DEVELOPMENT OF CANTONESE SPOKEN LANGUAGE CORPORA FOR SPEECH APPLICATIONS

W.K. LO, Tan LEE and P.C. CHING
Department of Electronic Engineering
Chinese University of Hong Kong, Hong Kong
Tel. +852 2609 8271, FAX: +852 2603 6868, E-mail: {wklo,tlee1,pcching}@ee.cuhk.edu.hk

ABSTRACT

In this paper, we will present the up-to-date status for the development of several large-scale Cantonese spoken language corpora. These corpora include speech data at different linguistic levels ranging from isolated syllable to continuous passage. This is the first ever effort in compiling a good collection of spoken language resources for research and development in Cantonese speech processing. Various considerations for specific applications have been taken into account during the design stage. This would ensure better usefulness and applicability of the collected speech data. Furthermore, for different targeted applications, different kinds of annotations are also provided to enhance the readiness of the resources.

1. INTRODUCTION

Cantonese is one of the major dialects in Southern China. It is also the mother tongue of most people residing in Hong Kong. Over the past decade, noticeable efforts have been spent on investigating the inherent acoustic properties as well as developing efficient and effective speech processing techniques for this dialect [1,2,3,4,5]. However, in order to achieve further advancement in this field of studies, a large-scale, carefully designed and publicly available set of spoken language corpora is desperately in need. Therefore, we have made initiating efforts to compile such kind of corpora for research and development purposes. This work can be regarded as a basic infrastructure building and is also a starting point for further contributions in collecting similar spoken language resources.

For other languages such as English [6], European languages [7] and Mandarin [8,9], many efforts have already been made in constructing speech databases with much fruitful results. Previous experiences on other languages are definitely valuable to our work. Based on these experiences, we have carefully designed the corpora to customize them according to the special characteristics of the language under consideration. Therefore, we would like to give a general overview on the phonology and phonetics of Cantonese before going further to discuss the details of the corpora.

1.1. Review on Cantonese Phonetics

Cantonese is one of the many Chinese dialects. It has similar structure as others. It is a monosyllabic language. Meaningful word, phrases and sentences are made up from the inventory of over 1600 syllables. However, the situation is complicated by the fact that the same syllable may refer to different characters carrying totally different meaning. This in fact makes the study of the dialect, and in fact the Chinese language, very challenging.

1.1.1. Lexical Tones

It is well known that Cantonese is a tonal language. Tones here carry lexical functions. That is, while syllables are made up from combination of different phonological segments, these syllables convey different meaning and/or correspond to different characters when carrying different tones. Traditionally, it is accepted that there are nine different lexical tones in Cantonese. These nine different tones are realized through the differences in the pitch level, pitch profile as well as the segmental duration of the syllables.

From another point of view, the nine different tones could be classified as six [10]. The three discarded tones are the entering tones. They are re-classified as one of those retained in the six-tone system. One advantage of this system is that segmental duration is ignored in lexical tone classification. In fact, the distinctive duration features of entering tones become less reliable in continuous speech. A rather indirect advantage of using the six tones system is that it makes phonemic transcriptions for many colloquial Cantonese syllables possible (e.g. entering tone uttered as high rising tone). Throughout this work, we shall adopt the six-tone system in phonemic transcriptions.

1.1.2. Syllables and Phonological segments

Phonologically, Cantonese speech is made up from strings of syllables. These syllables can be segmented into two types of phonological units : initial and final. In Cantonese, there are totally 19 initials and 53 finals. Out of these bases of initials and finals, around 600 valid base syllables are formed. If tone is also taken into account, there will be around 1600 common tonal syllables.

1.1.3. Phonetics and classifications of Cantonese

In Cantonese, all consonants are initials while not all initials are consonants. Some initials are semi-vowels or nasals. Table 1 sumarizes the manners of articulation for the Cantonese initials. For non-nasal initials, their manners of articulation include fricatives, affricates, plosives, glide, liquid as well as voiced-unvoiced and aspirated-unaspirated. Except for syllablic nasals, all Cantonese syllables contain at least one vowel in the final of the syllable. Generally speaking, Cantonese finals can be broadly classified into five categories: simple vowel, diphthong, vowel with nasal coda, vowel with stop coda and nasal only.

