Broadcast Audio and Video Bimodal Corpus Exploitation and Application

Zou Yu, Hou Min, Chen Yudong, Hu Fengguo, Fu Li
Dept. of Applied Linguistics, Presentation Art School
Communication University of China
Beijing 100024, P. R. China
{zouiy;byhoumin;bychenyudong;bushiwoshishui;red_fuli}@cuc.edu.cn

Abstract

The main purpose of this paper is the exploitation and application of an audio and video bimodal corpus of the Chinese language in broadcasting. It deals with the designation of the size and structure of speech samples according to radio and television program features. Secondly, it discusses annotation method of broadcast speech with achievements made and suggested future improvements. Finally, it presents an attempt to describe the distribution of annotated items in our corpus.

1 Introduction

Since the year of 2002, we've been engaged in setting up the Media Language Corpus aimed to provide the language resources for the researchers who are interested in broadcasting and television media language, for teachers and for researchers of presentation art. Up till now, we have established a 50 million word text corpus involving 40 million word television program text corpora and 10 million word radio program text corpora with 10 million annotated word corpora. The work of this paper is to introduce a ten-hour segmented and prosodic labeled broadcast audio & video bimodal corpus that we built just now.

Section 2 of this paper describes a method for selection of radio and television programs to record according to program features on radio and television stations. Recording conditions are proposed to record a quality spoken language corpus. Section 3 is dedicated to annotation methods. Section 4 shows the distribution of syllables, initials, finals and tones etc. Finally, section 5 contains the conclusion and outlines of our future work in this field.

2 Corpus Information

2.1 Corpus metadata

First of all, we have to select radio and television programs to record. Since a broadcast bimodal corpus should represent the real life usages of spoken language in radio and television, the differences between radio and television, the differences between central and local televisions, and the categories of programs should all together be taken into account during the process of collecting. The followings are the framework (`.wav` files & `.mpeg` files matched with `.txt` files) of head information (metadata) of broadcast audio & video bimodal corpus that has been collected:

No.: ...
Level: central, local, Hong Kong and Taiwan
Station: CCTV, CNR, Phoenix Television...
Style: monologue, dialogue, multi-style
Register: (hypogyny of monologue) presentation, explanation, reading, talk
(hypogyny of dialogue) two person talk show, three person talk show, multi-person talk show
Content: news, literature, service
Audiences: woman, children, elder...
Program: News probe, The first time...
Sub-program: ...
Announcer: ...
Gender: Male /female
Recording condition: Pinnacle PCTV pro card...
Sample rate/Resolution: 22 KHz/16bit...
Topic: ...
Time: xxxx-xx-xx

Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, pages 102–107,
Sydney, July 2006.

2.2 Corpus structure

The purpose of building the broadcast spoken language corpus is to provide the service for the research of broadcast spoken language, esp. for the contrastive studies of the prosodic features of different genres of broadcast language. Hence, the selections of samples of the corpus mainly involve monologues, dialogues or both. As the performing forms of radio and television programs are getting more and more diverse, it is very difficult to decide whether a program is a monologue or dialogue, because these two genres of programs often co-occur in one program. Furthermore, these kinds of programs are increasing their share of radio and television programs. Consequently, this kind of program is most frequent in the corpus. Table 1 displays the structural framework of the broadcast audio and video bimodal corpus.

Table 1 the structure of broadcast bimodal corpus

Style		Example
Dialogue	two person talk show / interview	Face to face...etc.
	three person talk show / interview	Behind the Headlines with Wen Tao...etc.
	multi-person talk show / interview	Utterly challenge...
Monologue	presentation	News...etc.
	explanation	Music story... etc.
	reading	Reading and enjoying... etc.
	talk	Tonight, Weather forecast... etc.
Multi-style		News probe, The first time...etc.

2.3 Recording & management information

All the recorded data are over the programs on radio and TV, that is, it is recorded directly from radio and TV programs by Pinnacle PCTV pro card to connect cable TV with our recording computers. The recorded speech data are saved as 22 kHz and 16bit, Windows PCM waveform, the video data are saved as MPEG or WMV format file by Ulead VideoStudio in a post-processing step. Every program or segment of programs is composed of three parts: `*.wav` data, `*.txt` data, and `*.mpeg/.wmv` data.

