First DIHARD Challenge Evaluation Plan

Version 1.1

February 2, 2018

1 Introduction

DIHARD is a new annual challenge focusing on “hard” diarization; that is, speech diarization for challenging corpora where there is an expectation that the current state-of-the-art will fair poorly, including, but not limited to:

- clinical interviews
- extended child language acquisition recordings
- YouTube videos
- “speech in the wild” (e.g., recordings in restaurants)

Because performance of diarization is highly dependent on the quality of the speech activity detection (SAD) system used, the challenge will have two tracks:

- diarization beginning from gold speech segmentation
- diarization from scratch

The results of the initial challenge will be presented at a special session at Interspeech 2018 in Hyderabad. For questions not answered in this document or to join the DIHARD mailing list, please contact `dihardchallenge@gmail.com`.

2 Objective

While state-of-the-art diarization systems perform remarkably well for some domains (e.g., conversational telephone speech such as CallHome), as was discovered at the 2017 JSALT Summer Workshop at CMU, this success does not transfer to more challenging corpora such as child language recordings, clinical interviews, speech in reverberant environments, web video, and “speech in the wild” (e.g., recordings from wearables in an outdoor or restaurant setting). In particular, current approaches:

- fair poorly at estimating the number of speakers (e.g., monologues are frequently broken into multiple speakers)
- fail to work for short utterances (<1 second), which is particularly problematic for domains such as clinical interviews, which contain many short segment of high information content
- deal poorly with child speech and pathological speech (e.g., due to neurodegenerative diseases)
- are not robust to materials with large amounts of overlapping speech or dynamic environmental noise with some speech-like characteristics

The goals of the inaugural DIHARD evaluation include:

- to create an evaluation set drawn from a diverse set of challenging domains
- to establish a baseline of performance for existing diarization technologies on this set
- to release the reference data and results for continued research after the evaluation to encourage further testing and development

3 Schedule

- Registration period – January 30 through February 23, 2018
- Dev set release – February 1, 2018
- Eval set release – February 23, 2018
- Interspeech abstract submission – March 16, 2018
- Interspeech paper submission – March 23, 2018
- Final system output due/system descriptions – March 31, 2018
- Interspeech 2018 special session – September, 2018

4 Data

4.1 Training data

DIHARD participants may use any data available to train their system, whether publicly available or not, with the exception of the following previously released LDC corpora, from which portions of the eval set are drawn:

- HCRC Map Task Corpus (LDC93S12)
- DCIEM Map Task Corpus (LDC96S38)

- MIXER6 Speech (LDC2013S03)

Portions of MIXER6 have previously been excerpted for use in the NIST SRE10 and SRE12 evaluation sets, which also may not be used.

All training data should be thoroughly documented in the system description document (see Appendix C) at the end of the challenge. For a list of suggested training corpora, please see Appendix B.

4.2 Development data

Speech samples with diarization and gold speech segmentation will be be distributed to registered participants and may be used for any purpose including system development or training. These samples consist of approximately 19 hours worth of 5-10 minute chunks[1] drawn from the following domains:

- *Child language acquisition recordings*
  Previously unexposed recordings of language acquisition in 6-to-18 month olds. The data was collected in the home using a LENA recording device as part of SEEDLingS.

- *Supreme Court oral arguments*
  Previously unexposed oral arguments from the 2001 term of the U.S. Supreme Court that were transcribed and manually word-aligned as part of the OYEZ project.

- *Clinical interviews*
  Previously unexposed recordings of Autism Diagnostic Observation Schedule (ADOS) interviews conducted at the Center for Autism Research (CAR) at Children's Hospital of Philadelphia (CHOP). ADOS is a semi-structured interview in which clinicians attempt to elicit language that differentiates children with social communication difficulties from those without (e.g., "What does being a friend mean to you?"). Note that in order to publish this data, it had to be deidentified by applying a low-pass filter to regions identified as containing personal identifying information (PII).

- *Radio interviews*
  Previously unexposed recordings of YouthPoint, a late 1970s radio program run by students at the University of Pennsylvania consisting of student-lead interviews with opinion leaders of the era (e.g., Ann Landers, Mark Hamill, Buckminster Fuller, Isaac Asimov). The recordings were conducted in a studio and are generally of quite good quality.

- *Map tasks*
  Previously **exposed** recordings of subjects involved in map tasks drawn from the HCRC Map Task Corpus (LDC93S12) and the DCIEM Map Task Corpus (LDC96S38).

[1] Excepting data drawn from VAST, which range from under 1 minute to more than 10 minutes.