LSHK	Manner of Articulation
/b/	Plosive, unaspirated
/d/	Plosive, unaspirated
/g/	Plosive, unaspirated
/p/	Plosive, aspirated
/t/	Plosive, aspirated
/k/	Plosive, aspirated
/gw/	Plosive, unaspirated, lip-rounded
/kw/	Plosive, aspirated, lip-rounded
/z/	Affricate, unaspirated
/c/	Affricate, aspirated
/s/	Fricatives
/f/	Fricatives
/h/	Fricatives
/j/	Glide
/w/	Glide
/l/	Liquid
/m/	Nasal
/n/	Nasal
/ng/	Nasal

Table 1. Manners of articulation for Cantonese initials

		/i/	**/u/**	**/p/**	**/t/**	**/k/**	**/m/**	**/n/**	**/ng/**
/aa/	/aa/	/aai/	/aau/	/aap/	/aat/	/aak/	/aam/	/aan/	/aang/
/a/		/ai/	/au/	/ap/	/at/	/ak/	/am/	/an/	/ang/
/e/	/e/	/ei/				/ek/			/eng/
/i/	/i/		/iu/	/ip/	/it/	/ik/	/im/	/in/	/ing/
/o/	/o/	/oi/	/ou/		/ot/	/ok/		/on/	/ong/
/u/	/u/	/ui/			/ut/	/uk/		/un/	/ung/
/yu/	/yu/				/yt/			/yn/	
/oe/	/oe/	/eoi/			/eot/	/oek/		/eon/	/oeng/
							/m/		/ng/

Table 2. Constructions of Cantonese finals from different phones (LSHK).

Altogether, Cantonese has seven long and four short vowels (short occurs only in diphthong or with coda). Based on the places of articulation, they could be broadly grouped into three categories : alveolar (/aa/, /a/, /e/, /i/, /oe/, /y/) , labial (/u/) and velar (/o/). Table 2 below summarizes the formation of Cantonese finals.

1.2. Overview of the Corpora

Having introduced the basic phonology and phonetics of Cantonese, readers should now be aware of the underlying complexity of this Chinese dialect. In order to facilitate different areas of applications, we have included speech data at different linguistic levels into the corpora: syllable, word, sentence and passage. (CUSYL, CUWORD, CUSENT, CUPASS respectively). In addition, there are also corpora of digit strings and selected navigation commands for simple and domain-specific product development purposes (CUDIGIT, CUCMD).

In the following sections, presentation will be made following the corpora development processes. We will first discuss the design of the corpora and then the process of data collection. After that, we will describe in detail the verification and annotation process of the collected data. We will also give a brief introduction on the organization of the speech and annotation data in the corpora. Finally, extensions for further development and a brief conclusion are given.

2. CORPORA DESIGN

There are many different kinds of spoken language corpus design and they are all targeted for different domains of applications. A well-designed corpus not only makes the database construction process simpler and easier, but also makes itself better suited to the specific applications and thus fosters fruitful results.

When collecting speech corpora, we have tried our best to ensure that they will be useful but yet there is still no guarantee that they will suit any particular task. On the contrary, we attempted to make each of the corpora applicable to a wider range of applications. In the following paragraphs, we will introduce the targeted area of applications and the underlining design paradigm for each of the corpora.

2.1. CUSYL

CUSYL is a syllable corpus covering 1806 different Cantonese tonal syllables from each of the participated speakers. The original goal is for use in syllable based concatenation synthesis. Almost all Cantonese speech utterances can be constructed from this inventory. Data from two male and two female speakers has been collected for this corpus.

一巴掌	jat1-**baa1**-zoeng2
掃把星	sou3-**baa2**-sing1
強行霸佔	koeng5-hang4-**baa3**-zim3

Figure 1. Section of the reading material in CUSYL

This corpus constitutes of hand-cut syllables of read speech. The syllables are recorded within meaningful carrier words to reduce pronunciation mistakes and to minimize syllable lengthening (as compared to speaking in isolation). Prolonged syllables, when being concatenated, will give annoying lengthy perception to the listener. In addition, the targeted syllables are positioned in the middle (where possible) of the words to further reduce this undesirable effect. Figure 1 lists a section of the reading materials from CUSYL. In summary, this corpus covers altogether 636 base syllables with 19 initials and 54 finals (with colloquial /et/ final).

