DEVELOPMENT OF A LARGE VOCABULARY SPEECH DATABASE FOR CANTONESE

P.C. Ching[1], *K.F. Chow*[1], *Tan Lee*[1], *Alfred Y.P. Ng*[2] *and L.W. Chan*[2]
[1] Department of Electronic Engineering [2] Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
Tel: (852) 2609 8271 Fax : (852) 2603 5558
{pcching,kfchow,tlee1}@ee.cuhk.edu.hk {ypng,lwchan}@cs.cuhk.edu.hk

ABSTRACT

This paper describes our recent work on developing a large vocabulary speech database for Cantonese. As a major Chinese dialect, Cantonese is spoken by tens of millions of people in Southern China and Hong Kong. It is very different from Mandarin or Putonghua in phonology, phonetics, vocabulary and grammatical structure. A speech database specially designed for Cantonese is urgently needed for the design, implementation and performance evaluation of various speech recognition systems. The proposed database contains a large number of speech utterances which include isolated syllables, polysyllabic words and phonetically rich sentences. It covers most of the intra-syllable and inter-syllable acoustic variations. We hope that this pioneer work will be beneficial and useful to facilitate future research activities in the related areas.

1 INTRODUCTION

Speech technology has acquired significant progress in real-world applications during the past decade. It aims at providing a more convenient and efficient way of human-machine communication. For the design, implementation and performance evaluation of speech processing systems, large amount of acoustic data are indispensable. There have been tremendous efforts on constructing speech database for English and many other western languages [1–3]. Similar work has also been carried out in Japan [4]. However, for Chinese languages, only a few small scale speech databases have been developed [5,6]. All of them are built for Mandarin or Putonghua, which is the "official" Chinese spoken language. However, there exist hundreds of different Chinese dialects. Amongst them, Cantonese is probably the most popular one and is spoken by tens of millions of people in the rapidly developing South China economic zone and Hong Kong. It is also widely used in many overseas Chinese communities. From both linguistic and acoustic points of view, Cantonese is significantly different from Mandarin. A speech database specially designed and collected for Cantonese is urgently needed to facilitate the fast growing research and development activities in multi-lingual speech communication.

This paper describes a pioneer study on the design and collection of a large vocabulary speech database for Cantonese. Basically, this database is intended for training and performance evaluation of syllable or sub-syllable based speech recognition system. The speech corpus contains all existing Cantonese syllables, 3,200 polysyllabic words and 2,400 sentences, This would cover most of the intra-syllable and inter-syllable acoustic variabilities in Cantonese. A large number of different speakers are involved to provide the essential data for speaker-independent applications.

2 CANTONESE PHONOLOGY

Like many other Chinese dialects, Cantonese is monosyllabic and tonal. Each Chinese character is pronounced as a single **syllable** with a specific **lexical tone**. Cantonese syllables have the general structure of C_1VC_2, where C_1 and C_2 are optional consonants and V is either simple vowel or diphthong. The combinations of C_1, V and C_2 are severely restricted. In fact, only four basic syllable types exist: V, C_1V, C_1VC_2 and VC_2. In the latter two cases, V can not be a diphthong. Furthermore, in the case of C_1V, if C_1 is a plosive or nasal, V can not be simple vowel /i/, /y/ and /u/. Traditional Chinese phonology treats each syllable as a combination of an **Initial** (**聲母**) and a **Final** (**韻母**), which correpond to C_1 and VC_2 respectively. According to such a classification, there are 20 Initials and 53 Finals in Cantonese (Table 1).

Cantonese is also well known of being rich in tones. Phonologically there are nine different tones in Cantonese while only four or five are defined in Mandarin. Figure 1 gives a sketch of the nine Cantonese tones and the syllable /si/ is used as an example to illustrate that different tones correpond to different Chinese characters. The tones are labeled from 1 to 9 and divided into two main categories: entering tones and non-entering tones. The six non-entering tones (1 – 6) are different from each other either by their relative pitch levels or pitch variation patterns. The three entering tones (7 –

9) occur exclusively together with final consonants /p/, /t/ and /k/ and therefore are very short in duration.

There are approximately 1,450 **tonal syllables** being used in contemporary Cantonese. If tone is not considered, the number of **base syllables** (Initial-Final combinations) is reduced to 580, which is still 40% more than that in Mandarin (about 400) [7, 8].

