Spontaneous Conversation Corpus CADCC

Li Aijun, Yin Zhigang, Wang Maolin
Xu Bo, Zong Chengqing **

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

1. ABSTRACT

Chinese Spontaneous Dialogue and Conversation Corpus (CADCC) has been collected and annotated. The general information and the orthographic and prosodic and segmental annotation system in this corpus are described in this paper. Some statistic results are introduced including the discourse topic number, the turn takings, and the paralinguistic phenomena.

2. PHONTANEOUS SPEECH CORPUS COLLECTION

Discourses include read speech and spontaneous speech and can then be categorized into monologues and conversations. Different forms of discourses have different attributes such as the number of people, topic, turn order and turn length.

We have collected a read discourse corpus ASCCD and a spontaneous monologue corpus CASS and made the phonetic transcription on segmental and prosodic tiers [2]. With the developments of the speech applications, a corpus of spontaneous conversations becomes more and more imperative. So we are collecting and annotating a spontaneous dialogue and conversation corpus CADCC which includes two subsets: a telephone corpus (set 1) and a conversation corpus recorded (set 2). The detailed corpus information is listed in table 1. The telephone subset was recorded through the hotel fixed lines so most of the conversation is about hotel reservation and some speakers have Chinese regional accents. The conversations in subset 2 were recorded in the ordinary rooms. The speakers recruited are colleagues or good friends themselves who speak Standard Chinese and have common interesting topics so that they can change their topics freely and naturally during an hour's conversation.

2. Orthographic transcriptions

All the recorded telephone and dialogue speech is transcribed orthographically with detailed spoken phenomena shown in table 2.

Here are some Chinese Character transcription examples:

This project is supported by National 973 high tech. and CASS foundations. This paper was read in Oriental COCOSDA'2001, held in Korea.

ex 1：B:**我傻** OV< B:**我印度人**;A:LA<LA>OV>

ex 2: A:LE<MO<**噢** LE> MO>;B:**那个就是** DS<**一一** DS>**锤子的事儿。**

Table 1. Corpus information

Specifications	Set 1 (telephone)	Set 2 (dialogue)
Content	Hotel reservation	No limitation
Total length	2 hours	14.2 hours
Speakers	More than 200 speakers	11 male and 11 female speakers
Regional Accent	√	×
Chinese Character transcription	√	√
Linguistic annotation	Prosodic, segmental, syntactic	Prosodic, segmental, syntactic
Sampling rate	8 KHz	16 KHz
Storage form	.wav files	.wav files

2.1 Discourse topic

In order to make acoustic analysis on conversational interaction [8], we annotated the topic shift points in the text. Here the topic is discourse topic rather than sentence topic. We find that discourse topics are hierarchically and nonlinearly organized, that is, a topic can be disrupted by another one and then continues.

The discourse topic number (sub-topic is excluded) for each pair of speakers and the duration information is shown in table 3.

Figure 1 gives the topic number distribution according to their lengths. The average length of each topic is 185.96 second disregarding the individual profiles such as speaking rate.

Table 4 is the accumulative frequency of topic occurrence from which we can observe that 80 percent of topics are less

than 260 seconds i.e. 4-5 minutes and 90 percent topics are less than 5-6 minutes.

Table 2. nonlinguistic / paralinguistic phenomena labels

NO.	phonomena		labels	
			start	end
1	Para-linguistic	lengthening	LE<	LE>
2		breathing	BR<	BR>
3		laughing	LA<	LA>
4		crying	CR<	CR>
5		coughing	CO<	CO>
6		disfluency	DS <	DS >
7		error	ER<	ER>
8		silence (long)	SI <	SI >
9		murmur/uncertain segment	UC <	UC >
10		modal/exclamation	MO <	MO >
11		smack	SM <	SM >
12		non-Chinese	NC <	NC >
13		sniffle	SN <	SN >
14		yawn	YA <	YA >
15		overlap	OV <	OV >
16		interjection	IN <	IN>
17		deglutition	DE <	DE>
18		hawk	HA <	HA>
19		sneezes	SE<	SE>
20		filled pause	FP <	FP>
21		trill	TR <	TR>
22	non-linguistics	noise	NS <	NS >
23		stable noise	TN <	TN >
24		beep	BP <	BP>

Table 3. Topic and duration information.