2.2. CUWORD

CUWORD is a collection of 2527 read words of syllable length ranges from 1 to 7 syllables. This corpus is designed for use in training and evaluation of speech recognition algorithms.

To ensure the usefulness of the speech data in recognition training, meaningful words are manually constructed to contain most of the Cantonese base syllables with suitable occurrence frequencies. The reading materials are extracted from a base inventory of 4055 words that cover all Cantonese base syllables with multiple occurrences. After that, a human assisted automatic process is employed to select from the base inventory those words containing syllables with fewer occurrences while discarding those contain syllables occur too often. The occurrence of each syllable is also bounded while those words contain rare syllable are kept intact. As a result, a total of 2527 words is kept in the corpus. The data is collected from 13 male and 15 female speakers. A section of the reading material is extracted and shown in Figure 2.

.

止咳化痰	zi2-kat1-faa3-taam4
比薩斜塔	bei2-saat3-ce4-taap3
水塔頂端	seoi2-taap3-deng2-dyun1
火辣辣	fo2-laat6-laat6
主僕	zyu2-buk6

.

Figure 2. Section of the reading material in CUWORD

There are altogether 1388 tonal syllables or 559 base syllables in this corpus. Amount them, the occurrence counts of each context free base syllable range from 1 to 39. The abundance of syllables is shown in Figure 3.

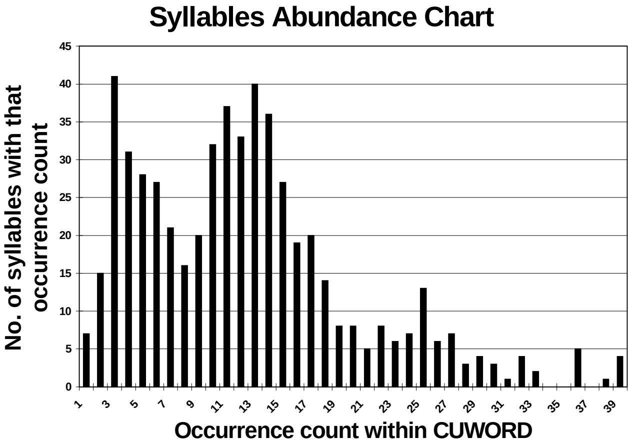

Figure 3. Syllable distribution in CUWORD

2.3. CUDIGIT

CUDIGIT is a read speech corpus of continuous Cantonese digit strings. The reading material is obtained by permuting the ten digit syllables in all combinations from single digit to four digits in sequence. Each of the speakers uttered a randomly selected portion of the permuted list together with a number of random generated long digit strings. Therefore, this digit corpus is guaranteed to have an enormous amount of data for digit recognizer training. In addition, since the reading material contains iterative lists of digit strings of different lengths, the corpus data will also be suitable for context dependent digit training. This corpus is targeted to have data collected from around 100 speakers

Section	Content	Partition	Amount
0	Calibration	1	10
I	Single digit	1	10
II	Double digit	1	100
III	Triple digit	5	200
IV	4-digit	50	200
V	Random 7-digit	Per speaker	20
VI	Random 8-digit	Per speaker	20
VII	Random 14-digit	Per speaker	10
	No. of string per speaker		570

Table 3. The reading materials of CUDIGIT corpus

Table 3 summarizes the material in the CUDIGIT corpus. Within section I, II, III, and IV, there are already 8050 samples for each of the context-free digit from every 50 speakers.

2.4. CUCMD

CUCMD is a small task specific word corpus. It is designed to include common navigation control commands. The corpus data is suitable for use in simple command control tasks even using word-based techniques. It is designed for simple voice controlled consumer products and the product developers could extract necessary materials from the corpus for their specific requirements.

.

左	zo2
右	jau6
中	zung1
中間	zung1-gaan1
中心	zung1-sam1
左o的	zo2-dit1

.

Figure 4. Section of the reading material in CUCMD

This corpus is collected together with CUDIGIT. The number of speakers will be the same. While the same 107 commands will be uttered by each speaker. Among these words, there are 72 different toned syllables or 67 different base syllables included. A section of the reading material is shown in Figure 4 for reference.