Zhao Shixia et al (2000) pointed out that the structure of a speech corpus consists of synchronized objects (text files, wav files, and annotated prosodic files), arranged in deep hierarchies (recording environment), and labeled with speaker-attribute metadata. Therefore, the managed objects of our broadcast bimodal corpus are integrated programs or segments of programs. All data are stored separately but have complex logical inter-relations. These inter-relations can be obtained through the description of the programs. Figure 1 displays the logical structure of the broadcast bimodal corpus.

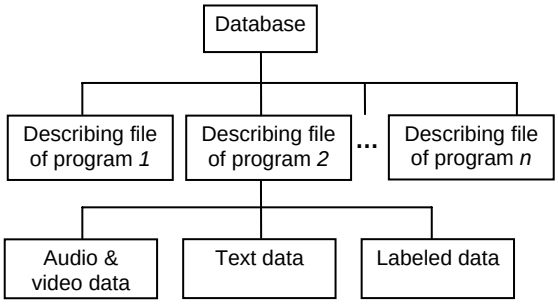

Figure 1 the logic structure of broadcast audio and video bimodal corpus

3 Annotation

Why should we annotate a corpus? An annotation is the fundamental act of associating some content to a region in a signal. The annotation quality and depth have a direct impact on the utility and possible applications of the corpus (Ding Xinshan 1998). The annotation of our corpus consists of transcription, segmental annotation, and prosodic annotation.

3.1 Transcription and segmentation

Transcription is primarily composed of *pinyin* transcription of Chinese characters. Besides, tones are annotated "1", "2", "3", and "4" after the syllable, the neutral tone is labeled "0"; final "ü" annotated as "v", and "üe" annotated as "ue", for example, "旅 (lǚ)" annotated as "lv3", "虐 (nüè)" annotated as "nue4".

In the utterance, compared with broken syllables, successive speech alters greatly, due to the influence of co-articulation, semantics and prosody. The purpose of segmental annotating is to annotate the altered phonemes in the syllables amidst the utterance. For instances, the voicing of some plosives (e.g. b, d, g); labial's influence on alveolar nasal (e.g. "-n" in "renmin" affected by the initial of "min" gradually change into "labionasal", demonstrating the similarities between alveolar nasal and labionasal initial in the frequency spectrum). In the places of unapparent pauses, the stop in the front of plosives esp. af-

fricates often vanishes, which are called the inexistence of silence.

We transcription and segmentation we used BSCA (Broadcasting Speech Corpus Annotator) which was designed by ourselves (Hu Fengguo and Zou Yu 2005). An annotated example is shown in Figure 2:

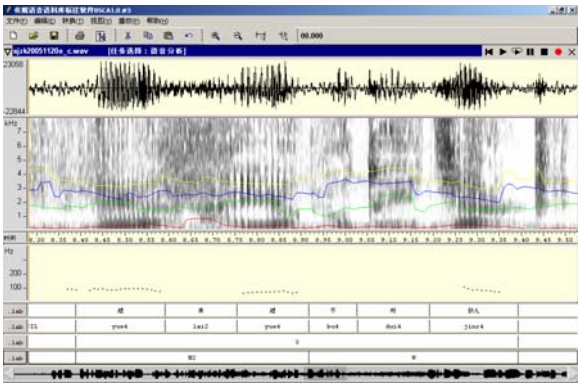

Figure 2 BSCA: a tool for annotation

3.2 Prosodic annotation tiers

Prosodic annotation increases the utility of a speech corpus. An annotated speech corpus can not only offer us a database for the research and exploration of speech information but can also enlarge our knowledge of speech and prosodic features through a visual and scientific method.

Prosodic annotation is a categorical description for the prosodic features with linguistic functions, in other words, annotation of the changes of tone, the patterns of stress, and the prosodic structure with linguistic functions. The prosodic labeling conventions are a set of machine-readable codes for transforming speech prosodies and rule conventions. Based on ToBI (Kim Silverman et al. 1992, John F. Pitrelli et al. 1994) and C-ToBI Conventions (Li Aijun 2002), according to the practical needs of broadcast speech language, the prosodic annotation mainly involves labeling the following parallel tiers: break index, stress index, and intonation construction tier (Chen Yudong 2004, Zou Yu 2004).

