Second DIHARD Challenge Evaluation Plan

Version 1.2

Neville Ryant[a], *Kenneth Church*[b], *Christopher Cieri*[a], *Alejandrina Cristia*[c], *Jun Du*[d], *Sriram Ganapathy*[e], *and Mark Liberman*[a]

[a]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA
[b]Baidu Research, Sunnyvale, CA, USA
[c]Laboratoire de Sciences Cognitives et Psycholinguistique, ENS, Paris, France
[d]University of Science and Technology of China, Hefei, China
[e]Electrical Engineering Department, Indian Institute of Science, Bangalore, India

June 18, 2019

1 Introduction

DIHARD II is the second in a series of diarization challenges focusing on "hard" diarization; that is, speaker diarization for challenging recordings where there is an expectation that the current state-of-the-art will fare poorly. As with other evaluations in this series, DIHARD II is intended to both (1) support speaker diarization research through the creation and distribution of novel data sets and (2) measure and calibrate the performance of systems on these data sets. The results of the challenge will be presented at a special session at Interspeech 2019 in Graz, Austria.

The task evaluated in the challenge is speaker diarization; that is, the task of determining "who spoke when" in a multispeaker environment based only on audio recordings. As with DIHARD I, development and evaluation sets will be provided by the organizers, but there is no fixed training set with the result that participants are free to train their systems on any proprietary and/or public data. Once again, these development and evaluation sets will be drawn from a diverse sampling of sources including monologues, map task dialogues, broadcast interviews, sociolinguistic interviews, meeting speech, speech in restaurants, clinical recordings, extended child language acquisition recordings from LENA vests, and YouTube videos. However, there are several key differences from DIHARD I:

- two tracks evaluating diarization of multi-channel recordings have been added; these tracks will use recordings of dinner parties provided by the organizers of CHiME-5
- the evaluation period has been lengthened (from 4 weeks to 16 weeks)
- Jaccard Error Rate replaces mutual information as the secondary metric
- baseline systems and results will be provided to participants

Participation in the evaluation is open to all who are interested and willing to comply with the rules laid out in this evaluation plan. There is no cost to participate[1], though participants are encouraged to submit a paper to the corresponding Interspeech 2019 special session. Accepted papers will be presented at the special session at Interspeech 2019 in Graz, Austria in September 2019.

[1]Access to the data used by tracks 1 and 2 is free to all participants. Access to the CHiME-5 audio data used by tracks 3 and 4 is free for not-for-profit organizations. All other users, regardless of use case, will be required to purchase a commercial license to the CHiME-5 data. For more details, see: `https://licensing.sheffield.ac.uk/i/data/chime5.html`.

For questions not answered in this document or to join the DIHARD mailing list, please visit the DIHARD website (`https://coml.lscp.ens.fr/dihard`) or contact `dihardchallenge@gmail.com`.

2 Schedule

- Registration period – January 30 through March 15, 2019
- Dev/eval set release – February 28, 2019
- Scoring server opens – March 12, 2019
- Baselines released – week of March 11, 2019
- Interspeech abstract submission – March 29, 2019
- Interspeech paper submission – April 5, 2019
- Camera-ready papers – July 1, 2019
- System descriptions due – August 16, 2019
- Interspeech 2019 special session – September 15-19, 2019

The deadline for submission of final system outputs corresponds to the Interspeech camera-ready paper deadline (July 1st, 2019 midnight Anywhere on Earth).

3 Task

3.1 Task definition

The goal of the challenge is to automatically detect and label all speaker segments in each recording session. Small pauses of <= 200 ms by a speaker are not considered to be segmentation breaks and should be bridged into a single continuous segment. A pause by a speaker is defined as any segment in which that speaker is not producing a vocalization of any kind. By vocalization, we mean speech, including speech errors and infant babbling, but also vocal noise such as breaths, coughs, lipsmacks, sneezes, laughs, humming or any other noise produced by the speaker by means of the vocal apparatus.

Two input conditions (single channel vs. multichannel) and two speech activity detection (SAD) conditions (reference SAD vs. system SAD) will be considered, yielding four possible evaluation conditions.

3.2 Input conditions

Two audio input conditions are considered:

- **Single channel** – In the single channel condition, systems are provided with a single channel of audio for each recording. Depending on the recording source, this channel may be taken from a single distant microphone, a single channel from a distant microphone array, a mix of head-mounted or array microphones, or a mix of binaural microphones.
- **Multichannel** – In the multichannel condition, each recording session contains output from one or more distant microphone arrays, each containing multiple channels. Participants should treat the arrays separately, producing one output per array. They are free to use as few or as many of the channels on each array as they wish to perform diarization. For instance, if the recording session contains 6

microphone arrays, each having four channels, participants are expected to produce 6 RTTM files, each containing the result of their diarization system for a **SINGLE** array.

The single channel and multichannel conditions use different data sets with the former drawing data from DIHARD I and the latter from CHiME-5. For more information about the construction and composition of the data, please see Section 5. For both conditions a development set will be distributed, which may be used for any purpose including system development or training.

