International Conference on Spoken Language Processing, pp. IV-724~727, Oct. 16-20, Beijing

THE PHONETIC LABELING ON READ AND SPONTANEOUS DISCOURSE CORPORA

Li Aijun, Chen Xiaoxia*, Sun Guohua*, Hua Wu*, Yin Zhigang*, Zu Yiqing*,*
*Zheng Fang**, Song Zhanjiang***

*Phonetic laboratory, Institute of Linguistics, CASS
** Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University
Email: Liaj@linguistics.cass.net.cn; fzheng@sp.cs.tsinghua.edu.cn

ABSTRACT

Read and spontaneous discourses are two different but very significant speech styles to be investigated. So phonetic labeling on read and spontaneous discourse corpora are made one is ASCCD, a 10 hours read discourse corpus and the other is CASS, a 4 hours spontaneous discourse corpus. First the principles and conventions of transcription are presented. Then, these two speech styles are compared from phonetic and syntactic point of view, including the statistic results of different phonetic units got from the annotated corpora.

INTRODUCTION

From the development of language, spontaneous speech is an archaic, common used and typical form of the language. In the past decades from 50s to 80s of the 20th century, we focus on read speech to do our research in three fields of acoustics, psychology and physiology. In the recent 10 years, the research on spontaneous speech are becoming more and more important for the speech applied technology and the associated theories. Spontaneous speech rather than read speech is one of the unresolved problems faced by many speech recognition systems.

The types of speech data can be characterized along the following continuum according to the degree of spontaneity: Reading tex t-> TODS -> Elicited -> Classroom -> News/Speech -> TV drama -> Talk shows -> Natural Discourse [4]. The more spontaneous the speech is the less we understand it.
The aim to make the phonetic labeling on read and spontaneous discourse corpora is (1) to get the segmented speech corpora for speech application technology. (2) to make prosodic labeling and syntactic labeling based on them. (3) to investigate and compare the phonetic features between these two corpora. So not only orthographic Pinyin for each syllable but also phonetic variability such as insertion and deletion are transcribed. Also SAMPA-C phonetic labeling convention for Standard Chinese is developed based on SAMPAX[1,2] .

Many differences exist between read and spontaneous speech in Chinese. This paper gives some illustrations and then depicts the results on read and spontaneous speech of phonetic segments. For example the duration distributions and the occurrence frequency and the accumulative probability of different segments are calculated. The results are compared between two speech corpora.

1. ASCCD -A READ DISCOURSE CORPUS AND CASS – A SPONTANEOUS DISCOURSE CORPUS

ASCCD was setup and recorded in the institute of linguistics, Chinese Academy of Social Sciences. It contains eighteen discourses which cover major discourse structures such as coherence relations as well as the phrasal structures. Each text contains 300-500 syllables and several paragraphs. Five male and five female speakers read this 18 discourses naturally. The speech signal is recorded in two channels: speech waveform and the glottal impedance waveform through Laryngograph [7].

CASS is originated from 19 cassettes provided by the Broadcast Station of TSINGHUA University (BSTHU), Beijing, China. The content in these cassettes includes lectures addressed by some teachers and invited talkers, colloquiums among some students, and so on. Most of the speech in the cassettes is causally given without paper preparation, thus it is natural and covers a lot of valuable spontaneous phenomena. The background noise and the spontaneity slow down the speech analysis precision and the labeling speed. (1-hour raw speech needs about 380 hours to be transcribed into orthographic syllables and SAMPA-C sequences with detailed pronunciation variants!). More than 6-hours' speech of higher acoustic quality is afterwards selected from the raw speech corpus and broken into smaller sentences having the duration of 2.5-4.0 seconds each.

Again, more than 3-hour casual speech of 7 speakers is manually picked out of the 6-hour sentences by discarding those whole-noise-sentences and keeping those sentences of relatively clearer utterances and of more casual speaking styles. [7]

2. THE PHONETIC LABELING ON ASCCD AND CASS

2.1 Segmental labeling convention – SAMPA-C and labeling principle

In order to provide the acoustic indicator of the sound variations, multilevel transcriptions are made. The labeling principle is :

(1) Adopting machine-readable phonetic alphabet and which is an open system for different dialects.
(2) Labeling the phonetic variability and spoken phenomena.
(3) High labeling consistency.

2.1 Labeling Consistency and Tiers

For CASS three layers are labeled including the syllable tier, demi-syllable (initial and final tier) and miscellaneous tier. In the syllable tier, *pinyin* and *tone* of each syllable is transcribed orthographically. In the demi-syllable tier, *initial and final* of each syllable are segmented and labeled using SAMPA-C convention[]. Sound variability such as assimilation and phoneme insertion and phoneme deletion is also transcribed on the semi-syllable level. Tones after tone sandhi or tonal variation are attached to the finals. In the miscellaneous tier the phenomena of spoken discourse are transcribed such as coughing and laughing and smacking.

