A CHINESE SPEECH DATABASE (PUTONGHUA CORPUS)

Wenxian Li, Yiqing Zu, Chorkin Chan

Speech Lab.
Dept. of Computer Science
Univ. of Hong Kong

ABSTRACT - In this paper, a Chinese speech database (Putonghua corpus) is introduced which has been constructed at HKU. It consists of isolated syllables, words, digit strings and sentences read by a total of 20 native speakers, ten females and ten males. No systematic effort in constructing a comprehensive Chinese speech database has been reported before, so this corpus has great importance to serve as a standard. It supplies a large scale, professionally built common test-bed for Putonghua recognizers.

INTRODUCTION

A Putonghua corpus named HKU93 was constructed at the Speech Laboratory of the Department of Computer Science, the University of Hong Kong in 1993-94. It consists of isolated syllables, words, digit strings and connected speech. A total of 20 native Putonghua speakers were employed to read prompting messages displayed on a monitor screen. All recordings were done in a quiet office with a microphone (National Cardioid Dynamic Microphone WM-333N IMP.600Ω) and a Sound Blaster 16 ASP A/D-D/A card (1) plugged into a PC486 sampling at a rate of 16129Hz.

Corresponding to each utterance, be it an isolated syllable or continuous speech, information stored includes the orthographic transcription (the Chinese characters in Big5 code) with the toned Pinyin symbols on the one hand and the digitized waveform on the other hand. These two kinds of information are stored in files of two categories, the 'txt' files which are text files for the former and the 'wfm' files which are binary files for the latter. Each utterance corresponds to a file in both categories, under the same file name but different file name extensions, viz., 'txt' or 'wfm'.

Utterance end points are determined automatically. Each utterance preceded and followed by a short period of silence was saved in a separate file.

The twenty speakers, ten females and ten males, employed to read the designed text material, are native Putonghua speakers of 20 to 39 years old. The only exception is a Hong Kong resident girl (speaker id 3f) who speaks with a mild accent.

This corpus of almost 3 GB of waveforms and their corresponding orthographic transcriptions and Pinyin symbols with tones is stored in three DDS-DAT tapes (60m, 1.4GB each), labeled as Tape 1 to Tape 3 respectively.

CONTENTS OF THE CORPUS

The text read by each speaker consists of six subsets. The first four subsets are isolated syllables (characters), words, digit strings, and strings of rhymed syllables which are common to everyone. The fifth subset is continuous speech, and each speaker read hundreds of unique text lines. The last subset is a set of retroflexed ending words, which is read by only two speakers. Despite the abundance of continuous utterances read by each speaker, certain phone to phone transitions between syllables are still missing. 5 of the speakers read extra words (included in the subset of continuous speech) to make sure all transitions of the last

phone of a syllable to the first phone (initial) of a syllable are covered.

- Isolated Syllables

  All Putonghua syllables in all tones which actually correspond to Chinese characters are read at least once. Since certain phones appear more abundantly than others, some syllables are repeated several times to make the collection more phonetically balanced.

- Words

  11 words of 2 to 4 syllables were selected in such a way that their pronunciations include all the Putonghua phones. Such a speech subset may be useful for speaker adaptation.

- Digit Strings

  16 digit strings of 4 to 7 digits each were designed to exhaust all the inter-digit triphones.

- Rhymed Syllables

  3 sentences of 7 characters (syllables) each were so designed that all syllables in the first sentence rhyme with /a/, those in the second rhyme with /i/ and those in the third rhyme with /u/. These sentences play a role similar to that of the E-set of the English alphabet which is an acid test to any speech recognizer.

- Continuous Speech

  Each speaker read hundreds of lines of text with unique contents.

- Retroflexed Ending Words

  A set of words with and without retroflexed endings have been read by 2 of the speakers. It dose not exhaust all the Putonghua syllables with retroflexed endings but consists of only those words with the last finals in the 4 groups listed below because they are likely to have retroflexed endings in normal Putonghua.

