The RATS Collection: Supporting HLT Research with Degraded Audio Data

David Graff, Kevin Walker, Stephanie Strassel, Xiaoyi Ma, Karen Jones, Ann Sawyer
Linguistic Data Consortium
University of Pennsylvania, Philadelphia, PA, USA
E-mail: {graff, walkerk, strassel, xma, karj, sawyera}@ldc.upenn.edu

Abstract

The DARPA RATS program was established to foster development of language technology systems that can perform well on speaker-to-speaker communications over radio channels that evince a wide range in the type and extent of signal variability and acoustic degradation. Creating suitable corpora to address this need poses an equally wide range of challenges for the collection, annotation and quality assessment of relevant data. This paper describes the LDC's multi-year effort to build the RATS data collection, summarizes the content and properties of the resulting corpora, and discusses the novel problems and approaches involved in ensuring that the data would satisfy its intended use, to provide speech recordings and annotations for training and evaluating HLT systems that perform 4 specific tasks on difficult radio channels: Speech Activity Detection (SAD), Language Identification (LID), Speaker Identification (SID) and Keyword Spotting (KWS).

Keywords: speech activity detection, language identification, speaker identification, keyword spotting, degraded audio channels

1. Introduction

The U.S. Defense Advanced Research Projects Agency (DARPA) established a program for Robust Automatic Transcription of Speech (RATS) in 2010, with a goal of developing human language technologies (HLT) for speech-based applications to analyze speaker-to-speaker communications over radio channels that evince a wide range in the type and extent of signal variability and acoustic degradation.

Prior to the RATS program, only a few speech corpora existed that involved relevant types of radio devices and channels (Godfrey 1994, Graff, Reynolds and O'Leary 1999), but each of these lacked several features required by the RATS program.

The RATS data specifications were designed to cover a diverse range of radio conditions, including varied combinations of transmitters and receivers, with varied orientations, configuration settings and interfering factors. In the conditions of interest, the signal-to-noise ratio (SNR) often falls below 10dB, and channel dependent distortions make it difficult to estimate SNR, or to relate this measure to human intelligibility of the signal.

Project data requirements were built around four specific HLT tasks: Speech Activity Detection (SAD), Language Identification (LID), Speaker Identification (SID) and Key Word Spotting (KWS). Each task would require a sufficient quantity of audio data and annotation to allow partitioning into training, development test, and two independent evaluation test sets.

For the SAD task, the original plan called for 1000 hours of transceiver audio for training and evaluation, with all transitions between speech and non-speech marked to a consistent level of accuracy. For LID, the collection would cover five target languages (32 hours per language) and 10 non-target languages (12 hours per language). For SID, the goal was to have 200 individual speakers in each of the 5 target languages, with 10 independent sample recording sessions per speaker. For KWS, only two languages would be addressed, with 100 words or phrases to be chosen from 100 hours of transcribed speech in each language; the target words/phrases would be selected based on frequency within the transcribed corpus, in order to yield an average overall hit rate of about 1 word/phrase per minute of speech. These goals evolved over the life of the program to reflect changing requirements for training data (for instance, it was determined that less significantly less SAD training data was needed to meet the program's goals) as well as challenges in data creation.

Three key design features allowed LDC's data collection for RATS to meet or exceed those goals: (1) configure a collection platform and processing pipeline that supports transmission, reception and capture on 8 independent radio channels simultaneously; (2) use clean, pre-recorded conversational speech data as input to the collection platform, and as input to the necessary annotation tasks for SAD, LID, SID and transcription for KWS; (3) develop processes to project and adapt the clean-audio annotations onto each of the degraded-audio radio channels, and apply automated metrics to validate the resulting sets of channel-specific annotations.

The remaining sections of this paper are organized as follows: Section 2 provides an overview of distinct RATS corpus releases to support each of the four research tasks; Section 3 covers technical details of the audio data pipeline: capturing audio simultaneously from multiple transmitter/receiver pairings, automatic alignment and quality assessment, and detection and treatment of various problems that emerged during the process; Section 4 describes annotation methods employed to provide both ground truth about the speech being transmitted and human assessment of the received signals; Section 5 explains automated editing of the manual annotations from source audio files to align them with the recordings from radio receivers: adjusting time offsets to account for variability in recording start-up and channel delays at the receivers, and marking portions speech regions where radio transmission was unexpectedly interrupted; and Section 6 gives a brief summary of HLT system performance targets and results to date in the RATS program, and general conclusions.

