

SPEECH CORPUS COLLECTION AND ANNOTATION

Li Aijun, Chen Xiaoxia, Sun Guohua, Hua Wu, Yin Zhigang, Zu Yiqing

Phonetic laboratory, Institute of Linguistics, Chinese Academy of Social Sciences
Liaj@linguistics.cass.net.cn

ABSTRACT

This paper will particularly introduce a read and a spontaneous speech corpus to show how to collect and annotate the task dependent speech corpora. Additionally, segmental labeling convention SAMPA-C and prosodic labeling convention C-ToBI are depicted. Finally, known and new results are given or compared for these two annotated corpora.

1. INTRODUCTION

Many speech corpora have been collected or designed for different research purposes in China. By the development of speech techniques, read and spontaneous discourse corpus are collected and annotated to investigate the relationship between sentences, 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 and the factors affecting sound variations and to make phonetic modeling of Chinese.

This paper we will give detailed description of two discourse corpora CASS --- a Chinese annotated spontaneous speech corpus and ASCCD- an annotated speech corpus of Chinese discourse¹.

2. CORPORA INFORMATION

Corpus collecting is task dependent. We collected two different corpora for different research purposes. The primary goal of ASCCD is to investigate the phonetic and prosodic and syntactic structure for speech synthesis. So the discourses are finely selected covering the discourse structure and the different written style as many as possible. While the aim to collect the CASS is for recognizing the spontaneous speech using phonetic model. So, the casual speech is collected which is recorded in the normal rooms.

2.1 CORPUS INFORMATION FOR ASCCD

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 these 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.

2.2 CORPUS INFORMATION FOR CASS

The speech in the CASS corpus was provided by the Broadcast Station of Tsinghua University (BSTHU), Beijing, China, from their audio archives. The recordings are of university lectures by professors and invited speakers, student colloquia, and other public meetings. The collection consists primarily of impromptu addresses, delivered in an informal style without prompting or written aids. As such the collection is a rich source of spontaneous speech phenomena and is well suited for pronunciation modeling. Except for the female speaker F03, all the speakers were very fast[2].

Table 1. Corpus Characteristics

Speaker ID and Gender	Speech rate (syllables/sec.)	Dialect Background
F03	3.89	Wu
F04	4.68	Wu
M01	5.31	HuNan or JiangXi
M02	4.94	AnHui or HuBei
M04	4.99	Shanxi/AnHui
M05	4.14	Northwestern
M12	4.06	North Wu

3. CORPUS ANNOTATION

Why annotate a corpus? Why it is important to annotate a corpus? At least there are three reasons mentioned in [21]: extracting information, re-usability and multi-functionality.

3.1 Segmental Labeling

3.1.1 Segmental Labeling Convention- SAMPA-C

Chinese PinYin is an effective way to transcribe Standard Chinese. But for some reasons, it is not an entirely one to one mapping to IPA. For example, "i" represents [i], [ɨ], [ɪ]. According to the international machine readable symbol system SAMPA [2], we give Chinese SAMPA convention for labeling continuous speech including sound variation phenomena such as centralization, reduction, insertion etc.. It's more flexible.

The labeling system was designed according to the following principles: (1) Use of the machine-readable phonetic alphabet. (2) Accurate transcription of phonetic variability and spoken phenomena. (3) High transcriber consistency.

SAMPA-C used here includes canonical symbols for consonants and vowels, the initials and finals, the retroflexed finals [after 10], and labels for tones, sound variability, and

¹ ASCCD is funded by 973 Grand and the USTC Ifly INFOTEK Co. LTD. and CASS is supported by American NSF

spontaneous phenomena. A detailed description of SAMPA-C is given in the proceeding of ICSLP2000 [3].

3.1.2 Labeling Principles and Tiers

The principles of formulating segmental labeling system are (1) Proper (2) Simplify (3) Corresponding.

