrase are statistically analyzed using the 52 dialogues as samples. [14] The results are the following: (1) the final syllable in MAP is longer than that in MIP if not considering tones, which is similar to English. The situation is much more complicated when tonal distinctions are considered. For example, the final syllable in MAP is shorter than that in MIP for tone 1 syllables. The tone 4 syllables are the shortest except the neutral tone syllable before the prosodic phrase boundary. (2) Pre-boundary lengthening in SC is more complicated than in English. Whether final lengthening exits or not is related to the tone of the final syllable tone in the phrase. (3) The average duration of MIP is similar in four types of sentence which is 0.6-0.7s. The duration of MAP is related to sentence type which is longer in **S** and **Q** (0.8-1s) and shorter in **I** and **E**(0.6-0.9s). (4) The syllabic number in MIP is 3-4 syllables; but it is variable in MAP, depending on sentence types: 5-8 in **S** and **Q,** and 3-5 in E and I.

2.3 Stress

Chao describes that the formation of stress in Chinese firstly involves expanding the tonal range and lengthening the tonal duration and secondly raising the air stream. SC is a language with lexical tones. Therefore its pitch cannot be freely used to mark stress [5]. The acoustic features for stress are duration, F0, spectrum tilting and amplitude whose correlations are rather complicated.

The contribution of the acoustic features for perceiving the prosodic word boundary in descending order is duration, F0, spectrum tilting and amplitude. [38] Besides, duration and F0 are compensatory. [34,37]

On the one hand, tonal range and register are changed by stress. On the other hand, they are affected by prosodic boundary. After making some statistic analysis on sentence stress by using 52 prosodically labeled dialogues [14], we found the following: the duration of stressed syllable is longer than that of unstressed syllable, but the neutral tone in the stressed word is not lengthened; the maximum F0 of the stressed syllable is higher except tone 3 for speaker A; the minimum F0 of the stressed syllable is higher except tone 3 and neutral tone; the F0 range of the stressed syllable is wider except tone 1; the tonal range is wider for stressed syllable; different speakers employ different strategies to make emphasis: speaker A produced the stress syllable by lengthening the duration, expanding the tonal range and descending the bottom line of F0 without raising the F0 top line; speaker B using lengthening the duration and expanding the tonal range and raising F0 top and bottom line.

Shen points out that intonation is used to realize the sentence stress. Sentence stress and rhythmic stress will modify the top line and bottom line of the tonal contours respectively. It is pitch rather than duration that contribute more to sentence stress. [27] The higher the stress category, the higher the pitch distribution curve. [37]. Sentential, phrasal and word stress correspond to 3 categorical pitch ranges.

3. THE COMPARATIVE STUDY FOR READ AND SPONTANEOUS SPEECH

In the previous section, what we introduced is all based on read speech. Our intention is to make a prosodic labeling system for spontaneous speech. Some of the results are still correct for spontaneous speech; others have to be improved. So this part will give some comparative results at the levels of segment, prosody and paralinguistic phenomena for read and spontaneous speech.

3.1 Prosody

(1) F0 variation

Speaker WJC read a piece of material transcribed from her spontaneous speech in corpus CADCC. [18] The read speech is then annotated prosodically. The max and min of the F0 of the major phrase are given in Fig. 5. It shows that the pitch range varies significantly in spontaneous speech than that in read speech. The pitch range is 24ST in spontaneous speech and 17ST in read speech. The F0 maximum points vary much more than F0 minimum points in spontaneous speech.

If the pitch range is categorized into 3 levels in read speech, the level must be more than 3 in order to annotate the real facts. We set up 2 more levels to express the extreme variation in spontaneous. Total 5 levels are: more expansion, expansion, neutral, compression and more compression. We find the "more expansion" usually corresponds to the attitudinal or emphasis stress, while the "expansion" to grammatical stress. Fig. 6a and 6b are examples of F0 contours for two sentences in spontaneous and read speech. The max pitch range of the phrase in Fig. 6a is 17.3 ST, while it is only 4ST in the second sentence in Fig 6b for spontaneous speech. The pitch range is rather stable in read sentences from 4ST to 8ST.

