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, USA

Linguistic Data Consortium

RATS Overview

- Robust Automatic Transcription of Speech (RATS) is a 3-year DARPA program
- Evaluating speech technologies in extremely noisy and/or highly distorted radio channels
  - Speech activity detection (SAD)
  - Language identification (LID)
  - Speaker identification (SID)
  - Keyword spotting (KWS)
- Levantine Arabic, Farsi, Urdu, Pashto and Dari
- Open eval on LDC-produced data (Phases 1-3)
- Closed eval on operational data (Phases 2-3)

Desired Data Characteristics

- Transactional, communicative, goal oriented speech
  - Density of talk, length of turns, turn-taking structure, amount of intervening silence resembling Ham radio or taxi driver radio chatter
- Variable radio channel transmission quality
  - Akin to quality found on air traffic control channels
  - With interference caused by multiple factors
    - Topographical, geological and environmental (e.g. humidity) variation
    - Manmade EMF/RF background radiation variation
  - Including squelch from push-to-talk devices
- Speech should be largely understandable by humans, but with some impairment of ability to
  - Detect or comprehend speech
  - Identify and/or distinguish between speakers, languages

Linguistic Data Consortium

Approach

- Build pipeline to simultaneously transmit, receive and capture audio on 8 independent radio channels
  - Channels designed to mimic operational environments
- Use clean, pre-recorded conversational speech as input to pipeline, and as input to annotation
  - Annotation on clean channel reduces cost, increases quality
- Develop processes to align channels and to project clean-audio annotations onto each degraded-audio radio channel
  - Requires extensive manipulation and validation

Input Data

- Existing data suitable for SAD, LID, KWS
  - NIST Speaker and Language Recognition test sets
  - CallFriend and Fisher Levantine Telephone Speech Corpora
  - Voice of America Broadcasts
- New telephone collection in 5 languages for SID
  - 6537 speakers recruited in Philadelphia and in country
  - Primarily unstructured conversations between friends/family or strangers
  - Some scenario-based sessions to elicit transactional, communicative, goal oriented speech
    - Collaborative games like “Twenty Questions”

Annotation on Input Data

- Annotation performed by native speakers using customized GUIs
- **SAD**: manually correct automatic speech/non-speech annotation
- **LID**: label short speech segments as target or non-target language
- **SID**: listen to (portions of) all recordings associated with one speaker ID and verify that it's the same person
- **KWS**: create time-aligned orthographic transcripts and/or convert existing Romanized transcripts to native orthography
  - Keywords selected post-hoc based on frequency
- Includes some independent dual annotation and post-hoc adjudication of system output

Multi-Radio Channel Collection System Design

- Input data is broadcast simultaneously over 8 radio channels
- Parallel, concurrent transmissions via HF, VHF, and UHF transceiver bank
- Remote listening post receiver bank captures these concurrent transmissions
- Transmitter/receiver pairings emulate conditions found in real-world radio communications
  - Manipulating RF signal strength, signal modulation, channel bandwidth, antenna efficiency, and reception parameters
  - Resulting in data impacted by RF interference, intermodulation, variations in noise floor, and competing transmissions
  - Affecting listener's ability to detect/understand speech, recognize language and speaker

Multi-Radio Channel Collection System Operation

- Transceiver bank, listening post placed at opposite ends of the LDC office suite, separated by about 50 meters
  - Effective radiated power (ERP) for transmitters set very low, to introduce desired degradation and to comply with regulatory constraints
- Process organized around “retransmission sessions”, consisting of
  - One side of a CTS conversation (5-30 minutes), or
  - Concatenation of short LRE test segments (2-5 minutes)
- System in operation around the clock for days or weeks at a time under database-driven program control, throughout 2012-2013

The image cannot currently be displayed.

LDC RATS Collection System Receive Station

Dual CPU, Quad Core, 8GB RAM, Eight 10K RPM, 146GB SAS Drives, Running Ubuntu 10.04 LTS. Drive system is configured so that audio is captured across four independent drives

Dell R710 Receiver Control Computer

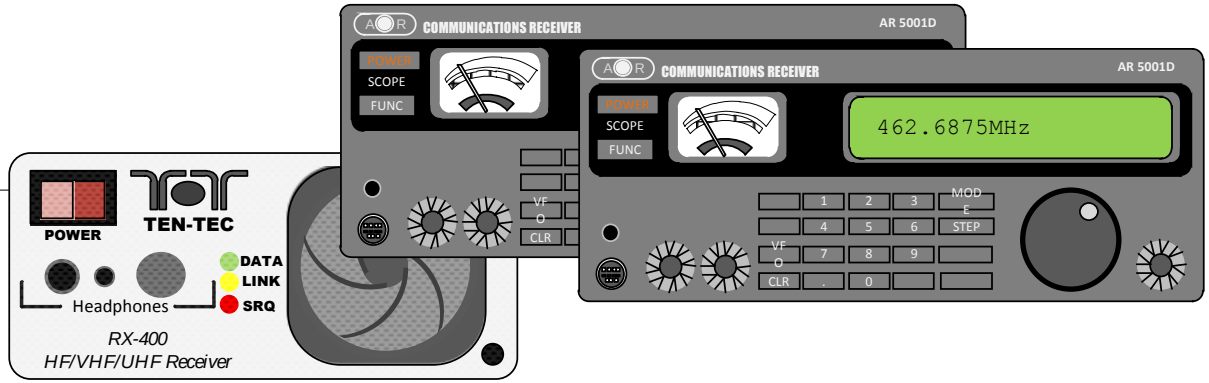
Digigram VX882e Audio Interface

PCI-Express Peripheral; Provides 8 Channels of Balanced Analog I/O and 8 Channels of AES/EBU Digital Audio.

All receivers equipped with RS-232 control ports are connected to an RS-232 to TCP/IP bridge

Comtrol DeviceMaster RTS

Three Wideband Receivers are used to collect UHF Narrow FM with different IF bandwidths.



