Design, Compilation and Processing of CUCall: A Set of Cantonese Spoken Language Corpora Collected Over Telephone Networks

W.K. LO, P.C. CHING, Tan LEE and Helen MENG

The Chinese University of Hong Kong

wklo@ee.cuhk.edu.hk, pcching@ee.cuhk.edu.hk, tanlee@ee.cuhk.edu.hk, hmmeng@se.cuhk.edu.hk

Abstract

The design and compilation of the CUCall telephone speech corpora is described in this paper. Speech database is an indispensable resource for research and development of state-of-the-art spoken language technology. These speech recognition systems rely greatly on a huge amount of well-designed and appropriately processed speech data for parameters training. On the other hand, as telephony applications are becoming more demanding and complicated, natural language interface is gaining more popularity than the traditional touch tone operation. Therefore, large telephone speech databases are required for such system building. Separate speech corpora are needed for telephone systems since there exist significant differences due to the channel difference. In this paper, we will describe the design and processing of a set of spoken language corpora for Cantonese that are collected over fixed line as well as mobile telephone networks. The corpora are intended as a versatile set of training data for general purpose application systems that adopt a statistical approach to spoken language processing. The designed set of corpora will be made up of over 1000 speaker calls.

1 Introduction

Speech data collected over telephone network is an essential resource for telephone based spoken language systems. The increasing penetration of remote system or service access over telephone networks has created a great driving force for collecting a huge amount of telephone speech data from a large speaker population and for different languages. Since the current state-of-the-art speech recognition techniques are statistically based, the availability of annotated data is particularly important. In general, the greater the amount and coverage of the data, the better the speech applications developed. In order to build a spoken language system over telephone network, the speech data has to be collected over telephone network and properly transcribed. The goal of this work[1] is to collect and compile a set of general purpose Cantonese telephone speech data from a large group of people of both genders. With the availability of this set of corpora, the rapid growth in the spoken language applications over telephone networks for the Cantonese speaking community is made possible.

Over the past decades, many telephone based spoken language systems have been developed with great success. They all take advantage of the existence of several spoken language corpora compiled in recent years. Examples include the Jupiter from MIT [25], HMIHY[2] from AT&T [17] and the European Union projects such as ACCeSS [27] and ARISE [28] etc. Nowadays, there are quite a large number of companies that make use of simple automatic telephone service systems to reduce the cost of employing human operators. Many of them have upgraded or wish to upgrade their touch-tone based system to speech enabled versions. It is obvious that continuous efforts are needed to enhance these services via speech technologies as much as possible.

For building telephone speech recognition systems, there has long been a great demand on Cantonese telephone speech data. This work is an initial effort to collect a set of Cantonese spoken language corpora over telephone network. It is targeted to provide some versatile data for public use. It aims to enrich the infrastructure for spoken language technology by providing the speech community with well-designed corpora in Cantonese. The compiled database will enable the integration of Cantonese speech technology to many of the existing telephone based interactive systems.

[1] http://dsp.ee.cuhk.edu.hk/speech/cucall.html

[2] How may I help you? is the service offered by AT&T.

1-1 Background

There has been much effort in spoken language corpora development over the past decades. These include the TIMIT [8], Resource Management [15], Wall Street Journal [14], Air Travel Information Service [16] etc. from the United States. In Europe, there are the EUROM1 [21] and SpeechDat [3] etc. They contain microphone data and telephone data as well. From the early adaptation of microphone corpora to network versions like NTIMIT [5] and the collection of real-world telephone data such as MACROPHONE [1], CALLHOME [30], SpeechDat [3], POLYPHONE [29] etc., there is an abundant amount of data available for the western languages. The availability of these telephone corpora has successfully helped drive the research and development of telephone-based speech technologies of these languages.

For Asian languages, there has been limited investment spent on corpora development. Much effort came from Japan, for example those reported in [6, 7, 13]. For Chinese language, speech database collection has only started relatively recently. More widely used databases include microphone speech corpora such as the USTC95 [19], HKU96 [2, 24], HKU99 [4], CMSC [22] and others [23]; and telephone speech corpora such as MAT-160 and MAT-2000 [18, 20, 31]. These telephone data become valuable resources for many voice-activated telephony applications development.

Among the many Chinese dialects, Cantonese is one of the most popular Chinese dialects used in the southern China. Development of spoken language corpora has just started within the past decade [9, 10, 11]. It began with some small-scale corpus collection for specific projects. There is great shortage in Cantonese speech corpora to drive the growth and advancement of Cantonese speech technologies.

