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

2. er_set

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

3. u_set

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 Of and Om 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, & 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										
361	а	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	cno	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

			_			00		2400		902	fa
780	e	80	ei	271	en	23	eng	2488	er	902 445	
430	fan	950	fang	2	far	531	fei	902	fen		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
		2	jiaor	1231	jie	3	jier	1112	jin	1702	jing
800	jiao	4	jir	1057	jiu	3742	ju	87	juan	422	jue
43	jiong		•	410	kai	326	kan	128	kang	176	kao
211	jun	65	ka			123		287	kong	294	kou
1062	ke	120	kei	138	ken		keng			189	
4	kour	195	ku	78	kua	147	kuai	124	kuan	815	kuang lai
4	kuar	118	kui	110	kun	124	kuo	167	la		
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
		144	ne	329	nei	120	nen	451	neng	120	ng
3	nar			80		58	niao	70	nie	30	nin
405	ni	923	nian		niang			121	nou	87	nu
100	ning	141	niu	2	niur	151	nong			322	
47	nuan	37	nue	59	nuo	410	nuu	89	0		ou
151											
	pa	343	pai	167	pan	147	pang	155	pao	6	par
218	pa pei	141	pen	195	peng	2	per	242	pi	313	pian .
	•	141 3	•	195 65	•	2	per pier	242 401	pi pin	313 291	pian ping
218	pei	141	pen	195	peng	2 2 202	per	242 401 2	pi pin puo	313 291 1766	pian ping qi
218 171	pei piao	141 3	pen piar	195 65	peng pie	2	per pier	242 401	pi pin	313 291 1766 287	pian ping
218 171 4	pei piao pir qia	141 3 411	pen piar	195 65 132	peng pie pou	2 2 202	per pier pu	242 401 2	pi pin puo	313 291 1766	pian ping qi
218 171 4 98 820	pei piao pir qia qing	141 3 411 864 53	pen piar po qian	195 65 132 333	peng pie pou qiang	2 2 202 172	per pier pu qiao	242 401 2 313	pi pin puo qie	313 291 1766 287	pian ping qi qin
218 171 4 98 820 3	pei piao pir qia qing quar	141 3 411 864 53 372	pen piar po qian qiong que	195 65 132 333 2	peng pie pou qiang qir qun	2 2 202 172 301	per pier pu qiao qiu	242 401 2 313 855	pi pin puo qie qu	313 291 1766 287 489	pian ping qi qin quan
218 171 4 98 820 3 165	pei piao pir qia qing quar	141 3 411 864 53 372 1819	pen- piar po qian qiong que ren	195 65 132 333 2 101 178	peng pie pou qiang qir qun reng	2 2 202 172 301 504 3	per pier pu qiao qiu ran rer	242 401 2 313 855 201	pi pin puo qie qu rang	313 291 1766 287 489 85	pian ping qi qin quan rao
218 171 4 98 820 3 165 143	pei piao pir qia qing quar re	141 3 411 864 53 372 1819 616	pen — piar po qian qiong que ren ru	195 65 132 333 2 101 178 20	peng pie pou qiang qir qun reng rua	2 2 202 172 301 504 3 53	per pier pu qiao qiu ran rer	242 401 2 313 855 201 886 99	pi pin puo qie qu rang ri	313 291 1766 287 489 85 244 42	pian ping qi qin quan rao rong run
218 171 4 98 820 3 165 143 136	pei piao pir qia qing quar re rou	141 3 411 864 53 372 1819 616 85	pen — piar po qian qiong que ren ru	195 65 132 333 2 101 178 20 173	peng pie pou qiang qir qun reng rua sai	2 2 202 172 301 504 3 53 808	per pier pu qiao qiu ran rer ruan san	242 401 2 313 855 201 886 99 82	pi pin puo qie qu rang ri rui sang	313 291 1766 287 489 85 244 42 82	pian ping qi qin quan rao rong run sao
218 171 4 98 820 3 165 143 136 253	pei piao pir qia qing quar re rou ruo se	141 3 411 864 53 372 1819 616 85 134	pen piar po qian qiong que ren ru sa	195 65 132 333 2 101 178 20 173 121	peng pie pou qiang qir qun reng rua sai seng	2 2 202 172 301 504 3 53 808 215	per pier pu qiao qiu ran rer ruan san sha	242 401 2 313 855 201 886 99 82 70	pi pin puo qie qu rang ri rui sang shai	313 291 1766 287 489 85 244 42 82 257	pian ping qi qin quan rao rong run sao shan
218 171 4 98 820 3 165 143 136 253 1100	pei piao pir qia qing quar re rou ruo se shang	141 3 411 864 53 372 1819 616 85 134 370	pen - piar po qian qiong que ren ru sa sen shao	195 65 132 333 2 101 178 20 173 121 468	peng pie pou qiang qir qun reng rua sai seng she	2 2 202 172 301 504 3 53 808 215 124	per pier pu qiao qiu ran rer ruan san sha shei	242 401 2 313 855 201 886 99 82 70 703	pi pin puo qie qu rang ri rui sang shai shen	313 291 1766 287 489 85 244 42 82 257 1107	pian ping qi qin quan rao rong run sao shan sheng
218 171 4 98 820 3 165 143 136 253 1100 2	pei piao pir qia qing quar re rou ruo se shang sher	141 3 411 864 53 372 1819 616 85 134 370 6299	pen piar po qian qiong que ren ru sa sen shao	195 65 132 333 2 101 178 20 173 121 468 3	peng pie pou qiang qir qun reng rua sai seng she shir	2 2 202 172 301 504 3 53 808 215 124 939	per pier pu qiao qiu ran rer ruan san sha shei	242 401 2 313 855 201 886 99 82 70 703 856	pi pin puo qie qu rang ri rui sang shai shen shu	313 291 1766 287 489 85 244 42 82 257 1107 71	pian ping qi qin quan rao rong run sao shan sheng shua
218 171 4 98 820 3 165 143 136 253 1100 2 103	pei piao pir qia qing quar re rou ruo se shang sher shuai	141 3 411 864 53 372 1819 616 85 134 370 6299 42	pen piar po qian qiong que ren ru sa sen shao shi shuan	195 65 132 333 2 101 178 20 173 121 468 3 115	peng pie pou qiang qir qun reng rua sai seng she shir shuang	2 2 202 172 301 504 3 53 808 215 124 939 294	per pier pu qiao qiu ran rer ruan san sha shei shou shui	242 401 2 313 855 201 886 99 82 70 703 856 89	pi pin puo qie qu rang ri rui sang shai shen shu	313 291 1766 287 489 85 244 42 82 257 1107 71 452	pian ping qi qin quan rao rong run sao shan sheng shua shuo
