GALE Phase 2 Arabic Broadcast News Speech Part 2

1. Introduction

GALE Phase 2 Arabic Broadcast News Speech Part 2 contains approximately 170 hours of Arabic broadcast news speech collected in 2006 and 2007 by the Linguistic Data Consortium (LDC), MediaNet, Tunis, Tunisia and MTC, Rabat, Morocco during Phase 2 of the DARPA GALE program.

Broadcast audio for the DARPA GALE (Global Autonomous Language Exploitation) program was collected at LDC's Philadelphia, PA USA facilities and at three remote collection sites: Hong Kong University of Science and Technology (HKUST), Hong Kong (Chinese); Medianet (Arabic); and MTC (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 conversation recordings in this release feature news broadcasts focusing principally on current events from the following sources: Abu Dhabi TV, based in Abu Dhabi, United Arab Emirates; Al Alam News Channel, based in Iran; Aljazeera , a regional broadcaster located in Doha, Qatar; Al Ordiniyah, a national broadcast station in Jordan; Dubai TV, based in Dubai, United Arab Emirates; Al Iraqiyah, a television network based in Iraq; Kuwait TV, a national television station based in Kuwait; Lebanese Broadcasting Corporation, a Lebanese television station; Nile TV, a broadcast programmer based in Egypt; Saudi TV, a national television station based in Saudi Arabia; and Syria TV, the national television station in Syria.

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 to be executed and their status; and it stores all audit judgments associated with a given recording.

For the GALE program, Medianet collected Arabic broadcast news (BN) and broadcast conversation (BC) programming from across the Gulf region using its internal system and LDC's portable broadcast collection platform installed in 2008. Among the sources collected by Medianet were Abu Dhabi TIV, Al

Arabiya, Al Baghdadya, Al Fayhaa, Al Forat, Al Hiwar, Al Iraqiyah, Al Manar, Al Ordiniyah, Al Sharqiya, Bahrain TV, Dubai TV, Kuwait TV, Oman TV, Qatar TV, Palestine Satellite Channel, Saudi TV and Tunis TV.

MTC collected Arabic BN and BC programming from Al Baghdadya, Alhurra, Al Maghribia, Arabiaa, Radio Sawa and Yemen TV using its internal collection system.

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 Medianet's Tunisian collection facility collected multiple streams of regional Arabic programming from various sources.

Further information about LDC's broadcast collection system can be found in Guidelines for Broadcast Audio Collection, Version 3.0, http://projects.ldc.upenn.edu/gale/task_specifications/Collection_Task_Specificationv3.0.pdf, LDC Broadcast Documentation – Version 1.0, http://projects.ldc.upenn.edu/gale/task_specifications/LDC_Broadcast_System_Documentation_v1.0.pdf and 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 in English.

Each remote collection site used a form of audit procedure based on the LDC model. Medianet generated English-language .xls reports for the Arabic programming it collected. Those reports contained one set of auditors' judgments for an entire program, including audio quality; genre; data format; percentage of Modern Standard Arabic; dialect type and percentage; topic; and comments. MTC generated English-language .xml and .html audit reports for the Arabic 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.

Further information about the audit procedure and LDC's Broadcast Audit Interface Tool can be found in Audit Procedure Specification, Version 2.0, http://projects.ldc.upenn.edu/gale/task_specifications/Audit_Procedure_Specificationv2.0.pdf.

4. Source Data Profile

This release contains 204 audio files. Following is a breakdown of files by source and distinct program:

Source	Program	Program ID	#Broadcasts	Total Hrs.
Abu Dhabi TV	Abu Dhabi News	ABUDHNEWS	35	36.4
Al Alam	News Report	NEWSRPT	23	23.9
Al Jazeera	News 15	NEWS15	21	11.3
Al Jazeera	Today's Harvest	TODHARV	22	22.8
Dubai TV	Dubai News	DUBAINEWS	14	8.7
Al Iraqiyah	Economic Report	ECONRPT	9	2.6
Al Iraqiyah	Iraq Today	IRAQTDY	2	1.1
Kuwait TV	News	NEWS	20	10.8
LBC	News	NEWS	8	8.3
Nile TV	Egypt Nightly News	EGYPNNSCO	12	12.5
Al Ordiniyah	Jordan Nightly News	JORDNNSCO	5	5.2
Saudi TV	Saudi Nightly News	SAUDNNSCO	15	8.1
Syria TV	News 25	NEWS25	18	18.7

5. Data Directory Structure

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

- Broadcast audio collection top directories

  /data

- Documentation directory

  /docs

6. Data File Description

6.1 Audio File Format

The audio files in this release are Waveform Audio File format (.wav), 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.wav

where -

- {SRC} is the source ID (e.g., CNN, VOA, etc.)

- {PRG} is the program ID (e.g., LARRYKING, etc.)

- {LNG} is the three-letter language ID defined in ISO639-3. ARB is Standard Arabic; CMN is Mandarin Chinese; ENG is English.

- 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 Arabic speakers audited every recording in this release.

All audio files were checked to be valid .wav 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