USTC95 --- A PUTONGHUA CORPUS

Ren-Hua Wang , Deyu Xia, Jinfu Ni, Bicheng Liu

University of Science & Technology of China
P.O.Box 4, Hefei, 230027 P.R.CHINA

ABSTRACT

A large Putonghua corpus is introduced, which is primarily designed to support research in Chinese speech recognition, analysis and recognition system evaluation. This corpus consists of four major sub-corpora corresponding to isolated syllables, multi-syllable words, sentences, and telephone speech. With an elaborate design, the corpus encompasses all the phones and mono-syllables, as well as the co-articulation effects in the Putonghua; besides, keeps as little redundancy as possible. This parsimonious corpus makes it possible to acquire acoustic-phonetic knowledge for isolated words recognition and continuous Chinese recognition, to provide speech data for training telephone speech recognizer, also to provide a common test base for the performance assessment of recognizer.

1. INTRODUCTION

The development of a comprehensive, standardized speech corpus is of great importance for speech processing, especially for speech recognition. Such a corpus is needed in order to acquire acoustic-phonetic knowledge, train and test speech processing algorithms, design robust recognition systems, meanwhile, common speech corpora and standardized performance assessment based on the corpora are the key parts in conducting benchmark tests to evaluate system performance and track the development of speech technology. In general, the recognition techniques seem to be language independent, but any recognition system cannot avoid, to some degree, having some strategies specially designed for the target language. So that many advanced countries in the world have been constructing their own large speech corpus suitable to mother tongue, such as corpus TIMIT, ATIS in USA[1], ATR, JEIDA in Japan[2], EUROM-0,EUROM-1 in EC etc. Chinese is one of languages spoken by most peoples in the world, due to involving vast territory and large population , even for the standard spoken Chinese commonly known as Putonghua or Mandarin, strong dialect accents are existed among the speakers from different regions. Therefore, it is particularly important to establishing a standard Putonghua corpus with a large number of speakers , various accents, age groups and sexes.

Considering the urgent need, a large Putonghua corpus is under construction with tremendous supports of national projects at the Speech Communication Laboratory of University of Science and Technology of China[3]. This corpus is primarily designed to support research in Chinese speech recognition, analysis and recognition system evaluation, Over the past few years, several sub-corpora corresponding to isolated syllables, multi-syllable words, connected digits, sentences and telephone speech have been built and made available to the general public on request. Based on the corpus, a national performance assessment of isolated word recognition system and continuous speech recognition system for Chinese is carried out under the auspices of the national 863 Hi-Tech project every year since 1991[4]. On the corpus of international status, researchers of spoken Chinese recognizers are qualified to exchange and compete with their colleagues in the world. It turns out to promote the research and development of speech recognition greatly in China. The corpus is described in more detail below.

2. CORPUS MATERIAL DESCRIPTION

2.1 Isolated Syllable Sub-corpus

Mandarin Chinese is a tonal language and there are basically four lexical tones and one neutral tone. Each Chinese syllable is characterized by three parts: the initial (Shengmu), the final (Yunmu) and the tone. The initial of a syllable can be one of the 21 consonants or empty. The final of a syllable is composed of a medial, a kernel vowel and a coda, whereas the medial and the coda can be empty. In Chinese Putonghua, 10 vowels can be as the kernel, only 3 vowels as the medial, two vowels and two nasals as the coda. Among them, some phone combinations are inhibited, thus there are totally 39 valid finals. Consequently, combining 22 initials (including empty initial) and 39 finals together can produce only 406 valid base-syllables disregarding the tone difference. Basically, Chinese syllables can be articulated in five different tones. Some syllable-tone combinations may not actually appear in speech and correspond to no Chinese characters. Thus there are totally about 1300 valid-toned syllables in Mandarin[5] .

