ELSEVIER

Speech Communication 36 (2002) 327–342

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Spoken language resources for Cantonese speech processing

Tan Lee [a,*], W.K. Lo [a], P.C. Ching [a], Helen Meng [b]

[a] *Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong*
[b] *Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong*

Received 28 January 2000; received in revised form 17 July 2000; accepted 3 October 2000

**Abstract**

This paper describes the development of *CU Corpora*, a series of large-scale speech corpora for Cantonese. Cantonese is the most commonly spoken Chinese dialect in Southern China and Hong Kong. *CU Corpora* are the first of their kind and intended to serve as an important infrastructure for the advancement of speech recognition and synthesis technologies for this widely used Chinese dialect. They contain a large amount of speech data that cover various linguistic units of spoken Cantonese, including isolated syllables, polysyllabic words and continuous sentences. While some of the corpora are created for specific applications of common interest, the others are designed with emphasis on the coverage and distributions of different phonetic units, including the contextual ones. The speech data are annotated manually so as to provide sufficient orthographic and phonetic information for the development of different applications. Statistical analysis of the annotated data shows that *CU Corpora* contain rich and balanced phonetic content. The usefulness of the corpora is also demonstrated with a number of speech recognition and speech synthesis applications. 

*Keywords:* Speech databases development; Chinese dialects; Chinese phonology and phonetics; Annotation of speech data; Applications of speech technology; Speech recognition; Text-to-speech synthesis

## 1. Introduction

Speech databases are the most important infrastructure for the advancement of state-of-the-art spoken language technologies. Properly recorded and annotated speech data are indispensable for the analysis and modeling of different sources of variabilities in human speech. Automatic speech recognition (ASR) algorithms based on statistical models require large amount of training data recorded from speakers with different genders, voice characteristics, regional accents, education backgrounds, etc. Most speech synthesis techniques also require pre-recorded speech segments for acoustic analysis and concatenation purposes, as well as continuous sentences that contain sufficient prosodic content. For the development of spoken dialogue systems, real conversational speech is needed to facilitate the modeling of the multifarious phenomena in human–computer interaction.

There have been tremendous efforts in creating large spoken language corpora for English and other major Western languages. Many of them

* Corresponding author. Tel.: +852-2609-8267; fax: +852-2603-5558.
*E-mail address:* tanlee@ee.cuhk.edu.hk (T. Lee).

0167-6393/02/$ - see front matter 
PII: S0167-6393(00)00101-1

were made available to general public as common platforms for systems development and evaluation. In the United States, the Advanced Research Projects Agency (ARPA) has supported the development of many corpora, including TIMIT (Lamel et al., 1986; Zue et al., 1990), Resource Management (RM) (Price et al., 1988), Wall Street Journal (WSJ) (Paul and Baker, 1992), Air Travel Information Service (ATIS) (Price, 1990), etc. The Linguistic Data Consortium (LDC) was established in 1992 to support and coordinate corpora development activities and has released more than 140 corpora in more than 20 languages to over 750 organizations worldwide (LDC, 2000). In Europe, a number of multilingual spoken language corpora have been developed with joint efforts from member countries in the European Union (EU). For instance, EUROM1, which covers 11 major languages, was designed for speech input/output assessment (Winski and Fourcin, 1994). SpeechDat has been created to support voice-activated applications over telephone networks with 20 regional variants of 14 major European languages (Höge et al., 1997; ELRA, 2000). Currently, the European Language Resources Association (ELRA) plays the role of a central repository for the creation, verification, and distribution of language resources in Europe (Choukri, 1999). Among the Asian countries, Japan has had the most organized work in speech database development (Kurematsu et al., 1990; Kuwabara et al., 1989; Ohtsuki et al., 1999).

Chinese is one of the major languages in the world. Spoken language technology for Chinese has become an area of hot pursuit in recent years and we have seen increasing efforts being devoted to enhance the research infrastructure in this area. Chinese is known to have many different dialects (Yuan, 1983). Among them, Putonghua (or Mandarin) is regarded as the official spoken language in both Mainland China and Taiwan. Existing Putonghua speech corpora include USTC95 (Wang et al., 1996), HKU96 (Zu et al., 1996; Chan, 1998), HKU99 (Huo and Ma, 1999), MAT (Tseng, 1995; Wang, 1999), CMSC (Wu, 1998) and others (Zhang, 1998).

Another prominent Chinese dialect is Cantonese. It is the mother tongue of the 60 million population in Southern China and Hong Kong. Cantonese is also commonly used in overseas Chinese communities in North America and Australia. Although Cantonese and Putonghua are both monosyllabic and tonal, there are significant differences between them at various linguistic levels, for examples, the inventory of syllable segments, tone system, lexical and grammatical structures (Matthews and Yip, 1994). It is often the case that monolingual Cantonese speakers and monolingual Putonghua speakers cannot communicate with their respective dialects. It is therefore necessary to develop spoken language technologies tailored for the Cantonese dialect. This paper describes the design and development of a set of large-scale spoken language resources for Cantonese. The speech databases, *CU Corpora*, which are the first of their kind, are intended to support research and development of speech and language technologies including Cantonese speech recognition and synthesis (Ching et al., 1994; Lee et al., 1995, 1999; Lee and Ching, 1999; Wong et al., 1999).

In the next section, the phonological and phonetic properties of the Cantonese dialect will be briefly described. The major differences between Cantonese and Putonghua will be addressed. The design considerations for various corpora will be explained in Section 3. The annotation process will be elaborated in Section 4, together with a statistical description of the phonetic content of the databases. In Section 5, each database will be evaluated in its targeted applications and some preliminary experimental results will be given.

## 2. Cantonese phonology and phonetics

### 2.1. Syllables

Like many other Chinese dialects, spoken Cantonese can be considered as a string of monosyllable sounds. Basically, each Chinese character is pronounced as a monosyllable. However, a character may have multiple syllable pronunciations, for example, the character 行 may be pronounced as [haŋ], [hɐŋ] or [hɔŋ], [1] among which

[1] These syllables are labeled with IPA symbols.

[hɐŋ] can be in two different tones. A syllable may also correspond to multiple characters. For example, the characters 司, 私, 施, 思, 師, 斯, 絲, 詩 all share the same pronunciation [si].

Table 1 gives the statistics on Cantonese syllables, in comparison with that of Putonghua. While the term "*base syllable*" refers to the tone-independent monosyllable units, "*tonal syllable*" refers to the syllable sounds that are acoustically and tonally distinct. Cantonese appears to have a richer inventory of syllable sounds than Putonghua. There are approximately 50% more base syllables in Cantonese than in Putonghua. In Table 1, two different information sources are provided to describe the Cantonese syllable inventory. The work described in this paper is mainly based on the publication from the Linguistic Society of Hong Kong (LSHK, 1997).

Cantonese is a spoken dialect. It does not have a standard written form on a par with standard written Chinese (Matthews and Yip, 1994). Written Cantonese is neither taught in schools nor used for official communication. When reciting Chinese text (e.g. newspaper), native Cantonese speakers seldom follow the original text content but usually substitute some of the words with typical colloquial expressions. In this work, we design the speech corpora based on formal written Chinese. However, colloquial sounds are frequently found in the actual recordings. This calls for an attentive annotation process as described later in Section 4.