2.5. CUSENT

CUSENT is a corpus of continuous Cantonese sentences collected from approximately 100 speakers. It is designed to be phonetically rich under various contexts. All Cantonese syllables, initials, finals and tones are included. Moreover, much attention has been paid to attain better occurrence count for intra-syllable (onset-nucleus) and inter-syllable (coda-onset) contexts. This will ensure an adequate amount of materials for context-dependent speech units modelling.

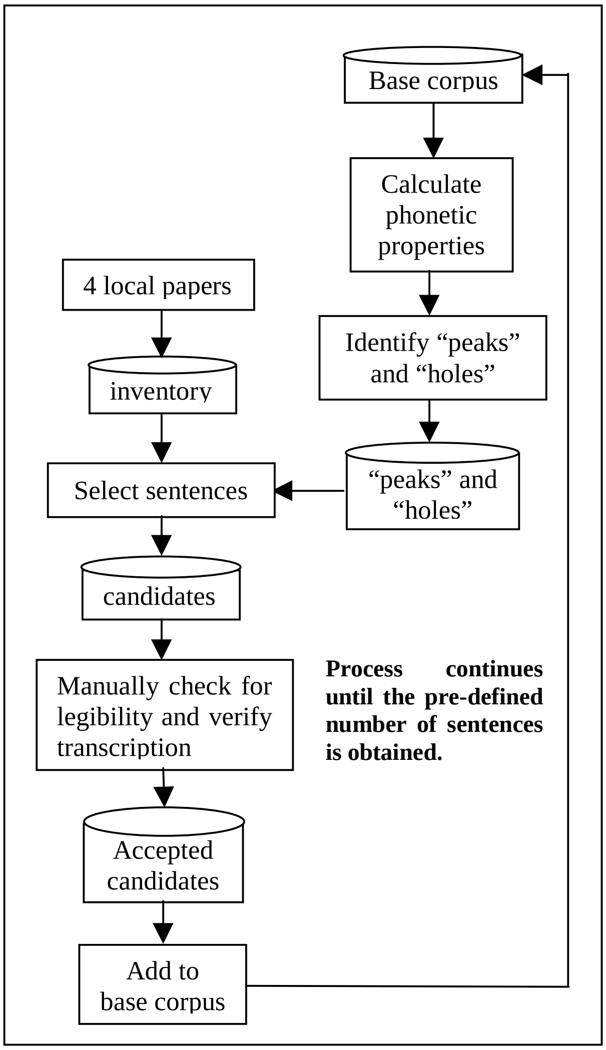
Base corpus
Calculate phonetic properties
4 local papers
Identify "peaks" and "holes"
inventory
Select sentences
"peaks" and "holes"
candidates
Manually check for legibility and verify transcription
Process continues until the pre-defined number of sentences is obtained.
Accepted candidates
Add to base corpus

Figure 5. Semi-automatic selection process for the CUSENT corpus material.

The sentence selection is carried out through a semi-automatic process. We collect sentences from four local newspapers and make a large inventory of sentences from them. First of all, a small amount of sentences is manually picked as the initial base corpus. Phonetic properties of the chosen sentences are then calculated. Those properties with extremely frequent occurrences are identified as "peaks" while those rarely occurred are identified as "holes".

After that, a certain amount of sentences is retrieved from the inventory and their phonetic properties are calculated separately. Sentences are marked as candidates if they fill any "holes" but do not hit any "peaks". Those marked candidates are then manually verified for legibility and transcription correctness. Accepted candidates will finally be added to the base corpus. Finally, phonetic properties of the modified base corpus are recalculated and new "peaks" and "holes" are re-identified. This process is repeated until the base corpus reached a pre-defined amount of sentences. The process is best illustrated in Figure 5.

Within the reading materials for the CUSENT, there are 50509 syllables having a total of 1349 different tonal syllables or 569 base syllables. For every 50 speakers, the average occurrence count of each tonal syllable is 37 and that of base syllable is 89.

2.6. CUPASS

CUPASS is a corpus of continuous read speech. The reading materials are sentences extracted from local newspapers. The passages are selected arbitrarily and screened solely on their legibility. The content covered includes general news, science, critics and journal articles.

