SPEECH CORPUS OF CHINESE DISCOURSE AND
THE PHONETIC RESEARCH

Li Aijun, Lin Maocan, Chen XiaoXia, Zu Yiqing, Sun Guohua, Hua Wu, Yin Zhigang, Yan Jingzhu

Phonetics Laboratory, Institute of Linguistics,
Chinese Academy of Social Sciences (CASS)
5 JianGuoMenNeiDaJie, 100732, Beijing, PRC
Tel. 86-01-65237408 Email: Liaj@linguistics.cass.net.cn

ABSTRACT

Speech corpus of Chinese discourse (ASCCD) was setup and annotated on segmental and prosodic and syntactic tiers. SAMPA-C and C-ToBI conventions are used for segmental and prosodic labeling. Sound variation such as assimilation, insertion and deletion are investigated on the labeled database. The prosodic research focuses on the sentence stress that involves the specification of relative prominence in prosodic structure,

INTRODUCTION

In traditional phonetic research, several sentences or phrases are often designed and read according to the researcher’s requirements; therefore some results are unreliable and insignificant, say nothing of being applied in speech application systems. In recent years, corpus based research sheds a new light on the phonetic study, meanwhile, it forces the speech corpus collecting and annotating hold pace with it.

Many speech corpora have been collected or designed for different research purposes in China. Here list some of them as followings:
① 863 recognition database designed by CASS [18]: 1500 sentences balanced with phonetic units such as syllables, diphones and triphones, 200 speakers. Phonetic segmentation on initials and finals is made for one male speaker.
② 863 database for phonetic research and synthesis designed by CASS[li]: including four subsets: a two level word database, a neutral tone and retroflexed syllable database, a monologue database and a dialogue database. The word or phrase database includes all tonal combinations and as many as intersyllabic triphones and diphones. The monologue database includes 18 sentence patterns and as many as triphones. Dialogue database have 52 topics which are posodically labeled.
③ Spoken dialogue database for domain specific application of travel and hotel information retrieval collected by NLPR of institute of Automation, CAS. [13]: 60 dialogues for travel and 100 dialogues for hotel reservation.
④ Spoken dialogue database for air flight information retrieval collected by Computer Department of THU.

The large speech database for corpus based synthesis system such as KD2000 synthesis system and “saying your say” recorded by Science and Technology university of China.

Corpus based phonetic research is done on the annotated 863 recognition and synthesis corpora in these aspects:
- . Tone and intonation [6]
- . Prosodic structure and the acoustic features of prosodic phrase [4,7,9]
- . Duration of phonetic units [17]

Most of these researches were made on isolated sentences that are far from the needs of the speech engineering. on the contrary, discourse corpus can provide us enough information to investigate the relationship between sentences ( through the information structure and coherence such as anaphora annotation ), the prosodic structure in discourse and the “mapping” rule to the syntactic structure, the intonational structure in discourse etc. For synthesis it can provide the prosodic model and stress model of discourse rather than isolated sentence to make the output voice more natural and more intelligible. For recognition it can be used to investigate the sound variation ( assimilation, vowel weak, consonant or vowel deletion and voiced of the unvoiced segments ) and the factors affecting sound variations and to make phonetic modeling of Chinese.

1. ASCCD – READ DISCOURSE CORPUS

To get good understanding of prosodic features and find the basic prosodic unit and the sound variation in continuous speech of Standard Chinese, we collected large amount of speech discourses. Eighteen texts which contain 300-500 syllables for each and which cover major discourse structures such as coherence relations as well as the phrasal structures were selected. Five male and five female speakers read this 18 discourses in sound proof recording room. The speech signal is recorded in two channels on DAT: speech waveform and the glottal impedance waveform through Laryngograph. Finally the digital data on DAT were transferred to WAV files through Sound Blaster Live and segmented into small
files according to the paragraphs of each text.

2. LABELING

2.1 Segmental Labeling

2.1.1 Segmental Labeling System

Segment labeling is a basic work to label segment in speech corpora labeling for Standard Chinese. Above segment labeling, we can give other labeling such as prosody labeling and syntactic labeling. With the labeled material we can do many work. It is worth discussing how to make segment labeling work well and how to make it acceptable. Chinese PinYin is an effective way to transcribe Standard Chinese. But for some reasons, it is not entirely one to one mapping to IPA. For example, “i” representing [i], [ɨ], [ɨ]. It is not easy to be as a machine-readable symbol system. According to international machine readable symbol system SAMPA [2], Zhu Weibin and Zhang Jialu have transcribed a symbol system with SAMPA for labeling syllable.[14, 15]. They give Chinese SAMPA symbols including consonant, vowel and tone charts according to Xu Shirong’s view. But it is not enough to label continuous speech whose representation is more complex than isolated syllable. There are sound variation phenomena in continuous speech such as centralization, reduction, insertion etc.. The labeling convention should be expanded and be flexible enough to annotate these variations.