### 2.2. Phonological units

Like in Putonghua and many other Chinese dialects, each Cantonese syllable is seen as the concatenation of two types of phonological units: *Initial* and *Final*, as shown in Fig. 1 (Hashimoto, 1972). There exist 19 *Initials* and 53 *Finals* in Cantonese, in contrast to 23 *Initials* and 37 *Finals* in Putonghua. The onset and the coda are optional segments in a Cantonese syllable.

#### 2.2.1. Initials

Table 2 lists the 19 Cantonese *Initials*. They are labeled using *Jyut-Ping*, a phonemic transcription scheme proposed by the Linguistic Society of Hong Kong (LSHK, 1997). In Cantonese, all consonants can be present as *Initials* while not all *Initials* are consonants. Some of the *Initials* are semi-vowels or nasals. Non-nasal *Initials* include

| *BASE SYLLABLE* | | |
|---|---|---|
| *Initial* | *Final* | |
| [Onset] | Nucleus | [Coda] |

Fig. 1. Structure of a Cantonese syllable. [ ] means optional.

Table 1
Syllable statistics for Cantonese and Putonghua

| | Cantonese | | Putonghua | |
|---|---|---|---|---|
| | LSHK (LSHK, 1997) | S.L. Wong (Wong, 1941) | CCDICT v3.0 (CCDICT, 2000) | HKU96[a] (Zu et al., 1996) |
| Total number of base syllables | 625 | 605 | 420 | 411 |
| Total number of tonal syllables | 1761 | 1622 | 1471 | 1370 |
| Average number of tones per base syllable | 2.8 | 2.7 | 3.5 | 3.3 |
| Average number of base syllable pronunciations per character | 1.1 | 1.1 | 1.6 | – |
| Average number of tonal syllable pronunciations per character | 1.2 | 1.1 | 2.0 | – |
| Average number of homophonous characters per base syllable | 17 | 13 | 31 | – |
| Average number of homophonous characters per tonal syllable | 6 | 5 | 8 | – |

[a] HKU96 is a Putonghua speech database. It was designed to cover all tonal syllables of Putonghua (Zu et al., 1996). The numbers given in the table were counted from the *pinyin* transcriptions being provided. The counts of character-syllable mapping for HKU96 are not considered as useful information because HKU96 is not dictionary by its nature.

Table 2
The 19 Cantonese *Initials* (labeled with the *Jyut–Ping* scheme)

| LSHK symbols | Manner of articulation | Place of articulation |
|---|---|---|
| [b] | Plosive, unaspirated | Labial |
| [d] | Plosive, unaspirated | Alveolar |
| [g] | Plosive, unaspirated | Velar |
| [p] | Plosive, aspirated | Labial |
| [t] | Plosive, aspirated | Alveolar |
| [k] | Plosive, aspirated | Velar |
| [gw] | Plosive, unaspirated, lip-rounded | Velar, labial |
| [kw] | Plosive, aspirated, lip-rounded | Velar, labial |
| [z] | Affricate, unaspirated | Alveolar |
| [c] | Affricate, aspirated | Alveolar |
| [s] | Fricative | Alveolar |
| [f] | Fricative | Dental-labial |
| [h] | Fricative | Vocal |
| [j] | Glide | Alveolar |
| [w] | Glide | Labial |
| [l] | Liquid | Lateral |
| [m] | Nasal | Labial |
| [n] | Nasal | Alveolar |
| [ng] | Nasal | Velar |

liquids, glides, fricatives, affricates, as well as plosives. In Putonghua, there are totally 23 *Initials*, many of which have similar counterparts in Cantonese. Putonghua has more variants in *Initials*, with the introduction of retroflexed fricatives and affricates.

### 2.2.2. Finals

Table 3 lists the 53 Cantonese *Finals* that can be divided into five categories: vowel (long), diphthong, vowel with nasal coda, vowel with stop coda and syllabic nasal. Except for the syllabic nasal, each *Final* contains at least one vowel element. Cantonese has 7 long vowels, 4 short vowels and 10 diphthongs. Based on the place of articulation, the vowels can be grouped into three broad classes: alveolar ([aa], [a], [e], [i], [oe], [y]), labial ([u]) and velar ([o]). The number of *Finals* in Putonghua is much less than in Cantonese. This is because Cantonese has six different consonant codas ([m], [n], [ng], [p], [t], [k]) but Putonghua has only two ([n], [ng]).

## 2.3. Lexical tones

Cantonese is a tonal language. Each syllable is associated with a specific lexical tone. Syllables with the same phonetic composition but different

Table 3
The 53 Cantonese *Finals* (labeled with the *Jyut–Ping* scheme)

| | | Coda | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Nil | [i] | [u] | [p] | [t] | [k] | [m] | [n] | [ng] |
| Nucleus | [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] | | | | [yut] | | | [yun] | |
| | [oe] | [oe] | [eoi] | | | [eot] | [oek] | | [eon] | [oeng] |
| | | | | | | | | [m] | | [ng] |

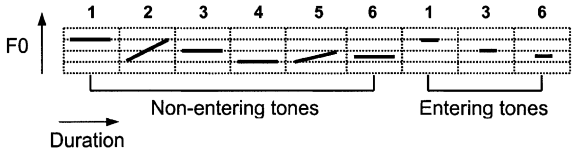


Fig. 2. Tones in Cantonese.

tones correspond to different written characters and convey different meanings. In Chinese languages, tone is basically a feature of pitch or F0 movement across the entire voiced portion of a syllable. Traditionally, Cantonese is said to have nine different tones. This is more complicated than Putonghua, in which five tones (including a neutral tone) are used. As shown in Fig. 2, the so-called entering tones have much shorter duration than the non-entering ones. However, each entering tone coincides with a non-entering counterpart in terms of F0 pattern. Thus in many transcription schemes including the LSHK one, only six distinctive tones are labeled (Fig. 2). This would not cause any phonological ambiguities because the entering tones are associated exclusively with stop codas [p], [t] and [k].

## 3. Corpora design and data collection

Our objective is to develop a wide spectrum of speech corpora to support the development of different applications. The corpora are divided into two types, namely application-specific and linguistics-oriented. The application-specific corpora are created to facilitate the development of some small-scale applications of common interest. The linguistics-oriented corpora contain a large amount of speech data at various linguistic levels: isolated syllables, polysyllabic words and continuous sentences. They are rich in phonetic content and suitable for the development of general-purpose applications like speaker-independent large-vocabulary continuous speech recognition (LVCSR) or text-to-speech (TTS) conversion.

There exist varieties of Cantonese in different geographical regions. In this work, we focus on "Hong Kong Cantonese" or "Cantonese as spoken in Hong Kong". Indeed, the "Hong Kong Cantonese" is considered to be increasingly influential and popular in recent years (Matthews and Yip, 1994). The majority of subjects recruited for our data collection were students at secondary or tertiary level. They grew up in Hong Kong and speak native Cantonese. Their age ranges from 14 to 30.

### 3.1. Recording conditions

All recordings were carried out in a reasonably quiet room. The speech data are expected to be clean and of good quality. Basically, the speakers did the recordings themselves without any supervision, although they had been given a few simple guidelines at the beginning. The recording hardware included a high-quality microphone, a pre-amplification mixer and a DAT recorder connected to a host computer via DATLink, which is a digital interface (Lo et al., 1998). The analog signal was first sampled at 48 kHz at the DAT and then passed immediately to DATLink where the data were down-sampled to 16 kHz. The digital data were then transferred to the computer and stored as binary data files. One advantage of collecting data in this way is that we can keep the high-quality speech data as computer files. Whenever channel or environmental artifacts are to be considered, the clean speech could be used to simulate the desired output speech by passing through appropriate modules emulating those effects. Since the *CU Corpora* are meant to be the first large-scale and general-purpose resources for Cantonese speech technology, they are designed to be as versatile as possible, so as to benefit many parties with diverse demands.