3.2.1 Break indices tier

Based on Cao Jianfen's (1999, 2001) categories of prosodic hierarchy structure combined with the practical needs of broadcast speech, we identified five break levels (0-4): 0 indicates the silence or the boundary of default internal syllables amidst the prosodic words. 1 stands for the boundaries of the prosodic words including the short breaks with silent pause and breaks with filled pause. The prosodic words are the fundamental prosodic units in broadcast speech. Simple prosodic words are composed of 1~3 syllables. Complex prosodic words normally contain 5~9 syllables, e.g., "Shang4hai3 he2zuo4 zu3zhi1" (i.e. the Shanghai Cooperation Organization). Break level 2 designates the boundaries of the prosodic phrases, most of which are apparent breaks with silent pause, and their patterns of pitch have also changed. Break level 3 represent the boundaries of intonational phrases, or the boundaries of sentences. Break level 4 stands for the boundaries of intonation groups, similar to the boundary of the entire piece of news in a news broadcast, or of a talker turn in dialogue. At indefinite boundaries, the code "-" is added after the numbers. The labels of the break tier occurring times are shown in table 2:

Table 2 the labels of the break tier occurring in 4 hours annotated corpora

Break index	Occurrence
1	1512
2	2998
3	1986
4	740

3.2.2 Stress indices tier

Stress is a significant prosodic feature. In training materials for broadcast announcers, emphasis is laid on labeling the stress on the basis of the purpose of the utterance, the pattern and rhythm of stresses, and the changes of emotions. Zhang Song's (1983) classification of nuclear stresses can be the guideline for broadcasting production and practice. However, there are some shortcomings in his classifications, for instances, the vague hierarchies between the sentences and discourses. This gets in the way of the formal description of the stresses by the computers. Nevertheless, his theories on the judgment of primary and minor stresses (i.e. non-stresses, minor stresses, primary stresses etc.) have some reference value for stress annotations, because distinguishing the hierarchies of stress is a crucial practical problem for annotation.

As to the problems with the hierarchies of stress, most of the experimental phonetics and speech processing researchers adopt Lin Maocan's (2001, 2002) classifications of stress hierarchies or some similar classifications. That is to say, the levels of stress include prosodic word stress, prosodic phrase stress, and sentence stress (or nuclear stress) in Chinese. According to real life broadcasting productions, this paper identi-

fies four categories of stresses in broadcast speech: the rhythm unit, the cross rhythm unit, the clause, and the discourse. Among them, the discourse stress often occurs at the place of an accented syllable, but they are relatively more important than the other sentence stresses. The labeling methods of all the ranks are listed as follows (Chen Yudong 2004):

Table 3 the stress levels in the stress indices tier

Ranks	Labels
Rhythm unit	1
Cross rhythm unit	2
Clause	3
Discourse	4

Table 4 the stress levels' mean of duration in 4 hours annotated corpora

Stress indices	Mean of duration. (seconds)	Variance
1	.585	.09628
2	.790	.19405
3	.728	.24882
4	.821	.29456

Furthermore, Zhang Song's (1983) other criteria for stress annotation (utterance purpose and emotion change), while perceptually important, are meta-linguistic or para-linguistic in character, and will therefore not be addressed in this paper.

3.2.3 Intonation construction tier

In line with Shen Jiong's view about intonation (Shen Jiong 1994), we found that the intonation construction tier is an important component of the annotation of discourses (Chen Yudong 2004). It can display the changes of sentence intonation structures. The annotation of the intonation construction is mainly to label the relationship of other syllables to the nuclear stress apart from prehead, dissociation etc. For example:

Table 5 the labels of the intonation construction tier occurring in 4 hours annotated corpora

Labels	Description	Occurrence
P	Prehead	794
H	Head	2980
N	Nucleus	2400
T	Tail	1600
W	Weak in stress	2321
D	Dissociation	527
Top	Topic	269
Conj	Conjunction	87

A sentence can have one nuclear stress, or multiple nuclear stresses.

Single nuclear stress: representing the fore-and-aft places of the nuclear stress, the steepness of nuclear stress, and the length of nuclear stress. Examples are listed as follows:

P-H-N-T;
P-H-H-N;
... ...

Among the above examples, long nuclear splitting type "H-N-T-H-N'-T", with the features of multi-nuclear "H-N1-T-H-N2-T" is greatly similar to multi-nuclear. However, "H-N-T-H-N'-T" differs from multi-nuclear in its dependent grammar unit.