Speakers	Genders	Total Topics	TOTAL Duration (sec.)	Sec. / per topic
SUNXI	MEN	42	6524.859	155.3538
ZHNGJ	MEN	35	4197.666	119.9333
XINGY	WOMEN	14	1953.768	139.5549
CHENX	WOMEN	13	3062.002	235.5386
DUYU	WOMEN	27	4002.587	148.244
SONGW	MEN	11	2911.886	264.7169
LIUJI	MEN	10	2718.652	271.8652
LVJIN	WOMEN	27	4177.172	154.7101
XUCHA	MEN	15	6221.138	414.7425
DURUI	MEN	32	5859.053	183.0954
TANJI	MEN	7	2531.122	361.5889

Fig 1. Topic distribution in length

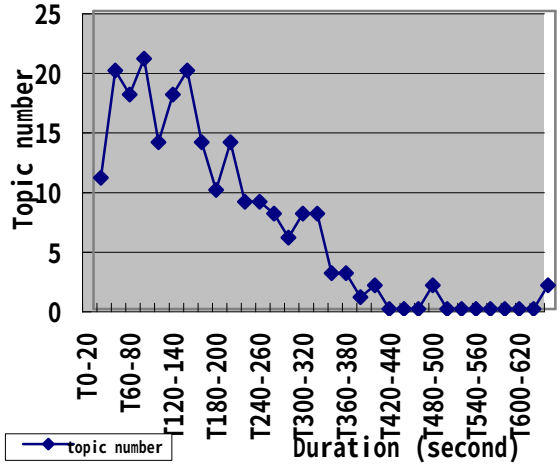

Table 4 Accumulative frequency of topic occurrence

Duration (Second).	Topic occurrence frequency
<160	60%
<200	72.3%
<260	83%
<320	93.6%
<620	100%

2.2 Paralinguistic and nonlinguistic phenomena

Some paralinguistic and nonlinguistic phenomena are statistically calculated in set 1 and set 2 respectively as shown in table 5. The phenomena with bold font are of top occurrences.

The highest numbers are for modal or exclamation words in two subsets, such as "**啊、呀、哇、吗、呢、吧**".

2.3 Turn taking

One of the important aspects in discourse analysis is the turn shifting mechanism, which we did some study using the CADCC corpus. The column on the left in Figure 2 represents the turn-shifting patterns between speakers A and B, and the column on the right represents the A-B-A patterns.

Table 6 shows the numbers of turns in the two sub-corpora. In set 1 there are 3256 turns and 315 overlapped turns and 40 'interjections' with 'overlap'. The overlapped turns account for almost 20.6% (（2*315+40）/3256) of the total turns. In set 2, there are 18608 turns and 986 overlapped turns account for almost 10.6% (986*2/18608)) of the turns.

	Phenomena	Lablels	Occurrce Times in set 1	Occurrce Times in set 2
1	lengthening	[LE]	**669**	**1302**
2	breathing	[BR]	43	**646**
3	laughing	[LA]	40	**1082**
4	crying	[CR]	0	0
5	coughing	[CO]	8	90
6	disfluency/error	[DS]	17	**4135**
7	noise	[NS]	**284**	**596**
8	silence (long)	[SI]	?	**685**
9	murmur	[UC]	**633**	**2223**
10	modal/exclamation	[MO]	**2057**	**8310**
11	smack	[SM]	80	**568**
12	non-Chinese	[NC]	11	10
13	sniffle	[SN]	5	175
14	yawn	[YA]	6	12
15	overlap	[OV]	**315**	**2904**
16	interjection	[IN]	68	**1337**
17	deglutition	[DE]	5	52
18	hawk	[HA]	9	40
19	sneezes	[SE]	6	2
20	beep	[BP]	**338**	0
21	topic start	[TP]	?	206

Table 5. Paralinguistic and nonlinguistic phenomena

From Table 6 we can see that there are more overlaps in the telephone corpus than the conversation corpus, which we speculate it's because the information exchanged in the telephone dialogue is relatively intense, or because the speakers are in haste, so that they respond before the other partner finishes speaking in order to save telephone fee.

Usually the speech recognition system regards the overlaps as noise and this phenomenon is not treated properly in the dialogue system. We do find the person-to-person dialogue is different from the person-to-machine dialogue system in that people cannot interrupt a machine.