Two HF receivers are used to capture Single Sideband and HF Narrow FM

2.4GHz Wide FM Is captured using the Vostek Receiver

900MHz FHSS is captured from a registered eXRS handset

One Wideband receiver is used to capture VHF Narrow FM

Radio Channel Map

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

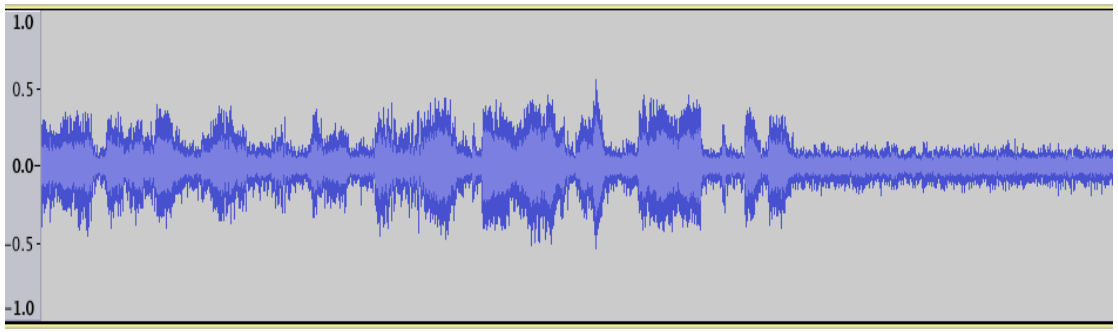
Both A and B are **UHF**, operating at 0.66 meter wavelength

Reference

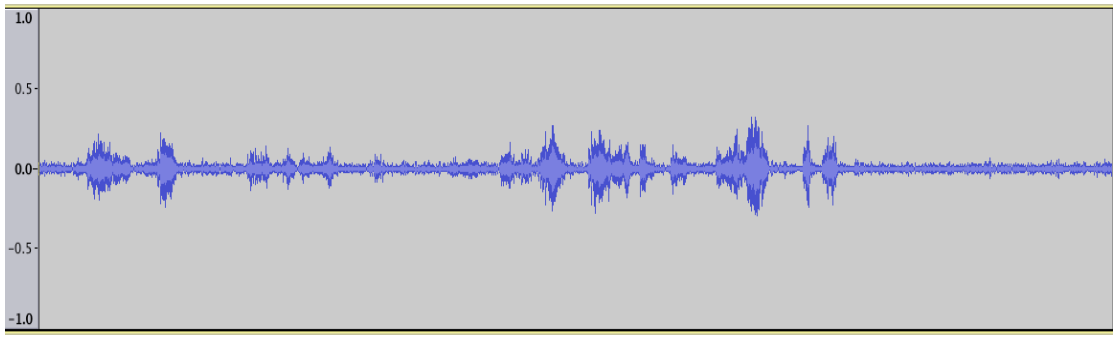
Channel A: up to 3kHz carrier deviation from center frequency, ERP of 4 watts. The receiver for Channel A is configured operate in **dual frequency** mode – one is tuned to the target frequency, the other is offset by 50KHz.

Channel B: up to 2.5KHz carrier deviation from center frequency, 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.

CH_A



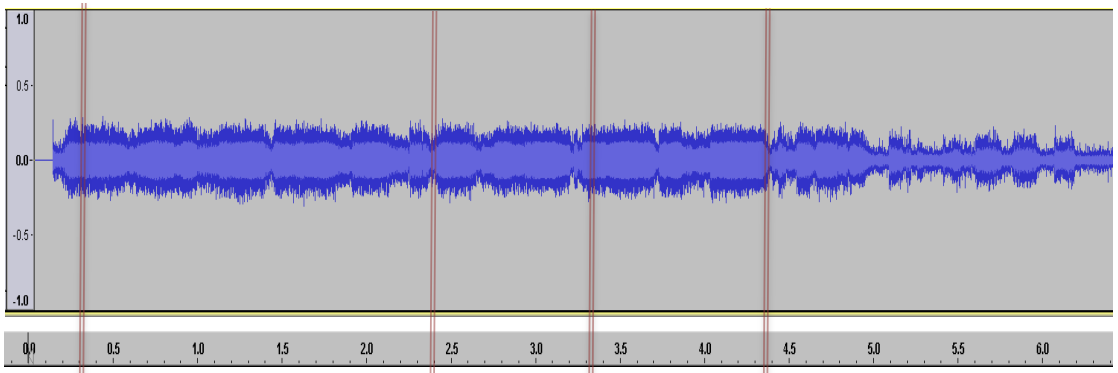
CH_B



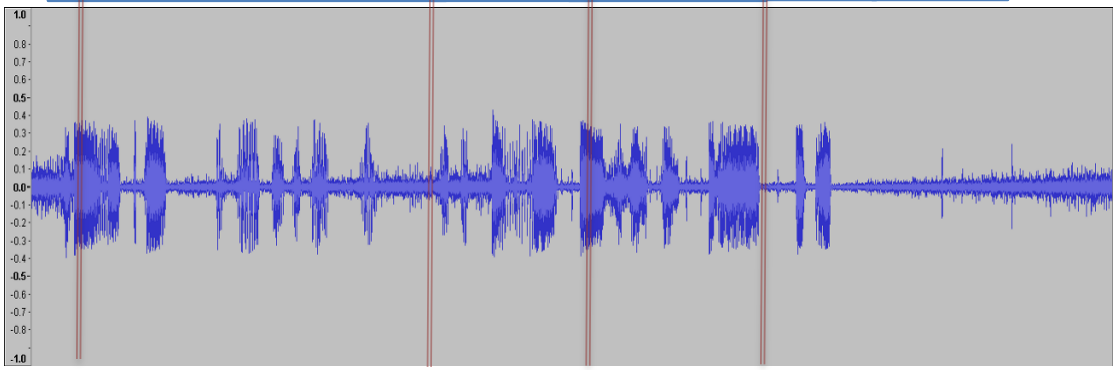
Channel D: HF, 11.41 meter wavelength, **Lower Side Band**. The target frequency of both the receiver and the transmitter drift over time, depending on the operational temperature of the equipment. This continuous shifting produces different degrees of **tonal shifting and distortion**.

Channel H: HF, 10.95 meter wavelength, **Narrow FM**. Longer wavelength allows signal to penetrate through obstructions; however, stray **EM interference** poses more of a problem than is found in the UHF systems.