- *Sociolinguistic interviews*
  Previously **exposed** recordings of sociolinguistic interviews drawn from the SLX Corpus of Classic Sociolinguistic Interviews (LDC2003T15).

- Meeting speech
  Previously **exposed** recordings of multiparty meetings drawn from the 2004 Spring NIST Rich Transcription (RT-04S) dev (LDC2007S11) and eval (LDC2007S12) releases.

- *Audiobooks*
  Previously unexposed single-speaker, amateur recordings of audiobooks selected from LibriVox.

- *YouTube videos*
  Previously unexposed recordings of web video collected as part of the Video Annotation for Speech Technologies (VAST) project.

All samples will be distributed as 16 kHz, mono-channel FLAC files. Where transcription exists and forced alignment is feasible, turn boundaries will be produced by refining the human marked boundaries with forced alignment by trimming of turn-initial/turn-final silence and splitting on silences > 200 ms in duration. Where forced alignment is not possible, manually assigned turn boundaries will be used. The reference speech segmentation is derived from the diarization turn boundaries by merging overlapping turns.

4.3 Evaluation data

The evaluation set consists of approximately 20 hours worth of 5-10 minute speech samples drawn from the same domains, though not necessarily the same sources, as the development set plus one novel domain: restaurant conversation. Details about novel domains or domains where the source is different than for the development set:

- *Sociolinguistic interviews*
  Previously **exposed** interviews recorded as part of MIXER6 (LDC2013S03). While these recordings have not previously been released with diarization or SAD, the audio data was released as part of LDC2013S03, excerpts of which were used in the NIST SRE10 and SRE12 evaluation sets.

- *Meeting speech*
  Previously unexposed recordings of multiparty meetings conducted at LDC in the Fall of 2001 as part of ROAR.

- *Restaurant conversation*
  Previously unexposed recordings of meetings held in restaurants recorded by LDC.

The domain from which each sample is drawn will not be provided during the evaluation period, but will be revealed at the conclusion of the evaluation.

4.4 File formats

For each recording, speech segmentation will be provided via an HTK label file listing one segment per line, each line consisting of three space-delimited fields:

- segment onset in seconds from beginning of recording
- segment offset in seconds from beginning of recording
- segment label (always "speech")

For example:

> 0.10 1.41 speech
>
> 1.98 3.44 speech
>
> 5.0 7.52 speech

Following prior NIST RT evaluations, diarization for recordings will be provided using RTTM files. See Appendix A for a description of the RTTM format.

5 Task and performance metrics

The goal of the challenge is to automatically detect and label all speaker turns in each audio recording. Small pauses of <= 200 ms by a speaker are not considered to be segmentation breaks and should be bridged into a single continuous segment. Vocal noises such other than breaths (laughter, cough, sneeze, lipsmack, etc), are considered to be speech for the purpose of this evaluation, though all other sounds are considered non-speech. Because system performance is strongly influenced by the quality of the speech segmentation used, two tracks will be supported:

- Track 1: diarization using gold speech segmentation
- Track 2: diarization from scratch

Systems submitted to the former track should use the provided reference speech segmentation for each file, which will allow for evaluation of the diarization component in isolation from the SAD component. Systems submitted to the latter track will work directly from the audio. All researchers are strongly encouraged to submit results to at least the first track.

System output will be scored by comparison to human reference segmentation with performance evaluated by two metrics:

- diarization error rate (DER)
- framewise mutual information (MI)

DER will be calculated using version 22 of the NIST *md-eval.pl* scoring script without collars and with explicit scoring of overlapping speech. Mutual information will be calculated by converting the system and reference diarizations into labelings of 10 ms frames of audio and computing the mutual information between these labelings in bits. For further details, please consult the official scoring tool at `https://github.com/nryant/dscore`. During the eval period, scores will be posted on an real-time online leaderboard at the challenge homepage.

6 Evaluation rules

The 2018 DIHARD challenge is an open evaluation where the test data is sent to participants, who will process the data locally and submit their system outputs to LDC via Zenodo for scoring. As such, the participants have agreed to process the data in accordance with the following rules:

- Investigation of the evaluation data prior to the end of the evaluation is disallowed.
- Automatic identification of the domain of the test utterance is allowed.
- During the evaluation period, each team may make at most two submissions per day per system. Additional submissions past the first two each day will be ignored.
- While most test data is actually, or effectively, unexposed, portions have been exposed in part in the following corpora:
  - HCRC Map Task Corpus (LDC93S12)
  - DCIEM Map Task Corpus (LDC96S38)
  - MIXER6 Speech (LDC2013S03)
  - NIST SRE10 evaluation data
  - NIST SRE12 evaluation data

  Use of these corpora is prohibited.
- Participants in the 2017 JSALT Summer Workshop would have had access to an earlier version of the following sources:
  - ADOS
  - SEEDlingS
  - YouthPoint

  Teams containing members who participated in JSALT will be allowed to submit systems, but their scores will be flagged on the leaderboard and in publications.

