A Very Large Scale Mandarin Chinese Broadcast Collection for the GALE Program

Liu Yi[1], Pascale Fung[1], Yang Yongsheng[1]

Denise DiPersio[2], Meghan Lammie Glenn[2], Stephanie M. Strassel[2], Christopher Cieri[2]

[1] Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology, Hong Kong
{eeyliu, pascale}@ece.ust.hk, ysyang@cs.ust.hk

[2] Linguistic Data Consortium,
University of Pennsylvania, U.S.A
{dipersio, mlglenn, strassel, ccieri}@ldc.upenn.edu

Abstract

In this paper, we present the design, collection, transcription and analysis of a Mandarin Chinese Broadcast Collection of over 3000 hours. The data was collected by Hong Kong University of Science and Technology (HKUST) in China on a cable TV and satellite transmission platform established in support of the DARPA Global Autonomous Language Exploitation (GALE) program. The collection includes broadcast news (BN) and broadcast conversation (BC) including talk shows, roundtable discussions, call-in shows, editorials and other conversational programs that focus on news and current events. HKUST also collects detailed information about all recorded programs. A subset of BC and BN recordings are manually transcribed with standard Chinese characters in UTF-8 encoding, using specific mark-ups for a small set of spontaneous and conversational speech phenomena. The collection is among the largest and first of its kind for Mandarin Chinese Broadcast speech, providing abundant and diverse samples for Mandarin speech recognition and other application-dependent tasks, such as spontaneous speech processing and recognition, topic detection, information retrieval, and speaker recognition. HKUST's acoustic analysis of 500 hours of the speech and transcripts demonstrates the positive impact this data could have on system performance.

1. Introduction

Large speech databases are a fundamental and important resource for spoken language processing and speech recognition. The rich variations in human speech can only be adequately analyzed and represented in properly recorded, annotated and processed speech data. Currently, most of the state-of-the-art automatic speech recognition (ASR) algorithms are based on statistical approaches, which require large volumes of training data representing speakers of various ages, accents, speaking styles, and channels to cover the diversity of human speech and speech environments. However, large-scale collections of Mandarin resources are often not widely available to speech recognition developers. This paper describes efforts by Hong Kong University of Science and Technology (HKUST) in partnership with the Linguistic Data Consortium (LDC) at the University of Pennsylvania to collect massive volumes of spoken Mandarin broadcast resources to support automatic speech recognition development within the DARPA GALE program.

2. Motivation

Much speech processing research has been done on English, Arabic, and many other languages. Such efforts were supported by speech corpora provided by LDC, European Language Resources Association (ELRA) and SpeechDat, among others. To support research on speech in Asian languages specifically, Japan has invested heavily in the development of different types of Japanese speech databases, including telephony, lecture, and broadcast speech (Ohtsuki, 1999). These databases have greatly facilitated the development of speech processing technologies, and many of them were published and released as standard development and evaluation resources.

Chinese is one of the most widely-spoken languages in the world, and Mandarin (Putonghua) is the official spoken language of mainland China, Hong Kong, Macau, and Taiwan. Mandarin speech recognition research has attracted great interest in recent years, particularly in the broadcast news (BN) and broadcast conversation (BC) domains. BC refers to call-in shows, roundtable discussions, and group debates, where participants speak casually, as in daily life.

Existing resources for Mandarin Chinese speech processing development include the 1997 Mandarin Broadcast News Speech (HUB4-NE), LDC98S73, released by LDC, is a BN speech corpus that is widely used for Chinese ASR tasks. This corpus consists of 30 hours of recorded broadcasts and transcripts that have been drawn from Voice of America (VOA), P. R. China Central Television (CCTV), and KAZN-AM, commercial radio based in Los Angeles, CA. CLDC-SPC-2003-001 is another widely-used BN database provided by Chinese LDC for speech recognition in mainland China[1]. In Hong Kong, the Chinese University of Hong Kong (CUHK) collected BN and BC data for speech recognition development in Cantonese, a primary dialect in South China (Lee, et. al., 2002).