CH_D

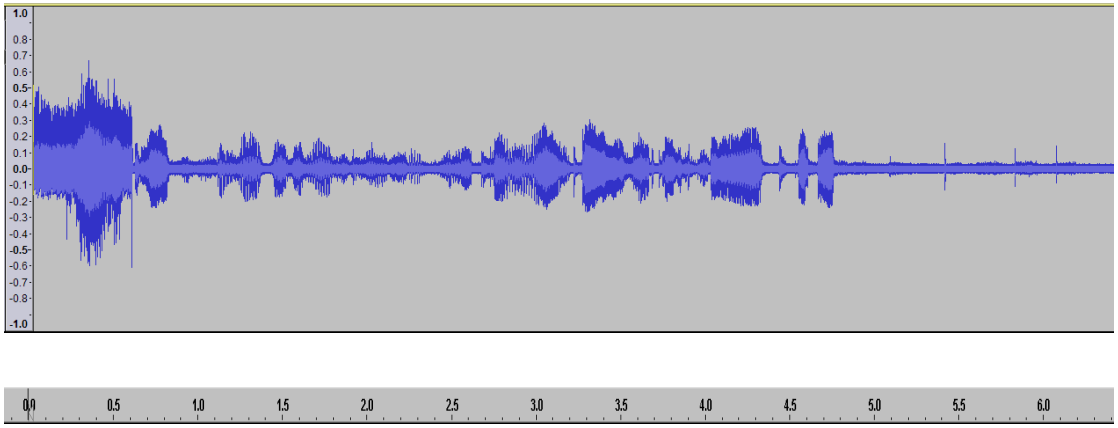


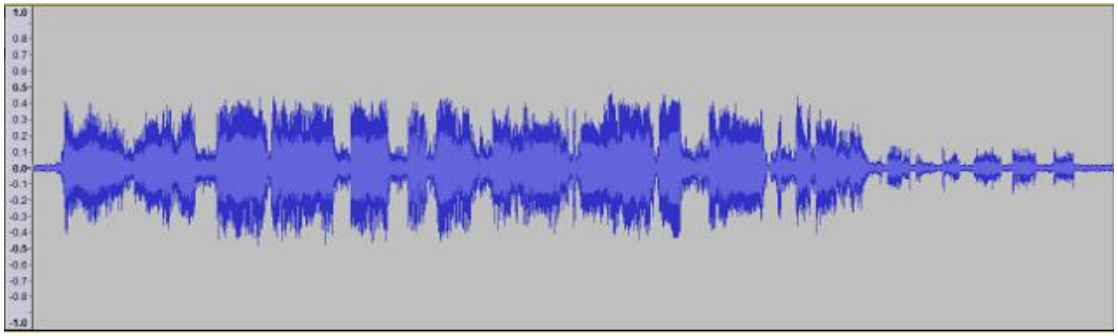
CH_H



Channel C: UHF, wavelength of 0.66 meters; receiver frequency offset 3khz relative to the transmission frequency; 10Khz IF Bandwidth setting. 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 E: VHF, wavelength of 2-meters, suffers from **diffraction, building penetration loss, and multipath loss**. The receiver is configured with 20-dB attenuation enabled, and with an IF of 12kHz.





UHF FHSS & Wideband FM Transceivers

Channel F: 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: UHF, 0.12 meter wavelength, **Wideband FM** , 5 watts ERP. This transmitter is designed to carry both video and audio – we are only using the audio input. The audio subcarrier uses up to 25kHz carrier deviation.

Human Intelligibility Study

- Is resulting data intelligible (with difficulty)?
- Signal-to-noise ratio (SNR) is inadequate metric
  - Two channels with equivalent SNR may differ significantly in terms of how much phonetic detail they preserve
- Study to assess intelligibility of data from each channel
  - Twenty native English-speaking judges listened to 96 unique recordings (12 segments * 8 channels)
  - Each segment judged on a 5-point intelligibility scale

Human Intelligibility Results

Channel	Description	Mean Rating	Stdev	Example
A	UHF, dual frequency	3.513157895	1.288650092	
B	UHF, tonal variation	3.364035088	1.440119133	
C	UHF, tonal distortion	3.881578947	1.129895382	
D	HF, lower side band	3.890350877	1.134673335	
E	VHF, multipath loss	**2.605263158**	1.360994849	
F	UHF FHSS	4.010526316	1.112647226	
G	UHF, Wideband FM	**4.745614035**	0.510875615	
H	HF, EM interference	3.48245614	1.335601672	