3.3 SAD conditions

Because system performance is strongly influenced by the quality of the speech segmentation used, two different SAD conditions are covered:

- **Reference SAD** – In the reference SAD condition, systems are provided with a reference speech segmentation that is generated by merging speaker turns in the reference diarization.
- **System SAD** – In the system SAD condition, systems are provided with just the raw audio input for each recording session and are responsible for producing their own speech segmentation.

3.4 Tracks

Together, the two input conditions and two SAD conditions yield four evaluation tracks:

- **Track 1** – Diarization from single channel audio using reference SAD
- **Track 2** – Diarization from single channel audio using system SAD
- **Track 3** – Diarization from multichannel audio using reference SAD
- **Track 4** – Diarization from multichannel audio using system SAD

Tracks 1 and 2 are identical to tracks 1 and 2 in DIHARD I and use the same data, though with improved annotation and additional development data (see Section 5). These tracks **DO NOT** contain any CHiME-5 data. Tracks 3 and 4 are new this year and consist exclusively of multi-person dinner party conversations taken from the CHiME-5 corpus.

All participants **MUST** register for at least one of track 1 or track 3 (diarization from reference SAD). Participation in tracks 2 and 4 is optional.

4 Scoring

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

- diarization error rate (DER)
- Jaccard error rate (JER)

4.1 Diarization error rate

Diarization error rate (DER), introduced for the NIST Rich Transcription Spring 2003 Evaluation (RT-03S), is the total percentage of reference speaker time that is not correctly attributed to a speaker, where "correctly

attributed" is defined in terms of an optimal mapping between the reference and system speakers. More concretely, DER is defined as:

$$\text{DER} = \frac{\text{FA} + \text{MISS} + \text{ERROR}}{\text{TOTAL}}$$

where

- *TOTAL* is the total reference speaker time; that is, the sum of the durations of all reference speaker segments
- *FA* is the total system speaker time not attributed to a reference speaker
- *MISS* is the total reference speaker time not attributed to a system speaker
- *ERROR* is the total reference speaker time attributed to the wrong speaker

Contrary to practice in the NIST evaluations, **NO** forgiveness collar will be applied to the reference segments prior to scoring and overlapping speech **WILL** be evaluated. For more details please consult section 6 of the RT-09 evaluation plan and the source to the NIST *md-eval* scoring tool[2].

4.2 Jaccard error rate

In addition to the primary metric we will score systems using Jaccard error rate (JER), a new metric developed for DIHARD. The Jaccard error rate is based on the Jaccard index[3], a similarity measure used to evaluate the output of image segmentation systems. An optimal mapping between reference and system speakers is determined and for each pair the Jaccard index is computed. The Jaccard error rate is then defined as 1 minus the average of these scores. While similar to DER, it weights every speaker's contribution equally, regardless of how much speech they actually produced.

More concretely, assume we have N reference speakers and M system speakers. An optimal mapping between speakers is determined using the Hungarian algorithm so that each reference speaker is paired with at most one system speaker and each system speaker with at most one reference speaker. Then, for each reference speaker ref the speaker-specific Jaccard error rate JER_{ref} is computed as:

$$\text{JER}_{ref} = \frac{\text{FA} + \text{MISS}}{\text{TOTAL}}$$

where

- *TOTAL* is the duration of the union of reference and system speaker segments; if the reference speaker was not paired with a system speaker, it is the duration of all reference speaker segments
- *FA* is the total system speaker time not attributed to the reference speaker; if the reference speaker was not paired with a system speaker, it is 0
- *MISS* is the total reference speaker time not attributed to the system speaker; if the reference speaker was not paired with a system speaker, it is equal to *TOTAL*

The Jaccard error rate then is the average of the speaker specific Jaccard error rates:

$$\text{JER} = \frac{1}{N} \sum_{ref} \text{JER}_{ref}$$

As with DER **NO** forgiveness collar will be applied to the reference segments prior to scoring and overlapping speech **WILL** be evaluated.

[2]Available as part of the Speech Recognition Scoring Toolkit (SCTK): `ftp://jaguar.ncsl.nist.gov/pub/sctk-2.4.10-20151007-1312Z.tar.bz2`. For DIHARD, we will be using version 22 of *md-eval*.

[3]`https://en.wikipedia.org/wiki/Jaccard_index`

JER and DER are highly correlated with JER typically being higher, especially in recordings where one or more speakers is particularly dominant. Where it tends to track DER is in outliers where the diarization is especially bad, resulting in one or more unmapped system speakers whose speech is not then penalized. In these cases, where DER can easily exceed 500%, JER will never exceed 100% and may be far lower if the reference speakers are handled correctly.

4.3 Scoring regions

In most cases the scoring region for each recording will be the **entirety** of the recording; that is, for a recording of duration 405.37 seconds, the scoring region will be [0, 405.37]. However, for a small subset of the recordings, personal identifying information (PII) has been removed from the recording, either by low-pass filtering or insertion of tones or zeroing out of samples. For these recordings, the scoring regions consists of the entirety of the recording minus these regions. In both cases the scoring regions will be specified by un-partitioned evaluation map (UEM) files, which will be distributed by LDC as part of the development and evaluation releases. Please see Appendix D for details of the UEM file format.