As we have known, the tonal register is categorized into 3 levels in read speech. They are still not enough to annotate the tonal register variation in spontaneous speech. So we add 2 more levels as follows: higher, high, neutral, low, lower. So the each pitch register will be 4ST for WJC. The height of the register will affect the F0 globe "grid" as used by Eva. Gärding, but it is not necessarily correlated with stress.

The top-line pattern of the prosodic phrase is rather complicated. The downstepping pattern of the top-line is zigzag rather than linear. [41]

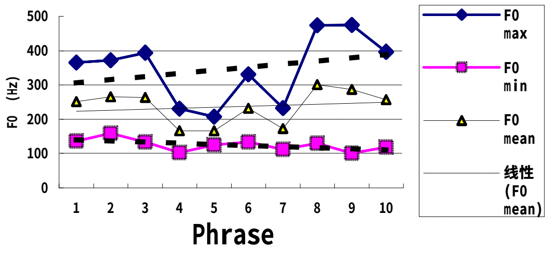


Fig 5a F0 max and F0 min of major phrase for spontaneous speech

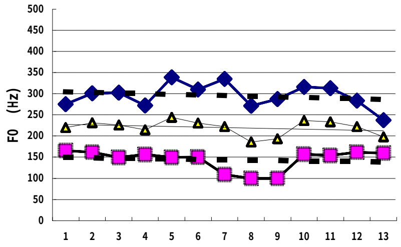


Fig5 b F0 max and min point for major phrase in read speech

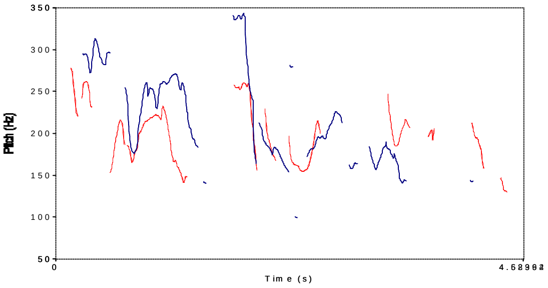


Fig6 a An example of F0 curves for read speech (light line) and spontaneous speech (black line) .The sentence is "其实所谓的压力呀，真的就是我我想是我给他找出来的"。

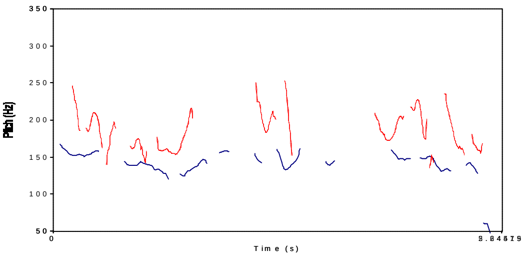


Fig 6b The second example: “**寄给我们室每个人，差不多寄一贺卡，后来给我寄一个**” (light line for read)“**寄给我们室每个人寄一贺卡，后来给我寄一个**” (black line for spontaneous)

Table 3, The pitch rang and the stress category

Pitch range	Domain of Assignment
Extremely wide	Emphasis stress for sentence stress
wide	Phrase level stress
Neutral	Word level stress
narrow	secondary word stress
Extremely narrow	Special attitude; unstressed word stress

(2) Prosodic structure

The hierarchy of the prosodic units in spontaneous speech is the same as in read speech, but the acoustic feature signaling the boundary is different as based on the perception test and analysis on ASCCD[13] and CADCC[18]. In read speech, the higher the boundary level is, the longer the silence between the boundaries, which does not necessarily occur in spontaneous speech. Other features such as F0 resetting, lengthening and tonal range of the adjacent boundary units are used to signal the break too. Additionally, the major prosodic boundary occurs less in fast turn-taking dialogues as in the switchboard sub-corpus than that in dialogue sub-corpus of CADCC.

(3) Stress structure

In our prosodic labeling system for read speech, we annotate word stress, phrase stress and prosodic group stress corresponding to the different prosodic units. But in spontaneous speech, we find more stresses are attitudinal or contrastive stressed. If the stress is highly related to the pitch range [25,38], which has five levels, there should be five levels of stress as well. This is shown in Table 3.