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

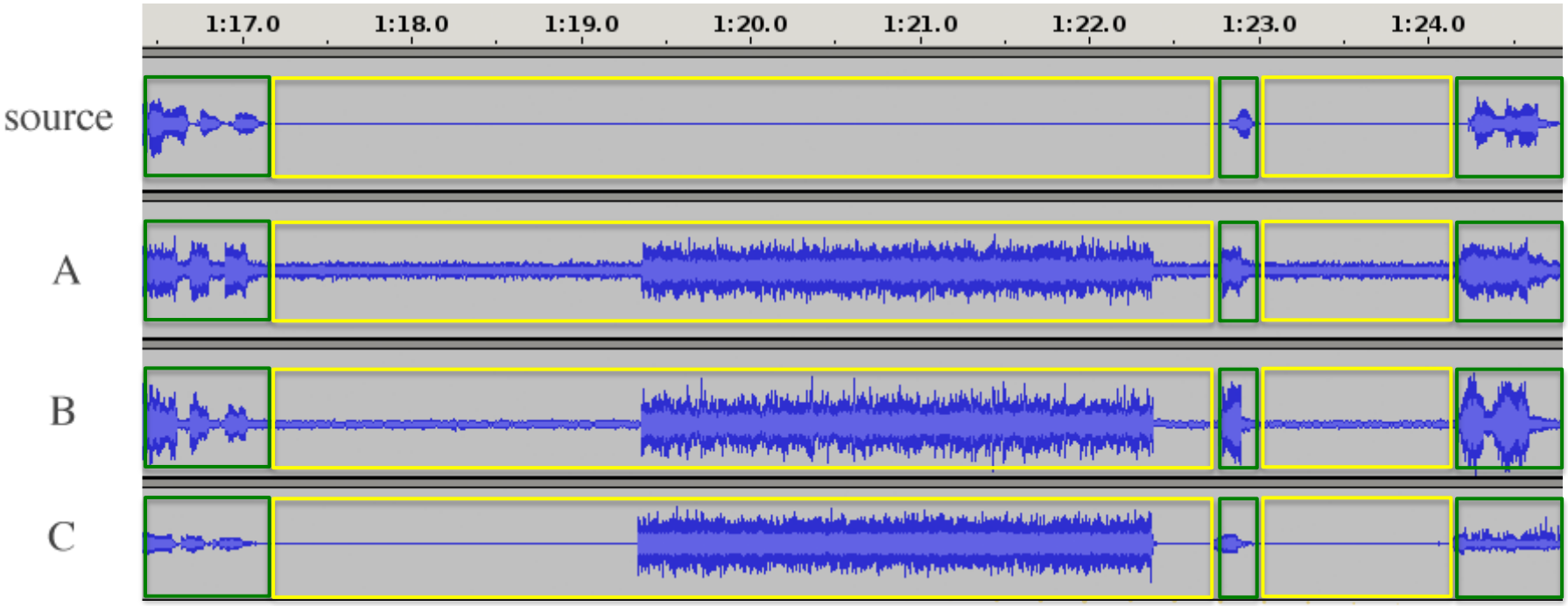
5 = All of the speech

Conclusion: Transmitted data is appropriately intelligible

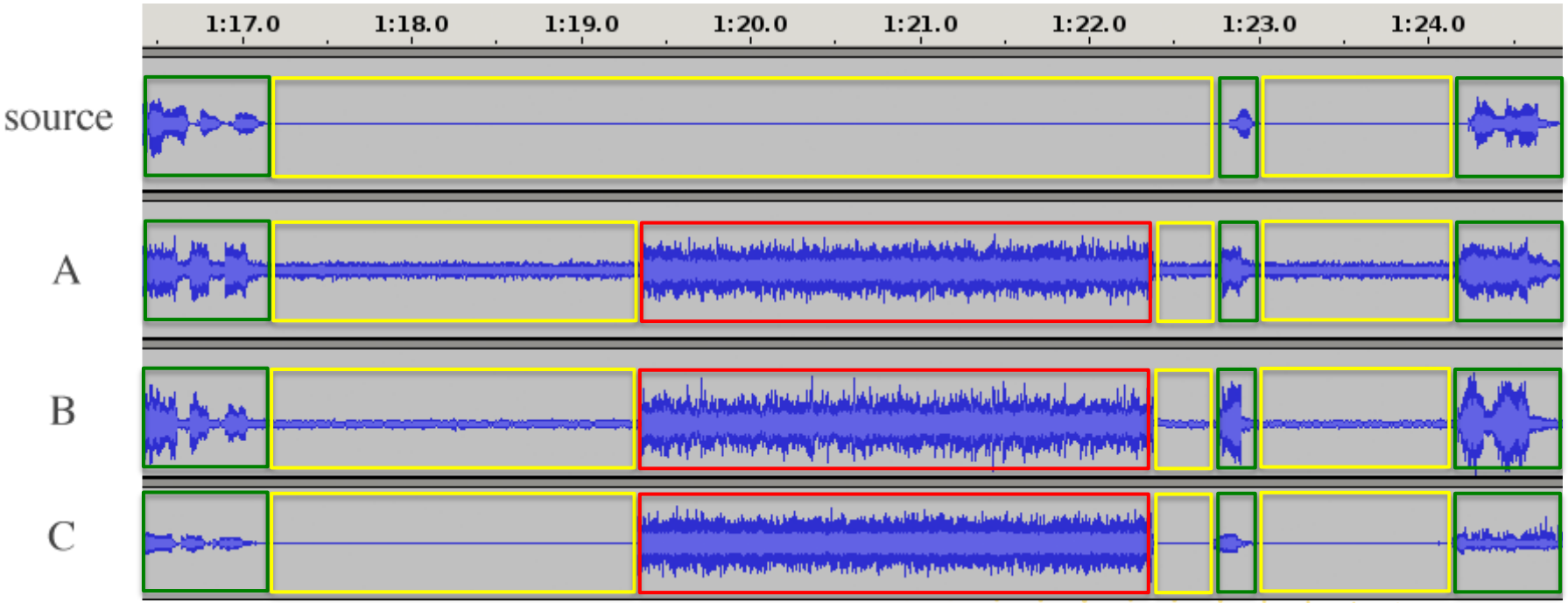
Post-transmission Processing

- After transmission, we have
  - Nine audio files
    - Clean source recording
    - Eight degraded channel recordings (A-H)
  - REF log: Indicates retransmission start time and source file parameters
  - VOX log: Timestamp for each voltage collector value transition, corresponding to push-to-talk dispatch commands
  - Reference annotation on clean channel
- We need to create
  - Accurate cross-channel alignment
  - Annotation on degraded channels, projected from clean channel