For CASS, the speech was transcribed in three tiers, which are syllable tier, semi-syllable tier, and miscellaneous tier. In the syllable tier, *pinyin* and *tone* of each syllable is transcribed orthographically. In the semi-syllable tier, the *initial* and *final* of each syllable is labeled using SAMPA-C. Segmentation boundaries are also provided in the semi-syllable tier. Sound variability such as phoneme change, insertion and deletion are also transcribed on the semi-syllable tier. Tones after tone sandhi, or tonal variation, are attached to the finals. In the miscellaneous tier, phenomena of spoken discourse are transcribed.

For ASCCD, PinYin and semi-syllable tier are labeled.

3.1.3 Transcriber Consistency

For CASS, to assess the agreement between annotators, 15 minutes of speech was transcribed in common by all four transcribers and their agreement was measured. The consistency of their transcriptions was measured in terms of number of transcriber pairs agreeing on the labeling of each particular segment; in these measurements.

The average transcriber agreement was found to be 84.23% in the semi-syllable tier when agreement on silence labels (including the closure segments before some initials) was counted. The average transcriber agreement at this Pinyin level is 86.12%, counting silences. Comparing Pinyin and semi-syllable agreement simultaneously, the transcriber agreement is 85.04%. The agreement was observed to be 88.92%, 85.88% and 87.14% without counting the silence labels for the above three situations.

The agreement in the semi-syllable tier and Pinyin tier, are reported individually and jointly in Table 2. Two measurements of agreement are given for each pair of transcribers. The number to the left of '/' is the percentage of labels in agreement counting silence labels, while the number to the right indicates percent agreement disregarding silence.

Table 2. Transcriber Consistency Measurements

Transcriber Pairs	SAMPA-C	Pinyin	Pinyin+ SAMPA-C
A-B	84.10 / 85.72	85.36 / 88.18	84.64 / 86.73
A-C	84.88 / 86.12	86.98 / 89.36	85.77 / 87.45
A-D	82.39 / 83.49	84.96 / 87.08	83.49 / 84.97
B-C	86.07 / 88.25	87.34 / 90.77	86.61 / 89.29
B-D	82.99 / 85.04	84.85 / 88.01	83.78 / 86.27
C-D	84.97 / 86.72	87.31 / 90.19	85.97 / 88.16

3.2 Prosodic Labeling

3.2.1 C-ToBI- Chinese Prosodic Labeling System

The phonetic features with functional significance in linguistics are phonologically labeled. Three parallel tiers are labeled for each sentence in our system:

- (1) Tone and intonation tier: tonal features and the change of

register and range are marked.

- (2) Break index tier: three kinds of breaks are tagged - minor phrase, major phrase and sentence break.
- (3) Stress/prominence tier: normal stress or contrast stress of each sentence is labeled.

The detailed labels and the description are given in 5.

3.2.2 Consistency For Prosodic Labeling

We checked the consistency on Break Index tier for each transcriber pair and 4 transcribers. The results are shown in Table 3. 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 3 The consistency checking results

transcriber pairs	consistency
S—L	73.66%
S—H	83.14%
S—C	71.97%
L—H	76.87%
L—C	90.04%
H—C	75.00%
total	78.00%

3.3 Syntactic Annotation

The annotated unit of syntax is sentence. The syntactic tree of each sentence is orderly coded and annotated in one dimension to represent the hierarchical structure and property of the syntactic structure as shown in fig.2 and fig. 3. Yu's symbols for POS are adopted [22]. For example, the label for “大哥大 (da4ge1da4)” is “S”/VP/VP/NP/11111” showing that it is the common boundary of five levels as n 11111, NP 1111, VP 111, VP11 and S' 1.

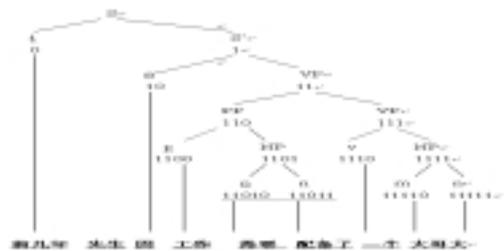


Fig.2 The orderly coded syntactic tree for sentence “前几年，先生因工作需要”.