[1] http://www.chineseldc.org/

Mandarin is one of the two source languages (the other being Arabic) identified for machine translation development and evaluation within the DARPA GALE program. Both Mandarin text and speech are required. Compared to available English broadcast resources that could be used for speech recognition research, there was insufficient Mandarin Chinese speech data to use for robust system development at the start of GALE. Most available Chinese language resources were in the BN or conversational telephone speech domains. HUB4-NE 1997, for example, contains a useful but limited amount of data, and is comprised of BN data recorded from VOA, CCTV in Mainland China and North American radio programming. The Topic Detection and Tracking project, as part of the DARPA TIDES program, provided several collections of Mandarin-specific broadcast programming, such as the TDT2 Mandarin Audio (LDC2001S93), and TDT3 Mandarin Audio (LDC2001S95) corpora. These corpora necessarily focused on the BN domain.

The GALE program aims much of its research at the challenges of unstructured data, which is representative of naturally-occurring material, broadcast conversations or discussions. Therefore, in partnership with LDC, HKUST undertook efforts to collect large volumes of Mandarin Chinese broadcast resources from a variety of programs, sources, speaker styles, and topics. A portion of the collection is also transcribed and annotated, to support Automatic Speech Recognition (ASR) development. The sections below describe the collection and transcription efforts in detail.

2.1. Linguistic challenges of Mandarin Chinese

Acoustically and phonetically, Mandarin is quite different from English and other European languages. The main differences include: (1) Chinese is monosyllabic; (2) Chinese characters are ideographic, words consist of one or several characters, and pronunciation is represented by the syllable; (3) different characters may share the same syllable, which is known as homophony.

Furthermore, Mandarin is a tonal language. There are five lexical tones (including neutral tone) in (Huang 1987; Lee et al., 2002). Each syllable is associated with a specific tone. The syllable with the same initial and final combination but with different lexical tones corresponds to different characters and has different meanings. Tones are a critical part of Chinese pronunciation and serve to differentiate meanings from characters of the same syllable. The pronunciation of Mandarin is represented by syllables. The structure of a syllable in Chinese is relatively simple: it consists of an initial and a final, or only the final. For standard Mandarin, there are around 1100 tonal syllables and 415 basic toneless syllables, 21 initials and 38 finals. Initials are very short in duration compared to syllables. There is a one-to-many mapping between syllable and characters. On average, each syllable translates to 17 commonly used characters.

Given these complexities, Mandarin pronunciation is very flexible in spontaneous, conversational speech. The recognition accuracy of BC is much lower than that of BN speech due to the effects of phonetic shifts, phone reduction, assimilation and duration changes (Liu and Fung, 2004). It has been demonstrated that the automatic recognition accuracy on BN speech is often over 85% while that of BC speech is only around 70% (Byrne, et al., 2001; Fung, et al., 2000).

Previous work has shown that high error rates in spontaneous speech recognition are due in part to poor acoustic modeling of pronunciation variations, compared to that of read and planned speech (Liu and Fung, 2004). A robust acoustic model trained using sufficient volumes of speech data with a diversity of pronunciation variations is a direct and efficient way to improve recognition accuracy for the BC domain. In the interest of continuing advancements in speech recognition, it is critical to have a large amount of speech data representing a variety of programs, resources, channels, speaking styles, speaking models, and topics, with good standardized transcripts.

3. Collection design and implementation

Several tasks are involved in the effort to develop large-scale Mandarin speech resources: data collection, program selection, transcription, and annotation. In partnership with LDC, HKUST has supported the DARPA GALE program by collecting broadcast resources, selecting subsets for transcription, and providing transcripts for those subsets, using guidelines developed by LDC. In 2009 “supralexical” annotation, described in Section 5.1, was added to these efforts.

In support of GALE Phase 4 in 2009-10, HKUST collected over 3000 hours of Mandarin Chinese broadcast programming via a cable TV and satellite transmission platform in China. We also provided collection reports showing recording channels, effective program duration, program types, etc. Collected files are manually audited in English for language, program and quality following guidelines provided by LDC. All information is processed and saved using both a web-based interface and a MySQL database.

