Fearless Steps Challenge Phase-3 (FS-III)

2021 Evaluation Plan

(Version 1.0)

Last Edited: February 1, 2021

Disclaimer: The Fearless Steps corpus is derived from a five-year NSF CISE funded project awarded to CRSS at the University of Texas at Dallas. UTDallas-CRSS established the hardware/software solutions to digitize and diarize 19,000 hrs of NASA Apollo data. All core Apollo data released as part of this challenge has been approved for public release by NASA Export Control. The full audio corpus is also available from UTDallas-CRSS. Any mention of or reference to organizations other than UTD is for information only; it does not imply recommendation or endorsement by

UTDallas-CRSS nor does it imply that the products mentioned are necessarily the best available for the purpose.

1 Introduction

The Fearless Steps Challenge 2021 Phase-3 (FS-3 2021) is the third edition in a series of Challenge tasks of the Fearless Steps Initiative hosted by Center for Robust Speech Systems (CRSS) at the University of Texas at Dallas. The goal of this Challenge is to evaluate the performance of state-of-the-art speech and language systems for large task oriented teams with naturalistic audio in challenging environments.

Researchers may select to participate in any single or multiple of these challenge tasks. Researchers may also choose to employ the FEARLESS STEPS corpus for other related speech applications. All participants are encouraged to submit their solutions and results for consideration to this ISCA INTERSPEECH-2021 special session.

While there is an extensive amount of audio (19,000 hrs), the core data for this FEARLESS STEPS challenge is drawn from 5 channels, over three Apollo-11 phases consisting of 100 hours. (see Section-4 for more details). A total of 80 hours of audio, human verified ground truth labels, and transcripts are provided for task system development. An additional set of 20 hours will be released for open test evaluation (see Section-4.7).

In an effort to foster research directions on naturalistic data, CRSS-UTDallas hosted 5 Challenge Tasks in 2019 as a part of the Inaugural Fearless Steps Challenge, with all participants being able to submit and present their systems at the Interspeech-2019. A similar format followed for the Fearless Steps Challenge Phase-02, with 4 Core Tasks with 2 sub-tasks for 3 tasks for a fully transcribed 100-hour corpus for the Interspeech-2020 Special Session. The proposed Phase-03 of the Challenge will include 10 additional hours of manually transcribed and previously unseen data. Researchers will be able to participate in 5 Core Speech and Language Technology tasks, with sub-tasks for each of the tasks (except SAD and SID) offered.

A 100 Hour Subset of the Apollo-11 Corpus revolving around three major events, namely: Lift Off; Lunar Landing; and Lunar Walking; are being made available, fully transcribed for all tasks, and released for Training, Development, and Evaluation of systems for the tasks that the researchers participate on 11th February 2021. In addition to this, additional 5 hours of previously unreleased Apollo-11 audio channel, and 5 hours of previously unreleased Apollo-13 Mission audio will be made available for participants to evaluate inter-channel and inter-mission system robustness. The Challenge Tasks will include Guidelines for the Tasks and baseline system results. For Fearless Steps Challenge Phase-03, CRSS-UTDallas are collaborating with **National Institute of Standards and Technology (NIST)** to host the Challenge using NIST web-resources. Challenge submissions and scoring for all participants will be made available through the NIST web-portal. Researchers will be able to use training and development transcripts for tasks other than their focus and use these transcripts to improve their systems.

The following Tasks designed for the Phase-03 Challenge are intended to advance research efforts not only in Speech Processing and Machine Learning, but also in Natural Language Understanding and Processing:

- **Task #1** Speech Activity Detection .... (**SAD**)
- **Task #2.a** Speaker Diarization Track 1 .... (**SD_track1**)
- **Task #2.b** Speaker Diarization Track 2 .... (**SD_track2**)
- **Task #3** Speaker Identification .... (**SID**)
- **Task #4.a** Automatic Speech Recognition Track 1 ...... (**ASR_track1**)
- **Task #4.b** Automatic Speech Recognition Track 2 ...... (**ASR_track2**)
- **Task #5.a** Conversational Hotspot Detection .... (**CONV_track1**)
- **Task #5.b** Extractive Summarization .... (**CONV_track2**)