  1. a_set

     /a/, /ua/, /ai/, /uai/, /an/, /uan/, /ian/

  2. er_set

     /(c)i/, /(ch)i/, /en/, /uen/, /eng/, /ueng/, /i/, /in/, /ing/, /yv/, /yun/

  3. u_set

     /e/, /u/, /ou/, /iou/, /ao/, /iau/

  4. ie_set

     /ie/

Here /(c)i/ is the front apical vowel and /(ch)i/ the back apical vowel. There is a total of 80 pairs of such words read in this subset. Table 1 lists the effects of the retroflexed ending on different syllables. Note that retroflexing a syllable ending with 'ng' causes the disappearance of 'ng' and nasalization of the vowel just before 'ng'. In this table SRE represents syllable without retroflexed ending, and PYWRE represents its pinyin with retroflexed ending.

CONVENTION OF FILE NAMES

On the tapes, all waveform files of the same speaker are stored in a separate directory, named as wvfrm[n][s], where n=0, ..., 9 and s=m/f. For example, the first female speaker's waveform directory is wvfrm0f, and the that of second male's is wvfrm1m. Likewise, all text files of the

same speaker are stored in another directory , named as text[n][s], where n=0, ..., 9 and s=m/f. For example, the first female speaker's text directory is text0f, and the second male's is text1m.

All files are named in a way to reflect the kind of utterances they are associated with. The files are named with 5 fields as follows:

⟨group_type⟩⟨speaker_id⟩⟨sex⟩⟨utterance_id⟩.⟨file_type⟩

The field widths in character are 2, 1, 1, 4, and 3 respectively.

- group_type :== is | cs | re | ds | wd | rs

  is stands for isolated syllables
  cs stands for continuous speech
  re stands for retroflexed ending words
  ds stands for digit strings
  wd stands for words
  rs stands for rhymed syllables

- speaker_id :== 0, 1, ..., 9
- sex :== m | f

  m stands for male
  f stands for female

- utterance_id :== 0001, 0002, ...
- file_type :== wfm | txt

  wfm stands for a waveform file
  txt stands for a text file

Only speakers 0f and 0m have retroflexed ending words files .

FILE FORMATS

There are two file formats, one for waveform files and the other for text files.

Format of waveform files (with extension 'wfm')

All waveform files are binary. Data are stored in a DOS format, i.e., a short integer is stored as

'low byte, high byte'

Each file consists of 2 blocks, the header block & the data block, where the header block has fields as follows

[header size][version number][precision][sampling Freq.][number of samples]

with field widths in bytes as 2, 2, 2, 2, & 4 respectively.

The data block is a sequence of integer (2 bytes) samples whose length is stored in the last four bytes of the head block (number of samples).

SRE	PYWRE	SRE	PYWRE	SRE	PYWRE	SRE	PYWRE
ban	bar	cha	char	cheng	cher	chi	chir
chuo	chuor	ci	cir	dao	daor	dian	diar
die	dier	di	dir	ding	dir	dou	dour
dui	duer	fa	far	fen	fer	gai	gar
gan	gar	ge	ger	hai	har	he	her
hua	huar	hui	huer	huo	huor	jiao	jiaor
jie	jier	jin	jir	jing	jir	kou	kour
kuai	kuar	lian	liar	men	mer	mian	miar
nao	naor	na	nar	niu	niur	pai	par
pan	par	pen	per	pian	piar	pie	pier
pi	pir	ping	pir	qi	qir	quan	quar
ren	rer	sheng	sher	shi	shir	si	sir
sui	suer	tan	tar	te	ter	tou	tour
tui	tuer	wa	war	wan	war	wei	wer
wen	wer	xian	xiar	xing	xir	xuan	xuar
ya	yar	ya	yar	ye	yer	yi	yir
yin	yir	ying	yir	you	your	yuan	yuar
yu	yur	zhe	zher	zhen	zher	zhun	zhuer
zi	zir	zuo	zuor				

Table 1. Effect of retroflexed ending on a syllable

Format of text files (with extension 'txt')

Each file consists of the orthographic transcription, the Pinyin symbols with tones of a recorded Putonghua utterance stored in two lines separated by an end-of-line symbol; one for the orthographic transcription and the other for the Pinyin and tones of the syllables.

Certain Chinese character do not have a Big5 code, in which case, blanks are used as its code instead.