Labeling principles

Based on these, we design SAMPA-C labeling system for Standard Chinese. Before we make the labeling system for continuous speech, we formulate some principles. The principles of formulating labeling system as following:

1. Proper: It is very important to give the most approximate IPA transcription for each segment in continuous speech for Standard Chinese.

2. Simplify: for every segment, we use a simple manner to transcript. For example, there are not voiced stop and voiced affricative consonants in isolated syllable in Standard Chinese. But they often occur in continuous speech. So, we just give voiced symbol “_v” in SAMPA-C not give voiced stop or voiced affricative.

3. Corresponding to SAMPA: We hope to give the precise mapping from segment’s IPA to SAMPA-C.

We have made a labeling system in syllable tier last year [1]. Now we make it in a continuous speech tier. What we rely on is Luo Changpei’s view [11] for consonant and vowel. For retroflex final, we rely on Wang Lijia’s result [10]. Then, we give diacritics for sound variation and give non-speech labels.

2.2 Prosodic Labeling

2.2.1 C-ToBI- Chinese Prosodic Labeling System

The phonetic features with functional significance in linguistics are phonologically labeled. Five principles of labeling are decided to guide us what to include and what to leave out

1. Labeling the tonal variation and intonation and stress and prosodic structure that have linguistic functions. So the tonal coarticulation between syllables is not labeled, but the tonal coarticulation caused by stress is labeled.

2. Prosody are quantitatively labeled and those qualitatively data are not labeled such as duration and amplitude.

3. Some uncertainty is permitted to avoid providing the wrong information for the user.

4. The transcriptions are machine-readable and easy to operate.

4. High inter-transcriber agreement.

This labeling system is the second version for discourse[5,7]. We think that the prosodic structure of SC is hierarchically organized from small to large constituent as syllable, prosodic word (PW), minor phrase (MIP), major phrase (MAP) and intonation utterance (IU). Prosodic word consists of one or more lexical words but with one word stress. Minor phrase consists of one or more prosodic words and bears one minor phrase stress. Major phrase consists of one or more minor phrase plus one major phrase stress. Intonation group consists of one or more major phrases plus one utterance stress. Five parallel tiers are labeled for each sentence in our system:

1. Orthographic tier: PinYin and tone number is annotated for each syllable.

2. Tone and intonation tier: tonal features and the change of register and range are marked.

3. Sentence function tier: four sentence types are annotated (interrogative, imperative, statement and exclamation).

4. Break index tier: three kinds of breaks are tagged - minor phrase, major phrase and sentence break.

5. Stress/prominence tier: normal stress or contrast stress of each sentence is labeled.

The detailed labels and the description are shown in Tab.1.

Table 1. labeling system

<table>
<thead>
<tr>
<th>tier</th>
<th>labels</th>
<th>description for C-TOBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>S, Q,</td>
<td>S: statement Q:</td>
</tr>
<tr>
<td></td>
<td>I, E</td>
<td>imperative; I:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>imperative; E:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>exclamation</td>
</tr>
<tr>
<td>5</td>
<td>stress</td>
<td>Five categories are</td>
</tr>
<tr>
<td></td>
<td>index:</td>
<td>labeled on stress</td>
</tr>
<tr>
<td></td>
<td>0-4</td>
<td>tier: 0-4 for non-break, PW, MIP, MAP and IU.</td>
</tr>
<tr>
<td>4</td>
<td>break</td>
<td>Five break categories:</td>
</tr>
<tr>
<td></td>
<td>index</td>
<td>0-4 for non-break, PW, MIP, MAP and IU.</td>
</tr>
<tr>
<td></td>
<td>0-4</td>
<td></td>
</tr>
</tbody>
</table>

Tonal and intonational features

Boundary tone

^ upstep
^ wide upstep
! downstep
!! wide downstep


Register shifting, register is shifted upward or downward, ( ) for the scope

Five categories for register shifting, register is shifted upward or downward, ( ) for the scope

R^ ( )
R^ ( )
R! ( )
R!!
& transitional tone
1-5 ? uncertainty

Table 2 The consistency checking results

<table>
<thead>
<tr>
<th>Transcriber pairs</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-L</td>
<td>73.66%</td>
</tr>
<tr>
<td>S-H</td>
<td>83.14%</td>
</tr>
<tr>
<td>S-C</td>
<td>71.97%</td>
</tr>
<tr>
<td>L-H</td>
<td>76.87%</td>
</tr>
<tr>
<td>L-C</td>
<td>90.04%</td>
</tr>
<tr>
<td>H-C</td>
<td>75.00%</td>
</tr>
<tr>
<td>Total</td>
<td>78.00%</td>
</tr>
</tbody>
</table>

2.2.2 Consistency For Prosodic Labeling

We checked the consistency on Break Index tier for each scriber pair and 4 transcribers. The results are shown in Table 2. We analyzed the results and found that the low consistency was mainly caused by the confusion of Break index 1 and 2 which provided another evidence that there is not a clear definition for word and phrase in Chinese.

