First DIHARD Challenge Evaluation Plan

Version 1.2

March 4, 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 fare 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 the performance of a diarization system is highly dependent on the quality of the speech activity detection (SAD) system used, the challenge will have two tracks:

- Track 1: diarization beginning from gold speech segmentation
- Track 2: 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:

- fare 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 segments 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 26, 2018
- Interspeech abstract submission – March 16, 2018
- Interspeech paper submission/final system outputs – March 23, 2018
- Final system descriptions – March 31, 2018
- Interspeech 2018 special session – September, 2018

The deadline for submission of final system outputs corresponds to the Interspeech deadline (March 23, 2018 midnight GMT).

4 Data

This first iteration of the DIHARD challenge selects data from multiple novel sources including previously unexposed data and data previously developed for purposes other than diarization. Annotations have been converted from other formats for some data sources and created anew for other sources.

4.1 Training data

DIHARD participants may use any data to train their system, whether publicly available or not, with the exception of the following previously released LDC corpora, from which portions of the evaluation 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 D) at the end of the challenge. For a list of suggested training corpora, please see Appendix C.

4.2 Development data

Speech samples with diarization and reference 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 annotation of 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. 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. Those tapes were later digitized and made available by Jerry Goldman of OYEZ.

- *Clinical interviews*
  Previously unexposed recordings of Autism Diagnostic Observation Schedule (ADOS) interviews conducted at the Center for Autism Research (CAR) at the Children's Hospital of Philadelphia (CHOP). ADOS is a semi-structured interview in which clinicians attempt to elicit language that differentiates children with Autism Spectrum Disorders from those without (e.g., "What does being a friend mean to you?"). All interviews were conducted by CAR with audio recorded from a video camera mounted on a wall approximately 12 feet from the location inside the room where the interview was conducted.

  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.

- *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, and Isaac Asimov). The recordings were conducted in a studio on open reel tapes and later digitized at LDC.

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

- *Sociolinguistic interviews*
  Previously **exposed** recordings of sociolinguistic interviews drawn from the SLX Corpus of Classic Sociolinguistic Interviews (LDC2003T15). These are field recordings conducted during the 1960s and 1970s by Bill Labov and his students in various locations within the Americas and the United Kingdom.

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

- Meeting speech
  Previously **exposed** recordings of multiparty (3 to 7 participant) meetings drawn from the 2004 Spring NIST Rich Transcription (RT-04S) dev (LDC2007S11) and eval (LDC2007S12) releases. Meetings were recorded 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[2].

- *Audiobooks*
  Previously unexposed[3] single-speaker, amateur recordings of audiobooks selected from LibriVox.

- *YouTube videos*
  Previously unexposed annotations of web video collected as part of the Video Annotation for Speech Technologies (VAST) project. This domain is expected to be particularly challenging as the videos present a diverse set of topics and recording conditions. Unlike the other sources, which contain only English speech, though not necessarily from native speakers, the VAST selections contain both English and Mandarin speech with half the selections coming from monolingual English videos and half from monolingual Mandarin videos.

All samples will be distributed as 16 kHz, mono-channel FLAC files.

4.3 Evaluation data

The evaluation set consists of approximately 21 hours worth of 5-10 minute speech samples drawn from the same domains and sources as the development set with the following exceptions:

- *Sociolinguistic interviews*
  Instead of SLX, previously **exposed** sociolinguistic interviews recorded as part of MIXER6 (LDC2013S03) are used. 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. The released audio comes from microphone five, a PZM microphone.

- *Meeting speech*
  For the meeting speech domain, previously unexposed recordings of multiparty (3 to 6 participant) meetings conducted at LDC in the Fall of 2001 as part of ROAR are used. All meetings were recorded in the same room, though with different microphone setups. A single centrally located distant microphone is provided for each meeting.

- *Restaurant conversation*
  The evaluation set includes a novel domain, unseen in the development set, consisting of previously unexposed recordings from LDC's Conversations in Restaurants (CIR) collection. These recordings consist of conversations between 3 to 6 speakers, all LDC employees, seated at the same table at a restaurant on the University of Pennsylvania campus. All recordings were conducted using binaural microphones mounted on either side of one speaker's head, whose outputs were then mixed down to one channel.

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.

[2]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.

[3]Unexposed in the sense that while the audio and text these segments were selected from are obviously online and available from LibriVox, they have not previously been released as part of a speech recognition corpus. In particular, care was taken to ensure that the chapters and speakers drawn from were not present in LibriSpeech.

4.4 Segmentation

Where transcription exists and forced alignment was feasible, initial segment boundaries were produced by refining the human marked boundaries with forced alignment by trimming of turn-initial/turn-final silence and splitting on pauses[4] > 200 ms in duration. Ideally, this segmentation was then checked and corrected by human annotators using a tool equipped with a spectrogam display. Where forced alignment was not possible, manually assigned segment boundaries were used. The reference speech-activity segmentation ("SAD") was derived from the diarization speaker-segment boundaries by merging overlapping segments, merging segments separated by less than 200 ms, and removing speaker identification.