Additionally, researchers will also be able to use any publicly available data to train (and develop/advance) their systems

2 Objectives

Traditionally, most speech and language technology concentrates on analysis of a single audio stream or channel with one or more speakers involved. The Apollo audio data represents 30 individual analog communications channels with multiple speakers in different locations working real-time to accomplish NASA's Apollo missions. For Apollo-11, this means each channel reflects a single communications loop (channel) that can contain anywhere from 3-33 speakers over extended time periods. While each channel has a primary function with a specific NASA Mission Specialist responsible, each of these channels are "loops", which contain core speakers working together plus speech from background conversations looped in at times, some reflecting Air-to-Ground (CAPCOM - Capsule Communicator) communications from the Astronauts. In addition, vast majority of the original Apollo Mission analog audio are all unlabeled, making application of speech technology a challenge. The inaugural phase of the Challenge (FS-1) will therefore emphasize the need to address various single channel speech tasks using unsupervised and/or semi-supervised speech algorithms. The Challenge Tasks for this session encourage the development of such solutions for core speech and language tasks on data with limited ground-truth/low resource availability, and serves as the "First Step" towards extracting high level information from such a task driven unlabeled corpus.

3 Tentative Schedule

Registration Period	**February 1 - March 7, 2021**
Data, and Evaluation Plan Release	**February 11, 2021**
Baseline Description and Results	**March 15, 2021**
System Submission Opens	**March 14, 2021**
Interspeech Paper Registration deadline	**March 26, 2021**
Interspeech Paper Submission deadline	**April 2, 2021**
Final System Submission Deadline	**March 31, 2021**
Final Results Announced for all Tasks	**April 1, 2021**
Interspeech 2019 Special Session	**August 30 - September 3, 2021**

4 Data Overview

4.1 Corpus Development

The Apollo 11 mission lasted 8 days 3 hours 18 minutes and 35 seconds. The entire communications between astronauts, flight controllers, and their backroom support teams inside NASA Mission Control Center (MCC) were continuously recorded using two 30-track analog reel-to-reel recording machines, namely Historical Recorder 1 (HR1) and 2 (HR2). By alternately changing the tapes, continuity was ensured without any loss of data; 29 of the 30 channels/tracks on the analog tape were used to record speech data with one channel recording the Mission Elapsed Time (MET) in an encoded IRIG-B format. The records stored by the United States National Archives and Records Administration (NARA) were used to digitize the original analog tapes, by designing a new read-head for the SoundScriber player (as shown in Fig. 1). The read-head was developed specifically with the aim of digitizing all the channels simultaneously, thus preserving the synchronicity of the data, enabling individual channels from each HR1 and HR2 to be indexed and stored separately.

Figure 1: (left): The SoundScriber device used to decode analog tapes, and (right): The UTD-CRSS designed read-head decoder retrofitted to the SoundScriber

Synchronous multichannel reconstruction of the entire mission is made possible by using the first channel of every tape which contains the MET. This data was stored and digitized initially at a 44.1 kHz sampling frequency, and later downsampled to 8 kHz for speech analysis. The recordings were saved as half-hour chunks per channel with their file names indicating the mission name, the historical recorder and channel the recording belongs to, followed by the start and end times as given by the mission elapsed time.

Communication Protocols: Specific protocols followed during the mission were imperative for ensuring successful communication. Knowledge of these communication characteristics can be leveraged to achieve better inferences through informed analysis. Fig. 2. shows a basic structure of the communication protocols. Only the Capsule Communicator (CAPCOM) could directly communicate with the astronauts, with the Flight Director (FLIGHT) in control of accessing all other channel loops. With multiple speakers joining in on these loops at various points in time, all personnel would address the channel owner by their assigned channel names. In fact, all backroom staff are present on multiple channel loops. Audio markers such as 'Quindar Tones' were used to infer communication with the astronauts.