### 3.2. Application-specific corpora: CUDIGIT and CUCMD

Automatic recognition of spoken digits or short commands has found a number of important practical applications, for examples, recognition of stock codes or credit card numbers and voice-controlled systems. It requires a large amount of speech data that contains digit strings of various lengths and specially selected command words, like "start", "stop" and "yes". Examples of such databases in other languages include TI46-Word (English alphabets and commands), TIDIGITS

(English digits), SPINA (German control commands), SPK (Italian digits) and many others (LDC, 2000; ELRA, 2000). Hence in our work, two application-specific speech databases have been created for Cantonese connected-digits and navigation/control commands, respectively.

CUDIGIT is a read speech database of Cantonese connected-digit strings. The base corpus was obtained by permuting the ten monosyllabic digits ("0" to "9") in all combinations of from a single digit to four digits in sequence (Lo et al., 1998). Each speaker was asked to utter a selected portion of the base corpus, plus a certain number of randomly generated long strings (Table 4). CUCMD is a task-specific word corpus for command/control tasks. The number of occurrences for each command word is large enough to facilitate word-level acoustic modeling. A selection of CUCMD material is shown in Fig. 3. The entire corpus contains 107 commonly used commands that were selected manually (Lo et al., 1998). A few colloquial alternative wordings are also included. The command words cover 72 different tonal syllables or 67 different base syllables. Each speaker was asked to read through all of the 107 commands once. For both CUDIGIT and CUCMD, a total of 50 speakers participated in the recording.

Table 4
Corpus materials in CUDIGIT (for each speaker)

| Section no. | Content | Count |
|---|---|---|
| 0 | Calibration | 10 |
| I | Single digit | 10 |
| II | Double digit | 100 |
| III | Triple digit | 200 |
| IV | 4-digit | 200 |
| V | 7-digit (random) | 20 |
| VI | 8-digit (random) | 20 |
| VII | 14-digit (random) | 10 |
| No. of utterances per speaker | | 570 |

| . | | (*English Translation*) |
|---|---|---|
| 左 | zo2 | left |
| 右 | jau6 | right |
| 中 | zung1 | middle |
| 中間 | zung1-gaan1 | in the middle |
| 左啲 | zo2-dit1 | a bit toward the left |
| . | | |

Fig. 3. Selected portion of the CUCMD corpus.

### 3.3. Linguistics-oriented corpora

While CUDIGIT and CUCMD are created for specific applications, the set of corpora described below are designed with a rather general scope. These corpora provide a wide coverage of Cantonese speech units (e.g., syllables and context-dependent phones). They are intended to facilitate the use of state-of-the-art speech recognition and speech synthesis technologies.

#### 3.3.1. CUSYL

CUSYL is a syllable corpus with an extended coverage of 1801 tonal syllables. While about 1760 of them are found in standard dictionaries (Wong, 1941; LSHK, 1997), the remaining are mostly common alternative pronunciations or colloquial sounds (Lo et al., 1998). CUSYL is intended for syllable-based TTS synthesis (Chu and Ching, 1998; Lee et al., 1999). With the extended coverage, the corpus can be used to generate Cantonese speech in almost unrestricted domains.

Two female and two male speakers were asked to read the entire corpus once. The syllables were recorded with meaningful carrier words so as to reduce pronunciation ambiguities and avoid undesirable pre-pausal lengthening. Whenever possible, the target syllables were embedded in the middle of utterances. Each recorded utterance was edited manually to extract the target syllable. Fig. 4 gives several examples of carrier words and the embedded target syllables in CUSYL.

#### 3.3.2. CUWORD

CUWORD is a read speech corpus of polysyllabic words. It is intended for the training and performance evaluation of syllable and sub-syllable based speech recognition algorithms (Chow et al., 1998). A base corpus of 4055 words was first created to cover most of the Cantonese base

| | | (*English Translation*) |
|---|---|---|
| 一巴掌 | jat1-**baa1**-zoeng2 | a slap |
| 掃把星 | sou3-**baa2**-sing1 | comet |
| 強行霸佔 | koeng5-hang4-**baa3**-zim3 | taking over by brute force |
| . | | |

Fig. 4. Example section in the CUSYL corpus.

| | | (*English Translation*) |
|---|---|---|
| 止咳化痰 | **zi2-kat1-faa3-taam4** | relieving cough |
| 比薩斜塔 | **bei2-saat3-ce4-taap3** | The tower of Pisa |
| 水塔頂端 | **seoi2-taap3-deng2-dyun1** | top of the water tower |
| 火辣辣 | **fo2-laat6-laat6** | red hot |
| 主僕 | **zyu2-buk6** | master and slave |
| . | | |

Fig. 5. Examples of words in the CUWORD corpus.

syllables. Despite the substantial coverage, the distributions of different syllables in the base corpus were fairly disproportionate. To refine the content, a human-assisted selection process was applied to keep the syllables with low frequency of occurrences and to discard the extremely frequent ones. As a result, a total of 2527 words were kept and a partial list of them is shown in Fig. 5 (Lo et al., 1998). On average each word consists of 2.8 syllables.

The 2527 words cover 1388 tonal syllables (78.8% of the LSHK inventory) and 559 base syllables (89.4%). The frequency of occurrences for each base syllable ranges from 1 to 39, as shown in Fig. 6. A total of 28 speakers (13 male and 15 female) participated in the recording of CUWORD. Each of them uttered the entire corpus once.

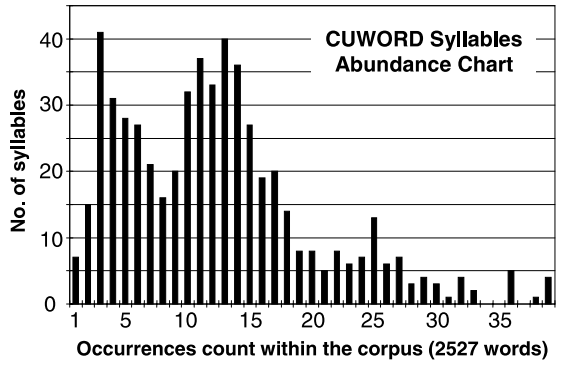


Fig. 6. Distributions of base syllables in the CUWORD corpus.

### 3.3.3. CUSENT

CUSENT is a read speech corpus of continuous Cantonese sentences. It is designed to be phonetically rich. All Cantonese *Initials*, *Finals* and tones are included and the syllable inventory is covered as much as possible. Moreover, much attention has been paid to the coverage of different intra-syllable (onset–nucleus) and inter-syllable (coda–onset) contexts. [2] This ensures an adequate range of linguistic data for context-dependent acoustic modeling using statistical techniques (Wong et al., 1999).