Initials (C_1	
Plosive:	
b	(p)
d	(t)
g	(k)
gw	(k^w)
p	(p^h)
t	(t^h)
k	(k^h)
kw	(k^{wh})
Approximant:	
l	(l)
w	(w)
j	(j)
Nasal:	
m	(m)
n	(n)
ng	(ŋ)
Fricative:	
s	(s/ʃ)
f	(f)
h	(h)
Affricate:	
dz	(ts/tʃ)
ts	(ts^h/$tʃ^h$)

Finals (VC_2			
Simple vowel:		on	(ɔn)
i	(i)	ong	(ɔŋ)
yu	(y)	am	(ɐm)
u	(u)	an	(ɐn)
e	(ɛ)	ang	(ɐŋ)
oe	(œ)	aam	(am)
o	(ɔ)	aan	(an)
aa	(a)	aang	(aŋ)
Diphthong:		*Vowel-stop:*	
ui	(ui)	ip	(ip)
ei	(ei)	it	(it)
eoi	(øy)	ik	(ɪk)
oi	(ɔi)	yut	(yt)
ai	(ɐi)	ut	(ut)
aai	(ai)	uk	(ʊk)
iu	(iu)	ek	(ɛk)
ou	(ou)	eot	(øt)
au	(ɐu)	oek	(œk)
aau	(au)	ot	(ɔt)
		ok	(ɔk)
Vowel-nasal:		ap	(ɐp)
im	(im)	at	(ɐt)
in	(in)	ak	(ɐk)
ing	(ɪŋ)	aap	(ap)
yun	(yn)	aat	(at)
un	(un)	aak	(ak)
ung	(ʊŋ)		
eng	(ɛŋ)	*Syllabic nasal:*	
eon	(øn)	m	(m)
oeng	(œŋ)	ng	(ŋ)

Table 1: Cantonese phonemes labeled using LSHK-CUEE system. The corresponding IPA symbols are also given (in parentheses) for the ease of reference

Throughout this study, a romanization system called LSHK-CUEE has been adopted to symbolize Cantonese phonemes. The LSHK system was originally proposed by the Linguistic Society of Hong Kong in 1993 [9]. LSHK-CUEE is a modified version of LSHK which provides a better representation of phonetic similarity. Using this system, each Cantonese syllable is labeled by 2 – 6 English alphabets (a – z) plus a single digit (1 – 9) signifying the tone. For example, the Chinese characters "象", "像" and "匠" can be transcribed as /dzoeng6/, which is formed by the Initial /dz/ and the Final /oeng/ with tone 6 (lower going).

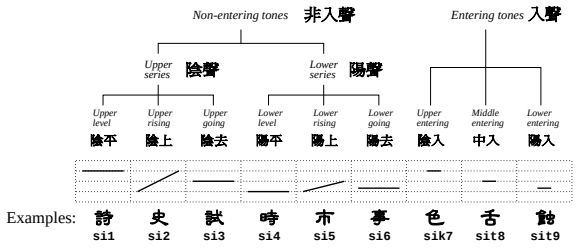

Figure 1: The nine Cantonese tones. The example syllables all have the same Initial /s/ and similar vowel nucleus /i/, but with different tones

3 CORPUS DESIGN

Corpus design is the first and the most important step to build a large vocabulary speech database. It involves the selection of appropriate text materials to be recorded. Conventionally, an efficient way of obtaining these materials is to extract them directly from generally accessible text databases like newspapers, books, etc. It is assumed that the extracted words or sentences are orally readable for ordinary speakers. However, this is not the case for Cantonese. As a regional dialect, spoken Cantonese is noticeably different from formal written Chinese although most Cantonese syllables have corresponding Chinese characters and vice versa. In many situations, it is difficult to recite written Chinese in fluent Cantonese speech [10]. On the other hand, a number of Cantonese words are really colloquial and can hardly be expressed in formal written Chinese. A typical example is shown below:

Spoken Cantonese	**keoy5** 佢	**dzou6 dzoeng2 soeng6 ngaat8** 做掌上壓	**go2 dzan6 si4** 個陣時	**dzing2 tsan1 dzek8 sau2** 整親隻手
Written Chinese	**ta1** 他	**zuo4 fu3 wo4 cheng1** 做俯臥撐	**de shi2 hou4** 的時候	**dzing2 tsan1 dzek8 sau2** 把手弄傷了
English Translation	He	do press-up exercise	moment	hurt his arm
	(He hurt his arm when doing press-up exercise.)			