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579	pei piao pir qia qing quar re rou ruo se shang sher shuai si	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2	pen — piar po qian qiong que ren ru sa sen shao shi shuan	195 65 132 333 2 101 178 20 173 121 468 3 115 255	peng pie pou qiang qir qun reng rua saii seng she shir shuang	2 2 202 172 301 504 3 53 808 215 124 939 294 136	per pier pu qiao qiu ran rer ruan san sha shei shou shui	242 401 2 313 855 201 886 99 82 70 703 856 89 349	pi pin puo qie qu rang ri rui sang shai shen shu shun	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437	pen — piar po qian qiong que ren ru sa sen shao shi shuan sir	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun	2 2 202 172 301 504 3 53 808 215 124 939 294 136 522	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou suo	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466	pi pin puo qie qu rang ri rui sang shai shen shu shun su	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169	pen — piar po qian qiong que ren ru sa sen shao shi sir sui	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao	2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou tar	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128	pei piao pir qia quar re rou ruo se shang sher shuai si suer tan teng	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7	pen — piar po qian qiong que ren ru sa sen shao shi shuan sir tang ter	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti	2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2 581	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou tar tian	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta te tiao	313 291 1766 287 489 85 244 42 257 1107 71 452 125 756 120 137	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tie
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743	pen — piar po qian qiong que ren ru sa sen shao shi sir sui	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti tou	2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2 581 4	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou suo tar tian	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta tiao tu	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tie tuan
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2	pei piao pir qia quar re rou ruo se shang sher shuai si suer tan teng	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199	pen — piar po qian qiong que ren ru sa sen shao shi shuan sir tang ter	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104	peng pie pou qiang qir qun reng rua saii seng she shir shuang song sun tao ti tou tun	2 2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2 2 581 4 174	per pier pu qiao qiu ran rer ruan san shei shou shui sou tar tian tour tuo	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta te tiao tu wa	313 291 1766 287 489 85 244 42 257 1107 71 452 125 756 120 137 156 415	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tie tuan wai
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer tan teng ting	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743	pen — piar po qian qiong que ren ru sa sen shao shi shuan sir sui tang ter tong	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti tou	2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2 581 4	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou suo tar tian	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203 829	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta te tiao tu wa	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tie tuan wai
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer tan teng ting tuer	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199	pen — piar po qian qiong que ren ru sa sen shao shi shuan sir sui tang ter tong	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104	peng pie pou qiang qir qun reng rua saii seng she shir shuang song sun tao ti tou tun	2 2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2 2 581 4 174	per pier pu qiao qiu ran rer ruan san shei shou shui sou tar tian tour tuo	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta te tiao tu wa	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141 1221	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tie tuan wai
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2 686	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer tan teng ting tuer wan	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199 534	pen — piar po qian qiong que ren ru sa sen shao shi shuan sir sui tang ter tong tui wang	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104 12	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti tou tun war	2 2 2 2 2 301 504 3 53 808 215 124 939 294 136 522 2 581 4 174 1865	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou suo tar tian tour tuo wei	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203 829	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta te tiao tu wa	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tie tuan wai