The creation of the sentence corpus was a *semi-automatic* process (Lo et al., 1998). First of all, a large inventory of Chinese sentences (containing over 98 million characters) was obtained from 5 local newspapers of Hong Kong over the period of March 1997–February 1998. The selection process started with a small number of manually selected sentences. As shown in Fig. 7, the corpus was

[2] Here "onset" refers to the phonetic unit that begins a syllable and "coda" refers to the one that ends a syllable. In many cases they correspond to the initial and final consonants in the syllable, respectively. In the case where the syllable begins or ends with a vowel segment, e.g. [aam1] or [hai6], the "onset" or the "coda" refers to that vowel segment.

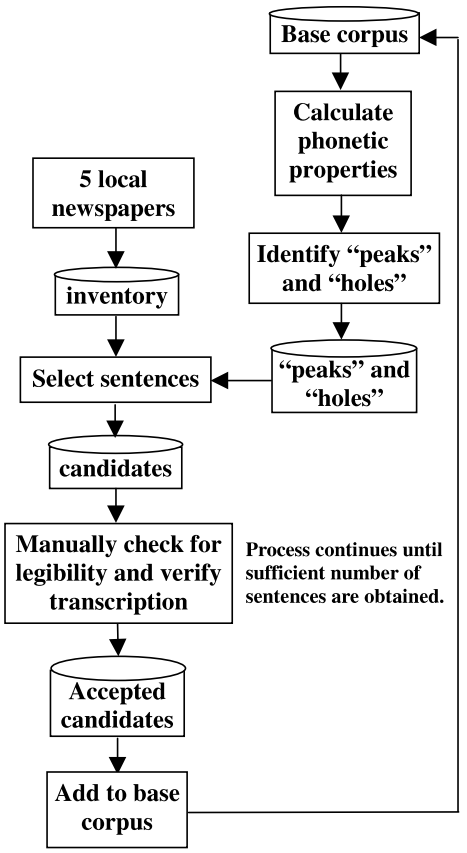


Fig. 7. The sentence selection process for CUSENT.

expanded by adding new sentences that led to improvement in the distributions of phonetic properties. There are about 2600 phonetic properties being considered, as shown in Table 5. Intra-syllable properties are subject to phonological constraints while inter-syllable ones are assumed unrestricted. Properties with extremely frequent occurrences in the current corpus were identified as "peaks" while those that rarely occurred were identified as "holes". A new sentence was said to improve the phonetic distribution and hereby marked as "candidate" if it filled in any "holes" but did not hit any "peaks". The marked candidates were manually verified for legibility and transcription correctness and those accepted were then added to the base corpus. Finally, the phonetic properties of the modified base corpus were re-calculated and a set of new "peaks" and "holes" were identified. This process was repeated until the corpus material comprised a pre-defined number of sentences. The overall process is outlined in Fig. 7.

It was observed that, at a later stage of selection, a number of survivable sentences might contain rarely used words. However, rarely used words or proper names were unnecessarily difficult to read. We relied much on subjective human judgement to discard "inappropriate" sentences, and eventually we could obtain a set of sentences with acceptable phonetic distribution.

As a result of this process, a base corpus of 5100 training sentences was created. Using the same approach, another 600 test sentences were selected separately for evaluation purposes. The training sentences were evenly divided into 17 groups, each containing 300 unique sentences. Each group of sentences were read by 4 speakers (2 male and

Table 5
Phonetic properties being considered in the design of CUSENT

| | Phonetic properties | Number |
|---|---|---|
| Syllabic components | *Initials* | 20 |
| | *Finals* | 53 |
| | Tones | 6 |
| Intra-syllable | *Initial–Final* combinations | 625 |
| | *Initial–Nucleus* combinations | 168 |
| | *Final–Tone* combinations | 303 |
| Inter-syllable | *Final–Initial* combinations | 1060 |
| | *Coda–Onset* combinations | 345 |
| | *Tone–Tone* combinations | 36 |
| | Total: | 2631 |

2 female). Thus, a total of 20,400 ($= 300 \times 4 \times 17$) training utterances were obtained from 68 ($= 17 \times 4$) speakers. The 600 test sentences were recorded from 6 male and 6 female speakers. Each speaker read 100 test sentences while each test sentence was read by 1 male and 1 female speaker. The total number of test utterances is therefore $600 \times 2 = 1200$. The populations of test speakers and training speakers are completely exclusive.

## 4. Post-processing and analysis of speech data

Post-processing of recorded speech is an important step in the compilation of a useful speech corpus. It involves either fully manual or semi-automatic verification and annotation for each utterance. The required level of annotation may differ greatly from one application to another. In this work, our goal is to support the most commonly used techniques in start-of-the-art speech technology and to minimize the additional efforts in applications development. For speech recognition purposes, the phonemic transcriptions of all utterances in the corpora were manually verified, whilst for the CUSYL corpus, pitch cycles were marked to facilitate pitch-synchronous modification of waveforms. In the following, we will briefly describe the annotation process for different corpora and then a statistical analysis that was performed on the annotated data.

### 4.1. CUSYL

Verification of CUSYL was done to check if the speakers had uttered the syllables correctly and accurately. By "accurately" we mean that the *Initial*, the *Final* and the lexical tone were all pronounced exactly as designated (Lo et al., 1998). Alternative pronunciations were not accepted in this case although they are quite common in daily conversation. Target syllables were excised manually from recorded waveforms. As a result, a total of 7204 monosyllabic utterances have been obtained from the 4 speakers.

Manual marking of pitch cycles was performed for the syllables from one male and one female speaker. For each utterance, the positions of pitch marks (in seconds) are stored orderly in a separate text file.

### 4.2. CUWORD and CUSENT

CUWORD and CUSENT are created to provide training and evaluation data for statistical acoustic modeling. The goal of annotation was to provide phonemic transcriptions of the actual recordings. This was carried out in two stages (Lo et al., 1998). In stage one, each recorded utterance was played-back to the human annotator. The annotator was instructed to indicate the incorrect, problematic or missing recordings as well as unclear or uncertain recordings, with respect to the original Chinese scripts. Commonly encountered errors include deleted or inserted characters, and alternative or colloquial pronunciations. In stage two, experts in Cantonese phonetics were asked to examine the problem utterances. They tried to resolve the problems by editing the respective scripts and/or the phonemic transcriptions according to what the speaker actually uttered. For a colloquial syllable that does not have a standard written counterpart, a Chinese character with the closest pronunciation was used in the orthographic transcription. In other words, the orthographic transcription is considered to be less accurate than the phonemic one. For CUWORD and CUSENT, respectively, about 11% and 26% of the utterances required manual editing of transcriptions.

If it was deemed impossible to transcribe the real recording or if a recording artifact was observed, the utterance would be discarded. For CUWORD and CUSENT, the percentage of discarded utterances were 0.7% and 0.1%, respectively.

As a result, each legitimate utterance is accompanied by a Chinese orthographic transcription (in BIG5 [3] code) and a verified phonemic transcription (in the LSHK scheme).

Table 6 summarizes the content and phonetic coverage of the annotated CUWORD corpus. It

[3] BIG5 is an encoding system for traditional Chinese characters. In this system, each character is represented as a 2-byte data. BIG5 is most commonly used in Taiwan and Hong Kong.