2. Overview of RATS Corpus Releases

The LDC has developed four corpora, one for each of the RATS research tasks, comprising clean source audio, the corresponding sets of 8 transceiver channels, and all channel-aligned annotations. Data for each research task is partitioned into training, development test and evaluation sets, including both test data for annual performance evaluations and additional (sequestered) progress data to measure performance improvements over the life of the program. Roughly 10% of the source audio was used in multiple tasks, but with distinct annotations suited to each task. (In some cases, a given source audio file was transmitted on two different occasions months apart, yielding different transceiver audio for each task.) Evaluation tasks in RATS are treated as independent, and so a given file may have been used as training data in one task and as development-test data in another. The status of each file with respect to a given data set or evaluation task is maintained in a comprehensive project database.

For each of the four research tasks, creation of the necessary transceiver audio data involved nearly continuous operation of the transceiver collection platform over a period of weeks or months (Walker and Strassel 2012). During these sustained collection periods, there were occasional hardware failures affecting some of the transmitter or receiver devices, causing dropout periods on certain channels, and the scheduling of collection and other activities was too tight to allow second attempts in many of these cases. Altogether, four of the eight channels were affected by varying periods of failure, yielding smaller quantities of data for these channels.

Set	Language	Source Files	Source Hours	Retrans Hours
Test	English	215	37.2	297.6
Test	Farsi	10	2.6	20.8
Test	Levantine Arabic	213	44.1	352.8
Test	Pashto	35	8.7	69.6
Test	Urdu	59	15	120
Train	English	605	104.9	839.2
Train	Farsi	29	7.5	60
Train	Levantine Arabic	573	117.7	941.6
Train	Pashto	117	29.9	239.2
Train	Urdu	140	35.2	281.6
Total		**1996**	**402.8**	**3222.4**

Table 1. SAD Source Audio Composition

The SAD corpus provides data in English, Urdu, Pashto, Farsi and Levantine Arabic; source audio has been drawn from LDC's Fisher English (Cieri et al. 2004, 2005) and Fisher Levantine Arabic (Maamouri et al. 2006a,b) corpora, plus new conversational telephone speech (CTS) data collected specifically for RATS. A total of 402.8 hours of source audio was processed for this task; the 8 transceiver channels yielded over 3222 hours of retransmitted audio. Table 1 shows a breakdown of content by language. Overall, roughly 45% of the audio content is speech.

The LID corpus focuses on five target languages: Levantine Arabic, Farsi, Dari, Pashto and Urdu. Non-target language data were drawn from audio data that had been used, either as training or test material, in the NIST 2009 Language Recognition Evaluation (LRE). Four of the 5 RATS target languages (all but Arabic) were represented in the NIST 2009 LRE data along with 19 other languages; these recordings were drawn from broadcast sources, particularly Voice of America, by selecting only portions of speech from narrow-band sources, such as reports or interviews conducted over the telephone during the broadcast. Other target-language audio for LID came from LDC's Callfriend Farsi (Canavan and Zipperlen 1996) and Fisher Levantine corpora, combined with new CTS data recorded for RATS. In order to improve the density of target language training data, transceiver audio from CTS sources was edited and concatenated to create 2-minute segments with relatively little non-speech content.

Set	Language	Source Files	Source Hours	Retrans Hours
Test	Dari	237	8.5	68
Test	Farsi	1009	35.1	280.8
Test	Levantine Arabic	878	29.3	234.4
Test	Mixed Non-Target	2470	161.1	1288.8
Test	Pashto	864	29.4	235.2
Test	Urdu	887	31	248
Train	Dari	133	4.9	39.2
Train	Farsi	399	14.6	116.8
Train	Levantine Arabic	3849	128.3	1026.4
Train	Mixed Non-Target	2690	141.6	1132.8
Train	Pashto	2581	86.5	692
Train	Urdu	1717	58.3	466.4
Total		**17714**	**728.6**	**5828.8**

Table 2. LID Source Audio Composition

To provide supplemental training material for LID, LDC also harvested 292 hours of recordings from the U.S. International Broadcasting Bureau (IBB) web sites, consisting of audio captures of VOA and similar broadcasts via radio receivers stationed at numerous locations around the world. This “found data” lacked clean source audio and was provided to RATS performers as-is, without retransmission, since the recordings already represented the kind of degraded radio channels of interest to the program.