4.2 Data Organization

Fig. 3. displays the overall Timeline of the Apollo-11 Mission. The Stages '1', '5' and '6' which were high impact mission-critical events were found to be ideal for the development of the 100-hour Challenge Corpus. With the quality of speech data varying between 0 and 20 dB SNR in this challenge corpus, the channel variability and complex interactions across all five channels of interest discussed in the previous section are mostly encapsulated in these 100 hours.

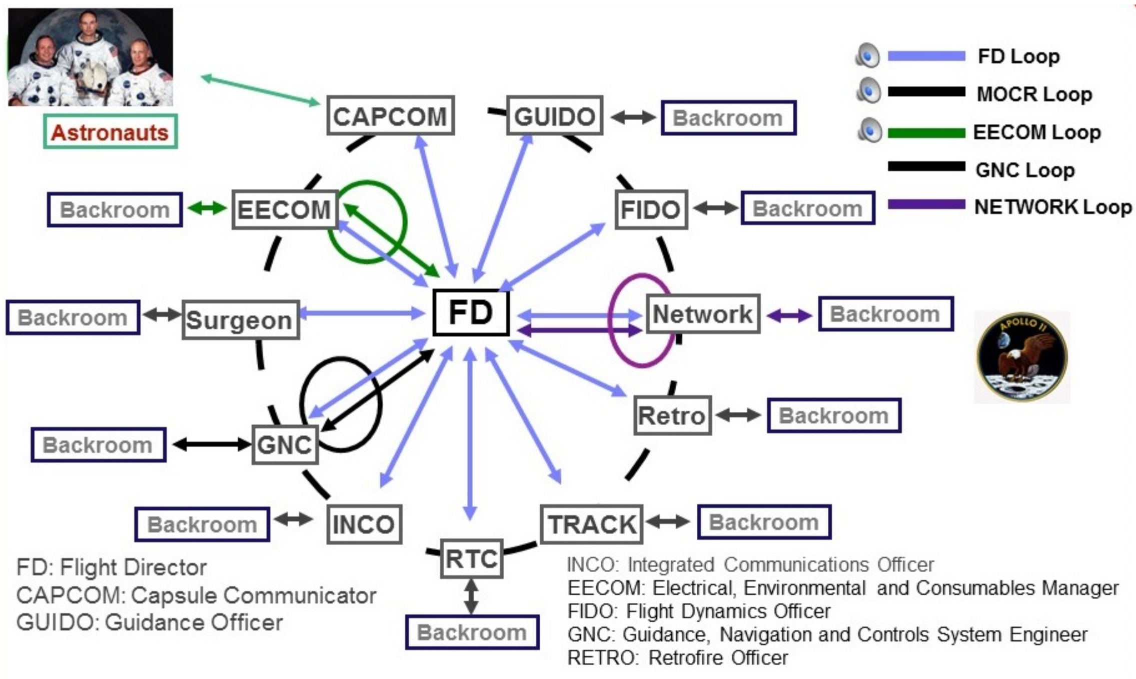

Figure 2: Apollo-11 communication overview. Ground communications between hundreds of flight controllers and their 'backroom' support staff are shown in the loop. The space-to-ground communications are linked to CAPCOM

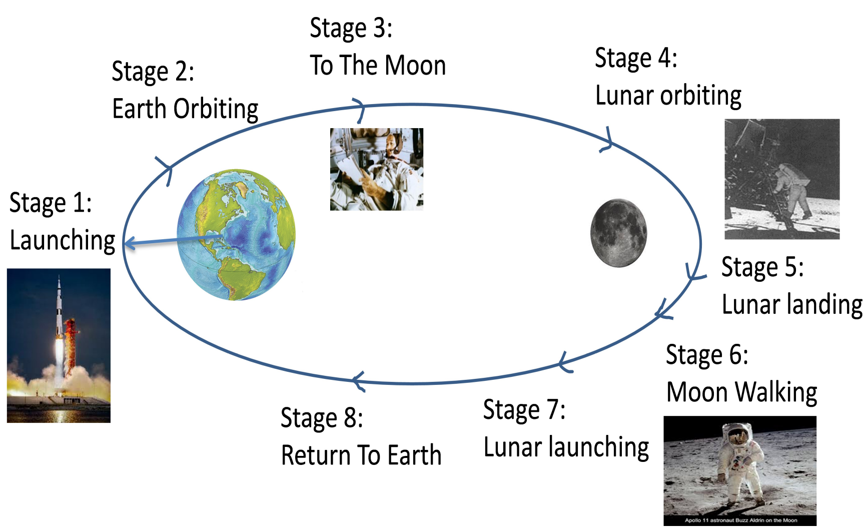

Figure 3: Overview of the timeline of Apollo 11 mission

These three major events the multichannel data is chosen from are:

- Lift Off (25 hours)
- Lunar Landing (50 hours)
- Lunar Walking (25 hours)

These landmark events have been found to possess rich information from the speech and language perspective. Out of the 29 channels, five channels of interest with the most activity over the selected events were chosen to select the data from:

- Flight Director (FD)
- Mission Operations Control Room (MOCR)
- Guidance Navigation and Control (GNC)
- Network Controller (NTWK)
- Electrical, Environmental and Consumables Manager (EECOM)

The personnel operating these five channels (channel owners/primary speakers) were in command of the most critical aspects of the mission, with additional backroom staff looping in for interactions with the primary owners of this channel.

4.3 Challenges with the Apollo Data

The Apollo data is unique and poses multiple challenges. Comprised of 19,000 hours of naturalistic multi-channel data, it is characterized by multiple classes of noise and degradation and several overlap instances over most channels. Most audio channels are degraded due to high channel noise, system noise, attenuated signal bandwidth, transmission noise, cosmic noise, analog tape static noise, noise due to tape aging, etc. Moreover, the noise conditions and signal-to-noise ratio (SNR) levels vary over a 25 dB range for separate channels, and in many cases, at different mission stages within each channel (See Tab. 2). Some channels even have the presence of babble noise depending on the location of the personnel in the MCC. For the Apollo missions, head-mounted Plantronics microphones were used, but some in-spacecraft recordings were made using fixed far-field microphones which also picked up the presence of environmental noise (e.g., glycol cooling pumps and thruster firings) that varied over time. These conditions severely degrade the performance of conventional speech activity detection algorithms. Due to the time-critical nature of the mission, multiple instances with rapid switching of speakers exist. In many cases, extremely short duration responses are common. As an example, the average duration for speakers during mission status updates are close to 0.5 seconds with as many as 15 speakers occurring in turns in a span of 10 seconds. This poses a serious challenge for speaker recognition and diarization systems. The speech density varies dramatically over time, depending on the channel, the stage of the mission, and issues encountered during the mission. Some instances show more than 20 active speakers at a time, carrying conversations for 15 minutes at a stretch, and some other instances show extended periods of silence, usually for hours. For astronauts, their vocal tract characteristics have been observed to have considerable changes through different stages of the mission. These factors render thresholding mechanisms and diarization system performances degraded. The conversational content in the missions was specific to the standards maintained by NASA for efficient air-to-air, air-to-ground, and ground communications. Using standard language models for this application can lead to misclassification of words detected, posing a significant challenge for speech recognition. Hence, there is a need for incorporating NASA specific vocabulary to existing language models. Analyzing unprompted speech has its own challenges. All the speech recorded in the corpus is unprompted, and hence subject to significant variations in speech characteristics for every speaker. These are some challenges that have shown to significantly degrade the quality of generalized SAD, ASR, SID, and Diarization models. Apollo specific application developments have shown to drastically improve system accuracy on these models. [1]

4.4 General Statistics

Due to the communication characteristics observed for the audio data, there is a presence of background conversation speech in most of the audio. Previous efforts have concentrated on analysis of only primary conversation speech, which is what is presented below. The distribution of total primary conversation speech content in each of the channels for every event has been given in Table 1, with the total content for each event provided in the final column, and over each channel provided in the final row.