A subset of the recorded speech data is manually transcribed with standard Chinese character transcription, adding restricted mark-up for speech phenomena in BC and BN programs. Transcription was performed according to Quick Rich Transcription (QRTR) guidelines provided by LDC, which focus on producing a verbatim transcript and dividing each speaker turn into a set of SUs, -- or Sentence Units – that have syntactic and semantic cohesion (LDC, 2008).

3.1. Programs

Collected programs are in the BN and BC domains. BC recordings include talk shows, roundtable discussions, call-in shows, editorials and other conversational programs that focus on news and current events. The collected recordings originate from sources broadcast from the People’s Republic of China and VOA radio programming. No more than 40% of the collection consists of BN programming. No more than 5% of dialectal Chinese (including Taiwanese Mandarin) is included within a particular program. Regional TV stations for the broadcast collection include Anhui TV,

Beijing TV, Dongfang TV, Fujian TV, Hubei TV and Jiangsu TV. CCTV programs comprise the bulk of the collection. The distribution of collected regional and national broadcast sources is shown in Figure 1 and Figure 2 below.

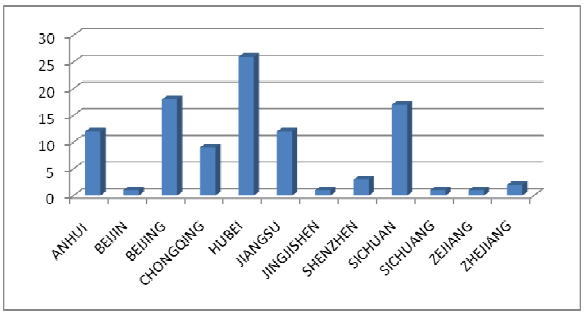

Figure 1. Collection distribution from regional TV programming

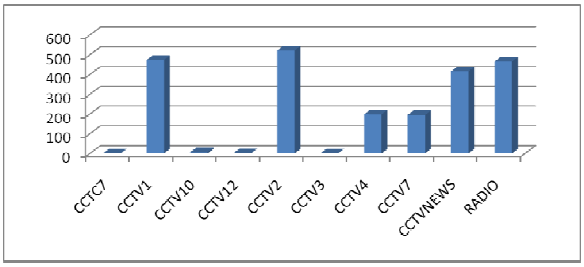

Figure 2. Collection distribution from CCTV and radio programming

3.2. Collection platform

The Mandarin Chinese broadcast collection described here is recorded through a cable TV and satellite transmission platform in China. The satellite used for data collection is Asiasat 3S. HKUST also collects programming using an internal recording system and a portable broadcast collection platform designed by LDC and installed at HKUST in GALE Phase 1. The platform produces recordings in MPEG-1 format @ 1.5Mbps which are then processed to extract the audio.

3.3. Speech styles and rates

Statistical criterion	Results
Total length	431.57 hours
Average utterance length	10.42 s
Average character numbers per utterance	36
Average speaking speed	3.48 syllables/per sec
Average utterance numbers per speaker in conversation	19

Table 1. Speaking rate information for HKUST's Mandarin broadcast collection

In order to collect natural BN and BC speech, especially for conversations in BC programs, we did not modify the collection data as it was received. The segmentation of programs and separation of video and audio signals are performed after the data is stored.

As mentioned in Section 2.1, speaking styles and rates in the current collection vary widely. Many of the pronunciation variations, which include phonetic shifts, phone reduction, and assimilation and duration changes, are also captured in the collection. The statistics for speaking style and rate of speech in the Mandarin broadcast collection are illustrated in Table 1.

4. Broadcast collection

4.1. Recording management

HKUST uses both automatic and operator-assisted recording approaches for data collection. An automatic data collection approach is used for programs with a fixed schedule, such as the CCTV programs "30 Minute News", "Military Reports", and "Across the Strait". The automatic collection system is installed in the data collection center in both mainland China and Hong Kong. HKUST engineers check the program collection schedule daily. Figure 3 shows HKUST's data collection support software.