Table 6
A summary of the content and syllable coverage of CUWORD

| | |
|---|---|
| No. of utterances | 70,277 |
| No. of syllable occurrences | 193,759 |
| No. of tonal syllables being covered | 1666 |
| No. of base syllables being covered | 620 |

can be seen that CUWORD has a rather extensive coverage of syllables. Although the basic corpus (2527 words) contains 559 base syllables, the annotated real recordings give a larger coverage of 620. This is due to the presence of some alternative pronunciations in the actual recordings. From speech recognition points of view, we are interested in not only how many units have non-zero frequency of occurrence but also how many of them have enough occurrences to facilitate statistical modeling. Thus, for a particular type of speech units, we define

$$N^{+}(\mathscr{F}) = \text{No. of units that occur more than } \mathscr{F} \text{ times in the corpus.}$$

Fig. 8 gives the plot of $N^{+}_{\mathrm{BS}}(\mathscr{F})$ to illustrate the distribution of base syllables in CUWORD. Nearly 400 base syllables (64% of LSHK inventory) occur more than 200 times in CUWORD, and over 100 of them have a minimum frequency of about 600. Syllable-based acoustic modeling is therefore possible for many small and medium vocabulary applications.

CUSENT was intended for context-dependent acoustic modeling at sub-syllable level. Thus we

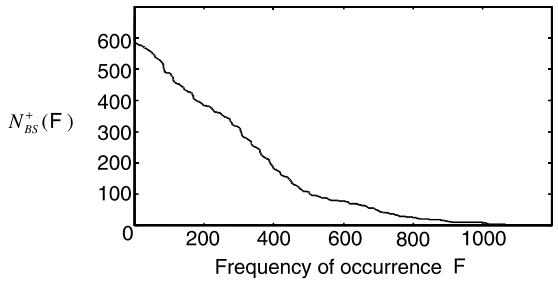


Fig. 8. Distribution of base syllables in the annotated CUWORD.

evaluate how well different phonetic units and contexts are being covered and distributed in the training and test data. The training data consist of over 20,000 utterances with average length of 11 syllables. The phonetic coverage of the annotated CUSENT corpus is summarized in Table 7. For the intra-syllable context, most (over 95% on average) of all legitimate *Initial–Final*, *Initial–Nucleus* and *Final–Tone* combinations are being covered in the training data. The training data also covers 97%, 94% and 100% of the inter-syllable *Final–Initial*, *Coda–Onset* and *Tone–Tone* combinations, respectively. The coverage of test data is not as good as that of the training data simply because the number of utterances is much smaller.

In Fig. 9, the distributions of context-independent *Initials* and *Finals* are illustrated by the plots of $N^{+}_{\mathrm{I}}(\mathscr{F})$ and $N^{+}_{\mathrm{F}}(\mathscr{F})$, respectively, and the exact frequency count for each of the six tones is given. The average frequencies of occurrence in the

Table 7
The content and phonetic coverage of CUSENT

| | | Training data | Test data |
|---|---|---|---|
| No. of utterances | | 20,378 | 1198 |
| No. of syllable occurrences | | 215,560 | 11,663 |
| No. of tonal syllables being covered | | 1584 | 972 |
| No. of covered Intra-syllable contextual units | *Initial–Final* combinations | 631[a] | 487 |
| | *Initial–Nucleus* combinations | 166 (99%) | 153 (91%) |
| | *Final–Tone* combinations | 261 (86%) | 234 (77%) |
| No. of covered Inter-syllable contextual units | *Final–Initial* combinations | 1033 (97%) | 826 (78%) |
| | *Coda–Onset* combinations | 323 (94%) | 262 (76%) |
| | *Tone–Tone* combinations | 36 (100%) | 36 (100%) |

[a] The manually transcribed CUSENT contains a few base syllables that were not covered in (LSHK, 1997). Most of them are due to pronunciation variation.

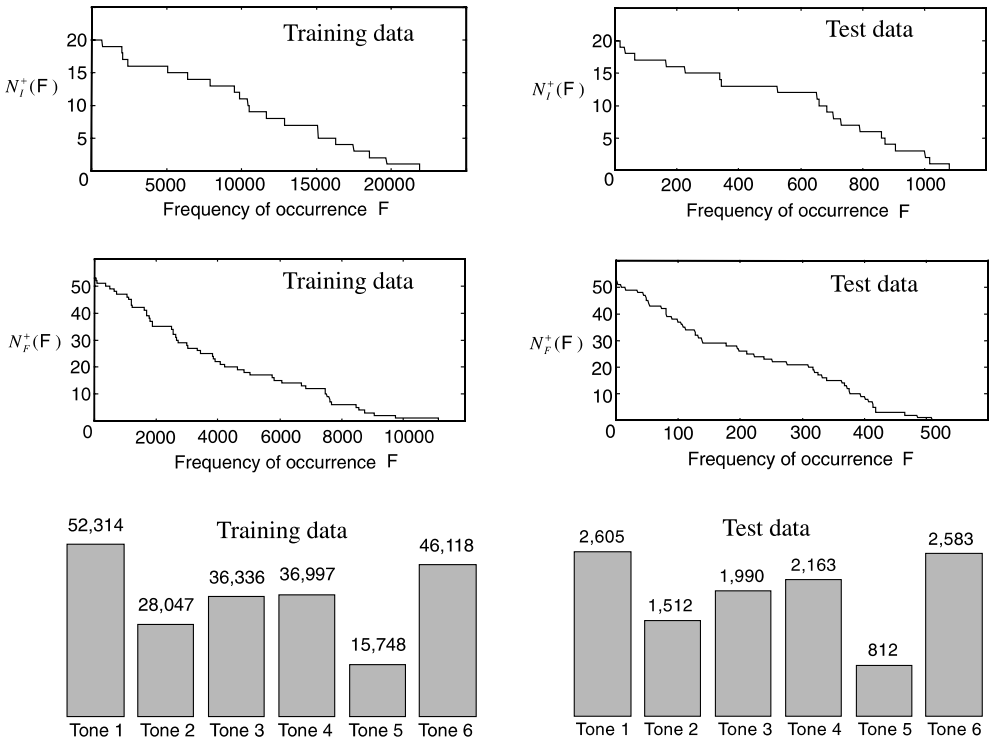


Fig. 9. Distribution of context-independent sub-syllable units in the annotated CUSENT.

training data are 10,000, 4000 and 36,000 for *Initials*, *Finals* and tones, respectively. Figs. 10 and 11 depict the distributions of various contextual units in the training and test data. For the intra-syllable *Initial–Final*, *Initial–Nucleus* and *Final–Tone* combinations, the average frequencies of occurrence in the training data are 321, 1268 and 807, respectively. For the inter-syllable contexts, the average frequencies of occurrence for *Final–Initial*, *Coda–Onset* and *Tone–Tone* combinations in training data are 184, 565 and 5421, respectively. Although the test data contains much less utterances than the training data, the distributions of phonetic units are very similar.

The frequencies of occurrence of individual units are far from even in both CUWORD and CUSENT. This is not beyond our expectation, considering that the composition of a Cantonese sentence is subject to a number of linguistic constraints and properties. For example, there are 101 (out of 1060) inter-syllable *Final–Initial* combinations that occur less than 10 times in the CUSENT training data. They mostly involve several rarely used *Finals* or *Initials*, namely [oe], [ot] and [kw]. The total frequency counts (context-independent) of these three units are 92, 60 and 729, respectively, which are well below the average.

### 4.3. CUDIGIT and CUCMD

The verification process of CUDIGIT also consisted of two steps. In the first stage, a trained annotator corrected obvious errors and indicated the remaining problematic data. Then in stage two, another (more experienced) annotator decided whether the problematic data should be kept or abandoned. If necessary, this annotator would also edit the waveform and cut out any undesirable noise segments. In this way, we maintained a high accuracy in annotation. Each utterance in CUDIGIT is also accompanied by a Chinese orthographic transcription and a phonemic transcription.