	EECOM	FD	GNC	MOCR	NTWK	Total
Lift Off	2.1	1.2	1.3	0.8	3.9	9.3
Lunar Landing	3.7	1.3	4.0	0.9	4.4	14.3
Lunar Walking	3.9	1.1	3.0	1.4	2.8	12.2
Total	9.7	3.6	8.3	3.1	11.1	35.8

Table 1: Total Speech Durations (hours) per Channel and Event

To make sure there is an equitable distribution of data into training, test, and development sets for the challenge tasks, we have categorized the data based on noise levels, amount of speech content, and amount of silence. Due to the long silence durations, and based on importance of the mission, the speech activity density of the corpus varies throughout the mission.

Tables 2 and 3 are a general analysis of the 100 hours, aiming to shed some light on the properties

[1] Fearless Steps Overview Paper Interspeech 2018

	EECOM	FD	GNC	MOCR	NTWK
SNR (Mean)	13.32	14.67	14.91	5.07	10.68
SNR (Std. Dev)	7.40	10.51	11.96	12.60	11.17

Table 2: Signal to Noise Ratio Statistics (dB SNR) per channel for Dev Data

	EECOM	FD	GNC	MOCR	NTWK
Avg. Num. Speakers	16	11	21	13	24
Avg. Speaker Dur.	23.04 s	28.74 s	25.18 s	22.36 s	17.12 s
Speaker Dur. (Std. Dev)	6.72 s	6.08 s	5.58 s	5.65 s	4.97 s

Table 3: Speaker Statistics for Dataset

of the data. Average number of speakers and the average and variation of speaker duration per 30 minute file of a channel is provided. In addition, the average and variation of SNR within each channel file is also displayed. Researchers will be provided with the channel information for the released data after the Challenge concludes.

4.5 Development Set

For all tasks with the exception of SID, the Dev set consists of a total duration of 20 hours and 10 minutes and consists of around 60% audio from clean channels and the other 40% is from degraded channels. For the SID task, a separate Dev set is provided. (See Section .5.3)

4.6 Training Set

Around 60 hours of audio data will be provided with manually transcribed labels for each task. Detailed information regarding the baseline systems and results will be released in a separate document. Researchers may use this data as they see fit.

4.7 Evaluation Set

Only the audio files will be provided for evaluation. The Eval set will consist of similar amounts of audio data of every channel, comprising of 20 hours in total. The helpful statistics about the Eval set are given in Tables 3 and 2.

5 Challenge Tasks

As an effort to motivate an initial research direction, five core speech and language tasks. These following Tasks are designed to advance research efforts not only in speech processing and machine learning, but also in natural language understanding. These five tasks include selected speech segments which would have the highest impact in terms of speech, text and language analysis. As the focus of this Challenge is mostly core speech tasks, speech content from both primary and background conversations is considered to form ground truth labels.

5.1 TASK#1: Speech Activity Detection (SAD)

The goal in the SAD task is to automatically detect the presence of speech segments in audio recordings of variable duration. A system output is scored by comparing the system produced start and end times of speech and non-speech segments in audio recordings to human annotated start and end times. Correct, incorrect, and partially correct segments will determine error probabilities for systems and will be used to measure a system's SAD performance by calculating the Detection Cost Function (DCF) value. The DCF is a function of false-positive (false alarm) and false-negative (missed detection) rates calculated from comparison to the human annotation that will be the reference for the comparison[2]. The goal for system developers will be to determine and select their system detection threshold, θ, that minimizes the DCF value.

5.2 TASK#2: Speaker Diarization (SD)

Speaker diarization has received much attention by the speech community, and while there are many available state-of-the-art systems for telephone speech, broadcast news and meetings, their performance does not translate to naturalistic speech in highly degraded noise environments. Some of the challenges diarization systems can encounter with the Apollo data are mentioned in Section 4.3. This challenge is focused on Diarization from scratch.

5.3 TASK#3: Speaker Identification (SID)

In addition to the issues faced by diarization systems, Speaker Identification system performance also relies on speech content per segment. Contiguous speech by a single speaker between 0.4 and 50 seconds have been observed in this data, and a significant portion of short utterances exist in the Corpus. With over 350 known speakers contributing in varying degree of content, the sample set of speakers is narrowed down to 218 speakers with at least 10 seconds of total speech content, that are distributed in the Dev and Eval Sets. Table 4. displays the necessary information.