Multi-nuclear stress: The two or more nuclear stresses in a multi-nuclear sentence take the patterns of like independent sentence intonation constructions, each with its own nucleus, preceded by a head and optional prehead, and followed by a tail. In other words, these relatively independent patterns already have the features of relatively independent intonation constructions, with the apparent features of "prehead, head, and nuclear ending". This kind of nuclear stress often occurs in relatively longer and more complex constructions. Intonation constructions can be labeled separately. A case in point is the contrastive sentence "zai4 wen3 ding4 de0 ji1 chu3 shang0, qu3 de2 bi3 jiao4 gao1 su4 de0 fa1 zhan3" (i.e. It got a comparative high-speed development on the stable conditions) that can be annotated as "H-N1-T, H-N2-T". For example:

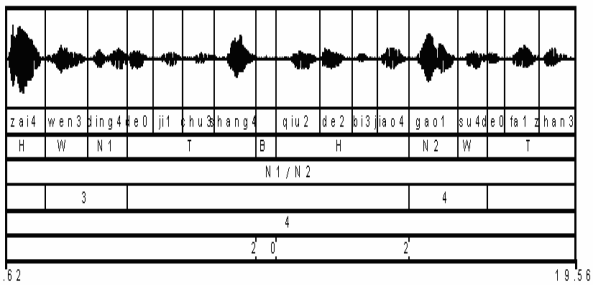

Figure 3 the contrastive sentence "zai4 wen3 ding4 de0 ji1 chu3 shang0, qu3 de2 bi3 jiao4 gao1 su4 de0 fa1 zhan3"(在稳定的基础上，取得比较高速的发展)

3.3 Other items of annotation

Some spoken language corpus can have some additional annotation information. For example, turn talking, paralinguistic and non-linguistic

information (e.g. spot, background music, coughing, sobbing and sneezing) and some hosts' accents (e.g. Shanghai accent) can be annotated in talk show corpus. There are 82 times of spot and 31 times of background music in 4 hours annotated data. Furthermore, some .wav files, .mpeg files can be annotated together for discourse analysis.

4 Distribution of annotated items

We conducted a statistic analysis of some annotated items using 4 hours of annotated data in our corpus.

The syllables (initials and finals) of the 20 top frequent occurring are given in Table 6. In addition to this, the duration and variance distribution for them are calculation shown as follows.

Table 6 the mean of duration and variance of the top 20 frequent occurring syllables

Syllable	Occurrence	Mean of duration. (seconds)	Variance
de0	1993	.1167	.00232
shi4	912	.2051	.00572
shi2	626	.2054	.00625
zai4	602	.1889	.00341
le0	540	.1325	.00334
ta1	442	.1765	.00461
bu4	423	.1492	.00267
guo2	404	.1673	.00328
yi4	398	.1656	.00350
zhong1	395	.1996	.00390
ren2	394	.1959	.00625
zhe4	386	.1499	.00317
you3	380	.1841	.00480
yi1	357	.1475	.00295
dao4	335	.1778	.00367
he2	309	.2078	.00687
wo3	287	.1704	.00755
men0	287	.1568	.00426
yi2	274	.1555	.00320
jiu4	250	.1724	.00332

Table 7 Distribution of initials (4 hours data)

Initials	Times	Initials	Times
b	1076	j	3136
p	443	q	1464
m	1636	x	2146
f	972	zh	2953
d	4635	ch	1112
t	1561	sh	3406
n	1085	r	895
l	2569	z	1705
g	2162	c	512
k	879	s	700
h	2071	?	6099

Table 8 Distribution of finals (4 hours data)

Finals	Times	Finals	Times	Finals	Times
a	1653	ian	1767	ua	229
ai	1909	iang	919	uai	136
an	1425	iao	773	uan	632
ang	1192	ie	838	uang	389
ao	1205	in	1175	uei	1317
e	5074	ing	1480	uen	368
ei	807	iong	128	ueng	3
en	1515	iou	1144	uo	1760
eng	1237	o	176	v	932
er	353	ong	1658	van	432
i	6856	ou	831	ve	474
ia	586	u	2533	vn	209

Table 9 Distribution of tones (4 hours data)

Tones	1	2	3	4	0
Occurrence	8948	9194	7401	14683	6134

The occurrence distribution of initial, final, and tone are calculated. These are shown in table 7, 8 and 9 respectively.

We also measured the mean duration and F0 of each tone in three speaking styles are listed in Table 10 and 11.