Capturing co-articulation in connected speech is certainly an important aspect of speech research in any language. However, capturing all the possible combinations of the speech units of a language is highly impractical because of the sheer size of such a corpus. Furthermore, the Chinese language employs mono-syllabic characters as building units hence it is not unacceptable to speak in isolated syllables although it is

unnatural. For that reason, including all the isolated syllables in the corpus has a special significance. First, it eases the problem of phonetic labeling which is important in the training phase of a speech recognizer. Secondly, there is still a large number of researchers interested in recognizing isolated Chinese syllables because of its relative simplicity to construct such systems. Thirdly, a sub-corpus of isolated syllables simplifies the derivation of phonetic information which is the jumping board to boot-strap the process of capturing co-articulation effects in connected speech.

Based on the above considerations, as a first step, we include in this isolated syllable sub-corpus 1264 valid-toned syllables. The frequency count of the initials and the finals in this sub-corpus is tabulated in reference [7].

2.2 Multi-syllable Words Sub-corpus

A compromise between isolated syllable and continuous speech is to speak in isolated words. There are at least 10^5 commonly used Chinese words, each is composed of one to several characters. There are at least 10^4 commonly used Chinese characters, all produced as mono-syllables. Two-syllable words account for 73.4% of the Chinese word vocabulary. This multi-syllable words sub-corpus consists of 3 groups, *viz.*, two-, three-, four-syllable words respectively. The objective of this sub-corpus is to reflect the co-articulation effects due to different syllable combinations and rhythm combinations. At the same time, the balance of the initials, the finals and the base syllables is also considered. On the one hand, one wants the sub-corpus to capture as much co-articulation effects as one can expect in the language. On the other hand, one also wants as little redundancy as possible in order to make the corpus concise and manageable. In particular, one wants the corpus flexible enough so that one can have it tailored to one's purpose.

The effects of co-articulation reflected by the syllable transitions can be described as different level of junctures in Putonghua[6], and the junctures are mainly divided into syllabic juncture and rhythmic juncture depending on the individual prosodic circumstance in the multi-syllable words. In order to achieve the above goal , we sort out 434 kinds of junctures in the classified combinations of Shengmu and Yunmu first, which are served as the most concise subset instead of all kinds of syllable transitions. Then combining with the grammatical structures we propose a frame, which offers the essential prosodic circumstances, for words selecting. Based on the frame patterns the required words are selected automatically from abundant words of candidate with the aid of computer. The more details can be found in the reference [7] . Following is a brief introduction by taking the three-syllable word subset as example. Since the tri-syllabic combinations are usually constructed by a sequence of mono-syllables (MS) followed by a bi-syllabic combination (BS), or vice versa, or sometimes by three MS in coordinate form, the various grammatical structures can be classified into 3 types : BS+MS, MS+BS and MS+MS+MS (8 types for four-syllable words) [8]. The juncture between syllables in the BS is defined as syllabic juncture (SJ), others are defined as rhythmic junctures (RJ). Figure 1 shows the corresponding frame patterns, where F stands for the Final and I stands for the Initial. Another important cue is tone combination, with 64 kinds of tone combinations (256 for four-syllable words) the frame turns into a 64x3 matrix, each element in the matrix define one kind of three-syllable words, which word should be taken depending on the statistical balances of such factors as the Initials , the finals, junctures, and tone combinations etc. Through an iterative optimization, 434 two-syllable words, 534 three-syllable words and 994 four-syllable words are selected from a large vocabulary of more than 50000 words, they constitute the multi-syllable words sub-corpus. The frequency count of the syllable transition classes in such designed sub-corpus is roughly tabulated in table 1, where the first column lists the last phones of previous syllables, and the first row lists the first phones of following syllables. The result shows that this sub-corpus represents a parsimonious subset of the multi-syllable words to capture the possible co-articulation effects in Mandarin Chinese.

2.3 Sentences Sub-corpus

Continuous speech recognition is the most important area of speech research. Although it is inherently more difficult than isolated-word recognition, considerable efforts have been made in recent years in China. Especially the voice dictation has been the main objective of speech recognition research in China, due to it's special significance for input of Chinese text to computer. The valuable voice dictation is required to be working on speaker independent, large vocabulary, continuous speech recognition. However it is impossible to make the system successful without the support of a large continuous speech corpus, which cover the basic acoustic-phonetic knowledge.