The KWS corpus contains only Farsi and Levantine Arabic, all drawn from CTS source data: CallFriend Farsi, Fisher Levantine, and new calls recorded for RATS. Full transcripts are provided for approximately 464 hours of data: 266 hours of Arabic in 1607 CTS recordings, and 198 hours of Farsi in 728 CTS recordings. Roughly 45% of audio content is speech. Transcripts are in UTF-8 (Perso-)Arabic script, with no diacritics.

Set	Language	Source Files	Source Hours	Retrans Hours
Test	Farsi	320	77.6	620.8
Test	Levantine Arabic	719	121.8	974.4
Train	Farsi	408	120.5	964
Train	Levantine Arabic	888	144.6	1156.8
Total		**2335**	**464.5**	**3716**

Table 3. KWS Source Audio Composition

The SID corpus consists entirely of CTS data newly recorded for the RATS program; this was necessary in order to establish “ground truth” with regard to speaker identification in each recording. The SID evaluation design called for 500 total speakers across 5 target languages, with each speaker making a minimum of 10 calls; 4 of the 10 sessions are used for speaker enrollment and 6 for testing, randomly sampled from the noisy channels. Recruited speakers reside primarily in the US, South Asia and the Middle East.

Set	Language	Source Files	Source Hours	Retrans Hours
Test	Dari	972	197.8	1582.4
Test	Farsi	640	134.7	1077.6
Test	Levantine Arabic	881	185.3	1482.4
Test	Pashto	2017	422.6	3380.8
Test	Urdu	1758	364.1	2912.8
Train	Dari	1167	236.1	1888.8
Train	Farsi	665	136.1	1088.8
Train	Levantine Arabic	689	143.8	1150.4
Train	Pashto	1911	387.3	3098.4
Train	Urdu	1904	388.1	3104.8
Total		**12604**	**2595.9**	**20767.2**

Table 4. SID Source Audio Composition

Despite the incentives provided for completing at least 10 calls, most recruited speakers dropped out after making a single call. It was therefore necessary for us to recruit over 6500 individuals to meet the collection goals; recruitment results varied across the five languages, ranging from a low of 642 speakers for Farsi to a high of 2013 for Pashto. All speakers making just one call were designated for use as training data, while speakers completing all 10 calls were used for test; speakers making between 2-9 calls were primarily used for training though some were used for devtest. Table 4 summarizes the amount of data recorded and retransmitted per language, while Table 5 provides details of the number of speakers making 1, 2-9 or 10+ calls for each language.

Set	Language	Speakers with		
		1 Call	2-9 Calls	10+ Calls
Test	Dari	0	18	82
Test	Farsi	0	16	46
Test	Levantine Arabic	0	33	68
Test	Pashto	0	38	155
Test	Urdu	0	25	151
Train	Dari	1040	65	0
Train	Farsi	530	38	2
Train	Levantine Arabic	583	48	0
Train	Pashto	1753	67	0
Train	Urdu	1675	104	0
Total		**5581**	**452**	**504**

Table 5. Speaker Yield by Language

3. Technical Details of Collection Protocol

3.1 Transmission Platform and Procedure

Nearly all source audio data used as clean input to the RATS collection platform and annotation consisted of CTS originally recorded from public telephone networks. (The only exception was the NIST LRE 2009 audio from VOA broadcasts used for the LID task, but these recordings had originally been selected for LRE by virtue of containing band-limited speech comparable to CTS.) Audio capture from the 8 transceiver channels was stored as 16-KHz, 16-bit sample data; for convenience, the source audio was also normalized to this format, both for quality assessment steps to be performed on the transceiver audio, and for final publication.

