ABSTRACT
The tempo and prosodic pattern in Chinese were investigated further in this paper. A special speech database, which was composed of three compound sentences and ten speakers (5 males and 5 females), was created. Three test sentences have different grammatical structures. The results show that: 1. The sentence prominence and the sentence intonation are invariant prosodic features; 2. The tonal pattern can be influenced by tone coarticulation and speech mood; 3. The effect of last constituent lengthening in relative duration in fast speech is not a kind of syntactic or semantic signal; 4. The pause signals a boundary of phrase and sentence and it will be compressed in fast speech.

1. INTRODUCTION
The role of prosody in both human speech communication and man-machine speech communication can not be overemphasized. Prosody carries both linguistic and paralinguistic information, which have to be properly processed by speech I/O systems. As the development of speech technology, the researchers and developers in this field pay more and more attention to model the prosodic features. And some duration modeling techniques were developed in both speech recognition [1] and speech synthesis [2].

Gårding and Zhang [3,4] have done some studies on the temporal and prosodic pattern in Chinese and Swedish since 1987. The testing material was two simple sentences for Chinese and the speakers were only two males. In order to gain an insight into this issue, we want to explore questions of that: what is speaker general or specific and what is sentence structure related prosodic features in temporal pattern. So that some sentences with different grammatical structures and different logic structures were designed and much more speakers were used in this study. In Sections 2 the creation of the speech database was described in details and the testing results were presented in Section 3. Finally some discussion were given in Section 4.

2. MATERIAL, RECORDING AND ANALYSIS

2.1 Speech Material
Three compound sentences were designed instead of the simple sentences used in previous work. In each sentence three primary vowels /i, u, a/, the consonants with different manners of articulation, and different lexical tones were included. There are two clauses (five syllables in each clause) in sentence 1 and sentence 2 and sentence 3 is the sentence 1 connecting with sentence 2. So it is easy to remember and to utter fluently. The two clauses in each sentence are identical in grammatical structure.

Sentence 1: “Ta qu Wuxishi, wo dao Heilongjiang.” (He goes to Wuxishi, I come to Heilonjiang.).

There is a subject and a predicate in each of the two clauses, which is similar with the simple sentences used in our previous work.

Sentence 2: “Shenyang Lvdashi, Guangzhou Neimenggu.”

There is no syntactic and logic structures in sentence 2, it is just a connexion of four city names. And the length of sentence 2 is the same with sentence 1. So that it is easy to do comparative study on the effects of semantic and syntactic structures on the temporal structures between sentence 1 and sentence 2.

Sentence 3: “Ta qu Wuxishi, wo dao Heilongjiang. Shenyang Lvdashi, Guangzhou Neimenggu.”

Sentence 3 is a connexion of sentences 1 and 2. No special instruction about how to read this sentence was given to the speakers.

2.2 Database Recording
A special speech database was created for this study. Ten speakers (5 males and 5 females) who speak standard Chinese—Putonghua uttered the test sentences nine times in three recording sessions. All recordings were made in an anechoic room by using a DAT tape recorder (Sony 2700A). The speech signal was recorded on the right channel while the vocal folds vibration signal—Laryngograph output Lx on the left channel. A condenser microphone (B&K 4165) was hanged 30cm in front of the speaker’s lips and there is a bar fixing on the hat of speaker to
help keeping the right distance. And there is another condenser microphone situated 1m from speaker’s lips which was connected to a sound level meter to monitor the speech level and the environment noise level.

The speech material issuing to the speaker was prompted on two monitors, one for speaker in the anechoic room and the other for operator in control room. Each speaker uttered each sentence three times in one recording session. And three recording sessions lasted three months. So there are total of nine renderings for each speaker and each sentence.

2.3 Acoustic Analysis

A software package—SFS (Speech File System) which was created by University College London was used for labeling and time-frequency domain analysis. The speech rate was measured at both global level and local level. At global level the speech rate was measured in the number of syllables per second i.e. the number of syllables in a sentence divided by the duration of the sentence. And the sentence-normalized scale is used for temporal structure study.

At local level a new measure called signed relative change rate of segment duration R(i) is adopted and the R(i) for fast speech Rf(i) and for slow speech Rs(i) are defined as

(1) \[ R_f(i) = \log \frac{D_n(i)}{D_f(i)} \]
(2) \[ R_s(i) = \log \frac{D_s(i)}{D_n(i)} \]

in which the Dn(i) is the duration of the segment i in normal speech, the Df(i) is that in fast speech and Ds(i) is that in slow speech.

In frequency domain the pitch contours were abstracted by using Laryngograph signal processing combined with speech wave processing. And syllable alignment is adopted for tonal pattern comparative study between different speech rates.

3. RESULTS

3.1 The temporal structure of sentences

The speech rate at global level was measured in two different methods. (1) Gross speech rate SR1 is the total amount of syllables in a sentence divided by the gross sentence duration in second. (2) Net speech rate SR2 is the total amount of syllables in a sentence divided by the net sentence duration, which equals to the gross duration minus the total duration of pauses in the sentence. The average speech rate of three tempi over three test sentences and ten speakers is listed in Table 1.

<table>
<thead>
<tr>
<th>Speed</th>
<th>Fast</th>
<th>Normal</th>
<th>Slow</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR1 syllables/s</td>
<td>6.01</td>
<td>3.31</td>
<td>1.74</td>
</tr>
<tr>
<td>Relative</td>
<td>1.82</td>
<td>1</td>
<td>0.526</td>
</tr>
<tr>
<td>SR2 syllables/s</td>
<td>6.59</td>
<td>3.91</td>
<td>2.03</td>
</tr>
<tr>
<td>Relative</td>
<td>1.69</td>
<td>1</td>
<td>0.519</td>
</tr>
</tbody>
</table>

Table 1. Speech rate in three tempi.