In 1997, the development of CUCorpora[3] [9, 11, 12] was initiated at the Chinese University of Hong Kong. CUCorpora is the first large-scale Cantonese spoken language corpora that are made available for public access. It is designed to cover both phonetically based content and common task oriented and application-specific content. The present work on telephone speech data compilation is a momentous extension of this effort. The vast variation of operator network protocols in Hong Kong[4] yet enrich the content of the

[3]http://dsp.ee.cuhk.edu.hk/speech

[4]Hong Kong has a large number of mobile network operators offering different kinds of network services using different protocols. This includes the GSM900, GSM1800, TDMA, CDMA.

corpora. Since the speakers will have to call our server to activate the data collection process, the resultant corpora are thus code named CUCall. The availability of the invaluable CUCall will undoubtedly nourish the booming technologies to a greater extent.

1-2 Paper organization

The paper is organized as follows. The design of the corpora materials will be described in detail first in Section 2. The design selection of the major parts of data will be elaborated. After that, actual collection process is presented. From the recording system setup down to the collection process, every detail of the process will be given. In Section 4, the post-processing of the captured data will be explained. The validation, transcription as well as the organization procedures are described. We will then provide some initial analysis on the designed corpora materials. Finally, conclusions are made in Section 6.

2 Corpora Design and Organization

The design of the CUCall has been based on our previous experience with CUCorpora. The concepts behind stay the same. Like CUCorpora, CUCall comprises of linguistically oriented and application-specific data. In CUCall, we take a step forward to include spontaneous conversations and short paragraphs data. These will altogether make up two major parts in the corpora:

1. Phonetically oriented continuous speech data that focus on:
   (a) coverage through carefully designed corpora materials; and
   (b) different speaking styles from short paragraphs and free form spontaneous conversation style.
2. Application-oriented short phrases and digit strings.

Figure 2 shows an overview of the organization of the CUCall telephone spoken language corpora.

2-1 Phonetically-oriented data

2-1-1 Phonetic coverage oriented

The phonetically oriented data in the CUCall is based on the design of the CUCorpora with some variations. This part of the data set is made up from sentences and short paragraphs. The materials for the sentences are based on the test and training materials of the CUSENT corpus in CUCorpora and the short paragraphs are excerpted from local newspapers.

Sentences The sentences are chosen to be phonetically rich in the sense that they constitute complete coverage of bi-phone class context. The selection of sentences was detailed in [9, 11]. It was implemented as a semi-automatic process where human intervention is included to decide on the readability of the automatically selected sentences.

Short paragraphs The short paragraphs attempt to emphasize more on the variations of the speaking behaviour and characteristics. For short paragraphs, the selection is solely based on the readability of the paragraphs without taking into consideration of the phonetic content. It aims to enrich the sentence data as well as provide data that bears very different speaking style. Table 1 shows the amount of data for each of these types.

Table 1: The number of reading materials for each type of the phonetically oriented data.

material	**number**
sentences	5719
short paragraphs	90

2-1-2 Speaking style oriented

For collecting speech data of different speaking styles, the design of CUCall included specifically short paragraphs and conversation parts.

Short paragraphs While the short paragraphs can enrich the phonetic coverage as mentioned in 2-1-1, the data collected in this part is believed to be very different from that of the stand alone sentences. There are many different speaking phenomena being

exaggerated when people reading a section of long text materials. These include correction, hesitation, breathing, long pause etc. Therefore, these recorded materials can also serve the purpose of representing another kind of speaking style in addition to enriching the phonetic content of the sentence corpus.

Spontaneous conversation In the CUCall corpora, a new type of speech data to be collected is the spontaneous conversation type of utterances. These data are collected with the aim to obtain the characteristics of various speakers when prompted to speak in an unprepared manner. There are expected delay, hesitation, correction and skipped words etc. In addition, there are also many colloquials, pronunciations and agrammatical sentences that will not be found in normal read speech. These will provide us with invaluable data for the study of the variation of speaking characteristics under different situations.

The design of "prompts" for this part of data collection has been carefully planned. It is implemented as a single round dialogue between the speaker and the system. Since the speakers are free to answer anything to the prompts, the phonetic content is uncontrollable. The major consideration here is to ensure that there is a high proportion of speakers capable of responding to the prompts. Due to the lengthy nature of the recording process, some speakers are expected to skip these prompts intentionally while some may be too enthusiastic to give very long answers. Several points are considered during the design of prompts:

1. The prompts must be simple enough that "spontaneous" response is possible. Calculation, memory recall or questions requiring accuracy are not suitable.
2. The prompts must have different answers from different speakers so as to increase the variations of the collected data. It would be even better if the same speaker will give different answers at different time.
3. The responses to the prompts may be either long or short.
4. Both for legal purpose and encouraging speakers to answer, the content of the answer must be irrelevant to privacy of the speakers.

Based on the considerations mentioned above, we have carefully designed six prompts. These prompts are carried out at the end of each of the collection sessions. It is done

this way because by that time, the speaker will be more familiar with the recording process. This will then reduce the probability of making mistake since unprepared types of responses are usually "error" prone. In Figure 1, the six prompts are listed for reference (in English translation, because of the colloquial nature of the Cantonese prompt, not all words are writtable in characters).