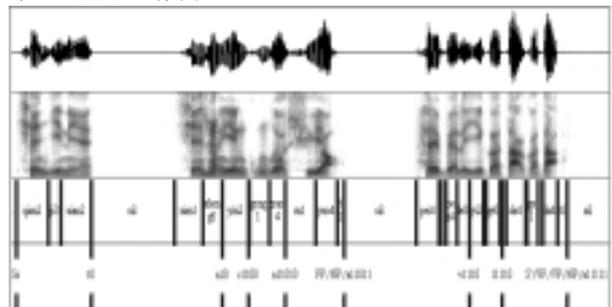


Fig. 3 Syntactic labeling for “前几年，先生因工作需要”.

4. READ AND SPONTANEOUS SPEECH – KNOWN AND NEW RESULTS

4.1. Syntax

In Chinese, Read and Spontaneous speech manifest quite differently in syntax. It is shown from the statistic results in [19,20] 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.

4.2. Spoken Phenomenon

Spoken phenomenon is another indicator to differentiate the read and spontaneous speech. Table 4 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 4. Spoken phenomena occurring in CASS

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

4.3 Sound Variation

In CASS, we labeled the sound variation such as insertion, deletion, pharyngealization, 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 5 and the feature matrixes of sound variation for initial and final are given in table 6 and 7.

Table 5. 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	

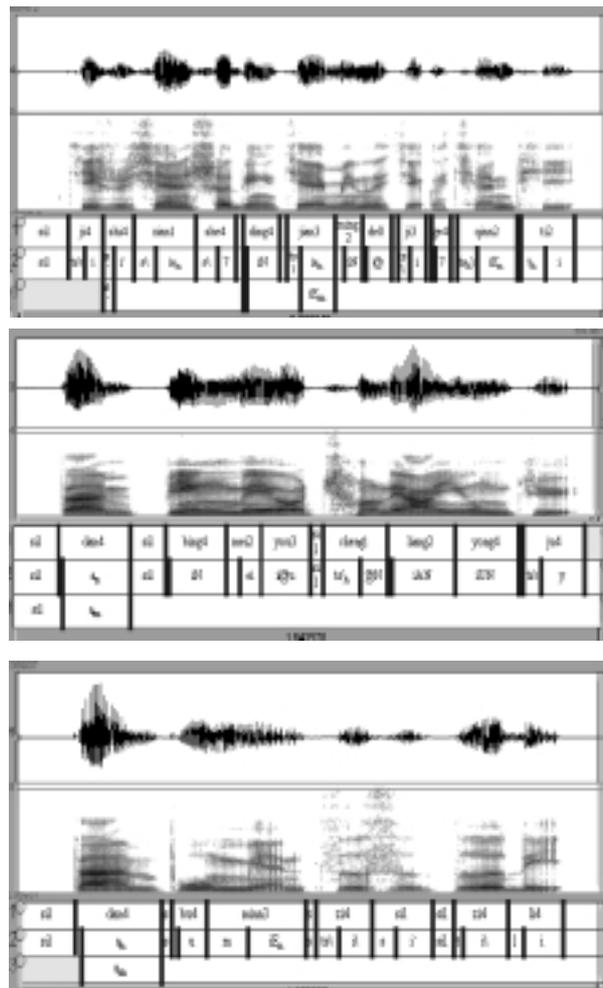


Fig. 4 [n] changes to [m] in different contexts: within a word ‘jian3ming2’ (top), across word boundary ‘dan4 bu4mian3’ and across a phrase boundary ‘dan4 bing4mei2you3’ (bottom).

In both corpora, 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. According to our study, the sound variation occurs often not only within one word or within one prosodic word but also between two words (fig. 4).

It is general that “shi[ʂɿ]” 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.

But the insertion and deletion are occurred more often in CASS. than that in ASCCD.

5. 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.