The transmitter and receiver systems were placed at opposite ends of the LDC office suite, separated by about 50 meters; effective radiated power (ERP) for the transmitters was set very low, both to induce a suitable degree of degradation at the receivers and to ensure compliance with regulatory standards. The transmission and recording of the 8 channels was carried out more or

Channel ID	Transmitter		Receiver		RF Band / Modulation	Transmission Protocol
	Make	Model	Make	Model		
A	Motorola	HT1250	AOR	AR5001/D	UHF / NFM	push-to-talk
B	Midland	GXT1050	AOR	AR5001/D	UHF / NFM	push-to-talk
C	Midland	GXT1050	TenTec	RX400	UHF / NFM	push-to-talk
D	Galaxy	DX2547	Icom	IC-R75	HF / SSB	push-to-talk
E	Icom	IC-F70D	Icom	ICR8500	VHF / NFM	push-to-talk
F	Trisquare	TSX300	Trisquare	TSX300	UHF / FHSS	PTT/hand-shake
G	Vostek	LX-3000	Vostek	VRX-24LTS	UHF / WFM	continuous
H	Magnum	1012 HT	TenTec	RX340	HF / AM	push-to-talk

Table 6. Radio Channel Configurations

less continuously, with the collection system operating around the clock for days or weeks at a time under database-driven program control, during the larger part of 2012 and 2013. The process was organized around "retransmission sessions." Each session involved a single source audio file, which was either one side of a CTS conversation (ranging between 5 and 30 minutes long), or a concatenation of four NIST LRE test segments (yielding a source file between 2 and 5 minutes long).

The eight radio channels are labeled A through H. Channels A and B are ultra high frequency (UHF) channels, operating at 0.66 meter wavelength. **Channel A** shows up to 3kHz carrier deviation from center frequency, with an ERP of 4 watts. The receiver for Channel A is configured operate in dual frequency mode – one is tuned to the target frequency, the other is offset by 50KHz. **Channel B** shows up to 2.5KHz carrier deviation from center frequency, with an ERP of 0.5 watts. The channel B receiver is configured to use a high level of noise reduction, which rejects off-channel interference but introduces tonal variations in the decoded audio.

Channels C, F and G are also UHF. **Channel C** has a wavelength of 0.66 meters, a receiver frequency offset 3khz relative to the transmission frequency, and a 10Khz IF Bandwidth setting. The carrier offset stresses the receiver's capability to stay locked on the transmit frequency. The tonal distortions found in audio from this channel are caused by the receiver FM detector continuously attempting to lock onto the transmit frequency. **Channel F** operates at the 900MHz ISM Band, FHSS, 0.33 meter wavelength. These transceivers execute 2.5 frequency hops per second. (As a point of reference, the Motorola DTR Handheld Transceiver Line hops 11 times per second, and the JTRS SINCGARS hops 111 times per second in FHSS mode.) **Channel G** operates at 0.12 meter wavelength, Wideband FM, and 5 watts ERP. This transmitter is designed to carry both video and audio; we use only the audio input. The audio subcarrier uses up to 25kHz carrier deviation.

Channels D and H are high frequency (HF) channels. **Channel D** operates at 11.41 meter wavelength, Lower Side Band. The target frequency of both the receiver and the transmitter drift over time, depending on the operational temperature of the equipment. This continuous shifting produces different degrees of tonal shifting and distortion. **Channel H** uses a 10.95 meter wavelength, Narrow FM. The longer wavelength allows signal to penetrate through obstructions; however, stray electro-magnetic interference poses more of a problem than is found in the UHF systems.

Finally, **Channel E** is very high frequency (VHF), operating at a wavelength of 2-meters, and suffers from diffraction, building penetration loss, and multipath loss. The receiver is configured with 20-dB attenuation enabled, and with an IF of 12kHz.

As source audio was played out over the transmission platform, it was distributed to the 8 transmitters as well as to a voice-activated relay (VAR) device, which would serve to control the state of the push-to-talk (PTT) controls on transmitters A-F and H; channel G used continuous transmission, while the transmitter on channel F used an additional built-in mechanism that ensures coordination of the operating modes on the two TSX300 devices.

Table 6 summarizes the equipment and settings for the transmitters and receivers across the eight channels. (See Walker and Strassel 2012 for a more complete description of the hardware setup and the methods used for calibration.)