Figure 3. Screenshot of the support software for program recording and management

HKUST also has an operator-assisted recording approach for approximately 1/3 of the collected programs for which the schedules change frequently. These programs are recorded from local cable TV at HKUST 's data collection center and offsite.

4.2. Recording hardware and software

LDC's portable broadcast collection platform is a TiVO-style DVR system capable of recording two streams of A/V material simultaneously; it supports analog CATV (NTSC and PAL) and FTA DVB-S satellite programming. The portable broadcast collection platforms weigh less than 30 pounds, have a footprint no larger than 60cm x 60cm x 10cm and contain scheduling software, diagnostic tools and remote control functionality (Walker, et al., 2010). The components of the portable broadcast collection platform are shown in Figure 4.

The recording software includes cable TV and satellite transmission platform-based software for automatic and operator-assisted recording approaches, automatic program management software,

communication software for different channels in BN and BC programs; separation software for video and audio separation and data saving; and ftp server communication software for data transmission.

Figure 4. Portable broadcast collection platform provided by LDC.

5. Transcription

The GALE research community requires large volumes of transcribed audio data in order to train their automatic speech recognition (ASR) systems. Transcripts conform to the Quick Rich Transcription (QRTR) specification provided by LDC. QRTR involves a verbatim, time-aligned transcript in Unicode (UTF-8) encoding, with minimal but useful markup, which includes time-aligned section boundaries, speaker turns, segmentation, sentence identification, speaker identification, etc. Transcript files are produced in a tab-delimited format, the native output of LDC's XTrans tool (Glenn, et al., 2009). A screenshot of XTrans is shown in Figure 5.

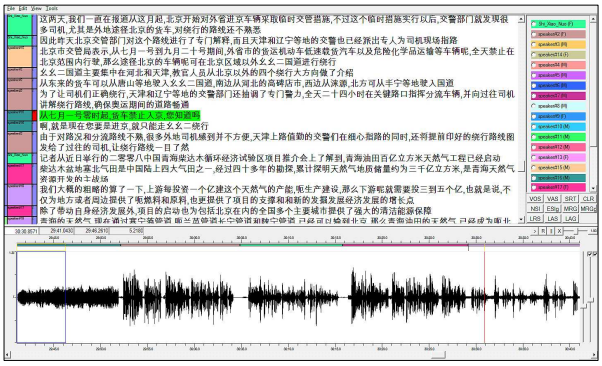

Figure 5. XTrans, LDC's transcription and speech annotation tool.

All transcripts are checked with a machine-aided semi-automatic method, whereby different transcribers use the Transcriber tool (see Figure 6), with special modifications to suit the team's needs (Barras, et al., 2001). Both XTrans and Transcriber are used to produce quick rich transcripts.

5.1. Speech annotation

Starting in Phase 4 of the GALE project, LDC and HKUST collaborated to produce supralexical annotation of existing transcripts. The supralexical annotation task includes marking disfluencies, speaker noises, named entities, bandwidth, and identifying regions of background noise and telephone speech.

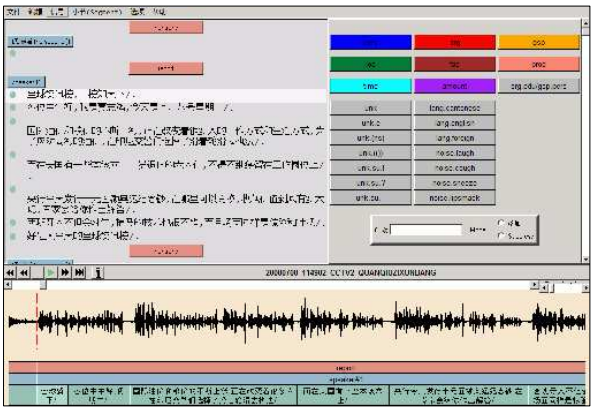

Figure 6. Transcriber interface for segmentation, transcription and time labeling

6. Collection analysis results

In order to help the reader better understand the collection, HKUST selected 500 hours of speech with transcripts and conducted an analysis of the acoustic properties of the subset, showing potential for significant improvement in system performance. We provide results of our acoustic analysis here.