Table 6. The sound variability features of initials in CASS (5samples)

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)		
[d <t>]	1999	2701	2709	7	4-t_h, 1-l		
[t <t_h>]	1460	1322	117	3	12-t, 6-t_v		
[g <k>]	3430	1231	2165	31	1(n)		
[k <k_h>]	745	687	38	4	13-k_v, 1-k, 1-t_h		

Table 7. The sound variability features of finals in CASS (5 samples)

finals in PinYin	total	unchanged	deletion	voicels	pharyngr ealization	breat hy	central ization	nasal icatio n	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,)
(zh)i	2747	2450	273	6					59(iI)
v	614	597	13	2					2 (i)
an	1326	1314	2	1					@_n-2, a_

REFERENCES

[1] Chen, Xiaoxia Zu, Yiqing & Li Aijun(1999) A cardinal labeling system for Standard Chinese, The fourth phonetics conference in China

[2] J. Wells, “Computer-coding the IPA: a proposed extension of SAMPA”, 2000, <http://www.phon.ucl.ac.uk/home/sampa/>

[3] Li Aijun, Chen Xiaoxia, Sun Guohua, Hua Wu, ect. “The phonetic labeling on read and spontaneous discourse corpora,” to appear in this proceeding.

[4] Li Aijun, “The Acoustic Analysis for Prosodic Phrase and Sentence Prominence of Chinese Dialogue”, The Proceeding of 4th National Conference on Modern Phonetics. Beijing, 1999.

[5] Li Aijun, ZuYiqing, Li Zhiqiang, ”A National Database Design and Prosodic Labeling For Speech Synthesis”, Oriental COCODA’99,Taipei

[6] Lin Maocan, “F0 Construction in Utterances of Standard Chines and its Founction”, The Proceeding of 4th National Conference on Modern Phonetics. Beijing, 1999.

[7] Li, Aijun (1998). Durational Characteristics of the Prosodic Phrase in Standard Chinese. *The Proceedings of the Conference on Phonetics of the Languages in China*, 65-68. HK

[8] Li, Zhiqiang (1997). “A pilot study on prosodic labeling”, *Proceedings of the 3th national Conference on computer intelligent interface and intelligent application.*

[9] Lin Maocan (1998) “The acoustic manifestation of prosodic phrase boundaries in Standard Chinese”, *Prof. of Conference on Phonetics of the Languages in China*, City University of Hong Kong.

[10] Lin, Tao & Wang, Lijia(1992), Textbook of Phonetics, Peking University press

[11] Luo, Changpei &Wang, Jun(1957) An outline of general phonetics, Science press

[12] Wang, Lijia(1992) The principle of phonology.

[13] Xu Bo, huang Taiyi ect. “A Chinese Spoken Dialogue Database and Its Application”, Oriental COCODA’99,Taipei.

[14] Zhu, Weibin & Zhang, Jialu (1997) Manual segmentation & labeling in Chinese speech database, The first China-Japan Workshop on Spoken Language Processing (CJSLP’97)

[15] Zhang, Jialu (1999) A SAMPA system for PUTONGHUA (Standard Chinese), Oriental COCODA’99, Taipei.

[16] Zu Yiqing, Li Aijun, Chen Xiaoxia, etc. “Continuous Speech Database: From isolated Sentence to Discourse”, . Oriental COCODA’99,Taipei.

[17] .Zu Yiqing, Chen Xiaoxia, “Syllable Lengthening and its Function in Spontaneous Speech”, The Proceeding of 4th National Conference on Modern Phonetics, Beijing, 1999.

[18] Zu Yiqing, “Text design for Continuous speech database of Standard Chinese”, Chinese Journal of Acoustics, Vol.18, No. 1, 1999.

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

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

[21] Roger Garside, Geoffrey Leech and Anthony McEnery, “Corpus Annotation”, Addison Wesley Longman Limited.

[22] Yu ShiWen, “现代汉语语法信息词典祥解”, TsingHua University Publishing House.