4.4 Scoring tool

All scoring will be performed using version 1.0.1 of *dscore*, which is maintained as a github repo at:

`https://github.com/nryant/dscore`

To score a set of system output RTTMs *sys1.rttm*, *sys2.rttm*, ... against corresponding reference RTTMs *ref1.rttm*, *ref2.rttm*, ... using the un-partitioned evaluation map (UEM) *all.uem*, the command line would be:

```
$ python score.py -u all.uem -r ref1.rttm ref2.rttm ... -s sys1.rttm sys2.rttm ...
```

The overall and per-file results for DER and JER (and many other metrics) will be printed to STDOUT as a table. For additional details about scoring tool usage, please consult the documentation for the github repo.

5 Data

5.1 Training data

DIHARD participants may use any publicly available or proprietary data to train their systems, with the exception of the following previously released corpora, from which portions of the evaluation set are drawn:

- DCIEM Map Task Corpus (LDC96S38)
- MIXER6 Speech (LDC2013S03)
- Digital Archive of Southern Speech (LDC2012S03 and LDC2016S05)
- any version of the SEEDLingS corpus, whether acquired via HomeBank or otherwise
- DIHARD I evaluation set

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 F) at the end of the challenge. For a list of suggested training corpora, please consult Appendix E.

5.2 Single channel data

The single channel input condition development and evaluation sets (used for tracks 1 and 2) consist of selections of 5-10 minute duration samples[4] drawn from 11 domains, each containing approximately 2 hours of audio. For most domains, the same source is used for both the development and evaluation set, though in some cases the development and evaluation sets use different sources; where the two sets draw from different sources, this is noted. For a detailed explanation of the domains and sources, please consult Appendix A.

5.2.1 Development data

The full composition of the single channel input condition development set, including domains, the sources drawn on for each domain, durations, and number of excerpts, is presented in Table 1.

Domain	**Source**	**Duration (hours)**	**# Recordings**
AUDIOBOOKS	LIBRIVOX	2.01	12
BROADCAST INTERVIEW	YOUTHPOINT	2.06	12
CHILD LANGUAGE	SEEDLINGS	1.92	23
CLINICAL	ADOS	2.18	24
COURTROOM	SCOTUS	2.08	12
MAP TASK	DCIEM	2.53	23
MEETING	RT04	2.45	14
RESTAURANT	CIR	2.03	12
SOCIOLINGUISTIC (FIELD)	SLX	2.01	12
SOCIOLINGUISTIC (LAB)	MIXER6	2.67	16
WEB VIDEO	VAST	1.89	32
TOTAL	-	23.81	192

Table 1: Single channel condition development set composition. For explanation of domains and sources, consult Appendix A.

5.2.2 Evaluation data

The full composition of the single channel input condition development set, including domains, the sources drawn on for each domain, durations, and number of excerpts, is presented in Table 2. Note that this set uses different sources than the development set for two domains:

- the MEETING domain draws from ROAR instead of RT04
- the SOCIOLINGUISTIC (FIELD) domain draws from DASS instead of SLX

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]Excepting data drawn from the WEB VIDEO domain, which range from under 1 minute to more than 10 minutes.

Domain	Source	Duration (hours)	# Recordings
AUDIOBOOKS	LIBRIVOX	-	-
BROADCAST INTERVIEW	YOUTHPOINT	-	-
CHILD LANGUAGE	SEEDLINGS	-	-
CLINICAL	ADOS	-	-
COURTROOM	SCOTUS	-	-
MAP TASK	DCIEM	-	-
MEETING	ROAR	-	-
RESTAURANT	CIR	-	-
SOCIOLINGUISTIC (FIELD)	DASS	-	-
SOCIOLINGUISTIC (LAB)	MIXER6	-	-
WEB VIDEO	VAST	-	-
TOTAL	-	22.49	194

Table 2: Single channel condition evaluation set composition. For explanation of domains and sources, consult Appendix A.

5.2.3 Segmentation

All reference diarization was produced at LDC by annotators using a tool equipped with a spectrogam display. Annotators were instructed to segment the recordings into labeled speaker turns, splitting on pauses $>$ 200 ms, where a pause by speaker "S" is defined as any segment of time during which "S" is not producing a vocalization of any kind, where vocalization is defined as any noise produced by the speaker by means of the vocal apparatus[5]. Boundaries were placed within 10 ms of the true boundary, taking care not to truncate sounds at edges of words (e.g., utterance-final fricatives or utterance initial stops). For some recordings (e.g., those from ROAR), close-talking microphones existed for each speaker; in these cases, segmentation was performed separately for each speaker using their individual microphone. Reference SAD was then derived from these segmentations by merging overlapping speech segments and removing speaker identification.