Does it mean that the study for overlaps is meaningless? We think that the psychological and cognitive mechanism in communication can be explained through the study of the frequently appeared overlaps. In actual interaction, the listener can make response before the speaker finishes his words, which means that people can get the complete idea after hearing the first part of a sentence, and this is quite useful to study this mechanism. In the long run, this is a project we must deal with in the phonetic study.

Table 6. Total turns in set 1 and set 2.

Turns		Numbers of turns
SET 1	A	1635
	B	1621
SET 2	A	9284
	B	9324

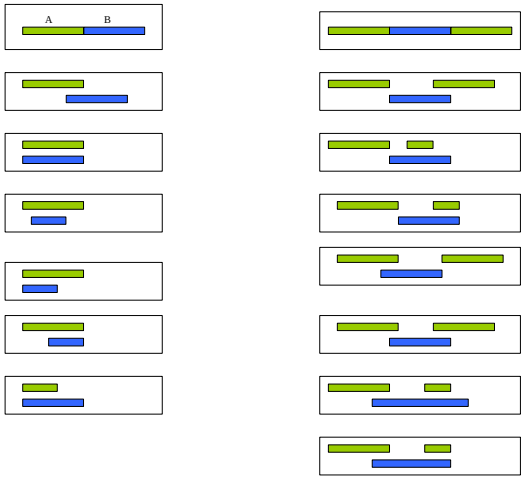

Fig. 2 The turn taking scheme in CADDC. A and B stand for the two speakers. The left column is for A-B, the right column is for A-B-A

3. PHONETIC ANNOTATION

3.1 Segmental annotation

Segmental labeling with sound variability is very useful not only for phonetic modeling of speech recognition, but also for sound change study. For example, the faster the speech rate is, the more sound changes there will be. Furthermore, segmental labeling is the basis to make other annotation, such as prosodic and syntactic labeling.

Praat [7] is used to make the segmental labeling. The speech is transcribed in a five-tiered annotation as followings:

- PY: **拼音层**, canonical pinyin and tones are labeled, based on the standard pronunciation of the word. Retroflexed syllable is labeled both on this tier and on the SY tier. "?" is marked for uncertain segment.
- SY: **声韵层**, initials and finals and sound variability such as phoneme change, insertion and deletion are also transcribed on this tier The symbols for sound changes are from SAMPA-C [1] shown in table 5 "?" stands for uncertain segments; ii for **[1]** ; iii for [ ]. Tonal variation and tone sandhi is labeled as well.
- MIS: **杂类层**, spoken or paralinguistic phenomena are transcribed. The symbols used are listed in table3.
- ACC: **口音层**, regional accents of each speakers are annotated on this tier.
- SM: **语句功能层：**Four kinds of sentence modes are

annotated on this tier. ‘S’ stands for statement, ‘Q’ for interrogative, ‘I’ for imperative and ‘E’ for exclamation. ‘B’ for some bitty or unfinished utterances occurred in spoken dialogue.

Consistency for segmental labeling is checked strictly within 2 ms time difference for any of two compared pairs for two transcribers. The results are as shown in table 7.

tiers	Consistency
PY	87 %
SY	81 %
AVERAGE	84 %

Table 7. The consistency of PY and SY tiers

3.2 Prosodic annotation

We have released two versions of prosodic labeling conventions, entitled C-ToBI 1.0. and C-ToBI 2.0 [3]. The latter one is developed with more prosodic research results on Standard Chinese and more efficient for discourse and spontaneous speech [4,5,6].

The phonetic features with functional significance in linguistics are phonologically labeled.

We think that the prosodic structure of Chinese is hierarchically organized, from small to large constituent as syllable, prosodic word (PW), minor phrase (MIP), major phrase (MAP) and breath group (BU). Eight parallel tiers are labeled for each sentence in our system: The detailed labels and the description are shown in Table 10.

Consistency on stress and break index tiers is checked for two transcribers using 5 minutes speech as shown in table 8 and 9.

We found that the turns in corpus Set 1 is much shorter than those in Set 2. Some of the turns in Set 2 are as long as a paragraph, so the distribution of the prosodic boundaries is similar to that of read speech. But in the telephone corpus, which is quite different from Set 2, the speakers tend to make quick response, so the turns are quite short and more closely linked, and there are few major phrases.