The primary focus of this challenge will be in-set identification of speakers with drastically varying duration of speech. A simple Top-5 accuracy metric to gauge system performance is mentioned in Section 5.6.

[2] https://www.nist.gov/itl/iad/mig/opensat

Table 4: General statistics for the SID task. The mean, median, minimum, and maximum values for cumulative speaker durations, and individual speaker utterances are all expressed in seconds.

Data set	**# Spkrs**	**Spkr. Duration (s)**			**Spkr. Utterances (s)**		
		mean	**median**	**(min , max)**	**mean**	**(min , max)**	**total**
Train	218	505.5	106.7	(6.89 , 11254.36)	4.03	(1.84 , 16.95)	27336
Dev	218	118.1	24.2	(3.13 , 2596.18)	4.04	(1.78 , 16.95)	6373
Eval	218	156.9	31.5	(3.19 , 3460.41)	4.04	(1.8 , 16.22)	8466

5.4 TASK#4: Automatic Speech Recognition (ASR)

The goal of the ASR task is to automatically produce a verbatim, case-insensitive transcript of all words spoken in an audio recording. ASR performance is measured by the word error rate (WER), calculated as the sum of errors (deletions, insertions and substitutions) divided by the total number of words from the reference. Sections of audio which could not be labeled by manual annotators will be provided in a separate folder. These sections will not be considered for scoring. This segment information will be updated on the official Challenge website.

5.5 TASK#5: Conversational Analysis (CONV)

Identifying salient events in a continuous audio stream is important. Its significance is even more for time-critical events such as the Apollo Missions. Identification of 'hotspots' in over 150,000 hours can help both STEM and non-STEM researchers in analysing high value content. We define the following critical events as 'hotspots'.

- Alarms
- Loss of Signal
- Acquisition of Signal
- mission critical status checks
- mission landings (earth-reentry and lunar)
- mission liftoffs (earth and lunar)
- Team switch
- Checkpoints

These hotspots are defined a conversational context. All mission control speech segments with the above mentioned events as primary topics of discussion between personnel are considered for this task.

Several studies have been motivated by the idea that involvement detection should help pinpoint important events in multi- party dialogue. The assumption is that parts of the meeting where participants are highly involved will be of interest to external viewers, and hence can act as cues in tasks such as meeting summarization. However, manual annotation of involvement is time consuming and costly. In order to deal with increasing amounts of multiparty meeting recordings becoming available, we would like to know whether automatically derived features can be used directly in detecting those noteworthy segments of meetings to include in a summary. Currently, involvement detection is generally treated as a supervised learning problem. However, studies differ in how involvement is annotated as ground truth, varying in domain (dialogue act vs conversation vs interval) and in who is actually involved (group vs individual). Work on automatic meeting summarization has generally focused on extractive summarization, i.e. creating summaries by selecting individual speaker segments/dialogue acts from the transcript. However, in multiparty dialogue, information may be distributed over several turns from different speakers. This diffusion of information causes difficulty when extraction is done with respect to single speaker segments.

5.6 Performance Metrics

We have followed NIST recommended standards for SAD, SD and ASR tasks, and have kept a simple Top-S accuracy measure for SID Task.

SAD

Four system output possibilities are considered:

1. True Positive (TP) – system correctly identifies start-stop times of speech segments compared to the reference (manual annotation),
2. True Negative (TN) – system correctly identifies start-stop times of non-speech segments compared to reference,
3. False Positive (FP) – system incorrectly identifies speech in a segment where the reference identifies the segment as non-speech, and
4. False Negative (FN) – system missed identification of speech in a segment where the reference identifies a segment as speech.