5.2.4 PII

A limited number of recordings from ADOS, CIR, and DASS contained regions carrying personal identifying information (PII), which had to be removed prior to publication. As systems have no way of plausibly dealing with these regions, they will not be scored and the relevant UEM files reflect this. The method used to de-identify these regions differs from source to source, with some opting to replace PII containing regions with a pure tone, while others used an approach based on low-pass filtering. Please see Appendix A for details about how PII was dealt with for each source.

5.2.5 File formats

All audio and annotations will be distributed via LDC. The audio will be distributed as single channel, 16 bit FLAC files sampled at 16 kHz, while reference speech segmentations will be distributed as HTK label files. In the case of the development set, a reference diarization will be provided, which will be distributed as Rich Transcription Time Marked (RTTM) files. For details regarding these file formats, please see Appendix B and Appendix C.

[5]For instance, speech (including yelled and whispered speech), backchannels, filled pauses, singing, speech errors and disfluencies, infant babbling or vocalizations, laughter, coughs, breaths, lipsmacks, and humming.

5.2.6 Differences from DIHARD I

While the single channel input condition development and evaluation sets are supersets of those used in DIHARD I, they exhibit several notable differences:

- the SEEDLINGS source from the CHILD LANGUAGE domain has been re-annotated from scratch to correct inconsistencies and outright errors present in DIHARD I
- the VAST source from the WEB VIDEO domain has been re-annotated from scratch to correct inconsistencies and outright errors present in DIHARD I
- the DIHARD I sociolinguistic interview domain has been split into two domains for DIHARD II:
  - SOCIOLINGUISTIC (FIELD) – sociolinguistic interviews conducted in the field
  - SOCIOLINGUISTIC (LAB) – sociolinguistic interviews conducted in a laboratory setting
- two additional hours of MIXER6 annotation have been added so that the SOCIOLINGUISTIC (LAB) domain is represented in both the development and evaluation sets
- two hours of new annotation for the previously unseen DASS source have been added so that the SOCIOLINGUISTIC (FIELD) domain is represented in both the development and evaluation sets
- two additional hours of CIR annotation have been added so that the RESTAURANT domain is represented in both the development and evaluation sets
- minor errors in the pre-processing scripts were corrected, which may result in small changes to the speaker segmentation for domains which did not undergo complete re-annotation
- all speaker ids and file ids were re-generated
- regions of recordings known to contain PII are no longer scored; see the UEM files distributed with the development and evaluation releases for each recording's scoring regions

5.3 Multichannel data

The multichannel input condition development and evaluation sets are drawn from the CHiME-5 dinner party corpus, a corpus of conversational speech collected during dinner parties held in real homes. Twenty parties were recorded, each lasting between 2 and 3 hours and having 4 participants: two hosts and two guests. The only constraints placed on these parties were that they last at least 2 hours and consist of three phases, each of which was held in a different location within the home:

- kitchen – where the meal was prepared
- dining – area the meal was eaten in
- living – location of post-dinner conversation/socializing

Participants were allowed to move freely between locations and speak on any topics they desired subject to the requirement that each phase lasted at least 30 minutes.

All parties were recorded using commercially available microphone arrays representative of those that might be found in an actual home or office environment. Within each home, 6 Microsoft Kinect devices (4 channel linear arrays) were distributed so as to ensure that each location was always captured by at least two arrays. This yielded 24 channels (6 arrays x 4 channels per array) of audio for each session.

For additional details regarding the recording setup, please consult the CHiME-5 website.

5.3.1 Development set

The DIHARD II multichannel input condition development set combines the CHiME-5 training and development sets and encompasses 45 hours of dinner parties from 18 homes.

5.3.2 Evaluation set

The DIHARD II multichannel input condition evaluation set is identical to the CHiME-5 evaluation set and consists of 5 hours of dinner parties from 2 homes.

5.3.3 Array synchronization

Due to a combination of clock drift and random frame dropping, the Kinects within each recording session exhibit massive desynchronization, both with each other and with the binaural recording devices worn by participants. This asynchrony worsens the further into a recording session one goes with the result that for some sessions, the median lag between one device and another is on the order of **SECONDS** by the time participants enter into the post-dinner socialization period. For this reason, each Kinect device is treated separately for the purpose of the evaluation, meaning that for each session participants should run their systems once per array, resulting in 6 RTTMs[6] (one per Kinect) per session.

5.3.4 Segmentation

The multichannel input condition evaluation set reference diarization was created manually by annotators at LDC using the same process described in Section 5.2.3. For each speaker, segmentation was performed from that speaker's binaural recording device. However, due to lack of synchronization between the the binaural recording devices and Kinects (see Section 5.3.3), this segmentation then had to be corrected for each array. The correction process used is identical to that used for CHiME-5 and consists of two stages[7]:

- every 10 seconds, estimate the current delay between the binaural recording device and the Kinect using normalized cross-correlation
- for each speaker turn, shift the boundaries established for the binaural recording device by the estimated delay

Due to time constraints, the LDC manual segmentation process could not be implemented for the development set. For the development set, all segmentations come from the turn boundaries established during CHiME-5 transcription.