While participants are encouraged to submit papers to the special session at Interspeech 2018, this is not a requirement for participation.

7 Evaluation protocol

7.1 Registration

To register for the evaluation, participants should email `dihardchallenge@gmail.com` with the subject line "REGISTRATION" and the following details:

- Organization – the organization competing (e.g., NIST, BBN, SRI)
- Team name – the name to be displayed on the leaderboard
- Tracks – which tracks they will be competing in

7.2 Data license agreement

One participant from each site must sign the data license agreement (available on the challenge website) and return it to LDC: (1) by email to `ldc@ldc.upenn.edu` or (2) by facsimile, Attention: Membership Office, fax number (+1) 215-573-2175. They will also need to create an LDC Online user account (`https://catalog.ldc.upenn.edu/signup`), which will be used to download the dev and eval releases

7.3 Zenodo registration

In order to submit system results performers will need to create an account with Zenodo (`https://zenodo.org/`).

8 Updates

Updates to this evaluation plan will be made available via the mailing list and the challenge website (`https://coml.lscp.ens.fr/dihard/index.html`).

9 Interspeech special session

The results of the challenge will be presented at a special session at Interspeech 2018, held September 2-6, 2018 in Hyderabad, India. Participants wishing to submit papers should do so through the Interspeech submission portal. Additional instructions will be provided once the Interspeech submission portal opens.

Appendix A: RTTM File Format Specification

Systems should output their diarizations as Rich Transcription Time Marked (RTTM) files. RTTM files are space-separated text files containing one turn per line, each line containing ten fields:

- Type – segment type; should always by "SPEAKER"
- File ID – file name; basename of the recording minus extension (e.g., "rec1_a")
- Channel ID – channel (1-indexed) that turn is on; should always be "1"
- Turn Onset – onset of turn in seconds from beginning of recording
- Turn Duration – duration of turn in seconds
- Orthography Field – should always by "<NA>"
- Speaker Type – should always be "<NA>"
- Speaker Name – name of speaker of turn; should be unique within scope of each file
- Confidence Score – system confidence (probability) that information is correct; should always be "<NA>"
- Signal Lookahead Time – should always be "<NA>"

For examples, please see the reference RTTM in the dev release (LDC2018E31).

Appendix B: Data Resources for Training

This appendix identifies a (non-exhaustive) list of publicly available corpora suitable for system training.

Corpora containing meeting speech
LDC corpora

- ICSI Meeting Speech Speech (LDC2004S02)
- ICSI Meeting Transcripts (LDC2004T04)
- ISL Meeting Speech Part 1 (LDC2004S05)
- ISL Meeting Transcripts Part 1 (LDC2004T10)
- NIST Meeting Pilot Corpus Speech (LDC2004S09)
- NIST Meeting Pilot Corpus Transcripts and Metadata (LDC2004T13)
- 2004 Spring NIST Rich Transcription (RT-04S) Development Data (LDC2007S11)
- 2004 Spring NIST Rich Transcription (RT-04S) Evaluation Data (LDC2007S12)
- 2006 NIST Spoken Term Detection Development Set (LDC2011S02)
- 2006 NIST Spoken Term Detection Evaluation Set (LDC2011S03)
- 2005 Spring NIST Rich Transcription (RT-05S) Evaluation Set (LDC2011S06)

Non-LDC corpora

- Augmented Multiparty Interaction (AMI) Meeting Corpus (`http://groups.inf.ed.ac.uk/ami/corpus/`)

Conversational telephone speech (CTS) corpora
LDC corpora

- CALLHOME Mandarin Chinese Speech (LDC96S34)
- CALLHOME Spanish Speech (LDC96S35)
- CALLHOME Japanese Speech (LDC96S37)
- CALLHOME Mandarin Chinese Transcripts (LDC96T16)
- CALLHOME Spanish Transcripts (LDC96T17)
- CALLHOME Japanese Transcripts (LDC96T18)
- CALLHOME American English Speech (LDC97S42)