Figure 1: The Cantonese prompts (with English translation) for spontaneous spoken response collection.

1. 請簡單介紹一吓你而家身處嘅環境，例如，你而家喺乜嘢地方，身邊有乜嘢人，有乜嘢事物等等。
2. 請講一吓你就讀過嘅學校，例如中學、小學，或者各樣嘅進修課程。
3. 請問，你而家係咪使用緊手提電話？
4. 請講一吓你居住嘅地區，同埋屋村名稱或者街道名稱。
5. 除咗廣東話之外，請問你仲會講乜嘢語言？
6. 請講一吓你最經常乘搭嘅交通工具，同埋所前往嘅目的地。

1. Would you please describe the environment of your recording, such as where are you, anybody nearby and anything happening?
2. Which schools have you been studying at? Such as primary and secondary school. Did you study other short courses of any kinds?
3. Are you using a mobile phone? *(this is intentional for a short yes-no answer)*
4. In which district of the city do you live? And what is the name of the estate or street?
5. Besides Cantonese, what other languages do you speak?
6. What kinds of transportation do you take the most frequently and where do you go?

2-2 Application-specific data

The CUCall corpora also contain digit strings as well as application-specific short phrases in some specific domains. The design of the digit corpus is similar to that of the CUDIGIT [9] corpus. In CUCall, the reading materials include all of the single digits together with some random generated long digit strings. This makes up a small-scale digit string

corpus collected over telephone network from a large number of speakers.

The short phrase materials are designed with reference to CUCorpora. Phrases are chosen from various reading materials including names of listed companies and their abbreviations, name of foreign currencies, district names and major housing estates in Hong Kong together with the navigation commands adopted from the CUCMD [9, 11] corpus. These phrases cover the financial domain, navigation commands, as well as major local places. They could be used when building command based speech applications for the related domains. Table 2 lists the amount of corresponding type of phrases.

Table 2: The amount of different types of phrases for the application-specific data.

material	**amount**
name of places (districts & housing estates)	228
listed companies	1085
foreign currencies	37
navigation commands	90
Total	**1440**

3 Data Collection Process

The data collection is facilitated by using an automatic call centre type telephone server system. The overall set-up is shown in Figure 3. This server system allows the speakers to call in and then read the provided materials. It is also equipped with the usual navigating features with a touch-tone telephone system.

3-1 Telephone Server

The telephone server is a cluster of computers with one file server and two computer telephony servers (see Figure Figure 4). The file server has a large 64 GB harddisk and is directly connected to the two telephony servers over a 100 Mbps isolated ethernet.[5] The

[5]This is intentionally set up to improve the security and robustness of the systems. The cluster of computer connected in their own network could eliminate the interference of possible network traffic from other irrelevant processes.

computer telephony servers are equipped with a Dialogic D/41-ESC four port telephony cards for telephone network connection.

There are eight ports available on the Dialogic D/41-ESC card, but only two ports are used. This is sufficient for our current scale of speech collection. Also, additional ports may be used as backup during system maintenance or occasional system breakdown. Furthermore, we can also even out the potential analog channel discrepancies among the different ports by intermittently changing the answering ports over the course of data collection.

3-2 Collection Process

The actual collection process was implemented in several steps:

1. Preparation of the reading materials;
2. Distribution of the reading materials;
3. Accepted speakers call to the telephone server.
4. Return of filled questionnaires from speakers.

Preparation of reading materials The reading materials are mixtures of phrases and sentences described in Section 2. Each part is randomly shuffled and printed out on paper. Every 10 to 30 successful calls will gives a complete set for that part of the corpora. In order to differentiate against different gender and different kinds of telephone networks, the reading materials are prepared and distributed in four parallel streams: male mobile, male fixed-line, female mobile and female fixed-line. At the end of each of the prompt sheets, there is a short questionnaire to enable the collection of information about the speaker's age group, telephone network operator (for mobile phone) or type of telephone (whether they are using extension line or direct line).

Distribution of prompt sheet The prepared prompt sheets are distributed through recruited agents. They pass the reading materials to candidate speakers. After recording, the speakers then return the prompt sheet with questionnaire duly completed to the agent and then the agent pass them back to us for processing. The adoption of an agent based distribution network allows an efficient collection process while we could indirectly control the speaker community by choosing appropriate agents.

Speakers call The speakers will make call to our telephone server at any time they so wish. The server would answer the calls whenever it is idle. The speakers are then requested to jot down a generated serial number for bookkeeping purpose. After that, our server program will prompt the speaker by the item numbers on the prompt sheet and then wait for the speakers' speech data with an automatic silence detector. After the speakers have read the prompted item (or time-out if the speakers do not say anything), the data is immediately stored on to the server's hard disk. This prompting process repeats until the last item is finished. The server then reminds the speakers to fill out the questionnaire and hang up subsequently.