3.2 Cross-Channel Alignment

In order for the transmitted data to be useful for research and evaluation, accurate cross-channel alignment is critical. The transmission process introduces several general and channel-specific alignment issues that must be accounted for. An initial quality-control (QC) step was to measure signal energy frame-by-frame over each transceiver channel. A transceiver was determined to have failed on a given recording session if the overall energy was low throughout, or if the difference between the minimum and maximum frame energy didn't exceed a specified threshold for the given channel.

Next, a custom implementation of cross-correlation analysis (Ellis, 2011) was used to compare each channel to the source audio, in order to establish the exact time offset between the beginning of the source audio file and the beginning of the transceiver recording. For example, the start-up of recordings on the receivers could occur several seconds before playback of the source audio, and the channels differed slightly in the amount of time delay induced during transmission and reception. The cross-correlation would provide a time offset value that could be added to the source annotations in order to align them properly relative to the beginning of each receiver recording. The analysis would also reveal any cases where alignment between channels was inconsistent or disrupted, due to failure of the transceiver hardware, problems with the analog-to-digital (A-to-D) recording system, or deviation in the A-to-D sampling clock rate.

When cross-correlation results indicated success, the computed time offset was used to extract putatively equivalent speech segments from the source and a given transceiver channel, and cross-correlation was run again on these segments; when the two extracted segments showed a strong correlation with a time offset within ±2 msec, this confirmed that channel's computed time offset relative to the source was correct.

4. Annotation Methods

An important aspect of the RATS corpus design involves performing annotation on the clean source audio (for accuracy and efficiency), then "projecting" those annotations onto the eight degraded audio transmission channels that have been aligned to the source channel using the procedure described above. The various annotation tasks are described in the sections that follow.

4.1 SAD

All RATS data was processed through LDC's automatic SAD system (Ryant 2013) to generate reasonably accurate segments indicating the presence of speech in the audio signal. Although analysis shows less than 10% difference between fully automatic and fully manual SAD labels at the frame level, training and test data designated for the SAD task was subject to additional manual annotation to increase the accuracy of the labeled data. Annotators used LDC's XTrans tool (Maeda et. al., 2008) to correct the automatically generated speech segments, adjusting endpoints and creating or removing segments as needed. Early versions of the SAD task specification categorized some human vocalizations like faint background speech and singing as non-speech, though later versions classified any human vocalization as speech. Non-speech in the recordings was left unsegmented and unlabeled. Experienced SAD annotators conducted a careful quality review on the first pass annotators' work.

4.2 LID

The LID annotation task was the simplest to execute. Accurate language labels already existed for a portion of the LID data, having been produced during the original collection effort (for instance, during LRE 2009). For unlabeled data, segments assumed to be in a given language were presented to native speakers of that language via a web-based GUI developed for this task. Annotators listened to each segment in its entirety then assigned one of four labels: *in the target language*, *not in the target language*, *unintelligible* or *non-speech*. A portion of the data was labeled by multiple annotators working independently to establish baseline human agreement and to monitor annotator performance. Each labeled segment contains about two minutes of speech, although the length of each audio file varies depending on the amount of intervening non-speech.

4.3 KWS

For the keyword spotting tasks, annotation consisted of selecting appropriate keywords from verbatim orthographic transcripts of Levantine Arabic and Farsi CTS, under the guidance of the RATS evaluation team. While there was sufficient existing transcribed data to support Levantine Arabic keyword selection, new transcription was required to generate sufficient data volumes for Farsi. In addition to new transcription, existing Farsi transcripts from the CallFriend Farsi corpus had to be adapted to the current task. The CallFriend Farsi corpus used a phonemic Romanized orthography whereas RATS called for the use of the native Perso-Arabic script. LDC converted the Romanized Farsi transcripts to Perso-Arabic script using a word list that mapped original transcript word forms to their Arabic-script correlates. This list didn't cover all the word forms in the original corpus, so additional annotation was required to produce fully Arabicized text. Annotators were presented with each Romanized token in full-sentence context, and were given a selection of automatically-generated Farsi tokens to choose from. If none of the automatically-generated forms was the correct match, annotators typed in the correct token in Perso-Arabic script. (The resulting augmented Farsi transcripts and corresponding speech have been published in LDC's catalog as LDC2014T01 and LDC2014S01). For the existing Levantine Arabic transcripts, a post-processing step removed any existing diacritics, as required by RATS.