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2 686 9	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer tan teng ting tuer wan wer xiang	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199 534 975 940	pen piar po qian qiong que ren ru sa sen shao shi shuan sir tang ter tong tui wang wo xiao	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104 115 115 115 115 115 115 115 115 115 11	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti tou tun war wu	2 2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2 581 4 174 1865 1063	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou tar tian tour tuo wei xi	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203 829 610	pi pin puo qie qu rang ri rui sang shai shen shu shun stu ta te tiao tu wa	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141 1221	pian ping qi qin quan rao rong run sao shan shua shua shua tei tei tuan wai weng xian
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2 686 9 1377 202	pei piao pir qia quar re rou ruo se shang sher shuai si suer tan teng ting tuer wan wer xiang xiong	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199 534 975 940 2	pen piar po qian qiong que ren ru sa sen shao shi shuan sir sui tang ter tong tui wang wo xiao xir	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104 12 1529 3	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti tou tun war wu xiar xiu	2 2 2 202 172 301 504 3 53 808 215 124 939 294 136 522 2 581 4 174 1865 1063 576	per pier pu qiao qiu ran rer ruan san sha shei shou shui sou tar tian tour tuo wei xi xi xie xu	242 401 2 313 855 201 886 99 82 70 703 856 89 1466 198 274 442 203 829 610 930	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta tiao tu wa wen xia xin xuan	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141 1221 1281	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tie tuan wai weng xian xing
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2 686 9 1377 202 562	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer tan teng ting tuer wan wer xiang xiong xue	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199 534 975 940 2 278	pen piar po qian qiong que ren ru sa sen shao shi shuan sir sui tang ter tong tui wang wo xiao xir xun	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104 12 1529 3 184 328	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti tou tun war wu xiar xiu ya	2 2 2 2 2 3 3 3 5 3 8 0 8 2 15 12 4 9 3 9 2 9 4 13 6 5 2 2 2 2 2 17 4 17 18 18 19 19 19 19 19 19 19 19 19 19 19 19 19	per pier pu qiao qiu ran rer ruan san shai shou shui sou tar tian tour tuo wei xi xie xu yan	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203 829 610 930 277	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta tiao tu wa wen xia xin xuan yang	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141 1221 1281 2	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tie tuan wai weng xian xing xuar yao
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2 686 9 1377 202 562 4	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer tan teng ting tuer wan wer xiang xiong xiong	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199 534 975 940 2 278 1175	pen piar po qian qiong que ren ru sa sen shao shi shuan sir tung ter tong tui wang wo xiao xir xun ye	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104 12 1529 3 184 328 3	peng pie pou qiang qir qun reng rua saii seng she shir shuang song sun tao ti tou tun war wu xiar xiu ya yer	2 2 2 2 2 3 3 1 5 3 5 3 8 0 8 2 1 5 1 2 4 9 3 9 2 9 4 1 3 6 5 2 2 2 5 8 1 4 1 4 1 6 1 6 3 6 6 6 7 6 7 6 7 6 7 6 7 6 7 6 7 6 7	per pier pu qiao qiu ran rer ruan san shai shou shui sou tar tian tour tuo wei xi xie xu yan yi	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203 829 610 930 277 503 13	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta te tiao tu wa wen xia xin xuan yang yin	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141 1221 1281 2 943 2750	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tei tuan wai weng xian xing xuar yao yin
218 171 4 98 820 3 165 143 136 253 1100 2 103 1579 2 432 128 354 2 686 9 1377 202 562	pei piao pir qia qing quar re rou ruo se shang sher shuai si suer tan teng ting tuer wan wer xiang xiong xue	141 3 411 864 53 372 1819 616 85 134 370 6299 42 2 437 169 7 743 199 534 975 940 2 278	pen piar po qian qiong que ren ru sa sen shao shi shuan sir sui tang ter tong tui wang wo xiao xir xun	195 65 132 333 2 101 178 20 173 121 468 3 115 255 76 232 640 374 104 12 1529 3 184 328	peng pie pou qiang qir qun reng rua sai seng she shir shuang song sun tao ti tou tun war wu xiar xiu ya	2 2 2 2 2 3 3 3 5 3 8 0 8 2 15 12 4 9 3 9 2 9 4 13 6 5 2 2 2 2 2 17 4 17 18 18 19 19 19 19 19 19 19 19 19 19 19 19 19	per pier pu qiao qiu ran rer ruan san shai shou shui sou tar tian tour tuo wei xi xie xu yan	242 401 2 313 855 201 886 99 82 70 703 856 89 349 1466 198 274 442 203 829 610 930 277 503	pi pin puo qie qu rang ri rui sang shai shen shu shun su ta tiao tu wa wen xia xin xuan yang	313 291 1766 287 489 85 244 42 82 257 1107 71 452 125 756 120 137 156 415 141 1221 2	pian ping qi qin quan rao rong run sao shan sheng shua shuo suan tai tie tuan wai weng xian xing xuar yao

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

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.