Questionnaire return After the agents have collected the prompt sheet, the serial number and questionnaire results are entered into our database for bookkeeping and analysis purposes. Up to this point the collection process is completed and the data are kept for later post-processing.

4 Post-Processing of Data

The most important part of a spoken language corpora development process is the post-processing of the collected speech data. The collected data need to be accurately annotated with necessary labels and organized properly for easy distribution and usage. Based on our previous experience from developing the CUCorpora [9], we have carefully designed the post-processing procedure for the telephone speech data. Figure 5 illustrates the general flow of the post-processing procedures.

Validation of the calls Among the large number of calls received, there is a small percentage of useless data. It may be due to the reason that the speakers give up reading after a short while, the recording environment is too noisy that the silence detector failed totally, or even the system broke down. Based on the serial numbers, we validate all of the calls by checking if there is reasonable amount of data being recorded. If the call is finished properly, the information of the speaker provided on the questionnaire is entered into our speaker database anonymously.

Phonemic transcription of the validated data A major effort in spoken language corpora development is annotation. This is the most important and labour intensive process. In our case, all of the validated data will be transferred using cassette tapes to our contracted professional transcribers. They will listen to the recording tapes and provide Cantonese phonemic transcriptions to all data or mark them as noise wherever applicable. Those successfully transcribed data will then be accompanied by the corresponding phonemic transcription when distributed.

Partitioning and distribution of the collected data The transcribed data will then be partitioned according to the different parts (e.g. digit strings, short phrases, sentences, spontaneous conversation etc.). The partitioned data will be organized into different directories according to different speakers. The phonemic transcription will also be provided in the form of LSHK[6] transcription symbols. These organized directories of speech data and transcription will be printed on to compact disk for distribution.

5 Data Analysis

In this section, some statistical information of the designed corpora reading materials will be presented. Although there are many expected discrepancies from the actual data that are collected, these statistics can still give an overview of the characteristics of the designed corpora. The discrepancies between the designed materials and the recorded data are mainly due to the reason that there are many speakers who read colloquial and 'lazy' pronunciations, mis-read of materials (e.g. insertion, deletion and substitution of words), and mis-use of the recording systems (e.g. start reading before the recording actually started, stop reading before all of the materials are read, etc.). These could only be analyzed after all data have been transcribed. Detailed statistical analysis of the actual collected data will be released after the information has been prepared.

Table 3 shows the basic information for different parts of the corpora. From this table, it can be observed that out of the 1600 common tonal syllables in Cantonese, the sentence materials have covered over 85% of the syllables. In the short paragraphs corpus, even though the tonal syllable coverage is not as high as that of the sentence recording, we are

[6]Linguistic Society of Hong Kong.

Table 3: Statistical information of the reading materials for the phonetically oriented and application-specific parts of the corpora.

Part	**# per speaker**	**# tonal syl.**	**# base syl.**	**syllable count**
Phonetically oriented corpora				
sentences	50 (out of 5719)	1399	579	4 to 31
short paragraphs	3 (out of 90)	768	418	23 to 120
Application-specific corpora				
1-digit string	10	N.A.	N.A.	N.A.
7-digit string	5	N.A.	N.A.	N.A.
8-digit string	5	N.A.	N.A.	N.A.
16-digit string	5	N.A.	N.A.	N.A.
phrases	48 (out of 1440)	562	344	2 to 8

expected to obtain speech data in the form of sentences of length ranging from 23 to 120 characters. These could give us a number of important and unique characteristics in long utterances.

Figure 6 gives another way to look at the properties of the designed reading materials in the sentences and paragraphs parts of the corpora. These are the frequency-of-frequency (FOF) scattered plots for the base and tonal syllables in these parts of the corpus. The FOF plots show the distributions of the occurrences of the syllables. From these figures, it is observed that the content of the corpora is reasonably distributed. While there are some frequently occurred syllables and also some rarely occurred syllables, the majority of the syllable occurrences lie in the middle range. This could then enable us to obtain a normal distribution for the syllables in these parts of the corpora.

For the application-specific corpora, information shown in Table 3 can give us an idea of what is being collected for the database. We have some randomly generated digit strings of various lengths. They should cover most of the common applications where digit strings are needed to be recognized. These may include getting identity card number, telephone number, credit card numbers etc. The 7-digit, 8-digit, 16-digit strings together with the single digits are targeted for these applications. However, since digit strings are so general that continuous digit string data can definitely be applied to other areas of applications.

The other application-specific data collected in this corpora are phrases of various kinds (see Section 2). The phrases from the various different domains are mixed and shuffled for each of the speakers so as to increase the variation in the collection data. From Table 3, it may be found that the acoustic coverage of the phrase part is not as good as that of sentences and paragraphs. Since these data are designed for use in the designated domains, phonetic coverage is not the major concern during corpus design. Nevertheless, the base syllable coverage for these phrase is not far deviated from the complete Cantonese syllable inventory.