After transcripts were finalized, keyword selection was done manually. The RATS evaluation team first prepared a list of candidate words and phrases extracted from the transcripts for each language. Native speaker annotators at LDC then reviewed the list to reject candidates that a) contained fewer than 3 syllables; b) were not likely to have been spoken words (e.g. transcript metadata artifacts like "coughing"); or c) represented a rare or archaic spelling variant.

4.4 SID

During collection of the RATS SID data, every call side was associated with a unique, persistent speaker ID number, making it possible to automatically establish baseline "ground truth" speaker labels which could be quickly verified through manual review of the calls. The manual audit was conducted in two stages. In the first stage, annotators listened to a portion of each call and made a general assessment of its overall recording quality; they also verified that the speaker sex and language matched expectations given the demographics reported for this speaker ID. In the second stage, all segments associated with a given speaker ID that passed stage one were simultaneously presented to a single native speaker annotator. The annotator listened to all segments to confirm that they were all from the same speaker. Speakers with ten or more calls were prioritized for auditing.

4.5 Adjudication

LDC annotators also performed a post-hoc review of system output after each annual evaluation, to identify cases where the ground truth annotation was incorrect (rare) or where the transmission process resulted in a segment whose quality was too low to permit fair evaluation (more common). The RATS evaluation team pooled system results for each of the four evaluation tasks and provided LDC with a prioritized list of segments for review. LDC annotators reviewed each segment in isolation using a customized web-based GUI. Each adjudication segment was judged for its intelligibility (e.g. "Does this segment contain intelligible speech?" and for its annotation status (e.g. "Is this keyword spoken in the segment?").

4.6 Intelligibility Judgments

To respond to performer concerns about the possibly low human intelligibility of the RATS transmissions produced by LDC, a final annotation task was conducted during the first phase of the program. This annotation did not pertain to any one particular RATS task, but rather affected all tasks equally, focusing on the collection system and the data it produced.

Each of the eight RATS transmission channels has distinct properties of acoustic distortion in the received signal; the most common measure of signal quality, the signal-to-noise ratio (SNR), was not adequate for assessing the relative intelligibility of the signal: two channels might have equivalent SNR but differ significantly in terms of how much phonetic detail they preserve.

In the process of configuring the collection platform, various channel settings were chosen based on informal judgment of the resulting signal; there were a few iterations in the initial stages of the collection to do a broader assessment of the recordings (by various RATS researchers as well as LDC staff), and adjust parameters accordingly; although this process was informal and unstructured, it nonetheless established a consensus for a configuration that would remain fairly stable throughout the collection of task-specific data sets.

In 2011 RATS performers requested that we conduct a formal study in order to get a more structured measure of human intelligibility for each channel. A set of English language utterances from a single male speaker were drawn from existing telephone speech data. Each utterance was approximately 20 seconds in duration and contained mostly speech. These source recordings were then transmitted over all eight channels on the RATS platform. Twenty native English-speaking judges were recruited to produce intelligibility judgments on the resulting segments. Each judge was presented with 96 channel recordings (12 samples from each of the 8 channels). Judges never heard the same utterance more than once. Segments were judged in isolation using a 5-point intelligibility scale:

I think I can understand…

1 Less than half of the speech
2 About half of the speech
3 Somewhat more than half of the speech
4 Almost all of the speech
5 All of the speech

The results of the intelligibility study are summarized in Table 7. A similar smaller-scale study was conducted for Farsi. The results of these studies suggest that most of the RATS data falls into the desired "noisy but understandable" range, though variance is perhaps larger than expected (even within a single channel). No significant changes to the channel configurations or transmission process were recommended as a result of the study.

