GALE Phase 4 Chinese Broadcast News Speech

1. Introduction

GALE Phase 4 Chinese Broadcast News Speech contains approximately 134 hours of Mandarin Chinese broadcast news speech collected in 2008 by the Linguistic Data Consortium (LDC) and Hong University of Science and Technology (HKUST), Hong Kong, during Phase 4 of the DARPA GALE program.

Broadcast audio for the GALE program was collected at LDC's Philadelphia, PA USA facilities and at three remote collection sites: HKUST (Chinese); Medianet, Tunis, Tunisia (Arabic); and MTC, Rabat, Morocco (Arabic). The combined local and outsourced broadcast collection supported GALE at a rate of approximately 300 hours per week of programming from more than 50 broadcast sources for a total of over 30,000 hours of collected broadcast audio over the life of the program.

The broadcast news recordings in this release feature news broadcasts focusing principally on current events from the following sources: China Central TV (CCTV), a national and international broadcaster in Mainland China; Phoenix TV, a Hong Kong-based satellite television station; and Voice of America (VOA), a U.S. government-funded broadcast programmer.

2. Broadcast Audio Data Collection Procedure

LDC's local broadcast collection system is highly automated, easily extensible and robust and capable of collecting, processing and evaluating hundreds of hours of content from several dozen sources per day. The broadcast material is served to the system by a set of free-to-air (FTA) satellite receivers, commercial direct satellite systems (DSS) such as DirecTV, direct broadcast satellite (DBS) receivers and cable television (CATV) feeds. The mapping between receivers and recorders is dynamic and modular; all signal routing is performed under computer control, using a 256x64 A/V matrix switch. Programs are recorded in a high bandwidth A/V format and are then processed to extract audio, to generate keyframes and compressed audio/video, to produce time-synchronized closed captions (in the case of North American English) and to generate automatic speech recognition (ASR) output.

The collection schedule is stored in a relational database using a Mysql database server. The database contains a history of all of the recordings that have been made; it has configuration and status information for all recorders; it has information about all receivers and associates specific programs of interest with the appropriate receiver; it contains a schedule of all recording jobs that need to be executed and their status; and it stores all audit judgments associated with a given recording.

HKUST collected Chinese broadcast programming using its internal recording system and a portable broadcast collection platform designed by LDC and installed at HKUST in 2006 (GALE Phase 1). Among the sources collected by HKUST for GALE are Anhui TV, Beijing TV, CCTV, Dongfang TV, Fujian TV, Hubei TV, Jiangsu TV and Voice of America (VOA) Mandarin.

LDC's portable broadcast collection platform is a TiVO-style digital video recording (DVR) system that records two streams of A/V material simultaneously. It supports analog CATV (NTSC and PAL) and FTA DVB-S satellite programming and can operate outside of the United States. It has a small footprint, weighs less than 30 pounds and can be transported as carry-on luggage. The portable platform deployed at HKUST's Chinese collection facility collected multiple streams of CCTV programming.

Further information about LDC's broadcast collection system can be found in LDC's Broadcast Collection System Data Sheet, http://www.ldc.upenn.edu/DataSheets/Broadcast_Collection_System_DS.pdf.

3. Broadcast Collection Audit Procedure

All broadcast data collected for GALE by LDC and by the remote collection sites managed by LDC were manually audited by Arabic, Chinese, and English speakers for language, program and quality. The broadcast auditing process served three principal goals: as a check on the operation of LDC's broadcast collection system equipment by identifying failed, incomplete or faulty recordings; as an indicator of broadcast schedule changes by identifying instances when the incorrect program was recorded; and as a guide for data selection by retaining information about a program's genre, data type and topic. LDC developed a Broadcast Audit Interface Tool to audit its local collection which presented auditors with three segments from each recording (beginning, middle and end) from which audit judgments were made.

Each remote collection site used a form of audit procedure based on the LDC model. HKUST generated English-language .xml and .html audit reports for the Chinese programming it collected. Those reports contained auditors' judgments from three portions of each program (beginning, middle and end), including whether a recording occurred, the audio quality, language, whether the correct program was recorded, the data type and topic.

4. Source Data Profile

This release contains 256 audio files. Following is a breakdown of files by source and distinct program:

Source	Program	Program ID	#Broadcasts	Total Hrs.
CCTV2	Economic 30 Minutes	ECON30MIN	40	20.7
CCTV2	News List	NEWSLIST	36	24.1
CCTV4	Daily News	DAILYNEWS	9	2.6
CCTV4	News3	NEWS3	23	12.4
CCTV7	Military News1	MILITARYNEWS1	36	15.3
CCTVNEWS	Evening News	EVENINGNEWS	21	7.3
Phoenix TV	Good Morning China	GOODMORNCN	8	4.3
Phoenix TV	From Phoenix to the World	PHNXWRLD	9	4.7
VOA	Current Events – AM	CURRENTEVENTSMORNING	10	10.2
VOA	Current Events	CURRENTEVENTS	19	15.6

VOA	International News and Finance	INTNLNEWSFINANCE	13	7.8
VOA	International News	INTNLNEWS	26	11.3

5. Data Directory Structure

The directory structure in this data release is organized as follows.

- Broadcast audio collection top directories

  /data

- Documentation directory

  /docs

- Broadcast audio collection top directories [names of these]

6. Data File Description

6.1 Audio File Format

The audio files in this release are FLAC compressed Waveform Audio File format (.flac), 16000 Hz single-channel 16-bit PCM files.

6.2 Audio File Names

The broadcast audio files in this collection follow LDC's defined naming convention for broadcast audio files.

{SRC}_{PRG}_{LNG}_YYYYMMDD_HHMMSS.flac

where -

- {SRC} is the source ID (e.g. CCTV1)
- {PRG} is the program ID (e.g., LEGALREPORT)
- {LNG} is the three-letter language ID defined in ISO639-3. CMN is Mandarin Chinese.
- YYYYMMDD is the data collection (broadcast) date.
- HHMMSS   is the start time of the program (HH is the hour in the 24-hour format)

7. Data Validation

Native Chinese speakers audited every recording in this release.

All audio files were checked to be valid .flac files.

The docs/CHECKSUM.md5 file contains MD5 checksums of all audio files in this corpus.

8. Copyright Information

Authors: Kevin Walker, Christopher Caruso, Kazuaki Maeda, Denise DiPersio, Stephanie Strassel