2.2.3 Break Index 4 In Discourse

Break index 4 indicates the prosodic group boundary. Most of these boundaries 4s are isomorphic with the syntactic sentence boundaries. It can contain one or several major phrases with F0 down stepping and reset one by one to a lowest point shown in Fig 1.

2.2.2 Consistency For Prosodic Labeling

We checked the consistency on Break Index tier for each scriber pair and 4 transcribers. The results are shown in Table 2. We analyzed the results and found that the low consistency was mainly caused by the confusion of Break index 1 and 2 which provided another evidence that there is not a clear definition for word and phrase in Chinese.

Table 2 The consistency checking results

<table>
<thead>
<tr>
<th>Transcriber pairs</th>
<th>Consistency</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-L</td>
<td>73.66%</td>
</tr>
<tr>
<td>S-H</td>
<td>83.14%</td>
</tr>
<tr>
<td>S-C</td>
<td>71.97%</td>
</tr>
<tr>
<td>L-H</td>
<td>76.87%</td>
</tr>
<tr>
<td>L-C</td>
<td>90.04%</td>
</tr>
<tr>
<td>H-C</td>
<td>75.00%</td>
</tr>
<tr>
<td>Total</td>
<td>78.00%</td>
</tr>
</tbody>
</table>

2.2.3 Break Index 4 In Discourse

Break index 4 indicates the prosodic group boundary. Most of these boundaries 4s are isomorphic with the syntactic sentence boundaries. It can contain one or several major phrases with F0 down stepping and reset one by one to a lowest point shown in Fig 1.

3 CORPUS BASED PHONETIC RESEARCH

3.1 Statistic Analysis For The Phonetic Segments

The occurrence distribution of syllable and demisyllable, the duration distribution of syllable and demisyllables are calculated. The duration of each syllable is coded or normalized according to each speaker’s average duration and the standard deviation.

3.2 Prosodic Research

The prosodic research here focus on the sentence stress that involves the specification of relative prominence in prosodic structure. Prosodic word and its prominence and prosodic phrase are examined in the perception experiment. It seems that the hierarchical stress in sentence spoken is one of intonational cues in Chinese. Tone and intonation in Chinese are two different phonological events in spoken sentence.

3.2.1 Perception Experiment

Speech material: 59 sentences taken from ASCCD of three speakers. Listeners are 20 naive and untrained listeners.

One listening test is to decide the boundary of prosodic phrase, another one is to decide what are the prosodic word and its prominence.

Prosodic words were sliced from each utterance by “gating equipment” in Kay Mul-speech model 3700. In each prosodic word, which syllable(s) sound acute and intense was judged by three listeners. The syllable(s) perceived with acute and intense refers to prominent part in prosodic word.

3.2.2. Acoustic Analysis

The data of F0 and duration (T) of each syllable in utterance were measured from the spectrogram made by Kay Mul-speech model 3700. Normalization of F0 and T are used in calculation.

\[ F_0 : J = 12 \times F_0 (\log F_0 / F_{0 \min})^2 (F_{0 \max} / F_{0 \min}) \]
The stress in prosodic word refers to the prominent part among the prosodic words contained. The utterance stress is the prominent part among phrases contained. Also, it is the stress in utterance that is the most acute and intense syllable in its prosodic phrases contained. In Standard Chinese, Stress has its hierarchical pattern.

For example, Figure 2 shows the normalized $E_0$ and $T$ of each syllable in the utterance “国际航空公司飞上海的航班因大雾取消了”。 “国际航空公司” and “因大雾” are major prosodic phrases, as they are separated with pause without silence. “国际航空公司” and “因大雾” are minor prosodic phrase, as they are separated by break without silence that is caused by the lengthening of “” in “国际航空公司”. Also, The $F_0$ manifestation in “国际航空公司” makes it a compound prosodic word. “国际航空公司” and “因大雾” are compound prosodic words, “国际” and “公司” are pure prosodic words.

In “国际航空公司”, “国际” is more acute and intense than “航空公司”; In “国际”,”国际” is more acute and intense; “国际” is also more acute and intense than “国际航空公司”. “国际” is more acute and intense than “航空公司”. “航空公司” is also acute and intense. The prominent part in prosodic word is those one or two syllables that are acute and intense in perception.