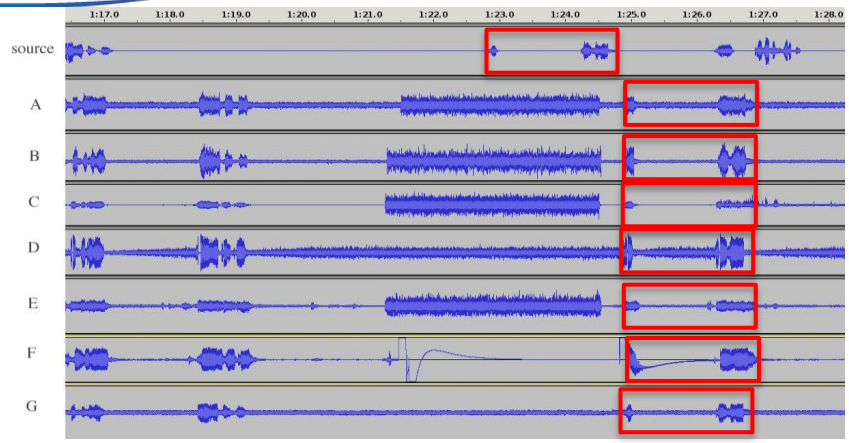
In a Perfect World



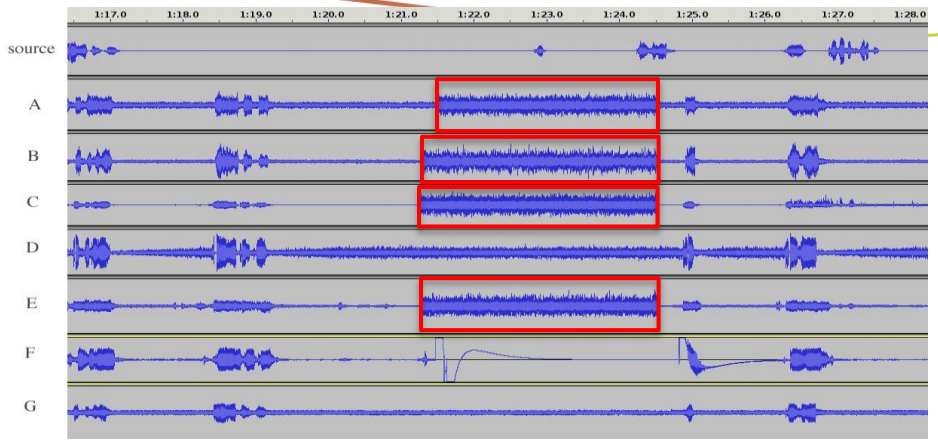
In a Perfect World



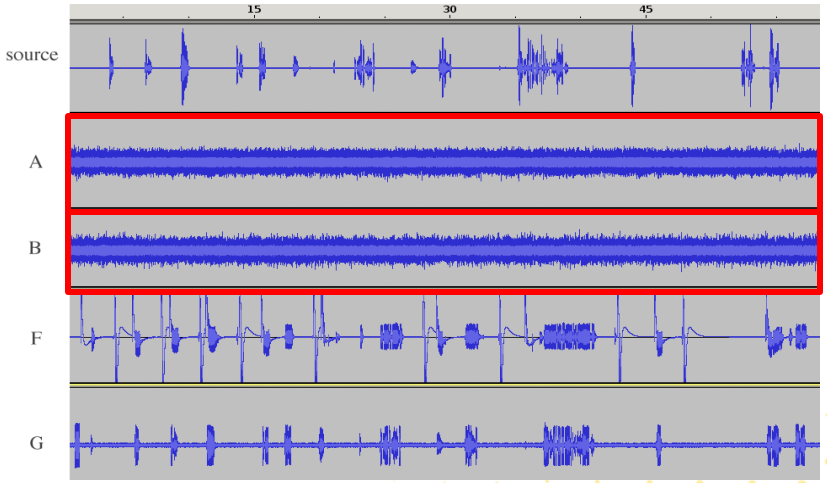
In the Real World



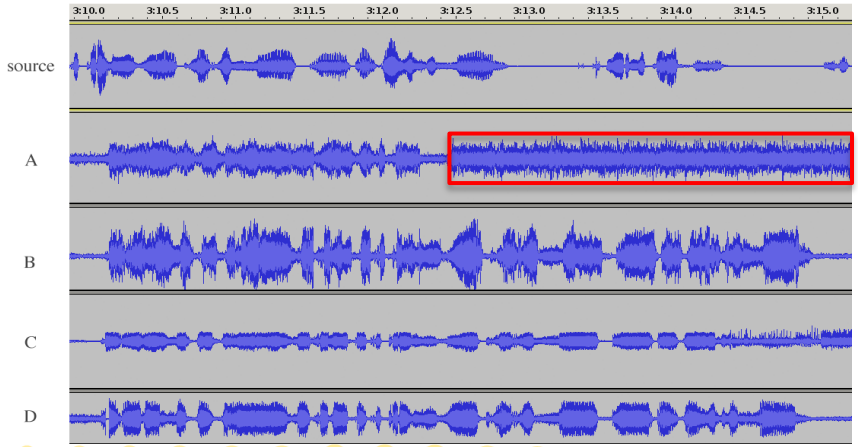
Channel-Specific Lag



Non-Transmission Region Drift



Full-File Transmission Failures



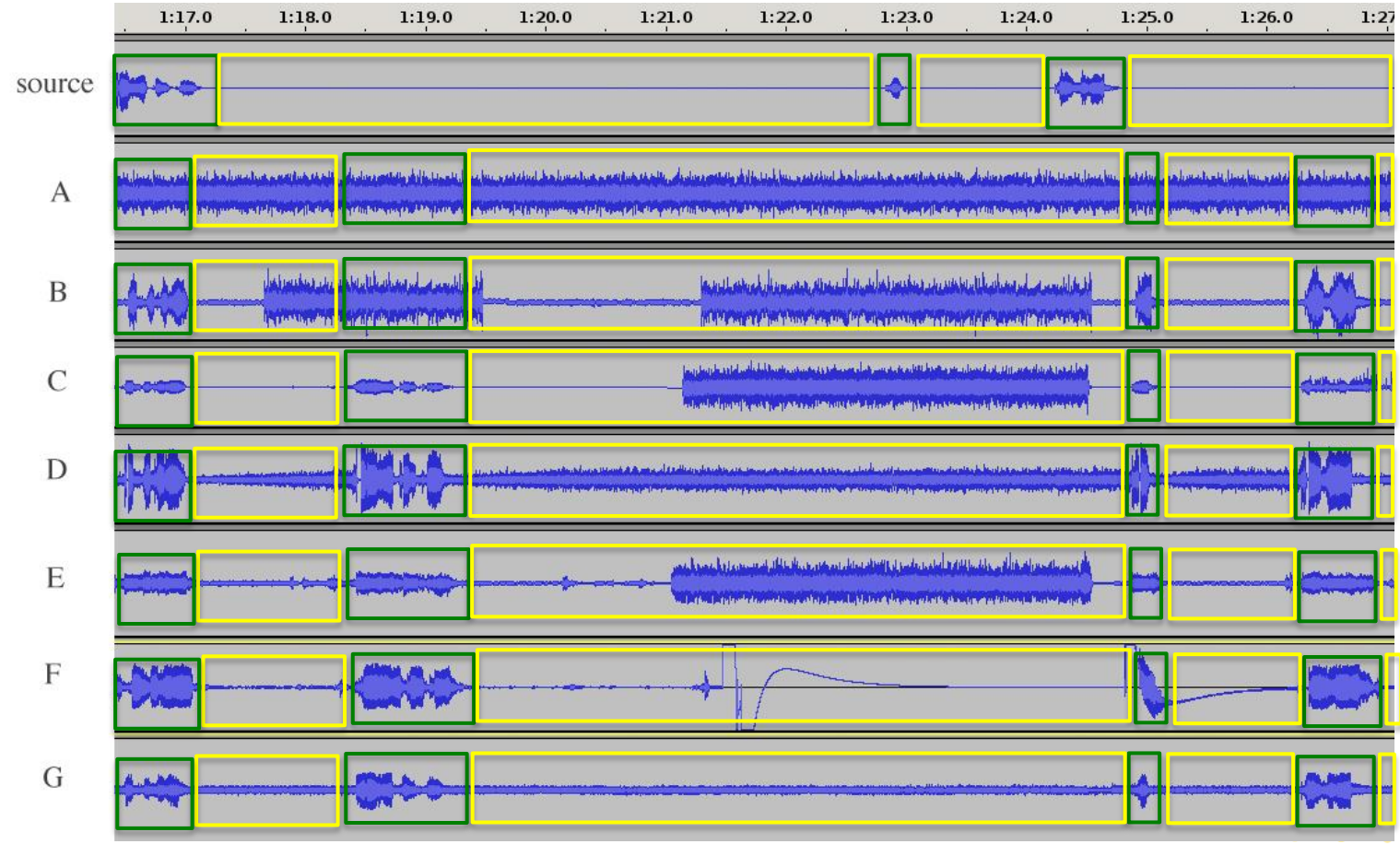
Channel-Specific Droputs

Channel Alignment Solutions

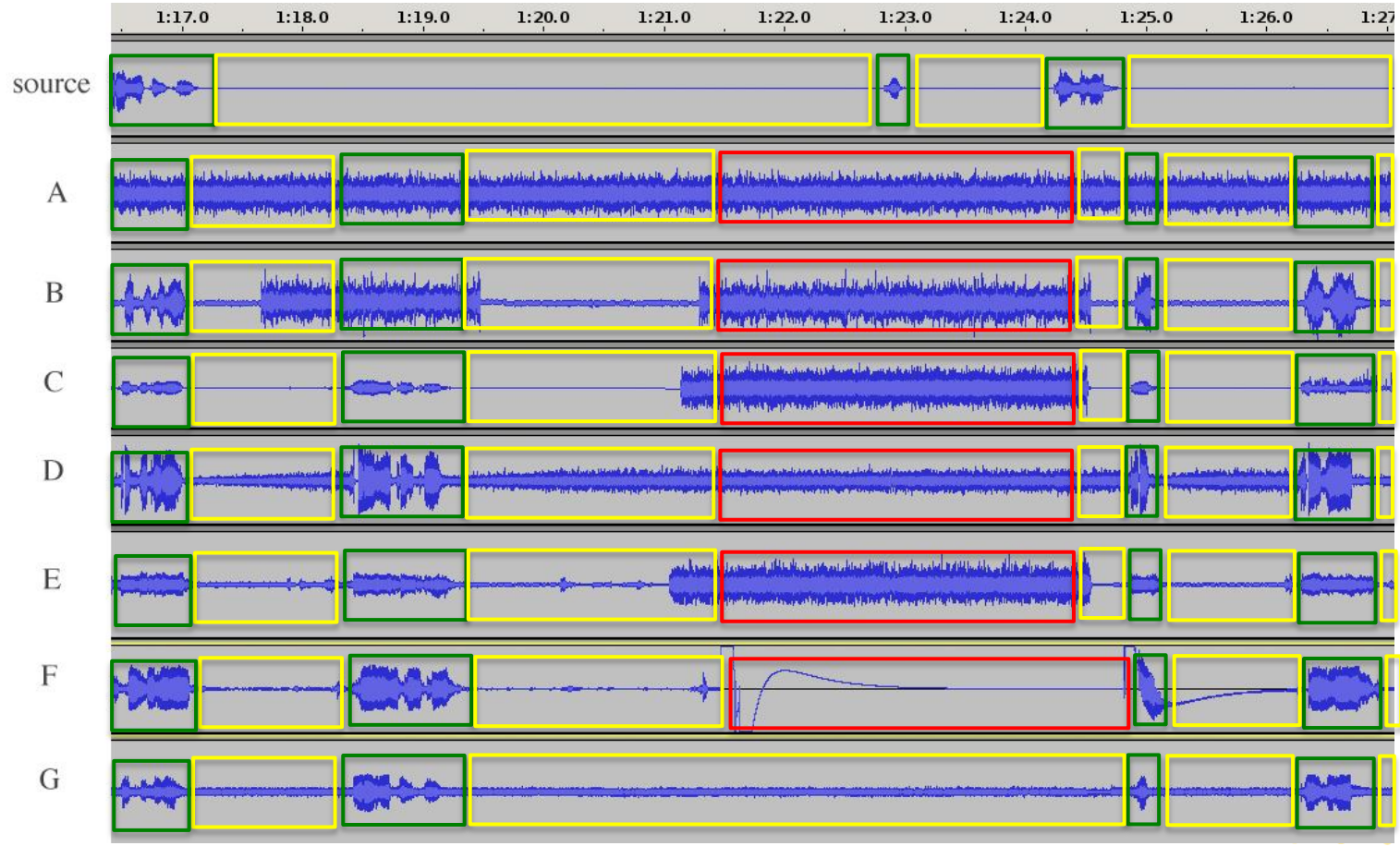
- Measure signal energy frame-by-frame over each transceiver channel to detect cases where
  - The overall energy is low throughout
  - The difference between minimum and maximum frame energy doesn't exceed channel-specified threshold
- Custom implementation of cross-correlation analysis* to compare each channel to source audio
  - Establishes time offset between start of source audio vs. transceiver recording
    - Offset value added to the source annotations so they are aligned relative to each channel recording
  - Also reveals cases of inconsistent alignment due to hardware failures or clock rate deviation
- Robust, channel-specific non-transmission detector* to detect short-duration dropouts