It can be seen that spoken Cantonese and written Chinese use very different words to express the same meaning, e.g. both "做掌上壓" and "做俯臥撐" mean "doing press-up exercise". Moreover, the word order is reversed in the last phrase "hurt his arm".

Therefore we have to be extremely careful in choosing text databases to avoide the inclusion of non-conventional or the exclusion of commonly used Cantonese speech. In the present study, we have identified the following useful resources:

(1) Recently published guide books for learning Cantonese;
(2) Chinese dictionaries edited specifically for Cantonese speakers;
(3) Part of two major newspapers in Hong Kong (including the pages of local and regional news, finance and economics, sports, information technology).

(1) & (2) are utilized to provide syllable or word level text data while running sentences are obtained mainly from newspaper articles. All selected materials are screened by a linguist expert as well as an ordinary Cantonese speaker to ensure that they are really contemporary spoken Cantonese.

The resultant corpus can be divided into three parts which are targeted for different applications.

Corpus I: Isolated Syllables

I-1 234 tone balanced syllables
Tone is carried by the Final of a syllable. The 234 syllables cover all existing combinations of the 53 Cantonese Finals and the 9 tones while the selection of Initials is arbitrary. This part of speech data can be utilized to develop and evaluate various tone recognition algorithms for Cantonese [11].

I-2 200 commonly used monosyllabes
The 200 syllables are selected based on two criteria: 1) they have high frequency of occurrences; 2) they can be used in isolation as meaningful monosyllabic words. The selected syllables include those describing number, time, actions, colors, directions, people, etc. The corresponding Chinese characters occupy about 40% of newspaper text. This part of speech data have been used to implement a small vocabulary recognition system of isolated syllables [12] or command dictation system.

Corpus II: Full Cantonese Syllabary

This speech corpus is designed to cover the entire syllabary of Cantonese which contains 1,450 monosyllables. Since a large number of these syllables would never be uttered in isolation, they are put into certain word context to make the recording process more realistic and efficient. For each Cantonese syllable, three different carrier words are designed, in which the target syllable occurs at the beginning, the end and the middle positions respectively. As a result, a total of 4,055 words (2 – 7 syllables) are being included in the corpus. Several examples are shown in Table 2.

Target syllable	Carrier words		
	Begin	End	Middle
saa1	saa1tin4 沙田	tsoeng4 saa1 長沙	hak7 saa1 waan1 黑沙灣
fuk1	fuk1 dzaap9 複雜	tsung4 fuk1 重複	pin1 fuk1 haap9 蝙蝠俠
m5	m5 jyut9 faa1 五月花	lok9 m5 落伍	haa6 m5 tsaa4 下午茶
dzaau2	dzaau2 tsam4 找尋	fung6 dzaau2 鳳爪	baat8 dzaau2 jyu4 八爪魚
lung4	lung4 dzau1 龍舟	hau4 lung4 喉嚨	siu2 lung4 baau1 小籠包

Table 2: A partial list of corpus II

Since Cantonese is a monosyllabic language, acoustic modeling at syllable or sub-syllable level is especially important for developing a large vocabulary speech recognition system. This full syllabary speech database provides multiple utterances for each Cantonese syllable to facilitate the training of HMM or NN based syllable models.

Corpus III: Phonetically Rich Words and Sentences

This corpus is designed for the development of speaker-independent continuous speech recognition system. It is required to have a broad phonetic coverage in terms of both context-dependent and context-independent variations. Considering the phonological characteristics of Cantonese, we utilize the following numerical figures to measure the phonetic content of the speech corpus.

1) No. of occurrences of each Initial;
2) No. of occurrences of each Final;
3) No. of occurrences of each intra-syllable Initial-Final combination (base syllables);
4) No. of occurrences of each intra-syllable Final-tone combination (there are 227 such combinations);
5) No. of occurrences of each inter-syllable Final-Initial combination (there are 1219 such combinations);

About 3,200 polysyllabic words (2 – 6 syllables per word) are selected mainly from word dictionaries for secondary school students in Hong Kong. Another 2,400 sentences ($\leq$ 25 syllables/sentences) have been extracted based on one year of newspaper text. The selection process is semi-automatic with the assistance of in-house developed Chinese processing softwares. First of all, the original newspaper text is partitioned into sentences by detecting the punctuations. The sentences are further converted into Cantonese pronunciations (LSHK-CUEE system) and the phonetic content

of each sentence is obtained accordingly. A small number of sentences are selected manually to form the initial set and its phonetic coverage. Then new sentences are added progressively and iteratively to expand the existing corpus. At each step, all candidate sentences are examined and sorted according to their potential contributions to improve the current phonetic coverage. The top 10 candidates, after manual screening of readability, would be included in the corpus.