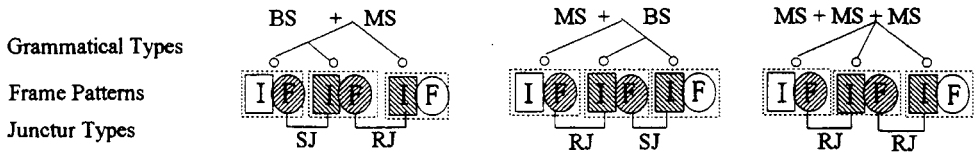

Figure 1 A sketch map showing the frame patterns for collecting three-syllable words.

Table 1 The frequency count of the syllable transition classes in mutil-syllable words

	b	f	m	p	d	l	n	t	g	h	k	j	q	x	r	zh	ch	sh	c	z	s	yi-	w-	yu-	er	*
-n/-ng	54	41	45	34	80	62	29	43	62	60	25	90	36	66	25	73	55	79	26	35	49	62	35	34	22	34
-i	55	35	47	37	55	76	33	42	54	60	29	57	49	60	28	52	47	57	21	37	32	60	38	35	29	41
-o/-u	42	33	35	27	46	63	26	45	39	40	30	64	39	52	28	46	30	56	25	41	32	41	32	28	28	32
-e/er	21	16	14	10	19	27	11	12	19	20	12	28	20	18	9	14	13	25	12	14	13	17	14	14	13	18
-u	9	7	5	4	7	9	4	5	8	9	4	7	5	8	6	5	9	6	6	6	4	9	5	7	3	6
-a	16	11	14	9	20	27	12	13	16	17	13	22	14	17	5	9	7	19	8	17	10	20	7	9	17	11
er	9	5	5	1	5	10	4	7	6	4	3	9	7	4	2	4	5	5	3	6	4	4	3	6	1	6

* The other syllables with initial empty as \a\,\e\,\ai\,\an\,\ao\,en,\ou\

This sentences sub-corpus is basically designed to serve continuous speech recognition. It is composed of 1000 sentences, which are elaborately selected out from the People's Daily, the most authoritarian daily paper in China. These sentences , lengths of two characters (syllables) to above 30 characters, cover basic sentence patterns, various tones and moods, prosodic features, and context relationships in the standard read Chinese, meanwhile keep the balance of phonetics. The data are being collected from more than 200 subjects with various accents, age groups and sexes. Since this sub-corpus is also designed for speech analysis, such as analysis of sentence intonation, pitch contour, and rules of rhyming etc. a few professional speakers are asked to read the materials more stable. .

2.4 Telephone Speech Sub-corpus

One of most valuable applications of speech recognition is a service over the telephone network. Because of the variations of telephone speech, such as among channel characteristics, handsets, speakers etc. a vast amount of telephone speech data is highly necessary for training of telephone speech recognition system. As a first step, the Sub-corpus is designed to include 12 isolated digits, several words of money units and control commands. Since a recognizer of connected digit strings will find many applications on the telephone service, this sub-corpus is also designed to include all the possible digit combinations in as few strings as possible. Each string cannot be too long otherwise the speakers will find them difficult to read naturally. We select a total of 37 such strings of 4 to 7 digits each. In order to cover the variations as far as possible, the data are being collected from 200 male and 100 female subjects living in the three major cities of China.

3. DATA COLLECTION

Different strategies of data collection are used for different sub-corpus. For the isolated syllables 12 subjects, 7 males and 5 females, are employed to read the texts mentioned above, different times in a sound-isolated recording booth. All the speakers are professional announcers and speak standard Putonghua. The average echo time of the recording studio is about 0.3 seconds and the sound-isolation ability is greater than 40 dB. The material was recorded using a DAT recorder, and originally digitized at a 48 KHz sampling frequency with 16-bit quantization and a signal to noise ratio greater than 50 dB. Two desktop microphones are used at different time and sites, *viz.*, DM-72LTD (PIONEER) microphone and CR1-6 microphone. They are positioned about 40 cm from the subject's lips, off-center at about 20 degree angle and out of the breath stream. After the recording, several staff members listen to the utterances carefully and those utterances with detected errors are re-recorded. After the recording session, the digitized speech data was transferred from the DAT tapes to a 486 PC, and simultaneously was down sampled to 16 KHz, and then segmented into files corresponding to individual utterances.

