AN APPLICATION OF SAMPA-C IN STANDARD CHINESE

Chen xiaoxia, Li Aijun, Sun Guohua, Hua Wu and Yin Zhigang*

ABSTRACT

Labeling segment is an important work in database building. This paper presents a labeling system for Standard Chinese named SAMPA -C. We give some charts: consonant chart, vowel chart, tone chart, retroflex final chart, sound variation chart and non-speech symbol chart. Then this labeling system is used in two corpora labeling. The result shows that the labeling system is suitable for Standard Chinese.

I. INTRODUCTION

SAMPA (Speech Assessment Methods Phonetic Alphabet) is a machine-readable phonetic alphabet. It was originally developed under the ESPRIT project 1541, SAM (Speech Assessment Methods) in 1987-89 by an international group of phoneticians, and was applied in the first instance to the European Communities languages such as Danish, Dutch, English, French, German, and Italian (by 1989); later to Norwegian and Swedish (by 1992); and subsequently to Greek, Portuguese, and Spanish (1993). Under the BABEL project, it has now been extended to Bulgarian, Estonian, **Hungarian, Polish, and Romanian (1996). Under the aegis of COCOSDA it is hoped to extend it to cover many other languages (and in principle all languages).**

These codes covers everything on the 1993 IPA Chart, including diacritics and tone marks, and is put forward as a proposed standard way to transmit IPA-transcribed material by e-mail and for similar purposes. It is an extension of the SAMPA standard, with which colleagues may be familiar. The most frequently used symbols are mapped onto single keystrokes in the ASCII range 33.126. Less frequently used symbols are mapped onto a single keystroke plus the backslash, \. Diacritics (other than those already catered for in SAMPA) are mapped onto a keystroke with a preceding underscore, _.

With SAMPA and X-SAMPA, we consider to get the same system in Standard Chinese, because it is significant work to label segment in speech corpora for Standard Chinese. With the labeled material we can do many research work. Hanyupinyin is an effective way to transcribe Standard Chinese. But it is not entirely corresponding to IPA. For example, "i" representing [i], [!Ÿ], [!¡]. It is not easy to be a machine-readable symbol system. According to international machine-readable symbol system SAMPA [1], Zhu Weibin and Zhang Jialu have transcribed a symbol system with SAMPA for labeling syllable. [2,3]. They give Chinese S**AMPA symbols** including consonant, vowel and tone charts according to Xu Shirong' s view. They label isolated syllable in a database. This is an important work for transcribing Standard Chinese. But it is not enough for Standard Chinese in continuous speech. We hope to label phonetic segment in continuous speech. The representation of continuous speech is more complex than isolated syllable. There are sound variation phenomena in continuous speech such as centralization, reduction, insertion etc. The detailed labeling must include them. We must formulate symbols to label them. Based on these, firstly, we give some rules for design the system and then we design SAMPA-C labeling system for Standard Chinese. We have made a labeling system in syllable tier last year [4]. Now we make it in a continuous speech tier. What we refer to is Luo Changpei' s view [5] for consonant and vowel. For retroflex final, we refer to Wang Lijia' s result [6]. Then, we give diacritics for sound variation and give non-speech symbols.

We have two-speech corpora, which are read speech corpora and spontaneous corpora in CASS. The first one includes 18 articles and 10 speakers. The materials are read in recording room with normal rate. The second is originated from 19 cassettes provided by the Broadcast Station of Tsinghua University (BSTHU), Beijing, China. 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. Those cassettes are then digitized into mono waveform at 16-bit precision and 16-kHz sampling rate through a standard Sound Blaster card onthe PC, resulting in the 1.5 GB raw speech database totally [7]. With Pinyin and SAMPA -C, we label the two corpora.

II. LABELING SYSTEM

2.1 The principles of labeling system are as follows:

(1) Compensative: It covers each phonetic segment entirely.

(2) Systematic: For one phenomenon, we use a consistence manner to transcript it. For example, there are not voiced stop and voiced affricate consonants in isolated syllable. But, for continuous speech, there are many voiced stops and voiced affricates. We just give voiced symbol "_v" to represent those consonants becoming voiced in SAMPA -C not give voiced stop or voiced affricate.

(3) Available: For most segments, we consider the real case in Standard Chinese. For example, we don' t use "_0" but "_u" to represent voiceless because "_0" is used in neutral tone.

2.2 Labeling system SAMPA -C

We give SAMPA -C as follows: consonant chart, vowel chart, retroflex final chart, sound variation chart and also non-speech chart.