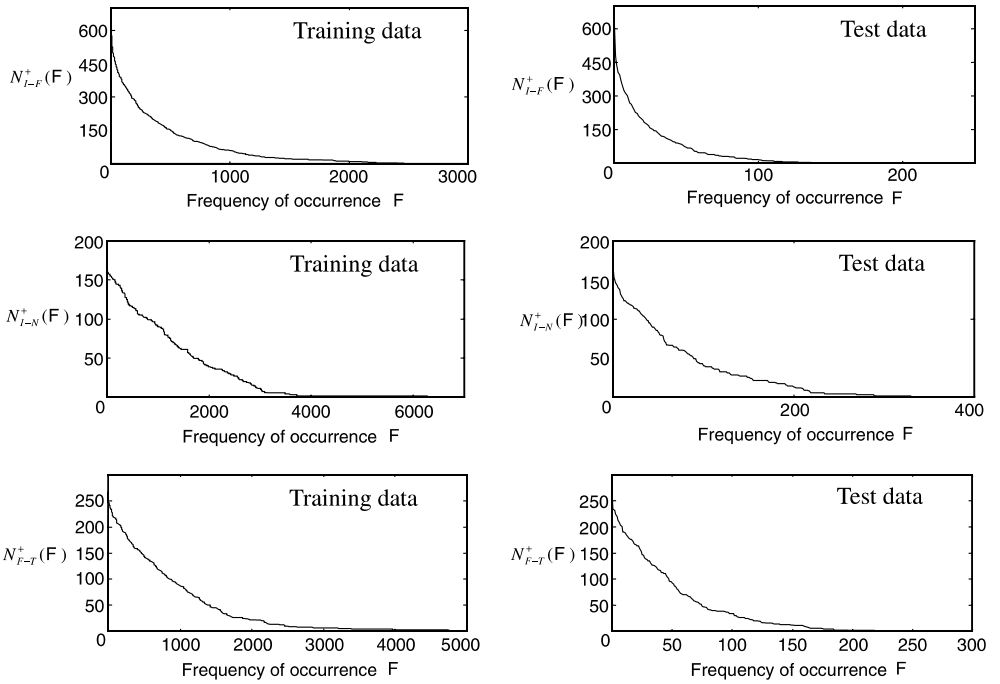


Fig. 10. Distribution of intra-syllable contextual units in the annotated CUSENT.

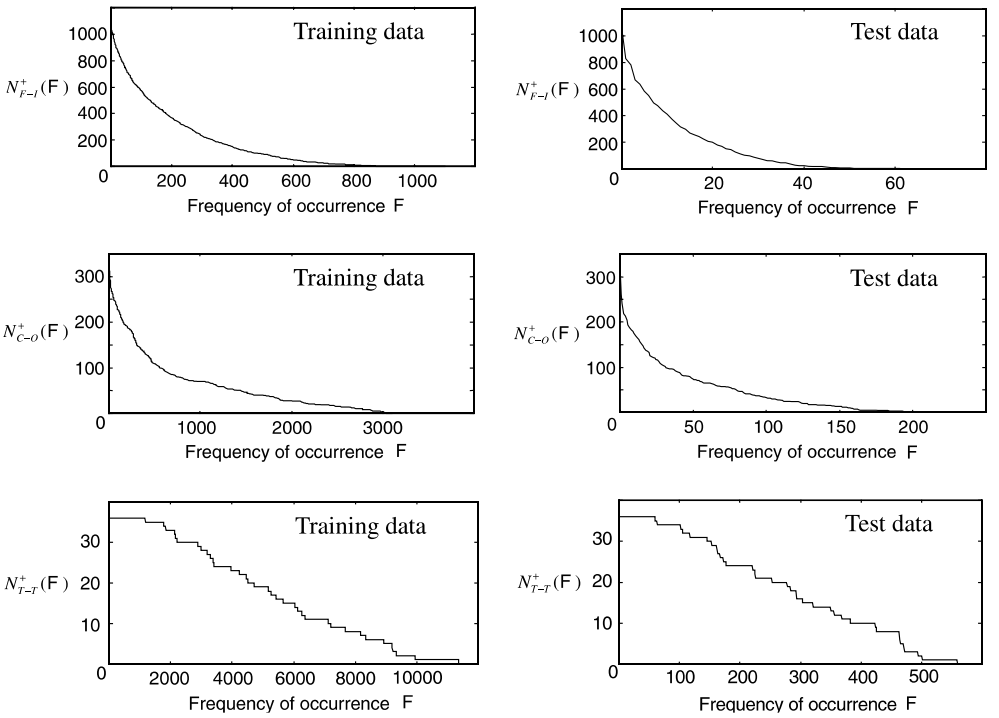


Fig. 11. Distribution of inter-syllable contextual units in the annotated CUSENT.

Table 8
Corpus code for the *CU Corpora*

| Corpus | Corpus code | No. of CD-ROMs |
|---|---|---|
| CUSYL | CS | 1 |
| CUWORD | CW | 5 |
| CUDIGIT | CD | 2 |
| CUCMD | CC | 1 |
| CUSENT | CN | 3 |

The content of CUDIGIT is rich enough for acoustic modeling for the vocabulary of digits. For each of the ten digits, the number of context-independent occurrences is around 10,000, which is suitable for acoustic modeling with statistical approaches.

Since CUCMD contains mostly short commands, the verification process was rather simple. The annotator listened to each utterance and thereby decided whether to keep or discard it.

### 4.4. Data organization and storage

As shown in Table 8, a specific corpus code is assigned to each of the five corpora described above. The corpora are stored separately and written onto CD-ROMs for distribution. Within each corpus, data from the same speaker are placed under the same directory. The directory name is composed of the corpus code, speaker code (00–FF) and gender (F or M). For example, the directory "CN3FF" holds the acoustic and annotation data for female speaker "3F" of CUSENT. Acoustic data and annotation data are in the sub-directories "DATA" and "ANNOTATE", respectively.

## 5. Preliminary evaluation

In this section, a series of preliminary experiments are described to show how the Cantonese speech corpora can be effectively used to develop various applications. The experimental results being reported can be taken as a reference for forthcoming research and development that are based on the *CU Corpora*. Some limitations of the corpora are also discussed.

### 5.1. CUSYL: TTS synthesis using TD-PSOLA

Time-domain pitch synchronous overlap add (TD-PSOLA) has been widely applied in TTS synthesis for many languages (Moulines et al., 1990; Bigorgne et al., 1993). It generates synthetic speech by concatenating pre-recorded sound segments. The original waveforms are modified in the time domain to match the prescribed duration and F0 targets. For Chinese languages like Putonghua and Cantonese, waveform modification is often applied at the syllable level (Chu and Lu, 1996; Chou et al., 1997). The CUSYL corpus is particularly suitable for TD-PSOLA based Cantonese TTS for the following reasons

1. It has an extensive coverage. The inclusion of colloquial and alternative pronunciations makes it possible to generate Cantonese as spoken in real life.
2. The tonal variations of a Cantonese base syllable, which exhibit distinct F0 patterns, are treated as different items in CUSYL. To match a particular F0 target, we can always find a template that requires the minimal degree of F0 modification.