As for the sentences sub-corpus and multi-syllable words sub-corpus, the data was collected in common office environment, where the noise level is controlled be less than 50dB. The material was recorded directly into 486 PCs by using the Sound Blaster card with 16KHz sampling rate and 16bit quantization, and an additional dynamic microphone CD1-40 (made in China) is used in sentences recording. Two groups of speakers, are age of 15 to 55 years old and speak standard Putonghua with different weak dialect accents, join in the recording. There are 80 speakers in the first group, half male and half female, each utters part of multi-syllable words. And the other group contains 200 speakers, each one is required to utter 500 sentences among the materials in the sentences sub-corpus. The posture of speaker is relative free, so that the collected data are more close to the practical circumstances.

In order to save disk space, the silence was removed from each utterance according to an automatic end-points detection procedure. After that, each utterance was verified in two ways. First, the digitized waveform was examined visually to determine if some portion of the utterance was incorrectly deleted by the chop program. If a significant portion of the utterance was deleted, the utterance was tagged and then was

recovered and chopped by hand. Second, each utterance was listened to. The listener noted ambiguous utterances and utterances that were incorrectly chopped.

Finally for the telephone speech data, a LINKON FC-3000 multimedia communication card was used as a platform to interface a 486PC and telephone network. Three hundreds speakers are of age 20 to 60 years old, they speak the materials in the telephone speech sub-corpus on telephone network by using a few assigned handsets. Recorded speech was sampled at 8KHz with 16 bit resolution. The digitized speech waveform was segmented into files, then verified just as same as other sub-corpus.

All waveform files are binary. The structure of files is compatible with TIMIT files. Data are stored in a DOS format.

4. SUMMARY

A Putonghua corpus -- USTC95 is introduced. This database, which is intended for use in analyzing spoken Chinese, designing and evaluating algorithms for Mandarin speech recognition, is being made available to provide: (1) a carefully structured research resource, and (2) benchmarks for performance evaluation to judge both incremental progress and relative performance. Both the digitized waveform data and the orthographic transcription data (including Pinyin symbols with tones) are provided by the medium of the CD-ROM, magnetic tape and other data cartridge etc. At present, the material alone amounts to approximately 10 Gigabytes (GB) of data. More speech data is now continuously being collected. Part of the database will also be phonetically labeled in future. A relational database system to manage the speech data more efficiently and conveniently is also developed.

5. REFERENCE

1. Pallet,D.S. "Speech Corpora and Performance Assessment in The DARPA SLS Program", Proc. of ICSLP 90, 1990.

2. Kurematsu, A. et al,"ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis", Proc. of ESCA Workshop, 1989.

3. Wang, R.H., Xia, D.Y. and Ni, J.F."The Development of A Chinese Voice Database for Machine Recognition", ACTA AUTOMATICA SINICA, Vol.18, N0.3, 1992.(in Chinese)

4. Wang, R.H. and Ni,J.F. , "Assessment of Chinese Speech Input System", ACTA AUTOMATICA SINICA, Vol.26, N0.4, 1994. (in Chinese)

5. Chen, Y.B. and Wang, R.H. SPEECH SIGNAL PROCESSING, USTC Press, 1990. (in Chinese)

6. Xu, Y. "Acoustic Phonetic Properties of Juncture in Putonghua", ZHONGGUO YUWEN, Vol.6, pp. 353-360, 1986. (in Chinese)

7. Wang, R.H. and Ni,J.F. "Designing Chinese Speech Corpus Based on The Juncture", COMPUTER APPLICATIONS AND SOFTWARE, Vol.11, N0.1, 1994. (in Chinese)

8. Wu, Z.J. "Tone-Sandhi Pattern of Quadro-Syllablic Combinations in Standard Chinese", RPR-IL(CASS), pp.1-13, 1988. (in Chinese)