5.3.5 PII

Portions of the CHiME-5 corpus contain PII or sensitive information, which had to be removed prior to publication. These regions (200-300 across the entire corpus) have been zeroed out in the released audio.

5.3.6 File formats

Audio will be distributed via University of Sheffield, while annotations will be distributed via LDC. All audio will be distributed as single channel, 16 bit WAV files sampled at 16 kHz. There will be one file per channel

[6] Excepting sessions S05 and S22 of the development set, which are missing arrays U05 and U03 respectively.
[7] For the precise implementation, see `https://github.com/chimechallenge/chime5-synchronisation`.

for each microphone array, yielding 16 files per recording session. Reference speech segmentations will be distributed as HTK label files (Appendix B), with one label file distributed per array. In the case of the development set, a reference diarization will be provided, which will be distributed as Rich Transcription Time Marked (RTTM) files (Appendix C).

6 Evaluation rules

The 2019 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 for scoring. As such, the participants have agreed to process the data in accordance with the following rules:

- While most of the test data is actually, or effectively, unexposed, portions have been exposed in part in the following corpora:
  - DCIEM Map Task Corpus (LDC96S38)
  - MIXER6 Speech (LDC2013S03)
  - Digital Archive of Southern Speech (LDC2012S03 and LDC2016S05)
  - NIST SRE10 evaluation data
  - NIST SRE12 evaluation data
  - DIHARD I evaluation sets
  - the SEEDLingS subset of HomeBank

  Use of these corpora is prohibited.
- Manual/human investigation of the evaluation data (e.g., listening, segmentation, or transcription) prior to the end of the evaluation is disallowed.
- Participants are allowed to use any automatically derived information (e.g., automatic identification of the domain) for the development and evaluation files.
- During the evaluation period, each team may make at most six submissions per day.
- Use of the evaluation server for per-recording hyperparameter tuning (e.g., attempting to establish the reference number of speakers in each recording by systematically altering clustering thresholds one recording at a time) is **EXPRESSLY** prohibited. We are being very generous compared to other machine learning competitions with our submission limits, so please do not abuse them. If teams are caught violating this rule, we will be forced to adopt stricter limits.

In addition to the above data processing rules, the participants agree to comply with the following general requirements:

- The participants agree to submit a system description document describing the algorithms, data, and computational resources used for all of their final systems (i.e., systems present on the leaderboard at the end of the challenge). These documents will be submitted at the end of the evaluation and should follow the format set forth in Appendix F.
- The participants agree to deposit the RTTM outputs of their final systems on Zenodo. At the conclusion of the challenge, the organizers will deposit an archive on Zenodo containing all system descriptions and final system outputs.

Failure to abide by these rules will be considered grounds for disqualification and will result in loss of access to the data, loss of access to the scoring server, and the removal of all existing submissions from the scoring server.

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 displayed on the leaderboard
- Tracks – which tracks they will be competing in

7.2 LDC 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 CHiME-5 data license agreement

LDC does not have permission to distribute the CHiME-5 audio data. Consequently, teams interested in participating in tracks 3 and 4 must obtain this data from University of Sheffield. Note that this applies to all interested teams, even those who participated in the CHiME-5 challenge. To do so, visit

`https://licensing.sheffield.ac.uk/i/data/chime5.html`

and select the appropriate license. Not-for-profit organizations should select the non-commercial license. All other organizations should select the commercial license, regardless of intended use for the data.

7.4 Baseline systems

Access to all baseline systems is provided via the challenge website.

7.5 Results submission

All system submissions will be done via an instance of CodaLab running on LDC servers. For instructions on how to register an account and submit results, see the challenge website.

8 Interspeech special session

The results of the challenge will be presented at a special session at INTERSPEECH 2019, held September 15-19, 2018 in Graz, Austria. Researchers wishing to submit papers should do so through the Interspeech submission portal. Additional instructions will be provided through the challenge website once the submission portal opens.

9 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`).