~~PinYin~~	IPA	SAMPA-C	PinYin	IPA	SAMPA-C
b	p	p	z	t s	ts
p	p Ḧ	p_h	c	t s Ḧ	ts_h
m	m	m	s	s	s
f	f	f	zh	t ©	ts`
d	t	t	ch	t©Ḧ	ts_h`
t	tḦ	t_h	sh	©	s`
n	n	n	r	¸	z`
(a)n	n	_n			
l	l	l	j	t »	ts\
g	k	k	q	t»Ḧ	ts_h\
k	k Ḧ	k_h	x	»	s\
h	x	x		?	?
ng	Ð	N			

Table 1: Consonant Chart for Standard Chinese

PinYin	IPA	SAMPA-C
a	§	A
o	o	o
e	°	7
i	i	I
u	u	u
ü	y	y
(zh)i	Ÿ	i`
(z)i	¡	i\
er	Ä	@`

Table 2: Vowel Chart For Standard Chinese

TONE	IPA	SAMPA-C	EXAMPLE
Tone 0	0	ba_0	ba0 吧
Tone 1	1	ba_1	ba1 巴
Tone 2	2	ba_2	ba2 拔
Tone 3	3	ba_3	ba3 把
Tone 4	4	ba_4	ba4 罢

Table 3: Tone Chart For Standard Chinese

NAME	PINYIN	IPA	SAMPA-C	EXAMPLE
opened	ar	a r	a`	par
	or	or	o`	mor
	er	°r	7`	ger
	(zh)i	Èr	i@`	zhir,shir
	(z)i	Èr	i@`	zir
	air	a r	a`	bair
	eir	Èr	@`	leir
	aor	§o r	Ao`	daor
	our	ou r	ou`	gour

	anr	a r	a`	ganr
	enr	Èr	@`	genr
	angr	a ‹r	a~`	gangr
	engr	È‹r	@~`	dengr
stretched	ir	iÈr	i@`	jir
	iar	i a r	ia`	iar
	ier	i E r	ie_r`	jier
	iaor	i§o r	iAo`	jiaor
	iour	i o u r	iou`	qiur
	ianr	i a r	ia`	jianr
	inr	iÈr	i@`	jinr
	iangr	ia ‹r	ia~`	liangr
	ingr	iÈ‹r	i@~`	ingr
	iongr	iu‹r	iu~`	xiongr
rounded	ur	ur	u`	gur
	uar	uar	ua`	guar
	uair	ua r	ua`	guair
	ueir	uÈr	u@`	gueir
	uanr	ua r	ua`	tuanr
	uenr	uÈr	u@`	lunr
	uor	uo r	uo`	luor
	uangr	ua ‹r	ua~`	kuangr
	uengr	uÈ‹r	u@`	uengr
	ongr	u‹r	u~`	kongr
protruded	ür	yÈr	y@`	yur
	üer	y E r	yE_r`	yuer
	üanr	y a r	ya`	yuanr
	ünr	yÈr	y@`	qunr

Table 4: Retroflex Final For Standard Chinese

NAME	IPA	SAMPA-C	EXAMPLE
nasalized	a ‹	~	e~
centralized	e ƒ	_’’	e_”
voiceless	n	_u	n_u
voiced	dŒ	_v	t_v
rounded	t ¦	_O	O_O
syllabic	\	=	M=
pharyngealized	Øt /	_?\	A_?\
silence		sil	sil
silence voiced		silv	silv

Table 5: Diacritics Chart For Standard Chinese

PHENOMENA	SAMPA-C
repairs	repair <...repair>
disfluencies	disfl <...disfl>
silences	silen <...silen>

laughing	**laugh<...laugh>**
coughing	**cough<...cough>**
breathing	**breath<...breath>**
crying	**cry<...cry>**
noise	**noise<...noise>**
lengthening	**leng<...leng>**
modal	**mod<...mod>**
murmur	**mum<... mum>**
smack	**smack<... smack>**

Table 6: Non-speech Chart for Standard Chinese

III. LABELING RESULT

Consistence is high: Using the labeling system, we segment and label the two corpora with Pinyin and SAMPA -C. Three tiers are given. The first tier is pinyin, the second is semi-syllable and the third is sound variation or other speaking phenomena. With manual work, we give a consistence test for labeler for nearly 15 minutes. It is about from 82.39% to 88.25%. The consistency is high. It shows that the labeling system is feasible. Most symbols are used during the labeling. The other result will be showed in another paper [7].

IV. DISCUSSION

We change some symbols in Standard Chinese. Next, we explain them as follows:

(1) Retroflex final is an important phonetic representation. We give final plus r as retroflex final.

(2) There is not voiced consonant in isolated syllable in Standard Chinese. But it is common that stop, affricate or fricative can be voiced. So, we give voiced symbol to represent the phenomenon in continuous speech, but not give voiced stop, voiced affricate or voiced fricative.

(3) The silence before stop and affricate often becomes voiced. It can be a long time. We just give a symbol “silv” to represent that duration.

(4) The neutral tone is a special tone in Standard Chines. We use “_0” as the symbol to consist with the other tones. So, for voiceless, we use “_u”. It is not consistence with SAMPA.

(5) For apical nasal “n”, we give two varieties according to their place in a syllable. As initial, it is showed with “n”. But as final, it is showed with “_n”.

REFERENCES

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

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

[3] Zhang, Jialu (1999), A SAMPA system for PUTONGHUA (Standard Chinese) Oriental COCOSDA’ 99 PROCEEDINGS

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

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

[6] Wang, Lijia (1992), The principle of phonology, YUWEN Press

[7] Li, Aijun etc. (2000), A phonetic labeling on read and spontaneous discourse corpora, ICSLP2000

[8] Philippe Blache & Daniel Hirst(2000), Multi-level annotation for spoken language corpora, ICSLP 2000