3.2 Other differences

Based on 20 hours of spontaneous speech and 10 hours of read speech, other differences are found for read and spontaneous speech. [18]

(1) Syllable occurring frequency: (omitted)

(2) Initial and final occurring frequency: (omitted)

(3) Paralinguistic and nonlinguistic phenomena: which are rich in spontaneous speech [18]. A very frequently used phenomenon is overlap, which presents the interactive complicated procedure, and needs to be interpreted with discourse analysis.

(4) Discourse topic: discourse topics (not sentence topic) vary extensively. 80 percent of topics are less than 4-5 mins and 90 percent topics are less than 5-6 mins.

(5) Segmental variability: We labeled the segmental variability such as Insertion, Deletion, Laryngrealization, Voicing, Devoicing, Nasalization, Rounding, Aspiration, Centralization and Phoneme Change in read and spontaneous speech corpora. It is found that the variability is 27.46% for initials and 12.02% for finals in spontaneous speech that occurs rather less in read speech except voicing and centralization. [15]

(6) Syntax: The most frequently used pattern in read speech is the “SPVO (Subject phrase + verb + object)”. In spontaneous speech it is the elliptical structure. [32]

4. C-TOBI: PROSODIC LABELING SYSTEM FOR CHINESE

The present labeling system is the third version for discourse and spontaneous speech. [12,14] Eight parallel tiers are labeled in C-ToBI.

① PinYin tier: Canonical Pinyin and tone for each syllable are labeled, e.g. 0,1,2,3 and 4 for different tones.

② Initial and final tier: the initials and finals of each syllable are annotated in terms of sound variation by using SAMPA-C labeling set [4,10]

③ Tone and intonation tier (T&I):

We do not have a theory that can be used directly in prosodic labeling for spontaneous speech. Intonation transcription is rather ambiguous in many Chinese prosodic labeling systems. In M-ToBI there is no detailed top and bottom line variation description and in our previous system there is no global description for intonation contour but detailed transcription for local tone and intonation. Now tonal register and range changing of intonation are both transcribed on intonation tier as shown in Table 4. Two examples are shown in Figs 8 and 9.

Table 4. Tone and intonation labels in C-ToBI(3.0)

Labels	Meaning
H-L, L-H, H-H L-L	Tonal features for tones
H, L	Neutral tone
L% H%	Boundary tones
&	Transitional tone
^, !, ^^, !!	^ upstep, ^^ wide upstep ! downstep, !! wide downstep
Re^() Re^^() Re!() Re!!()	For pitch register change, ( ) for the scope
Ra^() Ra^^() Ra!() Ra!!()	For global variation of pitch range, ( ) for the scope
H^^,H^,L!	For local variation of pitch range
%d	Beginning with a new pitch down-stepping
%r	Beginning with a new raised pitch
%l	Pitch reset with a level contour

③ Break index tier

Break is perceptually labeled for each syllable which is variable for the same text read by different speakers or at different time.

break index=0: the minimum break between syllables, usually breaks within a prosodic word;

break index=1：for prosodic word boundary;

break index=2：for minor prosodic phrase boundary；

break index=3：for major prosodic phrase boundary；

break index=4：for prosodic group boundary；

1p, 2p and 3p stand for the break produced by abnormal break like hesitation or some other acts like cough.

“?” for uncertainty.

④ Stress index tier：Stress indices=1-4 for hierarchical stresses corresponding to prosodic units, “@” is attached to the index for emphasis or contrastive stress.

⑤ Sentence function tier: four sentence types are annotated for interrogative, imperative, declaratives and exclamation sentences.

⑥Accent tier: Regional accent is labeled with acronym of the region, e.g. <SH,SH> for Shanghai accent and dialectic accent is annotated by codes used in [39].

⑦ Turn taking tier：Using <A A>,<B B>...to give the start and end point of each turn.