Appendix A: Single Channel Condition Domains and Sources

Domains

- *Audiobooks*
  Excerpts from recordings of speakers reading aloud passages from public domain English language texts. The recordings were selected from LibriVox and each recording consists of a single, amateur reader. Care was taken to make sure that the chapters and speakers drawn from were not present in LibriSpeech, which also draws from LibriVox.
- *Broadcast interview*
  Student-lead radio interviews conducted during the 1970s with popular figures of the era (e.g., Ann Landers, Mark Hamill, Buckminster Fuller, and Isaac Asimov). The recordings are selected from the unpublished LDC YouthPoint corpus.
- *Child language*
  Excerpts from day long recordings of infant (6 to 18 months) speech. All audio was recorded in the home using a LENA recording device, which consists of a vest worn by the child into which a microphone has been sewn. Because of their age, the child "speech" consists of a mixture of simplistic speech consisting of short utterances (possible very disfluent), babbling, laughing, crying, and diverse uncategorizeable non-speech vocalizations. Other speakers may be present in the recording, typically one or more parents, but also siblings, friends of siblings, aunts and uncles, and adult friends of the parents. Some of the recordings have quiet backgrounds, while others have radios or televisions playing. All recordings were taken from the SEEDLingS corpus.
- *Clinical*
  Recordings of Autism Diagnostic Observation Schedule (ADOS) interviews conducted to identify whether a child fit the clinical diagnosis for autism. ADOS is a roughly hour long semi-structured interview in which clinicians attempt to elicit language that differentiates children with Autism Spectrum Disorder from those without (e.g., "What does being a friend mean to you?"). The children included in this collection ranged from 12-16 years in age and exhibit a range of diagnoses from autism to non-autism language disorder to ADHD to typically developing. Interviews are typically recorded for quality assurance purposes; in this case, the recording was conducted using a ceiling mounted microphone. The recordings are selected from the unpublished LDC ADOS corpus.
- *Courtroom*
  Recordings of oral arguments from the 2001 term of the U.S. Supreme Court. The original recordings were made using individual table-mounted microphones, one for each participant, which could be switched on and off by the speakers as appropriate. The outputs of these microphones were summed and recorded on a single-channel reel-to-reel analogue tape recorder. All recordings taken from SCOTUS, an unpublished LDC corpus.
- *Map task*
  Recordings of speakers engaged in a map task. Each map task session contains two speakers sitting opposite one another at a table. Each speaker has a map visible only to him and a designated role as either "Leader" or "Follower". The Leader has a route marked on his map and is tasked with communicating this route to the Follower so that he may precisely reproduce it on his own map. Though each speaker was recorded on a separate channel via a close-talking microphone, these have been mixed together for the DIHARD releases. The recordings are drawn from the DCIEM Map Task Corpus (LDC96S38).
- *Meeting*
  Recordings of meetings containing between 3 and 7 speakers. The speech in these meetings is highly interactive in nature consisting of large amounts of spontaneous speech containing frequent interruptions

and overlapping speech. For each meeting a single, centrally located distant microphone is provided, which may exhibit excessively low gain. For the development set, these meetings are drawn from RT04, while for the evaluation set they are drawn from ROAR.

- *Restaurant*
  Informal conversations recorded in restaurants using binaural microphones. Each session contains between 4 and 7 speakers seated at the same table at a restaurant at lunchtime and was recorded from a binaural microphone worn by a designated facilitator; the mix of the two channels recorded by this microphone are provided. This data exhibits the following properties, which are expected to make it particularly challenging for automated segmentation and recognition:
  - due to the microphone setup, the majority of the speakers are farfield
  - background speech from neighboring tables is often present, sometimes at levels close to that of the primary speakers in the conversation
  - background noise is abundant with clinking silverware, moving chairs/tables, and loud music all common
  - the conversations are informal and highly interactive with interruptions and frequent overlapped speech

  All data is taken from LDC's unpublished CIR corpus.
- *Sociolinguistic field recordings*
  Sociolinguistic interviews recorded under field conditions. Recordings consists of a single interviewer attempting to elicit vernacular speech from an informant during informal conversation. Typically, interviews were recorded in the home, though occasionally they were recorded in a public location such as a park or cafe. The development set recordings were drawn from SLX and the evaluation set from DASS.
- *Sociolinguistic lab recordings*
  Sociolinguistic interviews recorded under quiet conditions in a controlled environment. All data is taken from the PZM microphones of LDC's Mixer 6 collection (LDC23013S03).
- *Web video*
  English and Mandarin amateur videos collected from online video sharing sites (e.g., YouTube and Vimeo). This domain is expected to be particularly challenging as the videos present a diverse set of topics and recording conditions; in particular, many videos contain multiple speakers talking in a noisy environment, where it can be difficult to distinguish speech from other kinds of sounds. All data is selected from LDC's VAST collection.

Sources

- *ADOS*
  ADOS is an unpublished LDC corpus consisting of transcribed excerpts from ADOS interviews conducted at the Center for Autism Research (CAR) at the Children's Hospital of Philadelphia (CHOP). All interviews were conducted at CAR by trained clinicians using ADOS module 3. The interviews were recorded using a mixture of cameras and audio recorded from a ceiling mounted microphone. Portions of these interviews determined by a clinician to be particularly diagnostic were then segmented and transcribed.

  Note that in order to publish this data, it had to be de-identified by applying a low-pass filter to regions identified as containing personal identifying information (PII). Pitch information in these regions is still recoverable, but the amplitude levels have been reduced relative to the original signal. Filtering was done with a 10th order Butterworth filter with a passband of 0 to 400 Hz. To avoid abrupt transitions

in the resulting waveform, the effect of the filter was gradually faded in and out at the beginning and end of the regions using a ramp of 40 ms.

- *CIR*
  Conversations in Restaurants (CIR) is a collection of informal speech recorded in restaurants that LDC originally produced for the NSF Hearables Challenge, an NSF-sponsored challenge designed to promote the development of algorithms or methods that could improve hearing in a noisy setting. It consists of conversations between 3 and 6 speakers, all LDC or Penn employees, seated at the same table at a restaurant near the University of Pennsylvania campus. Recording sessions were held at lunch time using a rotating list of restaurants exhibiting diverse acoustic environments and typically lasted 60-70 minutes. All recordings were conducted using binaural microphones mounted on either side of one speaker's head.