Regarding the amount of data in the corpora, a rough estimation has been made. Up to the time of writing, there have been over 1,000 successful calls received. These calls give a total of around 200 hours of data covering all sorts of acoustic events (speech, silence, noise, background etc). Among this volume of recording, the sentence, speaking style, short phrases and digit parts roughly contains 84, 40, 28 and 59 hours of recording. We are currently post-processing these data and they will be made available for public release in the near future when the data are processed.

6 Conclusions

In this paper, the design and data collection process for a telephone spoken language corpora is presented. Details about the post-processing and preliminary analysis of the data are given. Based on the previous experience in microphone speech data collection, this work is extended to collect telephone speech data so as to provide sufficient materials for the building of statistical spoken language systems. The corpora are again divided into two parts: phonetically oriented data and application-specific data. In this work, we have further extend our previous design to include also short paragraph for encompassing speaking characteristics when people reading long materials. Furthermore, we have also included some free-form open questions or prompts for obtaining speaking characteristics in spontaneous speech. Spontaneous speech presents new challenges to speech recognition and the collected data is a valuable resource for investigating possible solutions.

7 Acknowledgements

This project is developed with the support from the Innovation and Technology Fund (AF/96/99). We are grateful to industrial sponsors: Group Sense Limited and SmarTone Mobile Communication Limited. We would also like to thank the Hong Kong Blind Union for helping us transcribes the telephone speech data.

References

[1] J. Bernstein, K. Taussig, and J. Godfrey, "MACROPHONE, an American English telephone speech corpus for the polyphone project," *Proceedings of 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 81-84, 1994.

[2] C. Chan, "Design considerations of a Putonghua database for speech recognition," *Proceedings of the Conference on Phonetics of the Languages in China*, pp. 13-16, Hong Kong, 1998.

[3] H. Hoge, H.S. Tropf, R. Winski, H. van den Heuyel, R. Haeb-Umbach, and K. Choukri, "European speech databases for telephone applications," *Proceedings of 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1771-1774, 1997.

[4] Q. Huo, and B. Ma, "Training material considerations for task-independent subword modeling: design and other possibilities," *Proceedings of 1999 Oriental COCOSDA Workshop*, pp. 85-88, 1999.

[5] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," *Proceedings of 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 109-112, 1990.

[6] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357-363, Elsevier Science, 1990.

[7] H. Kuwabara, K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a large-scale Japanese speech database and its management sys-

tem," *Proceedings of 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 560-563, 1989.

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

[9] Tan Lee, W.K. Lo, P.C. Ching, and Helen Meng, "Spoken language resources for Cantonese speech processing," *to appear in Speech Communication*, Elsevier Science, 2001.

[10] W.K. Lo, K.F. Chow, Tan Lee, and P.C. Ching, "Cantonese databases developed at CUHK for speech processing," *Proceedings of the Conference on Phonetics of the Languages in China*, pp. 77-80, Hong Kong, 1998.

[11] W.K. Lo, Tan Lee, and P.C. Ching, "Development of Cantonese spoken language corpora for speech applications," *Proceedings of the First International Symposium on Chinese Spoken Language Processing*, pp. 102-107, Singapore, 1998.

[12] W.K. Lo, Helen Meng, and P.C. Ching, "Sub-syllabic acoustic modeling across Chinese dialects," *Proceedings of the Second International Symposium on Chinese Spoken Language Processing*, pp. 97-100, Beijing, 2000.

[13] K. Ohtsuki, T. Matsuoka, T. Mori, K. Yoshida, Y. Taguchi, S. Furui, and K. Shirai, "Japanese large-vocabulary continuous speech recognition using a newspaper corpus and broadcast news," *Speech Communication*, vol. 28, pp. 155-166, Elsevier Science, 1999.

[14] D. Paul, and J. Baker, "The design of the Wall Street Journal based CSR corpus," *Proceedings of the Fifth DARPA Speech and Natural Language Workshop*, Morgan Kaufmann, 1992.

[15] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," *Proceedings of 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 651-654, 1988.

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

[17] G. Riccardi, A.L. Gorin, A. Ljolje, and M. Riley, "A spoken language system for automated call routing," *Proceedings of 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1143-1146, 1997.

[18] C.Y. Tseng, "A phonetically oriented speech database for Mandarin Chinese," *Proceedings of 1995 International Congress of Phonetics Sciences*, vol. 3, pp. 326-329, 1995.

[19] R. Wang, D. Xia, J. Ni, and B. Liu, "USTC95-A Putonghua corpus," *Proceedings of the Fourth International Conference on Spoken Language Processing*, vol. 3, pp. 1894-1897, 1996.

[20] H.C. Wang, "Speech research infra-structure in Taiwan," *Proceedings of 1999 Oriental COCOSDA Workshop*, pp. 53-56, 1999.