6.1. Overview

At the character level, the 500-hour BN/BC batch contains 149,136 utterances and 5,416,492 characters in total (filled pauses and special mark-ups are not counted). 4,635 Chinese characters are used in the transcripts for these 500 hours. The auxiliary words "的" and "是" occur most frequently, which is in accordance with previous linguistic analysis (Huang 1987).

At the syllable (Pinyin) level, since one character corresponds to one syllable, the corpus contains 5,416,492 syllables and covers all 408 toneless base syllables. At the initial and final unit level, all 27 standard Putonghua initials (including zero initials) and 38 finals are covered. A summary of the contents, syllable and initial final coverage of the 500-hour collection described here is shown in Table 2.

Statistical criterion	Results
No. of utterances	149,136
No. of characters/syllables	5,416,492
No. of base syllable being covered	408
No. of standard initials being covered	27
No. of standard finals being covered	38

Table 2. Summary of the contents, syllable and initial final coverage of the 500-hour BN/BC subset.

6.2. Potential impact on speech recognition

From the perspective of speech recognition, we are not only interested in how many units have non-zero occurrence numbers but also in how many of them occur sufficiently frequently for robust acoustic model training.

Ideally, we would like to have sufficient samples of all acoustic units. In the 500-hour BN/BC dataset HKUST analyzed in more detail, we found that 99%, 94% and 83.5% of all base syllables occur more than 100 times, 200 times and 1,000 times respectively; and 36.3% of all base syllables have more than 10,000 occurrences. Therefore, syllable-based acoustic modeling would be possible for many small and medium lexicon applications using the Mandarin broadcast collection for system training.

Many state-of-the-art Chinese ASR systems use context-independent (CI) initial and final units instead of phonemes or phones as basic subword units for baseline acoustic model generation. Moreover, context-dependent (CD) acoustic modeling at the sub-syllable level, such as triphone, is widely used in ASR systems to achieve high recognition accuracy as well as good coverage of model complexity.

In Figure 7, we give the statistical distribution analysis of Chinese initials and finals of the 500-hour dataset. We can see that the distribution is in accordance with normal initial/final distribution; that is, the collection is phonetically balanced.

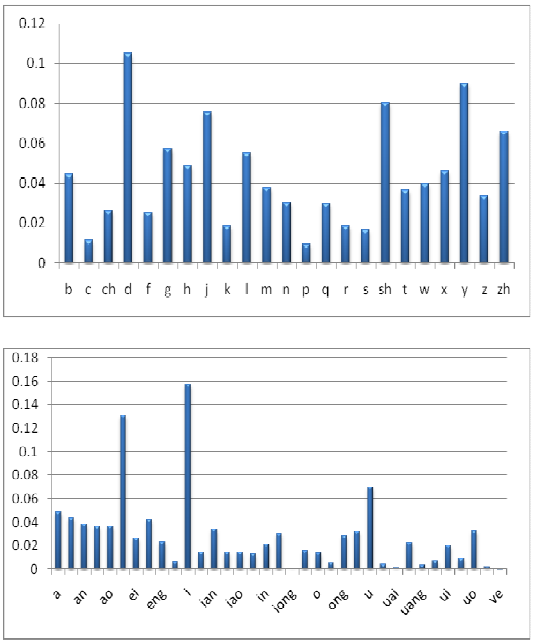

Figure 7. Distributions of Chinese initials and finals in the 500-hour BN/BC subset.

Table 3 shows detailed coverage of context-independent and context-dependent phonetic units. Since bi-phone level units as well as tri-phone level units are commonly used, both the intra-syllable contextual units and inter-syllable contextual units are considered.

These results show that all context-independent initial-finals are represented with sufficient samples. The same applies to bi-phone level intra-syllable units. The current Mandarin Chinese broadcast collection is an excellent resource for ASR system development using different sub-syllable units, as well as varied dictionary size for different applications. Furthermore, the high coverage of intra-syllable initial-finals and inter-syllable initial-finals means that the corpus is also suitable for robust tri-phone model generation and estimation.