  A limited number of regions from one recording were found to contain PII. These regions were de-identified using the same low-pass filtering approach as in ADOS

- *DASS*
  The Digital Archive of Southern Speech, or DASS, is a corpus of interviews (each lasting anywhere from 3 to 13 hours) recorded during the late 60s and 70s in the Gulf Coast region of the United States. It is part of the larger Linguistic Atlas of the Gulf States (LAGS), a long-running project that attempted to preserve the speech of a region encompassing Louisiana, Alabama, Mississippi, and Florida as well as parts of Texas, Tennessee, Arkansas, and Georgia. Each interview was conducted in the field by a trained interviewer, who attempted to elicit conversation about common topics like family, the weather, household articles, agriculture, and social connections. It is distributed by LDC as LDC2012S03 and LDC2016S05.

  Due to the nature of the interviews, they sometimes contain PII or sensitive materials. All such regions have been replaced by tones of matched duration. Unfortunately, this process does not appear to have been systematic, with the result that the type of tone (pure or complex), power, and frequency differs across the corpus.

- *DCIEM*
  The DCIEM Map Task Corpus (LDC96S38) is a collection of recordings of two-person map tasks recorded for the DCIEM Sleep Deprivation Study. This study was conducted by the Defense and Civil Institute of Environmental Medicine (Department of National Defense, Canada) to evaluate the effect of drugs on performance degradation in sleep deprived individuals. Three drug conditions (Modafinil vs. Amphetamine vs. placebo) were crossed with three sleep conditions (18 hours vs. 48 hours vs. 58 hours awake). During each session, subjects performed a battery of neuropsychological tests (e.g., tracking tasks, time estimation tasks, attention-splitting tasks), questionnaires, and a map task. All audio was recorded via close-talking microphones under quiet conditions.

- *LibriVox*
  LibriVox is a collection of public domain audiobooks read by volunteers from around the world. It consists of more than 10,000 recordings in 96 languages. Portions have previously appeared in the popular LibriSpeech corpus, though care was taken to ensure that DIHARD did not select from this subset.

- *MIXER6*
  Mixer 6 (LDC2013S03) is a large-scale collection of English speech across multiple environments, modalities, degrees of formality, and channels that was conducted at LDC from 2009 through 2010. The collection consists of interviews with 594 native speakers of English spanning 1,425 sessions, each roughly 40-45 minutes in duration. Each session contained multiple components (e.g., informal conversation styled after a sociolinguistic interview or transcript reading) and was captured by a variety of microphones, including lavalier, head-mounted, podium, shotgun, PZM, and array microphones. While the corpus was released without speaker segmentation or transcripts, a portion of the corpus

was subsequently transcribed at LDC. DIHARD II draws its selections from this subset.

- *ROAR*
  ROAR is a collection of multiparty (3 to 6 participant) conversations recorded by LDC as part of the DARPA ROAR (Robust Omnipresent Automatic Recognition) project in Fall 2001. While portions of this collection have previously been exposed during the NIST RT evaluations, all DIHARD data comes from previously unexposed meetings. The meetings were recorded at LDC in a purpose built room using a combination of lavalier, head mounted, omnidirectional, PZM, shotgun, podium, and array microphones. For each meeting, a single centrally located distant microphone is provided.

- *RT04*
  RT04 consists of meeting speech released as part of the NIST Spring 2004 Rich Transcription (RT-04S) Meeting Recognition Evaluation development and evaluation sets. This data was later re-released by LDC as LDC2007S11 and LDC2007S12. It consists of recordings of multiparty (3 to 7 participant) meetings held at multiple sites (ICSI, NIST, CMU, and LDC), each with a different microphone setup. For DIHARD, a single channel is distributed for each meeting, corresponding to the RT-04S single distant microphone (SDM) condition. Audio files have been trimmed from the original recordings to the 11 minute scoring regions specified in the RT-04S un-partitioned evaluation map (UEM) files[8].

- *SCOTUS*
  SCOTUS is an unpublished LDC corpus consisting of oral arguments from the 2001 term of the U.S. Supreme Court. The recordings were transcribed and manually word-aligned as part of the OYEZ project, then forced aligned and QCed at LDC.

- *SEEDLingS*
  SEEDLingS is a corpus of child speech collected at the University of Rochester. Excerpts from day-long recordings conducted in the home were selected, then segmented and transcribed by LDC.

- *SLX*
  SLX (LDC2003T15) is a corpus of sociolinguistic interviews conducted in the 1960s and 1970s by Bill Labov and his students. The interview subjects range in age from 15 to 81 and represent a diverse sampling of ethnicities, backgrounds, and dialects (e.g., southern Amercian English, African American English, northern England, and Scotland). While the recordings have good sound quality for field recordings (especially from that era), they were collected in a range of environments ranging from noisy homes (e.g., small children running around in the background) to public parks to gas stations.