Special thanks to Dan Ellis at Columbia

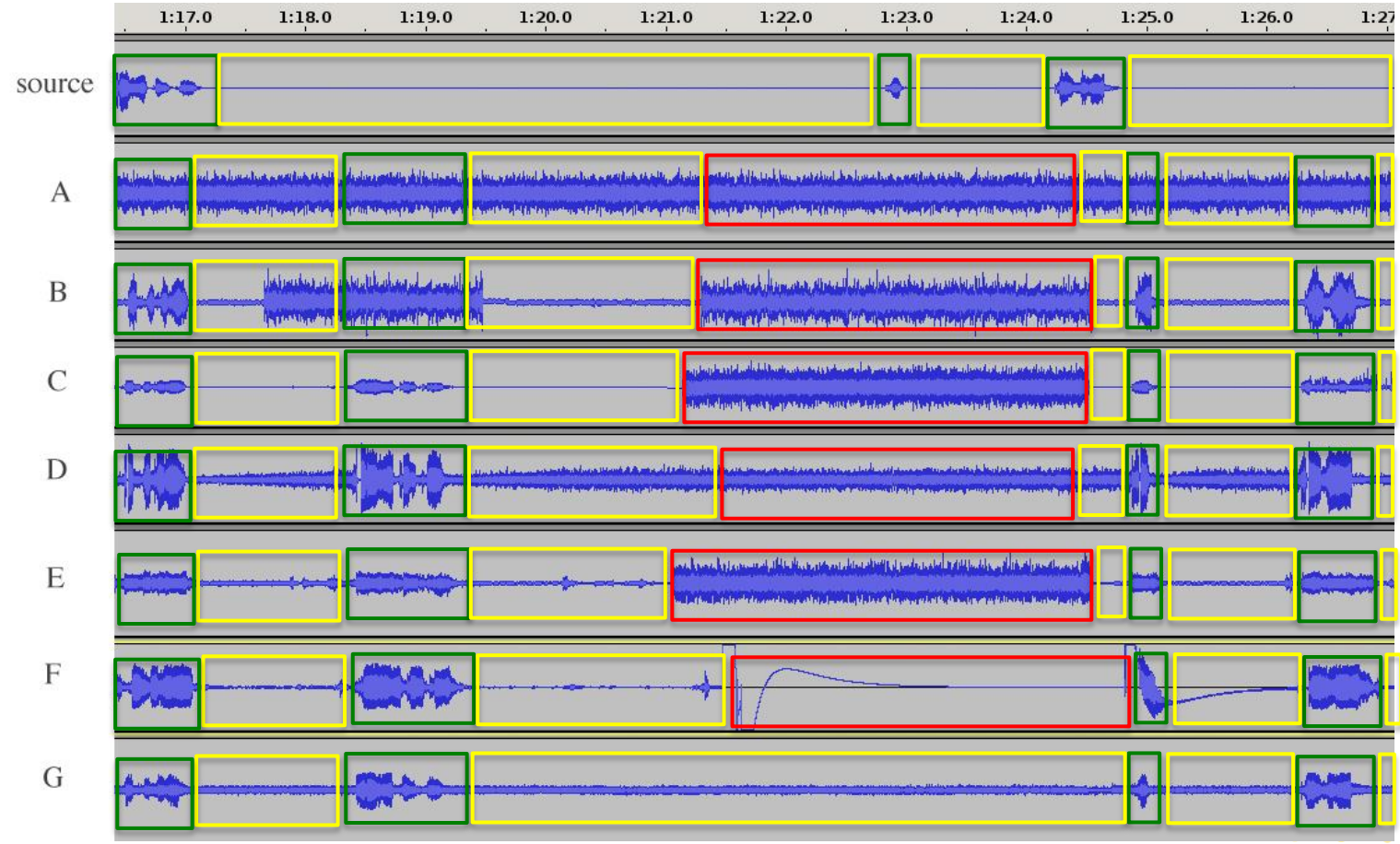
Annotation Projection



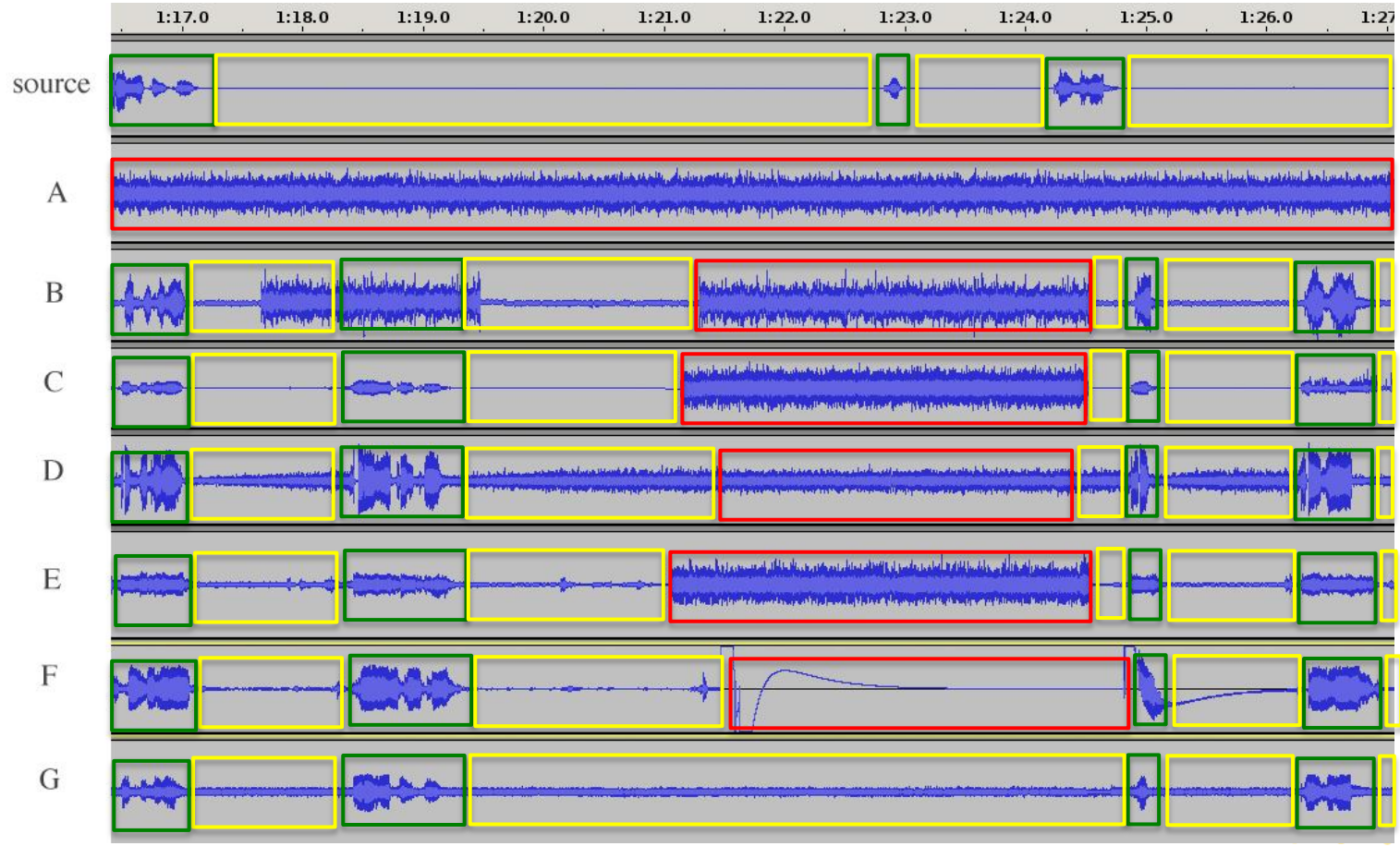
Annotation Projection



Annotation Projection

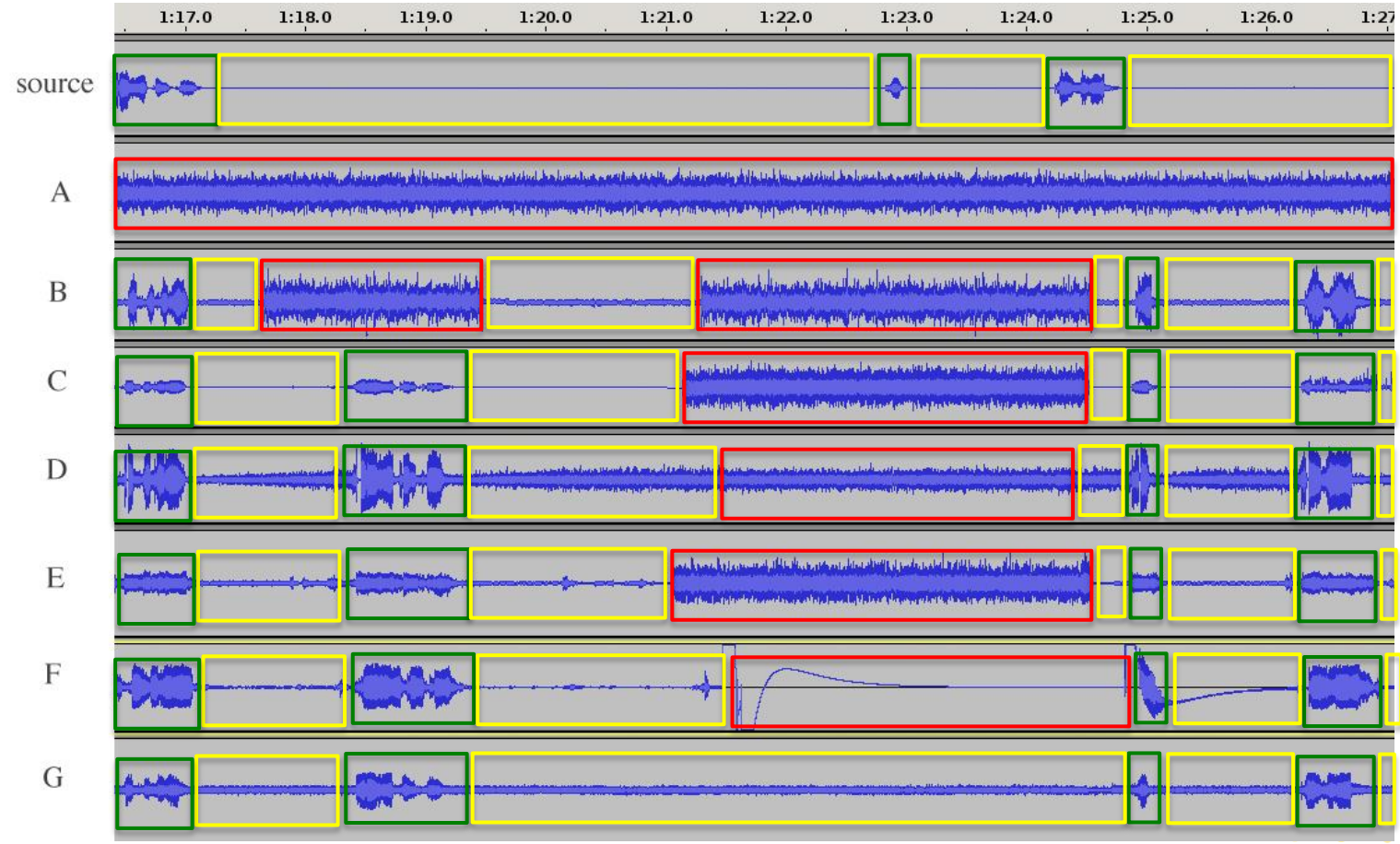


Annotation Projection



Linguistic Data Consortium

Annotation Projection



Release Preparation

- After channel alignment and annotation projection, prepare data for release
- Audio distributed on hard drive as flac-compressed, ms-wav format, 16-bit pcm, 1-channel, 16000-KHz sample rate
- Annotation and metadata distributed via web download
  - Annotation format is 12-field tab-delimited table, one row per transmission segment per channel
- Extensive validation and quality control prior to release
  - Manual spot checks on channel alignment, automatically-detected non-transmission regions
  - Additional package integrity checks by independent QC team
    - E.g. Verify flac decoding, sampling rate
    - E.g. Verify that all segments in annotation table have positive length
- Over 100 releases to RATS performers to date

Data Tally (1)

SAD Audio

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**

KWS Audio

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**

LID Audio

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**

Data Tally (2)

SID Speakers

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**

SID Audio

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**

Conclusions

- Accomplishments
  - Designed and deployed Multi-Radio Channel Collection Platform
  - Completed large-scale collection, retransmission and annotation in 5 challenging languages
    - Retransmitted over 3000 hours of data, yielding more than 16,000 hours of degraded signal broadcasts
    - Annotated over 1500 hours source data for SAD, LID, KWS, SID, Intelligibility and generated corresponding channel-specific annotation files
- Future Plans
  - Transmission over additional “novel channels” including new features
    - Greater distance between transmit/receive stations (up to several kilometers)
    - Include vocoded speech
    - Include recordings of environmental background noise, background speech, and audio from a wide range of communications systems sources (FAX handshaking, DTMF tones)
  - Publish RATS corpora in LDC catalog
    - SAD data set appearing in late 2014-early 2015
    - KWS data set will follow SAD