⑧ Miscellaneous tier: Paralinguistic and no-linguistic phenomena are annotated by symbols in [18].

5. DISCUSSION

We have manually annotated 10 hours of read speech and 7 hours of spontaneous speech on 7 tiers except intonation tier for corpora ASCCD and CADCC. The annotated corpora have been used for speech synthesis and recognition [9]. But as a robust and easily learnable system for Chinese dialects, some other information should also be labeled.

(1) Attitudinal intonation: the major function of intonation is to convey different emotion and paralinguistic information [23]. So the labeling for this information is necessary and important.

(2) Chinese dialects: tonal features for all Chinese dialects can be coded by Table 6 in an archaic system shown in Table 4. For special tone sandhi, “a” or “b” can be coded as the third number.

Another intention is to get an automatically prosodic labeling system which can get more detailed phonetic features for pitch range and register and the global pitch contour as in Wu’s system.

Acknowledgement: Many thanks to professor Wu and Li Zhiqiang who gave me a lot of help for this contribution.

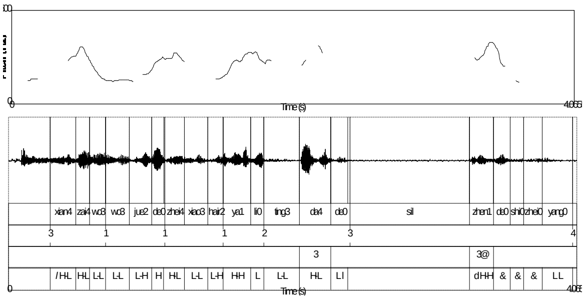


Fig.8 An example with pinyin, break stress and T&I tiers.

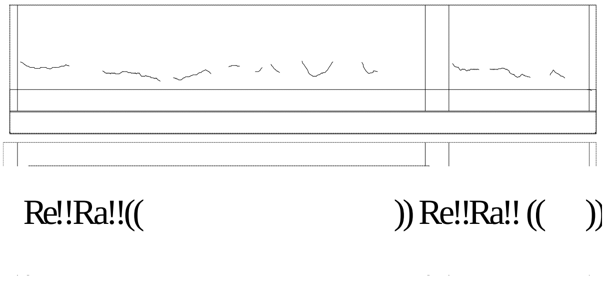


Fig.9 An example to show an extremely low register in a fast and low voice utterance.

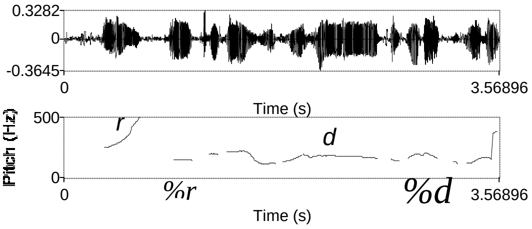


Fig10 an example of interrogative sentence with a raising intonation followed by an answer with a declining intonation contour.: “A:**啊？** B: K35 **明天北京到济南的** (A: What? B: Tomorrow from Beijing to Jinan, train K35.)”.

Table 6. Labels for Chinese tone (V for voiced initial, Vs for voiceless initial)

Tone	SAMPA-C	Tone	SAMPA-C
阴平 (Vs-Level)	_11	**阴去** (Vs-Departing)	_31
阳平 (V-Level)	_12	**阳去** (V-Departing)	_32
阴上 (Vs-Elevating)	_21	**阴入** (Vs-Entering)	_41
阳上 (V-Elevating)	_22	**阳入** (V-Entering)	_42
		轻声 (Neutral)	_0

6. REFERENCES

[0] Cao, J., Rhythm grouping and speech timing. *Proc. Of ICSLP’2000, Beijing, Oct. 17-20.*

[1] Chao,Y.R.,1933. A preliminary study of English intonation and its Chinese equivalents. *BIHP Supplement No.1.*

[2] Chao,Y.R.,1968. *A Grammar of Spoken Chinese.* University of California Press.

[3] Chao,Y.R.,1980. *YuYan WenTi*. Businesses Press.