- *VAST*
  The Video Annotation for Speech Technologies (VAST) corpus is a (mostly) unexposed collection of approximately 2,900 hours of web videos (e.g., YouTube and Vimeo) intended for development and evaluation of speech technologies; in particular, speech activity detection (SAD), diarization, language identification (LID), speaker identification (SID), and speech recognition (STT). Collection emphasized videos where people are talking with a particular emphasis on videos where the speakers spoke primarily English, Mandarin, and Arabic, which comprise the bulk of the corpus[9]. Portions of this corpus have been exposed previously as part of the NIST 2017 Speech Analytic Technologies Evaluation, the NIST 2017 Language Recognition Evaluation, NIST 2018 Speaker Recognition Evaluation, and DIHARD I.

- *YouthPoint*
  YouthPoint is an unpublished LDC corpus consisting of episodes of YouthPoint, a late 1970s radio program run by students at the University of Pennsylvania. The show had an interview format similar to shows such as NPR's Fresh Air and consisted of interviews between University of Pennsylvania students and various popular figures. The recordings were conducted in a studio on open reel tapes and later digitized and transcribed at LDC.

[8] In cases where the onset or offset of a scoring region was found to bisect a speaker turn, it was adjusted to fall in silence adjacent to the relevant turn.

[9] Eight languages are represented in total: Arabic, English, Mandarin, Min Nan, Spanish, Portuguese, Russian, and Polish.

Appendix B: Speech segmentation label files

For each recording, the reference 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

The segments in these files are guaranteed to be disjoint and to not extend beyond the boundaries of the recording session.

Appendix C: RTTM File Format Specification

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

SPEAKER CMU_20020319-1400_d01_NONE 1 130.430000 2.350 <NA> <NA> juliet <NA> <NA>

SPEAKER CMU_20020319-1400_d01_NONE 1 157.610000 3.060 <NA> <NA> tbc <NA> <NA>

SPEAKER CMU_20020319-1400_d01_NONE 1 130.490000 0.450 <NA> <NA> chek <NA> <NA>

Appendix D: UEM File Format Specification

Un-partitioned evaluation map (UEM) files are used to specify the scoring regions within each recording. For each scoring region, the UEM file contains a line with the following four space-delimited fields

- File ID – file name; basename of the recording minus extension (e.g., "rec1_a")
- Channel ID – channel (1-indexed) that scoring region is on
- Onset – onset of scoring region in seconds from beginning of recording
- Offset – offset of scoring region in seconds from beginning of recording

For instance:

CMU_20020319-1400_d01_NONE 1 125.000000 727.090000

CMU_20020320-1500_d01_NONE 1 111.700000 615.330000

ICSI_20010208-1430_d05_NONE 1 97.440000 697.290000

Appendix E: 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/`)
- CSTR VCTK Corpus (`https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html`)

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)
- Nautilus Speaker Characterization (LDC2018S17)
- SRI Speech-Based Collaborative Learning Corpus (LDC2019S01)

Non-LDC corpora

- AVA ActiveSpeaker (`http://research.google.com/ava/`)
- AVA Speech (`http://research.google.com/ava/`)

- LibriSpeech (http://www.openslr.org/12/)
- Speakers in the Wild (SITW) (http://www.speech.sri.com/projects/sitw/)
- VoxCeleb (http://www.robots.ox.ac.uk/~vgg/data/voxceleb/)
- VoxCeleb 2 (http://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html)

Appendix F: 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 PDF that jointly describes their final systems (i.e., those appearing on the leaderboard at the end of the challenge) in sufficient detail for a fellow researcher to understand the approach and data/computational requirements. In order to make the preparation and format as consistent as possible, participants should use the IEEE Conference proceedings templates:

`https://www.ieee.org/conferences/publishing/templates.html`

An acceptable system description should include the following information:

- Authors
- Abstract
- Data resources
- Detailed description of algorithm
- Results on the development set
- Results on the evaluation set
- Hardware requirements

System names used within this document should be consistent with those used on the leaderboard. If for some reason this is not possible, then a section should be included that provides a mapping between the two namespaces. If a large number of systems were submitted, not all must be included (e.g., tests of the CodaLab server, early baselines, abandoned approaches), but at a minimum the four best performing systems for each track should be described.

Section 1: Authors
Listing of people whose contributions you wish acknowledged. This section is optional, but is helpful to the organizers as any names listed in this section will be listed as co-authors for the Zenodo download containing the challenge results.

Section 2: Abstract
A short (a few sentences) high-level description of the system.

Section 3: 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 4: 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 only for tracks 2 and 4
- 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 5: Results on the development set
Report overall DER and JER on the development set when using the official scoring tool.

Section 6: Results on the evaluation set
Report overall DER and JER on the evaluation set. These results should be taken straight from the leaderboard.

Section 7: 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 GPUs used (model, single precision TFLOPS, memory)
- Total number of TPUs used
- Generations of TPUs used (e.g., v2 vs v3)
- Total available RAM
- Used disk storage
- Machine learning frameworks used (e.g., PyTorch, Tensorflow, CNTK)

System execution times to process the entire development set must be reported.