LDC: OPERATIONS

Broadcast Collection System

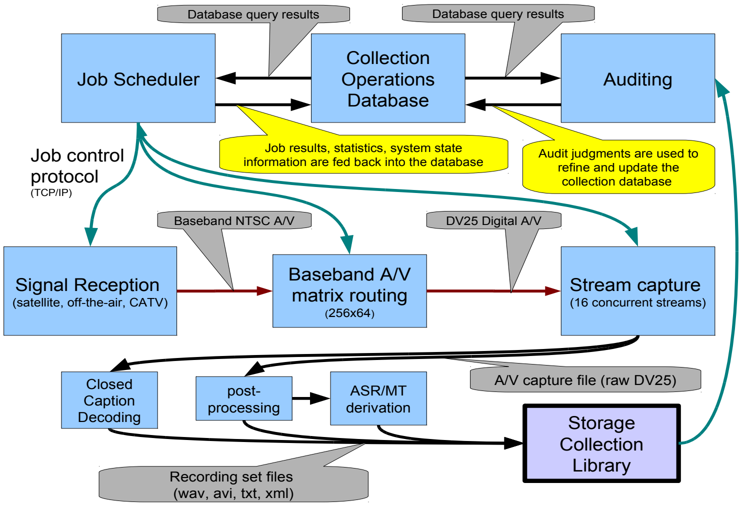
Collection System Overview

Broadcast news and conversation have provided source material to support multiple human language technologies over the last two decades. LDC has collected over 35,000 hours of broadcast data for technology development in fields such as continuous speech recognition, machine translation and information extraction. This material, a sizable portion of which is also annotated (e.g., transcribed, translated, treebanked), continues to be used in numerous common task evaluations and sponsored projects.

LDC's broadcast collection system represents a significant achievement in delivering volumes of high-quality broadcast data from multiple programming sources and geographic locations. Because it is robust, flexible and extensible, the system can be quickly deployed for virtually any type of broadcast collection.

From a simple monitor/VCR connection in 1998, the system has evolved to its present form – an array of antennae and other input sources, receivers, recording nodes and transcoding nodes supported by a MySQL database with associated closed caption decoders, automatic speech recognition (ASR) systems and local storage library. The cluster can log twenty four simultaneous audio/video (A/V) streams and process up to 1,000 hours of content daily.

System Operation


Database query results
Database query results
Job Scheduler
Collection Operations Database
Auditing
Job results, statistics, system state information are fed back into the database
Audit judgments are used to refine and update the collection database
Job control protocol (TCP/IP)
Baseband NTSC A/V
DV25 Digital A/V
Signal Reception (satellite, off-the-air, CATV)
Baseband A/V matrix routing (256x64)
Stream capture (16 concurrent streams)
A/V capture file (raw DV25)
Closed Caption Decoding
post-processing
ASR/MT derivation
Storage Collection Library
Recording set files (wav, avi, txt, xml)

LDC Broadcast Collection Functional Block Diagram

LDC designed the broadcast collection system to be modular, regularized and automated. All recording nodes are interchangeable, filenames and database fields follow consistent, formal rules and signal interconnects are also consistent. Humans audit the collected data and adjust the schedule as needed.

LDC Recording Lab, Philadelphia, PA, USA

The broadcast material is served to the system by a set of free-to-air satellite receivers, commercial direct satellite systems, direct broadcast satellite receivers and cable television feeds. The receivers feed into an A/V matrix switch so that any source can be routed to any receiver simply by changing an entry in the schedule.

Programs are recorded in a high bandwidth A/V format and are then processed to extract audio, to generate key frames and compressed A/V, to produce time-synchronized closed captions (for North American English programming) and to generate ASR output.

Broadcast news and broadcast conversation (talk shows) comprise the dominant genre of collected programming. Sources include Arabic, Chinese, English and Spanish global broadcast sources, among them, Aljazeera, Lebanese Broadcasting Corp. (Arabic); CCTV, New Tang Dynasty TV, Phoenix TV (Chinese); CNN, MSNBC/NBC (English); and Televisa, Univision (Spanish).

Signal Reception and Routing

Three steerable satellite dishes and a set of fixed dishes provide access to direct broadcast satellite service. Satellite signal reception is handled by a set of digital and analog satellite receivers. Each receiver is tuned to a specified transponder and the digital transmission for each channel is decoded into analog video and audio.

The system can route any signal from the receiver bank to any digitizer input in the digital video recorder cluster via the A/V matrix switch. The switch has 256 inputs and 64 outputs, and a single input signal can be distributed to multiple outputs simultaneously.

System Supervisor/Job Scheduler

A supervisor computer with a customized scheduling database drives all collection activity. From a dedicated server, the supervisor:

- polls the schedule for pending jobs
- marks jobs about to begin as "in progress"
- connects to a specified receiver and tunes it to the correct channel
- configures the A/V matrix switch to create a signal path from the receiver to the recorder
- connects to the digital video recorder
- polls for recording success, failure, timeout

A/V Matrix Switch Display

Utilities

The system relies on open-source software, including tools for signal capture and creating keyframes (**Dvgrab**); audio extraction and video compression (**mencoder**); digital signal compression and decompression (**x264**); audio and video streaming and conversion (**ffmpeg**); audio file conversion (**SoX**); and file transfer from recording nodes to file server (**rsync**).

Sponsored Broadcast Collections

Program	Language(s)	Volume
HUB4, 1996-1998	Chinese, English, Spanish	300 hours
TDT, 1998-2004	Arabic, Chinese, English	1000+ hours
TIDES, 2000-2005	Arabic, Chinese, English	10K+ hours
EARS, 2003-2005	English	10K+ hours
TRECVID, 2004-2006	Arabic, Chinese, English	400 hours
GALE, 2005-2011	Arabic, Chinese, English	20K+ hours

Note: Data may have been shared among programs.

Portable Collection Platform

LDC has designed a small-footprint portable collection platform that records two simultaneous A/V streams and supports a wide range of international broadcast standards. The portable platform and main collection system share a code base and hardware specification so that improvements to one system benefit the other. LDC uses the portable platform system at remote collection sites around the world.

Broadcast Language Resources

Most of the broadcast data collected by LDC is available through the Catalog, https://catalog.ldc.upenn.edu/. Browse the holdings by year, title, project, language or data source to find corpora of interest.