No. of covered context-independent CI units	*Initials*	21 (100%)
	Zero initials	6 (100%)
	Finals	38 (100%)
No. of covered CD intra-syllable units	*Initial-Final combinations*	408 (99.7%)
	Initial-Nucleus combinations	94 (100%)
No. of covered CD inter-syllable units	*Final-Initial combinations*	795 (99.6%)
	Coda-Onset combinations	42 (100%)
	Tone-Tone combinations	20 (100%)

Table 3. Phonetic coverage of 500-hour BN/BC batch.

7. Conclusions

The Mandarin Chinese broadcast collection and transcription efforts described here in support of the DARPA GALE program are among the largest and first of their kind. HKUST and LDC partnered on the collection infrastructure and implementation, processing, transcription, and annotation. The collection provides abundant and diversified samples for ASR development in the domains of BN, BC, and conversational speech; as well as for topic detection, information retrieval, or speaker recognition. In addition, the speech data and transcripts can be used jointly or separately for different purposes.

The resources described in this paper will be made available to the broader research community over time. Many resources, such as the HUB4-NE mentioned in Section 2, have already been distributed to LDC members and non-member licensees through the usual methods, including publication in LDC's catalog. Transcription specifications are available at http://projects.ldc.upenn.edu/gale/Transcription/, and LDC's transcription tool XTrans is freely available at http://www.ldc.upenn.edu/tools/XTrans.

8. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

9. References

1997 Mandarin Broadcast News Speech (HUB4-NE). Linguistic Data Consortium, Philadephia.

Barras, C., E. Geoffrois, Z. Wu, and M. Liberman. (2001)

"Transcriber: Development and use of a tool for assisting speech corpora production," Speech Communication, vol. 33, pp. 5-22, January 2001.

Byrne, W. et al., "Automatic Generation of Pronunciation Lexicons for Mandarin Spontaneous Speech", *Proc. ICASSP01*, 2001

Hoge, H., et al. (1997) "European speech databases for telephone applications." In Proceedings of the IEEE ICASSP, Vol. 3, pp. 1771-1774.

Huang, J.H. (1987) *Chinese Dialects.* Xiamen University Press. (Chinese version).

Fung, P. and W. Byrne, et.al. (2000) "Pronunciation Modeling of Mandarin Casual Speech", Final report at the ws00 of Johns Hopkins summer workshop, Aug.2000

Glenn, Meghan Lammie, Haejoong Lee, and Stephanie M. Strassel. (2009) "XTrans: a speech annotation and transcription tool." In Proceedings of Interspeech 2009, Brighton, UK.

Godfrey, J., et al. (1992) "SWITCHBOARD: Telephone Speech Corpus for Research and Development." In Proceedings of the IEEE ICASSP, vol. 1, pp. 517-520.

Lee, T., et al. (2002) "Spoken language resources for Cantonese speech processing"." Speech Communication 36, No.3-4, 327 - 342, March.

Linguistic Data Consortium. (2008) *Quick Rich Transcription (QRTR) Specification for Chinese Broadcast Data (XTrans-Format Version).* http://projects.ldc.upenn.edu/gale/Transcription/Chinese-XTransQRTR.V3.pdf

Liu, Y. and P. Fung. (2004) "State-Dependent Phonetic Tied Mixtures with Pronunciation Modeling for Spontaneous Speech Recognition" IEEE Transactions on Speech and Audio Processing, Vol.12, No.4, pp.351-364, July 2004.

Ohtsuki, K., et al. (1999) "Japanese large-vocabulary continuous speech recognition using a newspaper corpus and broadcast news." Speech Communication 28, 155-166.

TDT2 Mandarin Audio Corpus, LDC2001S93. (2001) Linguistic Data Consortium, Philadephia.

TDT3 Mandarin Audio LDC Catalog No.: LDC2001S95. (2001) Linguistic Data Consortium, Philadephia.

Walker, Kevin, Caruso, Christopher, DiPersio, Denise. (2010) "Large Scale Multilingual Broadcast Data Collection to Support Machine Translation and Distillation Technology Development." In Proceedings of LREC 2010, Malta.