- CALLHOME German Speech (LDC97S43)
- CALLHOME Egyptian Arabic Speech (LDC97S45)
- CALLHOME American English Transcripts (LDC97T14)
- CALLHOME German Transcripts (LDC97T15)
- CALLHOME Egyptian Arabic Transcripts (LDC97T19)
- CALLHOME Egyptian Arabic Speech Supplement (LDC2002S37)
- CALLHOME Egyptian Arabic Transcripts Supplement (LDC2002T38)
- Switchboard-1 Release 2 (LDC97S62)
- Fisher English Training Speech Part 1 Speech (LDC2004S13)
- Fisher English Training Speech Part 1 Transcripts (LDC2004T19)
- Arabic CTS Levantine Fisher Training Data Set 3, Speech (LDC2005S07)
- Fisher English Training Part 2, Speech (LDC2005S13)
- Arabic CTS Levantine Fisher Training Data Set 3, Transcripts (LDC2005T03)
- Fisher English Training Part 2, Transcripts (LDC2005T19)
- Fisher Levantine Arabic Conversational Telephone Speech (LDC2007S02)
- Fisher Levantine Arabic Conversational Telephone Speech, Transcripts (LDC2007T04)
- Fisher Spanish Speech (LDC2010S01)
- Fisher Spanish - Transcripts (LDC2010T04)

Other corpora
LDC corpora

- Speech in Noisy Environments (SPINE) Training Audio (LDC2000S87)
- Speech in Noisy Environments (SPINE) Evaluation Audio (LDC2000S96)
- Speech in Noisy Environments (SPINE) Training Transcripts (LDC2000T49)
- Speech in Noisy Environments (SPINE) Evaluation Transcripts (LDC2000T54)
- Speech in Noisy Environments (SPINE2) Part 1 Audio (LDC2001S04)
- Speech in Noisy Environments (SPINE2) Part 2 Audio (LDC2001S06)
- Speech in Noisy Environments (SPINE2) Part 3 Audio (LDC2001S08)

- Speech in Noisy Environments (SPINE2) Part 1 Transcripts (LDC2001T05)
- Speech in Noisy Environments (SPINE2) Part 2 Transcripts (LDC2001T07)
- Speech in Noisy Environments (SPINE2) Part 3 Transcripts (LDC2001T09)
- Santa Barbara Corpus of Spoken American English Part I (LDC2000S85)
- Santa Barbara Corpus of Spoken American English Part II (LDC2003S06)
- Santa Barbara Corpus of Spoken American English Part III (LDC2004S10)
- Santa Barbara Corpus of Spoken American English Part IV (LDC2005S25)
- HAVIC Pilot Transcription (LDC2016V01)

Appendix C: System descriptions

Proper interpretation of the evaluation results requires thorough documentation of each system. Consequently, at the end of the evaluation researchers must submit a full description of their system with sufficient detail for a fellow researcher to understand the approach and data/computational requirements. An acceptable system description should include the following information:

- Abstract
- Data resources
- Detailed description of algorithm
- Hardware requirements

Section 1: Abstract
A short (a few sentences) high-level description of the system.

Section 2: Data resources
This section should describe the data used for training including both volumes and sources. For LDC or ELRA corpora, catalog ids should be supplied. For other publicly available corpora (e.g., AMI) a link should be provided. In cases where a non-publicly available corpus is used, it should be described in sufficient detail to get the gist of its composition. If the system is composed of multiple components and different components are trained using different resources, there should be an accompanying description of which resources were used for which components.

Section 3: Detailed description of algorithm
Each component of the system should be described in sufficient detail that another researcher would be able to reimplement it. You may be brief or omit entirely description of components that are standard (i.e., no need to list the standard equations underlying an LSTM or GRU). If hyperparameter tuning was performed, there should be detailed description both of the tuning process and the final hyperparameters arrived at.

We suggest including subsections for each major phase in the system. Suggested subsections:

- signal processing – e.g., signal enhancement, denoising, source separation
- acoustic features – e.g., MFCCs, PLPs, mel fiterbank, PNCCs, RASTA, pitch extraction
- speech activity detection details – relevant for Track 2 only
- segment representation – e.g., i-vectors, d-vectors

- speaker estimation – how number of speakers was estimated if such estimation was performed
- clustering method – e.g., k-means, agglomerative
- resegmentation details

Section 4: Hardware requirements
System developers should report the hardware requirements for both training and at test time:

- Total number of CPU cores used
- Description of CPUs used (model, speed, number of cores)
- Total number of GPUS used
- Description of used GPUs (model, single precision TFLOPS, memory)
- Total available RAM
- Used disk storage
- Machine learning frameworks used (e.g., PyTorch, Tensorflow, CNTK)

System execution times to process a single 10 minute recording must be reported.