Table 10 Mean duration of tones in various speaking styles (seconds)

	T1	T2	T3	T4	T0
Presentation	.189	.199	.192	.180	.129
Reading	.338	.337	.324	.335	.277
Talk	.167	.173	.163	.163	.154

Table 11 F0 of tones in various speaking styles (Hz)

	Presentation	Reading	Talk
T1	162.78	158.86	207.37
min. of T2	126.39	134.46	168.73
max. of T2	147.27	155.34	180.94
range of T2	79.12	20.88	12.21
min. of T3	101.94	119.12	151.21
max. of T4	163.96	170.07	209.86
min. of T4	113.39	120.98	175.49
range of T4	50.57	49.09	34.37

To summarize, we conclude that the mean duration of tones of reading style is longer than that of presentation style; that of talk style is the shortest among three styles. As for the F0 of each

tone, the F0 and pitch range of presentation style is high and has big fluctuation; that of talk style is high and has small fluctuation. However, the F0 of tone 3 of presentation style is lower than that of reading and talk styles.

5 Further study

The broadcast audio and video bimodal corpus[1] is a presentation art-oriented corpus with radio and television news as its basis. This paper probes the development and compilation of broadcast audio and video bimodal corpus.

Firstly, on the collection of the corpus, what sort of audio and video corpus can represent the features of radio and television speech language? How can we auto-annotate the audio and video corpus? ...These are the problems that have always been bothering us.

Secondly, this corpus can be a platform for further research into non-accented or accented syllables, intonation construction, the prosodic functions of paragraphs and discourses, the emotions of speech, and genre styles.

Finally, we can statistically analyze the spectral and prosodic characteristics of various speaking styles by the corpora, such as presentation, reading and talk. All speaking styles would be synthesized based on the analysis results. This is also work for the future.

6 Acknowledgements

We would like to thank Prof. Wolfgang Teubert for his guidance and comments on this paper. I would also like to thank Mr. Daniel Zhang, Jan Van der Ven for their kind help.

References

Cao Jianfen. 1999. *Acoustic-phonetic Characteristics of the Rhythm of Standard Chinese*, In Proceedings of 4th National Conference on Modern Phonetics, Beijing, pp.155~159.

Cao Jianfen. 2001. *Phonetic and Linguistic Cues in Chinese Prosodic Segmentation and Grouping*, In Proceedings of 5th National Conference on Modern Phonetics, Beijing, pp.176~179.

Chen Yudong. 2004. *The Utterance Construction and Adjustment in Media Spoken Language*, PhD thesis, Peking University.

Ding Xinshan.1998. *Development and Research of Corpus Linguistics*, Contemporary Linguistics, 1: 4~12.

Hu Fengguo, Zou Yu. 2005. *The Design and Exploitation of Broadcasting Speech Corpus System*, In Proceedings of the Eighth Joint Seminar of Computational Linguistics (JSCL-2005), Nanjing, China, pp.521~527.

John F. Pitrelli, Mary E. Beckman, and Julia Hirschberg. 1994. *Evaluation of Prosodic Transcription Labeling reliability in the ToBI Framework*, In Proceedings of the 1994 International Conference on Spoken Language Processing(ICSLP), Yokohama, Japan, pp.123-126.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin Wightman, Patti Price, Janet Pierrehumbert, and Julia Hirschberg. 1992. *ToBI: A Standard for Labeling English Prosody*, In Proceedings of the 1992 International Conference on Spoken Language Processing(ICSLP), Banff, Alberta, Canada, vol.2, pp.867-870.

Li Aijun. 2002. *Chinese Prosody and Prosodic Labeling of Spontaneous Speech*, In Speech Prosody 2002 An International Conference, Aix-en-Provence, France.

Lin Maocan. 2001. *Prosodic Structure and F0 Declination in Sentence of Standard Chinese*, In Proceedings of 5th National Conference on Modern Phonetics, Beijing, pp.180~184.

Lin Maocan. 2002. *Prosodic Structure and Construction of F0 Top-Line and Bottom-Line in Utterances of Standard Chinese*, Contemporary Linguistics, 4: 254~265.

Shen Jiong. 1994. *Chinese Intonation structure and category*, Dialect, 4: 221~228.

Zhang Song. 1983. *Recitation*, Changsha: Hunan Education Press.

Zhao Shixia, Cai Lianhong, Chang Xiaolei. 2000. *Construction of Mandarin Corpus for Chinese Speech Synthesis*, Mini-Micro System, Vol.21 (3): 295~297.

Zou Yu. 2004. *Primary Research on Prosodic Labeling in Chinese News Broadcasting Speech*, In Proceedings of the 2nd Student Workshop on Computational Linguistics (SWCL2004), Beijing, pp.1-7.

[1] This research was supported by the National Working Committee on Language and Characters, project no. YB105-61A and Communication University of China, project no. BBU211-15.