Currently, there are twenty passages in which some of them were extracted sections from long articles. Speakers will be invited to record in a recording room with the passages displayed on-screen. They are left free to read the material in whatever style and manner they prefer. The target of this corpus is for use as training and testing material for real continuous speech. Since the reading materials in this corpus are not under tight control, the number of speakers will be kept around 20 or so.

The phonetic statistics of this corpus is that there are 18590 tonal syllables in the twenty passages. Among these materials, there are totally 993 different tonal syllables or 466 different base syllables.

3. DATA COLLECTION

All of the corpora data are collected in a closed silent recording room. The materials are expected to be good quality clean speech data. Basically, the speakers are left alone in the recording room to do the recording themselves without assistance or intervention. This is done so to make them comfortable for collecting natural and fluent speech.

The recording is done using high quality microphones. The signal passes through a pre-amplification mixer and is A/D converted at 48kHz, 16bit using DAT recorder. The sampled digital data is then passed immediately to DATLink where the data are down sampled to 16kHz.

The down sampled stream of data is then transferred through SCSI interface to computer and stored as disk files. The data collection set-up block diagram is given in Figure 6.

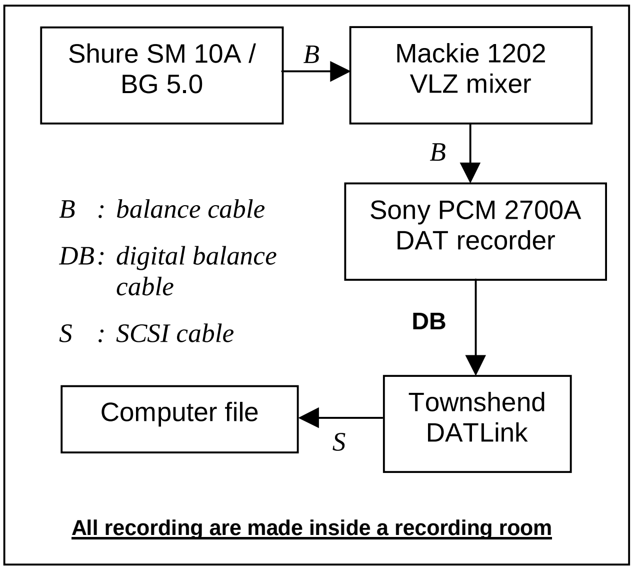

Figure 6. Recording set up for the collection of the corpora.

One of the advantages of collecting data in this way is that we can keep the high quality speech data as computer files. Whenever the channel effect or environmental effect is required to be taken care of, the same clean speech could be used to simulate speech under the desired effects by passing appropriate modules emulating those effects. Although the exact effect is slightly different, as the first large-scale spoken language corpora of its kind, we would rather having it as versatile as possible in order to benefit more parties.

4. VERIFICATION AND ANNOTATION

4.1. CUSYL

Verification of this corpus is done to check if the speakers utter the designated syllables accurately. In addition, the targeted syllables are also hand-cut out from the carrier words. Data annotation is provided as manual pitch marks for one male and one female speaker. In addition, each syllable is stored as separate files with the transcription given as filename.

4.2. CUWORD, CUSENT and CUPASS

Verification is done in two stages. Stage one is done by generally trained assistants. The goal is to mark out all incorrect, problematic, missing as well as doubtful data. The marked lists are passed to stage two.

Stage two of the verification process is performed by experts in phonetics. They will resolve all the marked problems and correct any mistakes. If deemed uncorrectable, the corresponding data will be discarded from the final distribution.

Moreover, the speech data will be accompanied by annotation data. These include orthographic transcriptions in BIG5 code for the corrected reading materials and verified phonemic transcriptions in LSHK scheme.

4.3. CUDIGIT

Verification process of CUDIGT is also conducted in two stages. During stage one, trained assistants will correct all incorrect data and mark those problematic or doubtful data. And then in stage two, other assistants will decide whether those marked data are to be kept or not. They will also hand-cut out those noise or unwanted signal preceding or following desired digit data.

This two-stage verification process can ensure the quality and usefulness of the corpus. It can also guarantee high degree of accuracy in the annotation of the data. Accompanying the speech data of this corpus are the corrected orthographic transcriptions in BIG5 code and verified phonemic transcriptions in LSHK scheme.