Because this was an unfunded pilot project, created under time pressure by volunteers, the full three-step workflow (transcription, alignment, checking and correction by human annotators) could not be implemented for all sources. The situation for each source is as follows:

- *ADOS*
  For the selections from "Autism Diagnostic Observation Schedule (ADOS)" interviews, the full workflow was implemented.
- *Conversations in Restaurants (CIR)*
  For the selections from "Conversations In Restaurants (CIR)", segments were derived from a careful turn-level transcription, without alignment and checking.
- *DCIEM*
  For the selections from the (Canadian) "Defence and Civil Institute of Environmental Medicine (DCIEM)" map task corpus, the full workflow was implemented.
- *LibriVox*
  For the selections from LibriVox audiobooks, the full workflow was implemented.
- *MIXER6*
  For the selections from sociolinguistic interviews conducted as part of MIXER 6, the full workflow was implemented.
- *ROAR*
  For the selections from meeting data collected at LDC in 2001 as part of the ROAR project, the full workflow was implemented.
- *SCOTUS*
  For the selections from 2001 U.S. Supreme Court oral arguments, the full workflow was implemented.
- *SEEDLingS*
  For the selections from child language recordings collected as part of the SEEDLingS project, segments were derived from manual segmentation done at LDC (with not entirely consistent guidelines). The evaluation set received an extra QC pass not performed for the development set and, consequently, should be of higher quality, though still imperfect.
- *VAST*
  For the selections from the "Video Annotation for Speech Technologies (VAST)" project, segments were derived from a careful turn-level transcription performed for that project, without additional alignment and checking.
- *YouthPoint*
  For the selections from YouthPoint radio interviews, the full workflow was implemented.

[4]For a given speaker, a pause is defined as any segment in which that speaker is not producing a vocalization. This includes breaths, but not coughs, laughs, or lipsmacks. In some cases, during the annotation process non-speech vocal noises were encountered that could not be accurately assigned to a speaker. All such segments have been omitted.

- *RT-04S*
  For the selections of meeting speech from LDC2007S1 and LDC2007S12, segments were derived from the original releases' RTTM files without any checking. These files have known issues such as overlapping turns, untranscribed speech, and speech that is inaudible on the distant microphones, which were not corrected.
- *SLX*
  For the selections of sociolinguistic interviews drawn from LDC2003T15, the full workflow was implemented.

We should also note that in cases where recordings from individual microphones were available, transcription and segmentation may have been done separately for each speaker using their individual microphone. This means that the reference RTTM may contain some segments that are inaudible, or nearly so, in the the released single-channel FLAC file, which may be taken from a single distant microphone. This affects MIXER6, ROAR, and (in the Dev set) RT-04S.

4.5 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 segments 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 other than breaths (e.g., laughter, cough, sneeze, and lipsmack), 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)
- frame-wise mutual information (MI)

During the eval period, scores will be posted on an real-time online leaderboard at the challenge homepage.

5.1 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 one-to-one 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[5].

5.2 Mutual information

We will also approach system evaluation from the standpoint of clustering evaluation, where both the reference and system segmentations are viewed as assignments of labels to frames of speech and a system's score is the mutual information in bits between its labeling and the reference labeling. More concretely, each segmentation will be converted to a sequence of 10 ms frames, each of which is assigned a single label corresponding to one of the following classes:

- non-speech
- non-overlapping speech by speaker$_i$
- overlapping speech by n speakers speaker$_{i_1}$, ..., speaker$_{i_n}$

where the sets of speakers are assumed disjoint for any pair of files. The contingency matrix between the reference and system labelings is then built and from this the mutual information computed according to:

$$\text{MI} = \sum_{i=1}^{R} \sum_{j=1}^{S} \frac{n_{ij}}{N} \log_2 \frac{n_{ij} N}{r_i s_j}$$

where

[5]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*.

- R is the number of reference clusters
- S is the number of system clusters
- n_{ij} is the number of frames assigned to the i-th reference cluster and j-th system cluster
- r_i is the number of frames assigned to the i-th reference cluster
- s_j is the number of frames assigned to the j-th system cluster
- N is the total number of frames

5.3 Scoring regions

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]. These regions will be provided to the scoring tool via un-partitioned evaluation map (UEM) files, which are plaintext files containing one scoring region per line, each line consisting of 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; ignored by scoring tool
- Onset – onset of scoring region in seconds from beginning of recording
- Offset – offset of scoring region in seconds from beginning of recording

For the dev set, the UEM will be available from the scoring section of the DIHARD website. The eval set UEM will be released alongside RTTMs at the conclusion of the evaluation.

5.4 Scoring tool

The official scoring tool is maintained as a github repo: `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) *dev.uem*, the command line would be:

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

The overall and per-file results for DER and MI (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.

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 and its addendum (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/`).

7.4 Submitting system outputs

For instructions on how to submit system outputs on the evaluation set, please consult the DIHARD website.

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 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 B: 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 C: 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)

Non-LDC corpora

- LibriSpeech (`http://www.openslr.org/12/`)
- VoxCeleb (`http://www.robots.ox.ac.uk/~vgg/data/voxceleb/`)

Appendix D: 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.