From Table 1 it can be seen that the speech rate in relative scale SR2<SR1, but in absolute scale it is contrary. That means that the pauses in fast speech were significantly compressed. Fig. 1 shows the normal speech normalized relative duration of segments averaged over ten speakers in three tempi. The black bar stands for the initial (Shengmu) and the white bar for the remained vocalic part—-the final (Yunmu).
some speakers the last pause can be disappeared in fast speech. It is worth noticing that the last constituent lengthening effect in fast speech is universal in different sentences and for different speakers. In order to show this effect clearly, the relative duration of the last word in the six clauses of three test sentences in three tempi is listed in Table 2.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Fast</th>
<th>Normal</th>
<th>Slow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wuaxishi</td>
<td>1.03</td>
<td>1.03</td>
<td>1.04</td>
</tr>
<tr>
<td>Heilongjiang</td>
<td>1.30</td>
<td>1.09</td>
<td>1.04</td>
</tr>
<tr>
<td>Lvdashi</td>
<td>1.10</td>
<td>1.09</td>
<td>0.984</td>
</tr>
<tr>
<td>Neimenggu</td>
<td>1.22</td>
<td>1.28</td>
<td>0.937</td>
</tr>
</tbody>
</table>

Table 2. The relative duration of the last word in clause of three test sentences in three tempi.

It is obvious that the relative duration of the word at the utterance (isolated sentence) final position should be lengthened, for example, the word Neimenggu was lengthened in both sentences 2 and 3. As for the word Heilongjiang it was only lengthened in sentence 1 but not lengthened in the case of sentence 3, where sentence 1 is followed by sentence 2. Generally, the relative duration of the last word in fast speech has increased by around 30% over normal speech.

3.2 The local speech rate in different tempi

In order to explore the temporal structure at segment level, a new measure for relative local speech rate was defined by the signed relative change rate of segment duration R(i). And the test results of sentence 1, as an example, were drawn in Fig. 2. The signed relative change rate of segment duration R(i) which are expressed in formula (1) and (2). The results of sentence 1, as an example, were drawn in Fig. 2. The signed relative change rate of segment duration Rf(i) and Rs(i).

It can be seen that the signed relative change rate of segment duration R(i) has a tendency towards to be symmetrical about normal speech for the fast and the slow. Some segments (most vowels) are more flexible and some others (most consonants) are not. And the pauses were significantly compressed in fast speech but not expanded in slow speech.

3.3 Tonal pattern and sentence intonation

In order to do time normalization at segment level, the vocalic part of a syllable was divided into six equal sections in duration. The F0 at the section boundary point of the nine renderings of a sentence for a speaker was averaged point by point. Then the normal -rate temporal pattern was used as a base and the tonal patterns in the three tempi were transposed to their respective segments via the syllable boundary alignment instead of the turning point alignment adopted in our previous work [3]. As an example, the results for two speakers, one male (on the top) and one female (on the bottom) were shown in Fig. 3.
generally higher and has smaller range. In the case of both speaker M05 and speaker F05, the number of turning points has changed with tempo. The tonal pattern of syllable “xi” (level tone) turned to be falling in three tempi and the tonal pattern of syllable “zhou” (level tone) to be rising in fast for speaker M05. As for speaker F05, the syllable “da” with falling tone turned to be rising tonal pattern and the syllable “zhou” (level tone) to be rising also in fast speech. In general the number of turning points is retained and they have rather fixed locations relative to the segments.

Comparing the pitch contours of different speakers we found that: (1) Different speakers use different strategies to manipulate the pitch contour to effect the declarative intonation; (2) Sentences 1 and 2 have different syntactic structures but have similar prosodic structures; (3) In sentence 3 the sentences 1 and 2 kept their original sentence intonation respectively, although there were some interaction between them in temporal structures.

4. DISCUSSION

Judging from this new material the following invariant prosodic features which were summarized in our papers [3,4] were reconfirmed: (1) The sentence prominence, (2) The sentence intonation, (3) The more relative duration is given to the last constituent in fast speech.

Generally with a shift of tempo the position of the tonal pattern to the segments has a certain leeway. In fast speech, however, the tonal pattern can be overridden by the tone coarticulation in certain context under different prosodic realization. For example, as mentioned above the syllable “zhou” appearing in between dipping tone syllable “guang” and falling tone syllable “nei” was changed into a rising tonal pattern in fast speech of both speakers M05 and F05. As for syllable “da” in fast speech, which appeared in between a dipping tone and a falling tone also, the pitch contour of “da” for speaker F05 was turned to be rising (maybe an atonic) but for speaker M05 it was not changed because he put a focus on “da”. On the contrary, Syllable “xi” (level tone) was changed into falling in all three tempi for speaker M05 but a little bit rising for speaker F05, because M05 focused on syllable “wu” and F05 on Syllable “xi” instead. It should be mentioned that the relative duration of these syllables, of which tonal pattern was changed, was almost not varied in three tempi. So that either position of the target points of a tonal pattern is somewhat adjusted to meet the requirement of tonal context in specific mood. The tone modification can only be decoded in proper tonal context under certain speech mood.

The sentence intonation is independent of speech rate and temporal structures of sentence although the tonal pattern of syllables was significantly influenced by tone coarticulation.

The pause, which is a kind of syntactic signal, plays an important role in shifting tempo.

ACKNOWLEDGEMENT
We wish to thank Prof. A. Fourcin and Dr. M. Huckvale for providing the SFS software package. This work was supported by the National Science Foundation of China.

REFERENCES