Table 8. The consistency for stress index tier

Stress Index	Meaning	Consistency
3	Major phrase stress	91.87%
2	minor phrase stress	90%
1	prosodic word stress	91.93%

Table 9. The consistency for break index tier

Break index	Meaning	Consistency
4	intonational group	100%
3	major phrase stress	66.67%
2	minor phrase stress	95.71%
1	prosodic word stress	93.55%
?	uncertain	91.93%

4. MORE INFORMATION

ASCCD is a read discourse corpus with phonetic annotation [2]. In order to make comparative study between spontaneous speech and read speech, we have collected the conversations in CADDC with the same speakers in ASCCD.

The phonetic annotation for set 1 has been finished and is carrying on for set 2.

The Web site of our lab is:
<http://www.cass.net.cn/s18_yys/yuyin/Default.html>

5. REFERENCES

[1] Chen, Xiaoxia Zu, Yiqing & Li Aijun), "A cardinal labeling system for Standard Chinese", *the proceeding of the 4th National Modern Phonetics*, edited by Lv shinan, published by JinCheng Publishing House,1999.

[2] Li Aijun, Chen Xiaoxia, et al (2000), "The phonetic labeling on read and spontaneous discourse corpora", *ICSLP'2000*.

[3] Li Aijun, etc.,"A national database design for speech synthesis and prosodic labeling of standard Chinese", In *proceedings of oriental COCOSDA'99*, TaiPei, TaiWan.

[4] Li Aijun, "The study on the phrasal and sentential accent", *the proceeding of the 4th National Modern Phonetics*, edited by Lv shinan, published by Jin Cheng Publishing House,1999.

[5] Lin Maocan, "Breaks and prosodic phrase in Standard Chinese", *Contemporary Linguistics, No.4,2000.* (In Chinese)

[6] Lin Maocan, "Hierarchical Stress and F0 Restructuring in Utterance of Standard Chinese---One of the Cues to Chinese", *Intonation proc. of SFSSLP’2000.*

[7] www.praat.org

[8] Gillian Brown and George Yule, Discourse analysis, Cambridge University Press,1983.

Table 10 C-ToBI 2.0

Tiers	Labels	Description
1 PinYin tier	ba1, ba2, ba3, ba4, ba0 ……	Canonical Pinyin and tone
2 Tone and intonation tier	H-L, L-H, H-H L-L, H, L,L% H%	tonal and intonational features, boundary tone
	^, ! ,^^, !!	^ upstep, ^^ wide upstep ! downstep, !! wide downstep . For pitch range
	R^ ( ) R^^() R! ( ) R!!	register is shifted upward or downward, ( ) for the scope
	&	transitional tone
4 Break index tier	break index: 0-4	five break categories. 0: syllable boundary within PW, 1: PW, 2:MIP,3: MAP and 4:IU .
5 Stress index tier	stress index: 0-4	five categories are labeled on stress tie. :0: for unstressed syllables, 1: PW,2: MIP,3: MAP and 4: IU
6 Sentence function Tier	S, Q, I, E B	S：statement Q：interrogative； I：imperative； E：exclamation B: for the bitty utterance
7 Miscellaneous tier	Laugh, Cry, Smack, noise …	< laugh :begin of laugh>laugh; end of laugh
8 Accent tier	SH, GZ …; SHR,GZR…	<SH: begin of Shanghai Accent >SH; end of Shanghai Accent; <SHR, >SHR Standard Chinese with Shanghai accent.
For 1-8	?	Uncertainty

口语对话语音语料库 CADCC

李爱军, 殷治刚, 王茂林, 徐波, 宗成庆*

中国社会科学院语言研究所， * 中国科学院自动化研究所

摘要

口语对话和朗读语篇的差别表现在句法、副语言学现象、音段和韵律等许多方面，这给口语对话的标注带来新的课题。本文介绍自然口语对话语音语料库 CADCC（Chinese Annotated Dialogue and Conversation Corpus）**和其文字转写、音段以及韵律标注。**CADCC **包括两个子库：电话对话库** set1 **和口语对话语篇库** set2**。其标注内容包括篇章话题、话轮、韵律和音段的标注。音段标注采用** SAMPA-C **标注系统，韵律标注采用** C-ToBI **标注系统。**

本文还详细报告了标注结果，如篇章话题的长度，口语话轮出现的模式，插入和叠接现象，韵律结构和朗读语篇的差异等等。