A Cantonese TTS system has been successfully developed using the CUSYL corpus (Chu and Ching, 1998; Lee et al., 1999). Together with a linguistic analysis front-end, the system can automatically convert Chinese text into either male or female speech with fairly good quality. However, as CUSYL contains only monosyllables, it does not facilitate the modeling of across-syllable juncture.

### 5.2. CUDIGIT: connected-digit recognition

A connected-digit recognizer has been trained with the CUDIGIT data. For each of the ten digits, a context-independent HMM with 6 emitting states was trained. About 80% of the utterances in CUDIGIT were used for training and the remaining 20% were used for performance evaluation. As shown in Table 9, the digit recognition accuracy is 91% with 4 Gaussian mixtures per state and the string accuracy is about 75%. One of the major sources of errors was due to frequent

Table 9
Performance of the connected-digit recognizer trained with CUDIGIT

| Number of Gaussian mixtures per state | Digit accuracy | String accuracy |
|---|---|---|
| 1 | 86.0% | 64.1% |
| 2 | 89.5% | 70.6% |
| 4 | 91.0% | 74.8% |

insertions of the digit "5". This digit is pronounced as a syllabic nasal [ng], which may be confused with and treated as part of the nasal coda in the digits "0" ([ling]) or "3" ([saam]). Special acoustic modeling techniques need to be developed to deal with this problem and improve the overall performance. Putonghua connected-digit recognition does not suffer from such a problematic digit and higher level of recognition performance has been attained (Chen and Soong, 1994; Zhang et al., 1999).

### 5.3. CUWORD: connected-syllable recognition

The CUWORD corpus has been evaluated in the task of connected-syllable recognition. Out of the 620 base syllables being covered, we selected 573 that occur at least 30 times and a context-independent HMM was trained for each of them. The number of states in the HMMs ranged from 3 to 8, depending on the phonetic composition of the syllables. The training data contained 54,842 polysyllabic utterances from 22 speakers (10 male and 12 female) while 14,896 utterances from the remaining 6 speakers (3 male and 3 female) were used for testing. As shown in Table 10, with 4 Gaussian mixtures for each emitting state, the base syllable accuracy was 69.0% (Chow et al., 1998).

Table 10
Performance of the connected-syllable recognizer trained with CUWORD

| Number of Gaussian mixtures per state | Syllable accuracy |
|---|---|
| 1 | 57.0% |
| 2 | 65.8% |
| 4 | 69.0% |

Provided that sufficient training tokens are available for each syllable, syllable-level acoustic modeling is desirable for Chinese languages, simply because the intra-syllable contextual effect is modeled explicitly. Although CUWORD was intended for syllable-level modeling, it seems not to be suitable for large-vocabulary applications. As shown in Fig. 8, for many base syllables, the number of training tokens is less than 100. Thus poor modeling of these syllables is anticipated.

### 5.4. CUSENT: continuous speech recognition

CUSENT is useful for large vocabulary continuous speech recognition (LVCSR) of read-style Cantonese speech. As shown in Tables 5 and 7, the design of CUSENT considered mainly the *bi-phone* type of contexts of *Initials* and *Finals*. Thus, in contrast to the syllable-level modeling using CUWORD, right-context-dependent *Initial* and *Final* (RCD-IF) models were trained with CUSENT. Each HMM had either 3 or 5 emitting states, depending on the phonetic composition of the unit. From the distribution plots in Figs. 10 and 11, it can be seen that a certain percentage of the contextual units do not have sufficient training data and a couple of them are even totally unseen in CUSENT. Thus, decision-tree based state clustering was applied to facilitate effective parameter sharing among different units (Young and Woodland, 1993). As a result, the 3025 RCD-IF models were built using about 1300 distinct states. The attained accuracy for base syllable recognition is shown in Table 11.

The design of CUSENT did not emphasize on the coverage of *triphone* contexts. The database is not a good choice for *tri-IF* modeling, i.e., mod-

Table 11
Performance of the continuous speech recognizer trained with CUSENT

| Number of Gaussian mixtures per state | Syllable accuracy |
|---|---|
| 4 | 71.7% |
| 8 | 73.1% |

eling of *Initials* and *Finals* with both left and right contexts being considered (Gao et al., 2000).

## 6. Summary and conclusions

In this paper, the entire development process for a series of large-scale Cantonese spoken language databases for speech processing has been described. These speech corpora are intended to support both application-specific as well as application-independent speech technologies, including recognition and synthesis. The application-specific corpora include commands (CUCMD) and digits (CUDIGIT). The general-purpose corpora, which are characterized by their rich phonetic contents, include syllables (CUSYL), words (CUWORD) and sentences (CUSENT). All recordings were made in a controlled environment, and underwent an elaborate annotation and verification process. This paper provides detailed statistics of the recorded data, which is important information for the design of a speech recognition/synthesis system. We have also reported the use of the corpora in the development of a TD-PSOLA TTS synthesis system using CUSYL, a connected syllable recognizer using CUDIGIT, a connected-syllable recognizer using CUWORD, and a large vocabulary continuous speech recognizer using CUSENT. It is expected that the CU Corpora will make a significant infra-structural contribution to the advancement of multi-lingual spoken language technologies.

CU Corpora are available for both research and commercial use. Detailed documentation and licensing procedures can be found at http://dsp.ee.cuhk.edu.hk/speech.

## Acknowledgements

This Cantonese database project is partially supported by the Industrial Support Fund awarded by the Industry Department, Hong Kong SAR Government, and also by a research grant from the Hong Kong Research Grants Council. The authors wish to thank the anonymous reviewers for their attentive reading and useful suggestions that help improving the quality of this paper.

## References

Bigorgne, D., Boeffard, O., Cherbonnel, B., Emerard, F., Larreur, D., Le Saint-Milon, J.L., Metayer, I., Sorin, C., White, S., 1993. Multi-lingual PSOLA text-to-speech system. In: Proceedings of 1993 International Conference Acoustics Speech and Signal Processing, Vol. 2, pp. 187–190.

CCDICT: Dictionary of Chinese Characters, Version 3.0, http://www.chinalanguage.com/CCDICT/, March 2000.

Chan, C., 1998. Design considerations of a Putonghua database for speech recognition. In: Proceedings of the Conference on Phonetics of the Language in China, pp. 13–16.

Chen, J.-K., Soong, F.K., 1994. An N-best candidates discriminative training for speech recognition applications. IEEE Transactions on Speech and Audio Processing 2 (1(II)), 206–216.

Ching, P.C., Lee, T., Zee, E., 1994. From phonology and acoustic properties to automatic recognition of Cantonese. In: Proceedings of 1994 International Symposium on Speech, Image Processing and Neural Networks, Vol. 1, pp. 127–132.

Chou, F.-C., Tseng, C.-Y., Chen, K.-J., Lee, L.-S., 1997. A Chinese text-to-speech system based on part-of-speech analysis, prosodic modeling and non-uniform units. In: Proceedings of 1997 International Conference on Acoustics, Speech and Signal Processing, Vol. 2, pp. 923–926.

Choukri, K., 1999. European Language Resources Association – history and recent developments. In: Proceedings of 1999 Oriental COCOSDA Workshop (invited paper).

Chow, K.F., Lee, T., Ching, P.C., 1998. Sub-syllable acoustic modelling for Cantonese speech recognition. In: Proceedings of 1998 International Symposium on Chinese Spoken Language Processing, pp. 75–79.