[21] R. Winski, and A. Fourcin, "A common European approach to assessment, corpora and standards," *in Advanced Speech Applications: European Research on Speech Technology*, K. Varghese, S. Pfleger, and J.P. Lefvre Eds., pp. 25-79, Spriner-Verlag, 1994.

[22] Y. Wu, "Chili Mandarin speech corpus," *Newsletter of ISCSLP98 Special Interest Group: Linguistic Database and Tools*, pp. 1-3, 1998.

[23] J. Zhang, "Notes on speech corpora of standard Chinese in China," *Newsletter of ISCSLP98 Special Interest Group: Linguistic Database and Tools*, pp. 4-5, 1998.

[24] Y.Q. Zu, W.X. Li, M.C. Ho, and C. Chan, "HKU96-A Putonghua corpus (CDROM version)," *HKU96 corpus*, Department of Computer Science, University of Hong Kong, Hong Kong, 1996.

[25] V. Zue, S. Seneff, J.R. Glass, J. Polifroni, C. Pao, T.J. Hazen, and L. Hetherington, "JUPlTER: a telephone-based conversational interface for weather information," *IEEE Transactions on Speech and Audio Processing*, vol. 8, is. 1, pp. 86-96, 2000.

[26] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, pp. 351-356, Elsevier Science, 1990.

[27] http://www.wcl.ee.upatras.gr/access/access.htm Automatic Call Center Through Speech Understanding System.

[28] http://www.compuleer.nl/arise.htm Automatic Railway Information Systems for Europe.

[29] http://www.icp.grenet.fr/ELRA/home.html, European Language Resources Association.

[30] http://www.ldc.upenn.edu, Linguistic Data Consortium.

[31] http://rocling.iis.sinica.edu.tw/ROCLING/MAT/index_cf.htm, Mandarin Across Taiwan corpus.

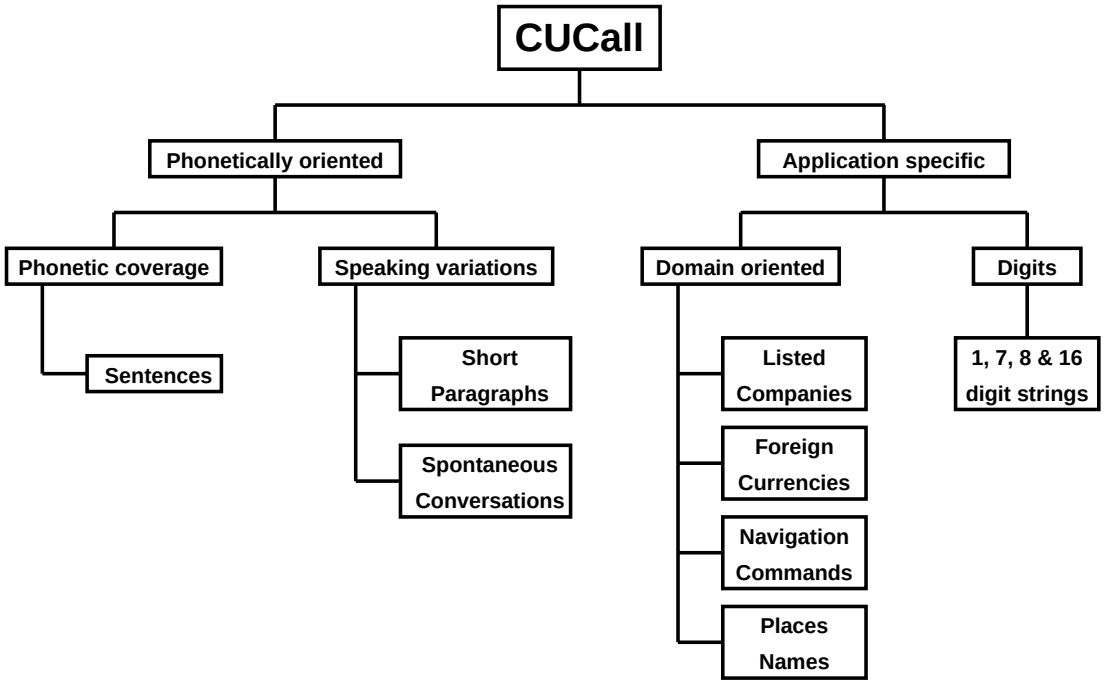


Figure 2: This is an overview of the organization of the CUCall telephone spoken language corpora for Cantonese.