Channel	Mean	Stdev
A	3.513157895	1.288650092
B	3.364035088	1.440119133
C	3.881578947	1.129895382
D	3.890350877	1.134673335
E	2.605263158	1.360994849
F	4.010526316	1.112647226
G	4.745614035	0.510875615
H	3.48245614	1.335601672

Table 7. Intelligibility by Channel

5. Annotation for Degraded Channels

In an ideal world the projection of annotation from the clean source channel onto the eight degraded transmission channels would be trivial given accurate cross-channel time alignments. In reality, a number of factors in the transmission process conspire to create challenges to accurate projection; this required some new techniques to analyze and post-process the data.

In addition to the eight audio files captured at the receiver, each transmission session produces a log file reporting the activity of the voice-activated relay. In the initial stage of collection, we established through manual review that one of the push-to-talk-mediated channels (E) showed a clear and consistent difference between "button on" and "button off" states; a minimal amount of signal processing was needed to map the VAR log entries to transitions for this channel in order to establish the exact alignment between timestamps in the log and positions in the audio. On channel F, which had a somewhat independent on/off behavior (due to the hand-shake layer in its design), the transition points caused distinctive transients in the audio, so a separate signal-processing tool was built to detect these transients. The collection database stored the timestamps for all "button-on" regions on each of the affected channels for each session.

Some transceivers had variable behavior with regard to sustaining their carrier signals during long periods of "button on" activity, resulting in audio dropouts on the degraded channels. An additional process (Ellis 2012) was devised to check these channels for non-transmitted regions within the time periods when the push-to-talk button was supposed to be engaged. Results were integrated with other information about the timing of button transitions, and with the transceiver channel time offsets, in order to accurately project timestamps from the source audio annotations onto each of the transceiver channels.

In SAD and KWS data, which has segment-based annotations for the speech regions in each source file, we created a separate annotation file for each receiver channel in each session by applying a stream-editing process to the corresponding source annotation file. The editing procedure used database queries to apply the following changes to the source annotations:

- Adjust all time stamps according to the channel's alignment offset relative to the source audio file.
- Subdivide speech and non-speech regions as needed to label non-transmitted (NT) portions, whether due to button-off transitions (VAR induced) or dropouts (transmitter induced).
- In KWS data, remove transcription text from speech segments affected by NT regions.

Regarding the last step, we did not have manual word-level time markings in any of the transcripts, so we could not reliably determine which words in a multi-word segment were affected by an NT region. In effect, NT within any portion of a KWS segment nullified the entire segment.

After this process, each segment has one of five values:

- S: a "button on" interval marked by annotators as containing speech
- NS: a "button on" interval marked by annotators as containing no speech
- T: a "button on" interval for which no manual SAD annotation exists
- NT: a "button off" interval
- RX: a "button off" interval not recorded by the log file but identified in post-processing

Figure 1 presents a visual depiction of the process. In the source audio waveform, the green boxes represent segments marked by annotators as containing speech while yellow boxes reflect non-speech segments. In the degraded channel waveforms A-G, green boxes contain "S" segments (button on + speech); yellow boxes are NS (button on + no speech); and red boxes are either NT or RX (button off as detected by logs or post-hoc analysis).

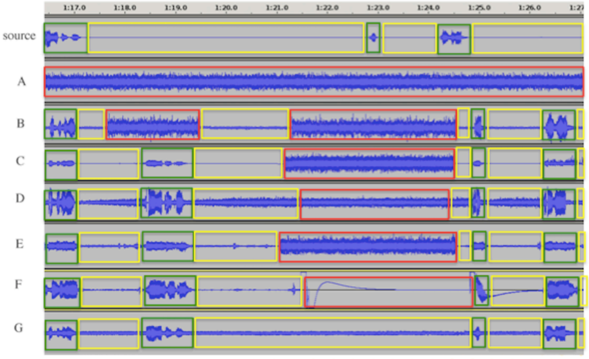

Figure 1. Annotation Projection

In LID and SID data, the core annotation is a single language or speaker label for each source file. The locations of speech, non-speech and NT regions were still important for training and testing, but it was sufficient to base the source annotation on automatic speech detection applied to the clean signal. Again, a separate annotation file was created for each channel, to reflect the channel-specific dropouts and their different delays relative to the source audio.