Chu, M., Ching, P.C., 1998. A hybrid approach to synthesize high quality Cantonese speech. In: Proceedings of 1998 International Conference Acoustics, Speech and Signal Processing, Vol. 1, pp. 277–280.

Chu, M., Lu, S., 1996. A text-to-speech system with high intelligibility and high naturalness for Chinese. Chinese Journal of Acoustics 15 (1), 81–90.

European Language Resources Association (ELRA), 2000. http://www.icp.grenet.fr/ELRA/.

Gao, S., Lee, T., Wong, Y.W., Xu, B., Ching, P.C., Huang, T., 2000. Acoustic modeling for Chinese speech recognition: a comparative study of Mandarin and Cantonese. In: Proceedings of the 2000 International Conference on Acoustics, Speech and Signal Processing, Istanbul, Turkey, May 2000, Vol. 3, pp. 1261–1264.

Hashimoto, O.-K.Y., 1972. Studies in Yue Dialects 1: Phonology of Cantonese. Cambridge University Press, Cambridge.

Huo, Q., Ma, B., 1999. Training material considerations for task-independent subword modeling: design and other

possibilities. In: Proceedings of 1999 Oriental COCOSDA Workshop, pp. 85–88.

Höge, H., Tropf, H.S., Winski, R., van den Heuvel, H., Haeb-Umbach, R., Choukri, K., 1997. European speech databases for telephone applications. In: Proceedings of 1997 International Conference on Acoustics, Speech and Signal Processing, Vol. 3, pp. 1771–1774.

Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., Shikano, K., 1990. ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Communication 9, 357–363.

Kuwabara, H., Takeda, K., Sagisaka, Y., Katagiri S., Morikawa, S., Watanabe, T., 1989. Construction of a large-scale Japanese speech database and its management system. In: Proceedings of 1989 International Conference on Acoustics Speech and Signal Processing, Vol. 1, pp. 560–563.

Lamel, L., Kassel, R., Seneff, S., 1986. Speech database development: design and analysis of the acoustic-phonetic corpus. In: Proceedings of DARPA Speech Recognition Workshop, pp. 100–109.

Lee, T., Ching, P.C., 1999. Cantonese syllable recognition using neural networks. IEEE Transactions on Speech and Audio Processing 7 (4), 466–472.

Lee, T., Ching, P.C., Chan, L.W., Mak, B., Cheng, Y.H., 1995. Tone recognition of isolated Cantonese syllables. IEEE Transactions on Speech and Audio Processing 3 (3), 204–209.

Lee, T., Meng, H.M., Lau, W., Lo, W.K., Ching, P.C., 1999. Micro-prosodic control in Cantonese text-to-speech synthesis. In: Proceedings of the Sixth European Conference on Speech Communication and Technology, Vol. 4, pp. 1855–1858.

Lo, W.K., Lee, T., Ching, P.C., 1998. Development of Cantonese spoken language corpora for speech applications. In: Proceedings of 1998 International Symposium on Chinese Spoken Language Processing, pp. 102–107.

Linguistic Data Consortium (LDC), 2000. Various resources on http://www.ldc.upenn.edu.

Linguistic Society of Hong Kong (LSHK), 1997. Hong Kong Jyut Ping Characters Table, (粵語拼音字表) Linguistic Society of Hong Kong Press (香港語言學會出版).

Matthews, S., Yip, V., 1994. Cantonese: A Comprehensive Grammar, Routledge Grammars.

Moulines, E., Emerard, F., Larreur, D., Le Saint Milon, J.L., Le Faucheur, L., Marty, F., Charpentier, F., Sorin, C., 1990. A real-time French text-to-speech system generating high-quality synthetic speech. In: Proceedings of 1990 International Conference on Acoustics, Speech and Signal Processing, Vol. 1, pp. 309–312.

Ohtsuki, K., Matsuoka, T., Mori, T., Yoshida, K., Taguchi, Y., Furui, S., Shirai, K., 1999. Japanese large-vocabulary continuous speech recognition using a newspaper corpus and broadcast news. Speech Communication 28, 155–166.

Paul, D., Baker, J., 1992. The design of the Wall Street Journal-based CSR corpus. In: Proceedings of the Fifth DARPA Speech and Natural Language Workshop, Morgan Kaufmann.

Price, P., 1990. Evaluation of spoken language systems: The ATIS domain. In: Proceedings of the Third DARPA Speech and Natural Language Workshop, Morgan Kaufmann.

Price, P., Fisher, W.M., Bernstein, J., Pallett, D.S., 1988. The DARPA 1000-word resource management database for continuous speech recognition. In: Proceedings of 1988 International Conference on Acoustics, Speech and Signal Processing, pp. 651–654.

Tseng, C.-Y., 1995. A phonetically oriented speech database for Mandarin Chinese. In: Proceedings of 1995 International Congress of Phonetic Sciences, Vol. 3, pp. 326–329.

Wang, H.-C., 1999. Speech research infra-structure in Taiwan – from database design to performance assessment. In: Proceedings of 1999 Oriental COCOSDA Workshop, pp. 53–56.

Wang, R., Xia, D., Ni, J., Liu, B., 1996. USTC95 – a Putonghua Corpus. In: Proceedings of 1996 International Conference on Spoken Language Processing, Vol. 3, pp. 1894–1897.

Winski, R., Fourcin, A., 1994. A common European approach to assessment, corpora and standards. In: Varghese, K., Pfleger, S., Lefvre, J.-P. (Eds.), Advanced Speech Applications: European Research on Speech Technology. Springer, Berlin, pp. 25–79.

Wong, S.L., 1941. A Chinese Syllabary Pronounced According to the Dialect of Canton. Chung Hwa Book Co, Hong Kong (see electronic version at http://www.arts.cuhk.edu.hk/Lexis/Canton).

Wong, Y.W., Chow, K.F., Lau, W., Lo, W.K., Lee, T., Ching, P.C., 1999. Acoustic modeling and language modeling for Cantonese LVCSR. In: Proceedings of the Sixth European Conference on Speech Communication and Technology, Vol. 3, pp. 1091–1094.

Wu, Y., 1998. Chili Mandarin speech corpus. In: Newsletter of ISCSLP'98 Special Interest Group: Linguistic Database and Tools, pp. 1–3.

Young, S.J., Woodland, P.C., 1993. The use of state tying in continuous speech recognition. In: Proceedings of the Third European Conference on Speech Communication and Technology, Vol. 3, pp. 2203–2206.

Yuan, J., 1983. Hanyu Fangyan Gaiyao (A Precise of the Chinese Dialects). Wenzhi Gaige Chubanshe, Beijing (in Chinese).

Zhang, B., Liu, J.P.G., Wang, W.S.-Y., 1999. A higher performance Mandarin digit recognizer. In: Proceedings of the Fifth International Symposium on Signal Processing and Its Applications, Brisbane, Australia, August, pp. 629–632.

Zhang, J., 1998. Notes on speech corpora of standard Chinese in China. In: Newsletter of ISCSLP'98 Special Interest Group: Linguistic Database and Tools, pp. 4–5.

Zu, Y.Q., Li, W.X., Ho, M.C., Chan, C., 1996. HKU96 – a Putonghua corpus (CD-ROM version). Department of Computer Science, University of Hong Kong.

Zue, V., Seneff, S., Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech Communication 9, 351–356.