For ASCCD, also three layers are labeled including the syllable tier, demi-syllable (initial and final tier) and sound variability tier. The first two tiers are transcribed canonically and the third tier gives the sound variability by using SAMPA-C .

Four transcribers transcribe the CASS and ASCCD by first checking the consistency between each other and the joint agreement on PinYin and Demisyllable tiers. The observed consistency for 15 minutes speech of CASS is given somewhere in this proceedings [3].
The consistency for ASCCD is above 95% which is very high on PinYin and demisyllable tiers.

3. READ AND SPONTANEOUS SPEECH – KNOWN AND NEW RESULTS

3.1. Syntax

In Chinese, Read and Spontaneous speech manifest quite differently in syntax. It is shown from the statistic results in [5,6] that the most frequently used clause in read speech is the “SPVO (Subject phrase + verb + object)”, while in spontaneous speech is the elliptical clause. The natural unit in discourse is not what has been assumed in syntactic theories, it should based one the prosodic segments.

3.2 Spoken phenomenon

Spoken phenomenon is another indicator to differentiate the read and spontaneous speech. Table 1 is the spoken phenomena occurring in 4,000 spontaneous sentences in the annotated spontaneous corpus of 4 hours of CASS. These spoken phenomena seldom exist in read speech except lengthening. So in ASCCD there is not a miscellaneous tier to labeling the spoken phenomena.

Table 1. Spoken phenomena occurring in 4,000 spontaneous sentences

No.	Spoken phenomena	Occurring times
1	Lengthening	409
2	Breathing	401
3	Laughing	40
4	Crying	0
5	Coughing	65
6	Disfluency	230
7	Noisy	627
8	Murmur/uncertain	567
9	Modal / exclamation	1511
10	Smacking	40
11	Not Chinese	18

3.3 Sound variation

We labeled the sound variation such as insertion, deletion, pharyngrealization, voiced, voiceless, nasalization, more round, more aspirated or breathy, centralization and phoneme change in read and spontaneous speech corpora respectively and found that the sound variability is 27.46% for initials and 12.02% for finals in spontaneous speech which is head and shoulders above that in read speech.
Insertion and the contexts for insertion in CASS are listed in Table 2 and the feature matrixes of sound variation for initial and final are given in Table 3 and 4.
The factors that causing the sound variability are discussed in other papers in this proceeding. []

Table 2. Insertions in CASS

Insertions	Count	Context
(N+)	13	-ng+a0 -> N+a0; ai-> N+ai
(m+)	11	-an + m
(n+)	1	
(t_h+)	8	Before k, d j

(x+)	3	
(z`+)	36	(zh)i+a-> ra
(N_h+)	4	(N_h+)+a0
total	76	

Table 3. The sound variability features of initials in CASS (10 samples)

Initials in PinYin <sampac>	occurrence	Unchanged	voiced	deletion	Phoneme change (Num. and the phones in SAMPA-C)	Aspirated / breathy	voiceless
[b <p>]	2318	1171	1117	8	7 (m)		
[p<p_h>]	324	305	17				
[m <m>]	2101	2092		3		1	4
[f <f>]	666	560	102	4			
[d <t>]	1999	2701	2709	7	4-t_h, 1-l		
[t <t_h>]	1460	1322	117	3	12-t, 6-t_v		
[n <n>]	2355	2347		5	1(m)		
[l <l>]	1915	1909		3	1(n)		
[g <k>]	3430	1231	2165	31	1(n)		
[k<k_h>]	745	687	38	4	13-k_v, 1-k,1-t_h)		

Table 4. The sound variability features of finals in CASS (10 samples)

fianls in PinYin	total	unchanged	deletion	voiceless	pharyngrealization	breathy	centralization	nasalication	Phoneme change (number and the phone in SAMPA-C)
a	2366	2276	10	15	7	18			33(7-10,@-9,AN-8,ei-3,AU-1,7~-2)
o	99	89	1						8(e) 6(7) 2(@)
e	6566	1915	28	11	3		4455	1	8 (@_n-6, aI-1, i\-1,)
i	4299	4230	44	19					y-1
(z)i	582	562	19	1					
(zh)i	2747	2450	273	6					59(i1)
u	2089	2051	24	12				1	
v	614	597	13	2					2 （i）
ai	1753	1746	1	3					
an	1326	1314	2	1					@_n-2, a_"

3.4 Distribution of Segmental unites

We made the statistic analysis using 3 hours annotated data in CASS and two speakers (a male and a female) annotated data in ASCCD.
The syllables and demisyllables (initials and finals) information are given in Table. In addition to this, the duration distribution for syllables and demisyllables are calculation shown in Fig 1 and Fig 2.

Observed the top 20 frequent occurring syllables, we found that 13 of them are same except that “wo zhei en me zuo ni ne” in CASS do not appear in the top list of ASCCD. But these syllables are often called “spoken words” which occur more frequently in spontaneous speech than that of in read speech.

For the top 20 frequent occurring demisyllables, we found that 14 of them are same except that “zh k n m l ei” in CASS do not appear in the top 20 of ASCCD.
Also the covering speed for syllables and demisyllables is given in table 5.