SOME CORPUS STATISTICS

The following is a list of frequency counts of each syllable in this Corpus (2).

count	pinyin	count	pinyin	count	pinyin	count	pinyin	count	pinyin	count	pinyin
361	a	416	ai	584	an	95	ang	185	ao	846	ba
543	bai	570	ban	159	bang	670	bao	2	bar	782	bei
347	ben	136	beng	578	bi	409	bian	455	biao	221	bie
103	bin	494	bing	317	bo	2171	bu	46	ca	615	cai
240	can	95	cang	180	cao	147	ce	126	cen	266	ceng
386	cha	107	chai	224	chan	882	chang	209	chao	2	char
394	che	305	chen	930	cheng	2	cher	729	chi	2	chir
240	chong	190	chou	1227	chu	81	chuai	345	chuan	206	chuang
62	chui	178	chun	61	chuo	2	chuor	1195	ci	2	cir
304	cong	121	cou	117	cu	64	cuan	85	cui	166	cun
151	cuo	1293	da	536	dai	804	dan	409	dang	1111	dao
4	daor	5726	de	131	dei	120	den	402	deng	942	di
20	dia	784	dian	206	diao	14	diar	78	die	2	dier
384	ding	4	dir	35	diu	617	dong	10756	dou	2	dour
459	du	253	duan	2	duer	652	dui	1353	dun	639	duo

780	e	80	ei	271	en	23	eng	2488	er	902	fa
430	fan	950	fang	2	far	531	fei	902	fen	445	feng
3	fer	127	fo	180	fou	1117	fu	85	ga	327	gai
423	gan	162	gang	547	gao	15	gar	1174	ge	189	gei
215	gen	255	geng	2	ger	1053	gong	231	gou	449	gu
115	gua	88	guai	759	guan	301	guang	229	gui	50	gun
1289	guo	89	ha	584	hai	248	han	151	hang	17691	hao
5	har	909	he	133	hei	284	hen	152	heng	2	her
34	hm	264	hong	766	hou	442	hu	934	hua	136	huai
405	huan	262	huang	3	huar	12	huer	1033	hui	165	hun
774	huo	2	huor	2550	ji	1158	jia	1090	jian	602	jiang
800	jiao	2	jiaor	1231	jie	3	jier	1112	jin	1702	jing
43	jiong	4	jir	1057	jiu	3742	ju	87	juan	422	jue
211	jun	65	ka	410	kai	326	kan	128	kang	176	kao
1062	ke	120	kei	138	ken	123	keng	287	kong	294	kou
4	kour	195	ku	78	kua	147	kuai	124	kuan	189	kuang
4	kuar	118	kui	110	kun	124	kuo	167	la	815	lai
205	lan	165	lang	390	lao	1137	le	242	lei	143	leng
1576	li	25	lia	447	lian	519	liang	321	liao	3	liar
219	lie	366	lin	727	ling	609	liu	131	lo	193	long
193	lou	481	lu	79	luan	71	lue	166	lun	225	luo
341	luu	20	lve	40	m	394	ma	271	mai	258	man
110	mang	454	mao	205	me	779	mei	521	men	183	meng
3	mer	263	mi	453	mian	143	miao	2	miar	73	mie
273	min	751	ming	20	miu	282	mo	188	mou	585	mu
413	na	110	nai	427	nan	95	nang	164	nao	3	naor
3	nar	144	ne	329	nei	120	nen	451	neng	120	ng
405	ni	923	nian	80	niang	58	niao	70	nie	30	nin
100	ning	141	niu	2	niur	151	nong	121	nou	87	nu
47	nuan	37	nue	59	nuo	410	nuu	89	o	322	ou
151	pa	343	pai	167	pan	147	pang	155	pao	6	par
218	pei	141	pen	195	peng	2	per	242	pi	313	pian
171	piao	3	piar	65	pie	2	pier	401	pin	291	ping
4	pir	411	po	132	pou	202	pu	2	puo	1766	qi
98	qia	864	qian	333	qiang	172	qiao	313	qie	287	qin
820	qing	53	qiong	2	qir	301	qiu	855	qu	489	quan
3	quar	372	que	101	qun	504	ran	201	rang	85	rao
165	re	1819	ren	178	reng	3	rer	886	ri	244	rong
143	rou	616	ru	20	rua	53	ruan	99	rui	42	run
136	ruo	85	sa	173	sai	808	san	82	sang	82	sao
253	se	134	sen	121	seng	215	sha	70	shai	257	shan
1100	shang	370	shao	468	she	124	shei	703	shen	1107	sheng
2	sher	6299	shi	3	shir	939	shou	856	shu	71	shua
103	shuai	42	shuan	115	shuang	294	shui	89	shun	452	shuo
1579	si	2	sir	255	song	136	sou	349	su	125	suan
2	suer	437	sui	76	sun	522	suo	1466	ta	756	tai
432	tan	169	tang	232	tao	2	tar	198	te	120	tei
128	teng	7	ter	640	ti	581	tian	274	tiao	137	tie
354	ting	743	tong	374	tou	4	tour	442	tu	156	tuan
2	tuer	199	tui	104	tun	174	tuo	203	wa	415	wai
686	wan	534	wang	12	war	1865	wei	829	wen	141	weng
9	wer	975	wo	1529	wu	1063	xi	610	xia	1221	xian
1377	xiang	940	xiao	3	xiar	576	xie	930	xin	1281	xing
202	xiong	2	xir	184	xiu	640	xu	277	xuan	2	xuar
562	xue	278	xun	328	ya	865	yan	503	yang	943	yao
4	yar	1175	ye	3	yer	4838	yi	13	yin	2750	yin
764	ying	6	yir	124	yo	602	yong	3222	you	2	your
1796	yu	1209	yuan	2	yuar	647	yue	319	yun	2	yur