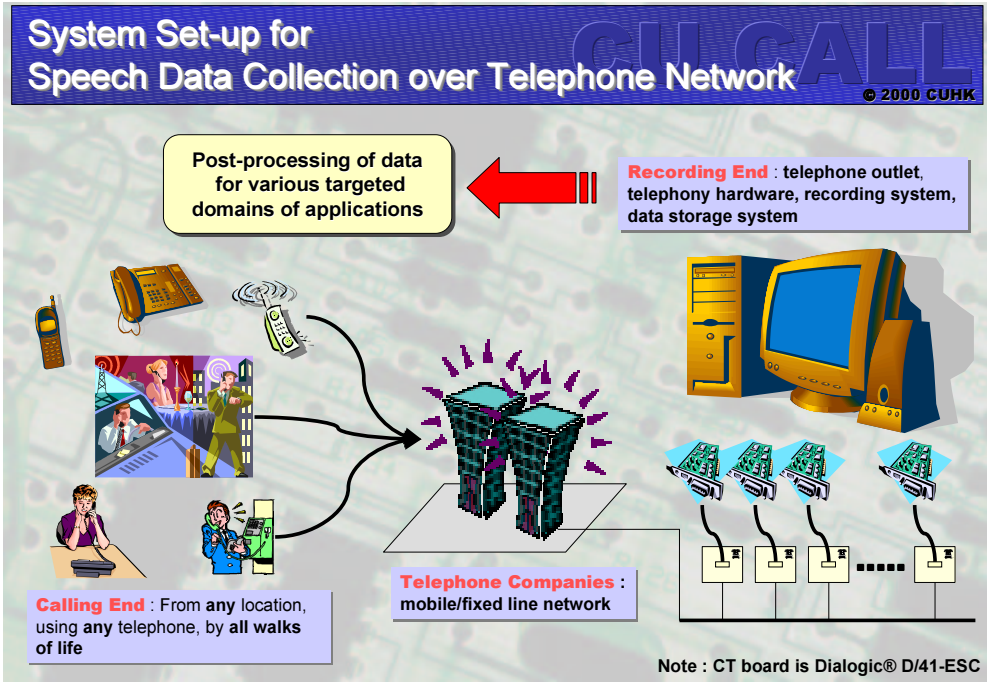


Figure 3: The data collection process for the CUCall corpora over the telephone networks.

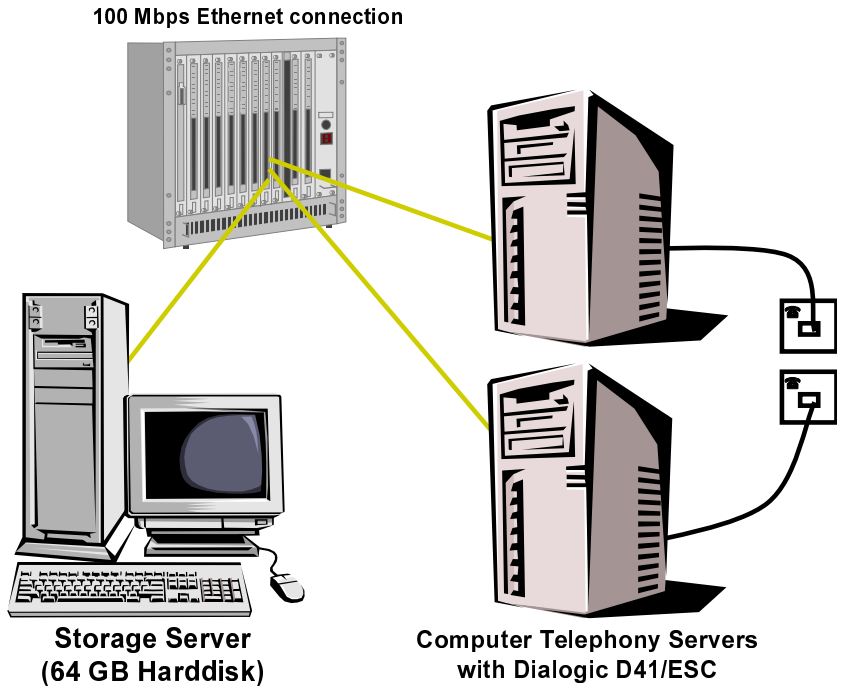


Figure 4: The telephone server setup for corpora data collection.