The stress in prosodic word refers to the prominent part in the stress phrase in prosodic phrase is the most prominent part among the prosodic word contained. The utterance stress is the prominent part among phrases contained. So, “” and “” are the stress in each prosodic phrase, because the $F_0$ range is wider than that in others. The stress in this utterance may be in “”. In Standard Chinese, stress has its hierarchical pattern as shown in Figure 3.

### 3.2.3. Conclusion

Chinese is a tone language; Tone is lexical specified. $F_0$ in syllables can be varied to different extent, even to lose its identity, due to the effects of tone sandhi and the perturbation by $F_0$ coarticulation. The variations in $F_0$ of syllables are the events that are due to the intersyllabic action, of course, $F_0$ coarticulation across adjacent syllable. However, rise or down of $E_0$ register and expansion or contraction of $E_0$ range is caused by utterance, it is the events that are due to utterance level. It seems to us that tone information has been differentiated from stress pattern in utterance. The hierarchical stress may be one of cues to Chinese intonation.
“国际航空公司飞上海的航班因大雾取消了”

3.3 Sound Variation

3.3.1 Sound variation

It is known that sound will change in continuous speech because of context or prosodic structure. There are some important phonetic phenomena which cause sound variation. Retroflex (儿化) and neutral tone (轻声) are the main parts. The others are insertion, deletion, assimilation, reduction, metathesis and other variation [11]. The further study for Chinese is Lin.[10]. In continuous speech, modal word such as “a” can change because of preceding phone “ia, ua, na, ra”. Retroflex often cause vowel centralizing or nasal deletion. Its representation is a changed sound with retroflex. There are 38 retroflex final in Standard Chinese. Neutral tone causes vowel reduction or deletion, for example, “dou4fu0 [ tòufu ]”. The vowel of “fu” often delete. “dong1xi0 [ dōngxì ]”. The vowel of “xì” is reduction.

The reason resulting in sound variation is complex. For free sound variation, context is the main reason. Coarticulation often occurs in continuous speech. Assimilation, deletion and reduction are the main representation. Prosodic structure is another important aspect. Sound variation often occurs in a prosodic word. Lexical and syntactic structures can affect sound variation too.

3.3.2 Sound variation in read speech

We transcribed the ASCCD with sound variation tier. The highest occurrence is that voiceless consonant becomes voiced consonant. It often occurs when the consonant is placed between two vowels or between a vowel and a voiced consonant. It is common that the structure of utterance is the sequence of CVCV or CVNCV. The structure of syllable in Standard Chinese is (C)V(N). C and N may not exist but V must exist in Standard Chinese. So, C is always adjacent with vowel or nasal. When syllables are produced fluently, it is easy to produce voiceless consonant as a voiced consonant. We found that nearly every voiceless consonant can become voiced. But the most frequent occurrence is unaspirated stops and unaspirated affricatives and fricatives. Aspirated stops and affricatives do not change that often.

The other evident one is assimilation. That is an anticipatory coarticulation, which is an important feature for Standard Chinese. For example, “tian1 an1men2 [tian an mÈn]”, the apical nasal [n] in “an” becomes bilabial nasal [m]; “fenbei [fenpei]” and “jian3ming2 [ jiàn mìng ]” is the same thing (Fig 4, top). According to our study, the sound variation occurs often not only within one word or within one prosodic word but also between
two words. For example, “dan4 bu4 mian3 zi4 si1 zi4 li4 但不免自私自利”, “但” and “不” are two words from acoustic representation. There is a 29ms silence between the two syllables. The final coda [n] of the first syllable changes to [m](Fig 4 middle). Another example is “dan4 bing4 mei2 you3cheng1 liang2 yong4 ju4 huo4 you3 ke4 du4 de0 rong2 qi4 但并没有称量用具或有刻度的容器” at the bottom of the Fig 4 ., there is 123ms silent pause between “但” and “并”. The nasal [n] of “但” becomes [m]. This kind of change is called free change because it does not affect the meaning of the syllable. It is general that “shi[ ]” for “是” deletes final or initial.

The reason is that the articulator for consonant and vowel of the syllable is the same. When it is not emphasized, it is easy to be left only consonant or vowel. But the duration existing may be longer than normal.

Fig. 4 [n] changes to [m] in different contexts: within a word ‘jian3ming2’ (top), across word boundary ‘dan4 bu4 mian3’ and across a phrase boundary ‘dan4 bing4 mei2you3’ (bottom).

4. SUMMARY

SAMPA-C has also been used to annotate CASS -a spontaneous speech corpus [3]. The different between read and spontaneous speech is compared in another paper [3]. We don’t discuss the syntactic labeling or the information structure or the anaphora labeling here for the limited space. However, how to annotate the prosody and syntax for spontaneous discourse is a very important research work to carry on.

REFERENCES