6. Research Results and Conclusions

The RATS program was ambitious not only in corpus creation, but also in performance goals for HLT system development. At project start-up, DARPA set targets for miss/false-alarm metrics in all four tasks across three annual evaluations. The scheduling of evaluations was also ambitious, creating tension relative to the time required to create the corpus. Training and development test data were provided to researchers incrementally as they became available; some data sets (like Farsi KWS and SID) were incomplete at the point of the first phase evaluation. Significant performance improvements were made between Phases 1 and 2 in part due to the availability of complete training and devtest data. For the SAD task, Phase 2 goals were exceeded by all systems, while for LID the Phase 2 targets were exceeded by each team's primary system at most segment durations. Results for the KWS and SID tasks were more variable.

Some aspects of radio channel audio which are of noted interest to the sponsor have not yet been addressed by the RATS data collection due to resource constraints. In particular, varying the relative locations of transmitters and receivers, and recording while one or both devices are in motion, are factors for which we considered and presented collection plans, but it was ultimately impractical to fit these into time frame and funding available for the project. These issues, along with a wider sampling of the potential diversity in other factors affecting radio quality, remain open for future work. We would also hope for progress in establishing a quantifiable metric for speech intelligibility that can be applied in this domain.

Data creation for the RATS project led to a wide range of novel challenges in corpus development. The summary presented here necessarily omits a considerable number of detailed problems that arose from unanticipated modes of failure or unintended consequences of the design and implementation in our collection and annotation pipelines. But the basic strategy, which had been proven effective in previous smaller speech corpora (such as the various degraded-signal versions of the TIMIT corpus), yielded a major success overall.

As RATS concludes in mid-2014, we are adding the program's corpora to the LDC Catalog. The first of these is the RATS Speech Activity Detection Corpus, which is slated for publication in late 2014 or early 2015.

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20016. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition Services Directorate.

References

Canavan, A., Zipperlen, G. and Graff, D. (2014). CALLFRIEND Farsi Second Edition Transcripts (LDC2014T01). Linguistic Data Consortium, Philadelphia, PA.

Canavan, A. and Zipperlen, G. (1996). CALLFRIEND Farsi Speech Corpus (LDC1996S50). Linguistic Data Consortium, Philadelphia, PA.

Canavan, A. and Zipperlen, G. (2014). CALLFRIEND Farsi Second Edition Speech (LDC2014S01). Linguistic Data Consortium, Philadelphia, PA.

Cieri, C. et al. (2004, 2005). Fisher English Training Speech, Part 1 (LDC2004S13), Part 2 (LDC2005S13). Linguistic Data Consortium, Philadelphia, PA.

Ellis, Dan (2011). SKEWVIEW - Tool to visualize timing skew between files. Columbia University, New York, NY. http://labrosa.ee.columbia.edu/projects/skewview/

Ellis, Dan (2012). FINDNTS - Tool to locate NT (no transmission) regions in audio. Columbia University, New York, NY. http://labrosa.ee.columbia.edu/projects/findNTs/

Godfrey, J. (1994). Air Traffic Control (LDC1994S14). Linguistic Data Consortium, Philadelphia, PA.

Graff, D., Reynolds, D. and O'Leary, G. (1999). Tactical Speaker Identification Speech Corpus (TSID) (LDC1999S83). Linguistic Data Consortium, Philadelphia, PA.

Maamouri, M. et al. (2006a). Levantine Arabic QT Training Data Set 5, Speech (LDC2006S29). Linguistic Data Consortium, Philadelphia, PA.

Maamouri, M. et al. (2006b). Levantine Arabic QT Training Data Set 5, Transcripts (LDC2006T7). Linguistic Data Consortium, Philadelphia, PA.

Maeda, K. et al. (2008). Annotation Tool Development for Large-Scale Corpus Creation Projects at the Linguistic Data Consortium. LREC 2008: 7th International Conference on Language Resources and Evaluation, Marrakech, May 28-30.

Ryant, N. (2013). LDC HMM Speech Activity Detector. Linguistic Data Consortium, Philadelphia, PA. https://github.com/Linguistic-Data-Consortium/ldc_sad_hmm

Walker, K. and Strassel, S. (2012). The RATS Radio Traffic Collection System. Odyssey 2012: The Speaker and Language Recognition Workshop, June 25-28.