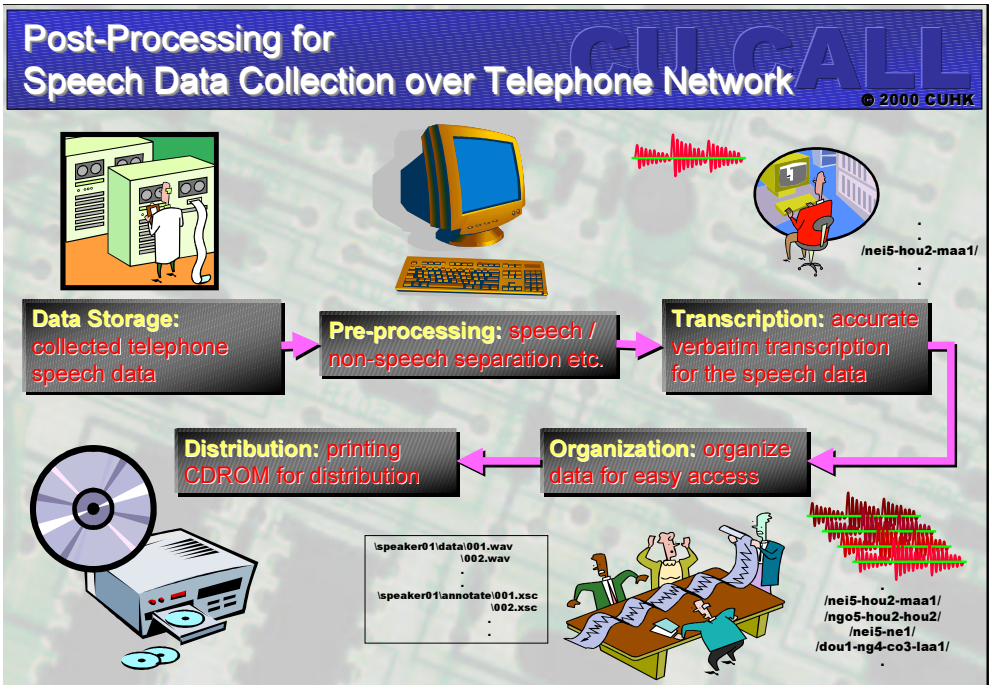
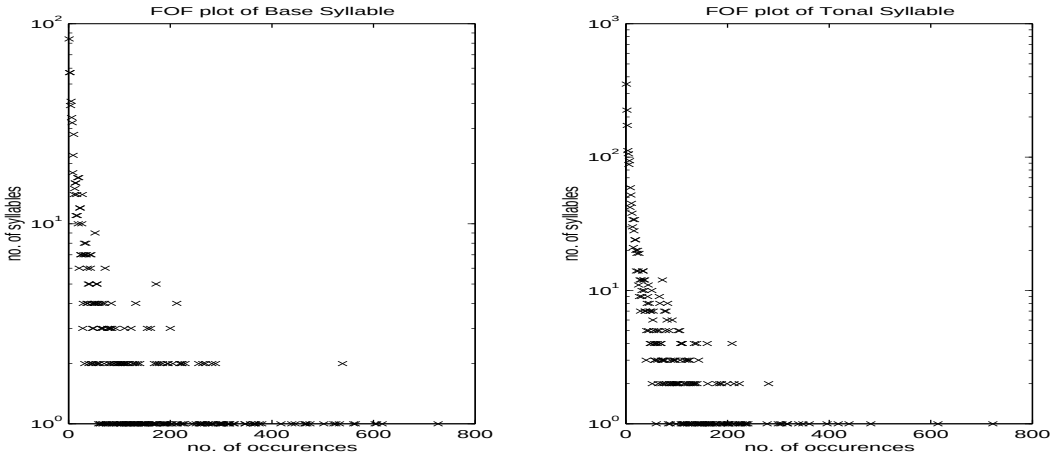
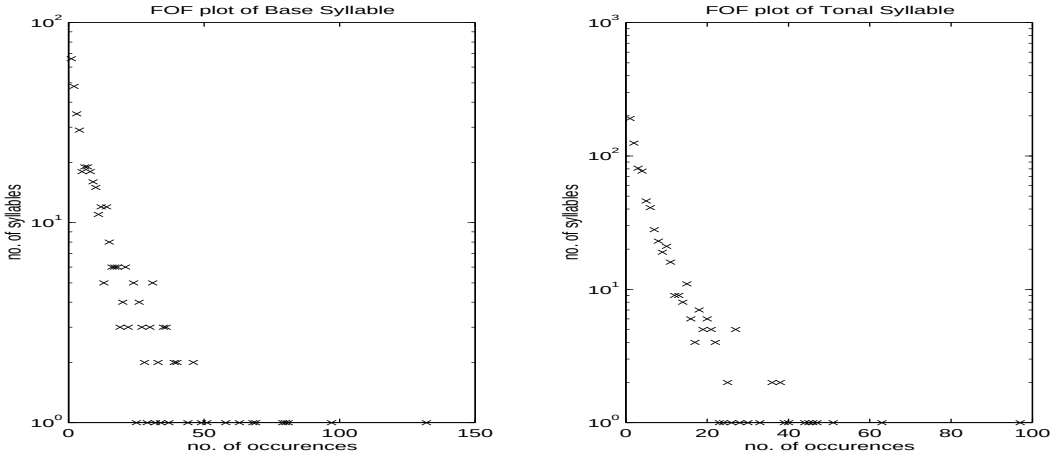


Figure 5: Data post-processing for the CUCall corpora.



(a)



(b)

Figure 6: Scatter plots showing the frequency-of-frequency statistics for syllables in (a) the sentence and (b) the paragraph reading materials.