[4] Chen Xiaoxia, Li Aijun, etc , An Application of SAMPA-C for Standard Chinese. *ICSLP2000*, Beijing.

[5] Duanmu, San , Chinese Rhythm, *Modern Linguistics.* No.4,2000.

[6] Feng, Shengli, Prosodically determined distinctions between word and phrase in Chinese. in *ZhongGuoYuWen* No.1 2001

[7] Feng Shenli, C*hinese prosody, morphology and syntax.* Peking publish house,1997.

[8] Guo Jinfu, *A survey on the Chinese tone and intonation.* Beijing Language Institute Publish house.

[9] Hu Weixiang,et.al, The detection of the prosodic boundary and its recognition for Standard Chinese. *NCMMSC6,2001*.

[10] J. Wells, *Computer-coding the IPA: a proposed extension of SAMP.* 2000, http://www.phon.ucl.ac.uk/home/sampa/

[11] Li Aijun et.al.,2001. Spontaneous speech corpus CADCC and the Phonetic Research. *The 5th National Modern Phonetics Conference.* In Chinese

[12] Li Zhiqing, 1997. A Preliminary Study of the Prosodic Transcription System for Standard Chinese. Paper presented at The Third National Conference of Intelligence Interface and Its Application. Zhangjiajie. (in Chinese)

[13] Li Aijun, Lin Maocan, ChenXiaoxia, et.al., Speech corpus of Chinese discourse and the phonetic research. *ICSLP’2000*.

[14] Li Aijun, ZuYiqing, Li Zhiqiang, A National Database Design and Prosodic Labeling For Speech Synthesis. *Oriental COCOSDA'99*,Taipei.

[15] Li Aijun, Zheng Fang, etc.CASS: A phonetically transcribed corpus of spontaneous speech. *Report of Phonetic Research*, 2000.

[16] Lin Maocan, 2000. Break in utterance and the prosodic phrase in Standard Chinese. *Contemporary Linguistics*, 2001.2

[17] Lin Maocan and Yan Jinzhu,2000, prosodic structure and the stress levers-one of the feature of the Chinese intonation. in the *proceedings for celebrating 80 years old birthday of professor Li Rong*.

[18] Liu Yabin and Li Aijun, 2001. A comparative study between read and spontaneous speech. *NCMMSC6*.

[19] Luo ChangPei, Wang Jun, *The outline of Phonetics of Standard Chinese*, Sciences Publish House,1957.

[20] Mary E. Beckman & Gayle M.Ayers,1994.*Guidelines for ToBI Labeling. Manuscript*, Ohio State University.

[21] Marjorie K.M. Chan, http:/ /deall.ohio-state.edu/chan.9/ MToBI.htm

[22] Pitrelli, John, Beckman, Mary, & Hirschberg, Julia.1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework, *ICSLP1994*, vol. 1, pp. 123-126.

[23] D. Robert. Ladd, 1996. *Intonational phonology.* Cambridge University Press.

[24] Selkirk, Elisabeth.1984. *Phonology and syntax: the relation between sound and structure*. Cambridge, Mass.: MIT Press.

[25] Shen, Jiong, The study for the Chinese sentence stress, in *YuWen YanJi.* vol.3 1994.

[26] Shen Jiong, A pilot study on Chinese intonation classification and its labeling. in *YuYanWenZi YingYong*,1998, No.1.

[27] Shen Jiong, Chinese intonation structure and category. in *Dialects*, No.4,1994.

[28] Shen, Jiong, The tonal range and intonation. in *Beijing YuYin ShiYanLu*, edited by Lin Tao, Peking Univ. Publish house, 1985.

[29] Shih, Chilin,1997. Mandarin third tone sandhi and prosodic structure. In Wang Jialing and Norval Smith (eds.), *Studies in Chinese Phonology.* 81-123. Berlin: Mouton de Gruyter.

[30] Shih, Chilin,1997.Declination in Mandarin, *ESCA workshop on Intonation: Theory, Models and Applications.* Athens Greece.

[31] Silverman, Kim; Beckman, Mary,et.al.,1992. ToBI: a standard for labeling English prosody. *Proceedings of the 1992 International Conference on Spoken Language Processing*, vol. 2, pp. 867-870.