120	za	1845	zai	120	zan	111	zang	387	zao	362	ze
125	zei	128	zen	169	zeng	121	zha	85	zhai	405	zhan
495	zhang	393	zhao	1506	zhe	143	zhei	467	zhen	786	zheng
7	zher	2875	zhi	1634	zhong	266	zhou	878	zhu	52	zhua
60	zhuai	323	zhuan	303	zhuang	2	zhuer	109	zhui	125	zhun
116	zhuo	1760	zi	6	zir	246	zong	241	zou	323	zu
70	zuan	414	zui	96	zun	1862	zuo	2	zuor		

SUMMARY

A large corpus of Putonghua has been built at HKU. This corpus is primarily designed for research in speech recognition and perhaps natural language understanding also. The corpus encompasses all the mono-syllables of Putonghua as well as all the phones and phone-phone transitions. Beside, It includes a large amount of read continuous utterances. This corpus is being labeled phonetically by means of a speech recognizer. With such a corpus, ambitious projects in speech research and system development can be launched.

NOTES

(1) Due to inadequate insulation, this card tends to pick up a weak background noise at about 5KHz. The user of this corpus is advised to consider removing such background noise by DSP techniques.

(2) Pinyin ending with an 'r' are syllables with a retroflexed ending.

REFERENCES

A.Kurematsu, K, Takeda, H. Kuwabara and K. Sgikano, (1989) *ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis* Proc, of ESCA Workshop, pp. 2.3.1-2.3.4.

C. Chan & K.K. Lee, (1992) *Construction of a Mandarin Corpus* Proc. of Int. Computer Symposium, pp. 933-937, TaiChung.

K. Shirsi, H. Fujisaki and S. Itahashi, (1989) *Speech Database Projects in Japan — Present and Future* Proc of ECSA Workshop, pp. 2.4.1-2.4.4.

L.F. Lamel, R.H. Kassel and S. Seneff (1986) *Speech Database Development: Design and Anslysis of the Acoustic-Phonetic Corpus* Proc. Speech Recognition Workshop, DARPA, pp. 100-110.

Mei-yuh Hwang, Hsiao-wuen Hon and K.F. Lee, (1989) *Modeling Between-word Coarticulation in Continuous Speech Recognition* Proc. Eurospeech 89, pp. 5-8, Paris

X.Q. Chen, C.L. Li, F.Y. Mo and S.N. Lu (1987) *A Chinese Language Speech Database,* Proceedings of Conference on Speech, Communication and Image Processing, pp.127-129.