4.4. CUCMD

Verification of CUCMD is a one-phase process. Trained assistants will only decide whether to keep the data or not. For data annotation, orthographic transcriptions in BIG5 code and phonemic transcriptions in LSHK scheme are provided.

5. DATA ORGANIZATION

In order to facilitate the research and development of those working on Cantonese speech processing, the corpora data will be distributed in electronic form through CDROM. All speech data will be accompanied by appropriate annotation data. The following paragraph will summarize the format and provisional naming scheme of the data in the corpora.

Speech data format	NIST SPHERE
Sampling rate	16kHz
Precision	16 bit per sample
Orthographic transcription	BIG5 code
Phonemic transcription	LSHK scheme

Table 4. Format of the data and annotation

The data layout structure is simple. All data obtained from the same speaker within a corpus will be placed under the same directory. While the directory name is determined by the corpus code, speaker code and gender.

Table 5 shows the corpus code. For speaker code, it will be in two-digit hexadecimal form that is followed by either ‘M’ or ‘F’ for speaker gender. For example, data for a male speaker in CUWORD may be put under the directory CW06M.

Corpus	Corpus code
CUSYL	CS
CUWORD	CW
CUDIGIT	CD
CUCMD	CC
CUSENT	CN
CUPASS	CP

Table 5. Corpus code for the set of corpora

6. FUTURE EXTENSIONS

Since these corpora are high quality clean speech, extension works on them are possible and useful. For example, we may pass the data over telephone lines or mobile phone connections for collecting telephone and mobile phone versions of the data. User may even pass the data over a particular channel that they are interested for their own requirements.

7. CONCLUSION

In summary, the collection of a large-scale Cantonese spoken language database for speech processing has been launched. A large amount of verified data will be available for research and development in Cantonese speech synthesis and recognition. Currently, part of the corpora are ready while the remaining are still under construction.

With these speech data resources, we hope that more in depth investigation could be fostered and fruitful results might be achieved to further the development of speech technology.

6. ACKNOWLEDGEMENT

This Cantonese database project is partially supported by the Industrial Support Fund awarded by the Industrial Department and also by a RGC research grant from the Hong Kong Research Grant Council.

REFERENCES

1. Tan LEE, P.C. CHING, L.W. CHAN, Y.H. CHENG and B. MAK, “Tone Recognition of Isolated Cantonese Syllables”, *IEEE Trans. on Speech & Audio Processing*, Vol. 3, No. 3, pp. 204-209, 1995.
2. Tan LEE, P.C. CHING and L.W. CHAN, "Isolated word recognition using modular recurrent neural networks", *Pattern Recognition*, Vol.31, No.6, pp.751 - 760
3. Min CHU and P.C. CHING, “A Hybrid Approach to Synthesize High Quality Cantonese Speech”, *Proc. of ICASSP98*, Vol. 1, pp. 277-80, Seattle, 1998.
4. K.F. CHOW, Tan LEE and P.C CHING, “Sub-syllable acoustic modeling for Cantonese speech recognition”, *Proc. Of ISCSLP98*, Singapore, 1998.
5. W.K. LO, K.F. CHOW, Tan LEE and P.C. CHING, “Cantonese databases developed at CUHK for speech processing”, *Proc. of the Conference on Phonetics of the Lang. in China*, pp. 77-80, Hong Kong, 1998.
6. Various resources distributed by Linguistic Data Consortium, http://www.ldc.upenn.edu
7. Databases for the Creation of Voice Driven Teleservices, http://www.phonetik.uni-muenchen.de/SpeechDat.html/ by joint efforts in European Union.
8. Chorkin, CHAN, “Design considerations of a putonghua database for speech recognition”, *Proc. of the Conference on Phonetics of the Lang. in China*, pp. 13-16, Hong Kong, 1998.
9. H.S. WANG, “Design and implementation of Mandarin speech database in Taiwan”, *Proc. of EUROSPEECH95*, Vol.1, pp.875-7, Madrid, 1995.
10. Linguistic Society of Hong Kong (LSHK), *Hong Kong Jyut Ping Character Table (粵語拼音字表)*, Linguistic Society of Hong Kong Press (香港語言學會出版).