4 DATA COLLECTION

Corpus I is designed for small scale applications and, therefore, only 5 male and 5 female speakers are needed. Each of them provides 3 full sets of speech data, which contains about 1,300 utterances. The recording was carried out with Gradient Desklab 216, using 16-bit A/D converter at 10 KHz sampling rate.

For Corpus II, 15 male and 15 female speakers are involved. All of them are university students, aged from 18 to 23 years old. Each speaker has to read the entire corpus and 4,055 utterances are collected. The total recording time is about 12 – 15 hours.

Since corpus III is intended for speaker-independent applications, the speech data would be obtained from a large number of speakers with different background and characteristics. Currently, a total of 120 speakers are involved. Their ages vary from 12 to 50 years old. In this case, each speaker is asked to read 150 polysyllabic words and 30 sentences. The recording takes 45 – 60 minutes for each subject.

Speech data for Corpus II & III are initially recorded into a high-quality DAT (Sony PCM-2700A) using a desk-mounted uni-directional microphone (Shure BG 5.0) in a moderately quiet room. Raw speech data are kept in tapes for future use. The DAT is connected via a digital audio interface (*DAT-Link+*) to a Sunsparc workstation. The raw data are down-sampled to 16 KHz. The non-speech part as well as recording artifacts are identified and discarded. Each utterance is stored separately in a binary file which begins with an NIST format 1024-byte header.

5 TRANSCRIPTION AND LABELING

The recorded utterances are verified and transcribed by trained personnels (senior undergraduate students from linguistic departments). For Corpus II and III, manual labeling is performed on the speech waveform to give the boundaries of syllables, Initials and Finals. Furthermore, a selected part of utterances in Corpus will be labeled in more detail (up to allophone level) by experienced spectrogram reader. All labeling work is carried out on Sunsparc workstations with Entropic ESPS/waves+ software tools.

The whole project is labour intensive and still require a tremendous effort before completion. However, we wish that this pioneer work will be useful for further research in the area of speech processing, specifically automatic recognition and natural language understanding of Cantonese.

ACKNOWLEDGEMENT

This work is partly supported by a Research Grant from Hong Kong Research Grants Council (RGC).

REFERENCES

[1] V. Zue, S. Seneff and J. Glass, "Speech database development at MIT: TIMIT and beyond", *Speech Communication*, Vol.9, pp.351 – 356.

[2] T. Robinson *et al*, "WSJCAM0: A British English speech corpus for large vocabulary continuous speech recognition", *Proc. ICASSP-95*, Vol.1, pp.81 – 84.

[3] K. Varghese, S.Pfleger and J.P. Lefevre (Eds.), *Advanced Speech Applications: European Research on Speech Technology*, Springer-Verlag, 1994.

[4] A. Kurematsu *et al*, "ATR Japanese speech database as a tool of speech recognition and synthesis", *Speech Communication*, Vol.9, pp.357 – 363.

[5] J. Sun *et al*, "A Chinese speech database", *Acta Acustica*, Vol.16, No.6, pp.466 – 471.

[6] H.-C. Wang, "Design and implementation of Mandarin speech database in Taiwan", *Proc. EUROSPEECH-95*, Vol.1, pp.875 – 878.

[7] P.C. Ching, Tan Lee and Eric Zee, "From phonology and acoustic properties to automatic recognition of Cantonese", *Proc. ISSIPNN-94*, Vol.1, pp.278 – 283.

[8] Tan Lee, *Automatic Recognition of Isolated Cantonese Syllables Using Neural Networks*, PhD Thesis, The Chinese University of Hong Kong, May 1996.

[9] Linguistic Society of Hong Kong, *Newsletter*, No.15, 1993.

[10] S. Matthews and V. Yip, *Cantonese: A Comprehensive Grammar*, Routledge, London, 1994.

[11] Tan Lee, P.C. Ching, L.W. Chan, B. Mak and Y.H. Cheng, "Tone recognition of isolated Cantonese syllables", *IEEE Trans. SAP*, Vol.3 No.3, pp.204 – 209.

[12] Tan Lee and P.C. Ching, "A Neural Network Based Speech Recognition System for Isolated Cantonese Syllables", to be presented in ICASSP-97, Munich, April 1997.