Table 5. Distribution of syllable and demisyllable

speech styles	read (two speakers) (sil included)	spontaneous (without sil)
syllable occurrence num. (without tone)	382	375
syllable occurrence times (without tone)	23399	47104
Initials and finals	175	342

occurrence num.		
Initials and finals occurrence times	38411	87104

Table 5. The covering speed (the number is the position in the sorted corpora)

covering percent		spontaneous	read
Syllable	50%	38	24
	80%	109	106
	100%	375	382
demi-syllable	50%	18	19
	80%	38	51
	100%	342	175

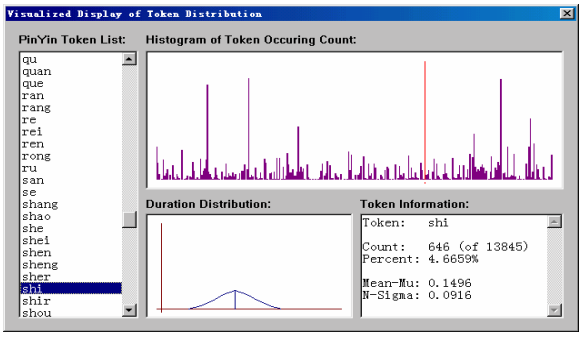

Fig 1. The distribution of occurrence and duration for each syllable (CASS)

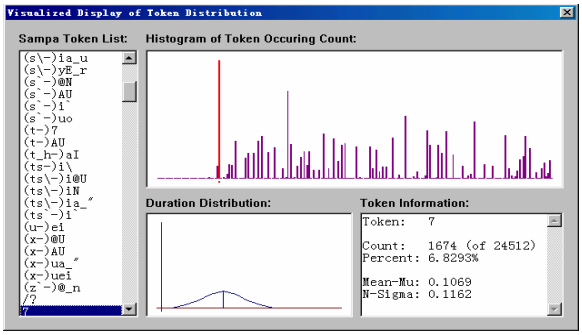

Fig 2. The distribution of occurrence and duration for each demisyllable in SAMPA-C (CASS)

4. WORK TO DO

Some phonetic or linguistic issues were encountered in the transcription work such as how to transcribe the tones of the modal function words as "呢 (ne)" and "吧(ba)". How to decide the phonetic variability of "shi"—when it should be the deletion of initial "(s`-)i`"and when it should be a voiced "sh" with the deletion of final "s`_v(i`-)".

The difference between these two speech types is not only on grammar, sound variability, spoken phenomena, but also on other aspects such as prosody. By now the prosodic labeling information for ASCCD is carrying on but nothing has been done for CASS. So this should be left for another paper.

We found that most of the speakers talked with accent or in dialect in many applied systems. All these corpora are not big enough or sufficient enough or scientific enough in data collecting to investigate the syntax and the phonetics of spontaneous speech especially for Chinese dialects. So the Institute of Linguistics of CASS are making their efforts to collect and setup a huge speech corpora referred to as "Spoken Corpora of Modern Chinese ". It includes a 1000 hour spontaneous corpus with different typical social interaction, a dialect spoken corpus with 50 thousand utterances on 3 dialect spots, a dialectic accent corpus on 3 spots. All these three corpora will be transcribed to texts and annotated phonetically and syntactically.

The corpus-based phonetic research for spontaneous speech will focus on the following aspects:

. sentence types of spontaneous speech
. spoken discourse structure
. the relationship between the syntactic structure and the prosodic structure
. cognition research on spontaneous phenomena.
. phonetic and syntactic feature of dialects and dialectic accent.

REFERENCES

[1] Chen Xiaoxia, Li Aijun, etc , "An Application of SAMPA-C for Standard Chinese". to appear in the proceedings of Icslp2000.

[2] John Wells, "Computer-coding the IPA: a proposed extensions of SAMPA". Unpublished notes, Department of Phonetics and Linguistics, University College London, http://www.phon.ucl.ac.uk/home/sampa/home.htm

[3] Li Aijun, Zheng Fang, William Byrne, Pascale Fung etc., "A Phonetically Transcribed Corpus of Mandarin Spontaneous Speech", in this proceeding of (ICSLP), Oct. 2000, Beijing.

[4] Li-Chiung Yang, Intonational Structure of Mandarin Discourse, Dissertation of Graduate Schools of Georgeton University, 1995.

[5] Luo Zhensheng, Yuan Yulin, "ji4 suan4 ji1 shi2 dai4 de0 han4 yu3 he2 han4 zi4 yan2 jiu1", TsingHua University Publishing House.

[6] Tao HongYing, Units in Mandarin Conversition. John Benjamine Publishing Company, Amsterdam/ Philadilphia, 1996.

[7] Zu Yiqing, Li Aijun, Chen Xiaoxia, etc. "Continuous Speech Database: From isolated Sentence to Discourse", Oriental COCOSDA'99, Taipei.