[32] Tao Hongying,1996. *Unit in Mandarin Conversation Prosody, Discourse, and Grammar.* John Benjamins Publishing Company Amsterdam, The Netherlands / Philadelphi USA.

[33] Wang Hongjun, Chinese prosodic word and prosodic phrase. *Zhongguo Yuwen*, No.6, 2000.

[34] Wu Zhongji,2000. From Traditional Chinese Phonology To Modern Speech Processing--Realization of Tone and Intonation in Standard Chinese, *ICSLP'2000*.

[35] Wu Zhongji, Wang Renhua, Liu Qingfeng,1997. Toward a project of All- phonetic-labeling-text (APLT) for TTS synthesis of Spoken Chinese. *the first Chinese-Japan workshop on Spoken language processing*.

[36] Xu Yi, 1997. What can tone studies tell us about intonation? *ESCA workshop in Intonation: Theory, Models and Applications*, Athens Greece.

[37] Xu, Jieping,1999. *Research on Chinese prosodic characteristics based a large speech database.* Ph.d thesis of Institute of Acoustics, CAS.

[38] Zhong Xiaobo, *Stress perception on Standard Chinese and its acoustic feature.* Ph.d thesis of Institute of Psychology 2000. (in Chinese)

[39] Language Board of China,2000. *Difang Putonghua Diaochayuan Shouce*.

[40] Zheng QiuYu, The major cues of prosodic structure in running speech. 2001, *NCMMSC6*, China

[41] Wang Maolin & Lin Maocan,2001.Intonational phrase and its pitch patten. *NCMMSC6*, ShenZen, China.

[42] Zu Yiqing, Chen Xiaoxia,1999. Segmental durations and lengthened syllables. *ICPHS'99*,1:277.

汉语口语的韵律研究和标注

李爱军

摘要

韵律标注研究是随着言语应用发展而在 90 年初开展的。较早提出的是英语的韵律标注系统 ToBI(for Tones and Break Indices)[13,14]，旨在标注英语的语调模式和其他韵律信息，提供韵律标注的通用标准。一方面，韵律标注可以为语音基础研究提供共享数据库；另一方面，从语音应用技术方面来看，韵律标注已经是基于数据驱动语音合成系统的十分关键的技术，对合成音库的韵律标注的准确性和科学性，直接影响合成系统的自然度。同时，口语语音识别也越来越需要利用韵律标准信息，提高识别率、提高语音理解和语音翻译的正确率。

为此，很多语种都建立了韵律标注系统，如日语 J-TOBI、韩语 K-TOBI、希腊语 GR-TOBI、德语 G-TOBI、台湾国语 M-TOBI、西班牙语 SP-ToBI、广东话 Cantonese-TOBI、泛汉语 Pan-Mandarin ToBI 等等。早在 1996 年，我们承担了国家 863 语音合成语料库的设计和韵律标注课题，制定了针对朗读话语的韵律标注系统 C-ToBI 1.0，随着语音研究逐步深入到朗读和独白口语语篇，我们提出了韵律标注系统 C-ToBI 2.0。那么标注对象扩大到口语以后，汉语的韵律标注系统又有新的挑战。

自然口语不是按照实验室语音（Lab Speech）库中所提供的经过精心校对的文本进行控制发音或重复发音的。自然口语语音的韵律受到交际和修辞功能的影响，这些语境因素相当可变，目前我们还不甚了解，也不知道如何控制。因此，有很多种类的自然口语语音主要的差别不光是与实验室语音不同，就是自然口语之间也有差别。所以对自然口语的精细标注，难度增大了很多。

在对几个语音语料库进行分析之后，我们总结了普通话口语和朗读语音之间在音段和韵律方面存在的差异，并据此，修订韵律标注系统，使得 C-ToBI 3.0 也使用于普通话口语韵律的标注。

做后对进一步的工作和存在的问题，如口气和情感标注、重音的标注、各种方言声调的标注等问题进行了讨论。