SAD error rates represent a measure of the amount of time that is misclassified by the system's segmentation of the test audio files. Missing, or failing to detect, actual speech is considered a more serious error than misidentifying its start and end times. A 0.5 s collar, a "buffer zone", at the beginning and end of each speech region will not be scored, taking into account inconsistencies in human annotations. If a segment of non-speech between collars is not 0.1 s or greater, then the collars involved are expanded to include the less-than 0.1 s non-speech. For example, no resulting non-speech segment with a duration of just 0.099 s can exist. Similarly, for a region of non-speech before a collar at the beginning of the file or a region of non-speech after a collar at the end of the file, the resulting non-speech segment must last at least 0.1 s or else the collar will expand to include it. In all other circumstances the collars will be exactly the nominal length. Figure 4: Shows how a <0.1 s non-speech segment (0.09 s) is added to the collars and not used in scoring, and illustrates the relationship between human annotation, the scoring regions resulting from application of the collars, a hypothetical system detected output, and the resulting time intervals from the four system output possibilities.

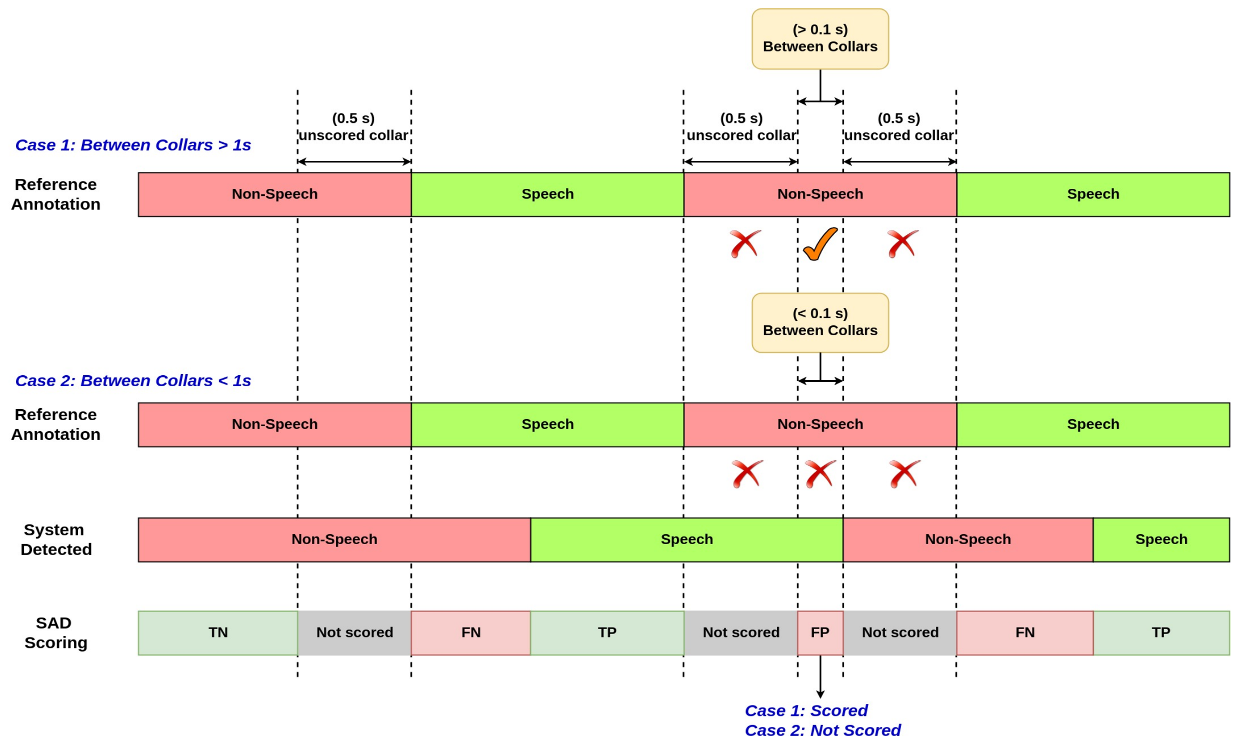

Figure 4: SAD Scoring and Collar Information

The scoring collars also help compensate for ambiguities in noisy channel annotation. Non-speech collars of half a second in length will define those regions that will not be scored. As can be seen, with collars applied to the annotation, parts of system-detected non-speech and potentially speech are not used in scoring. Below illustrates an example of a system detected output and the resulting scoring zones relative to the annotation with 0.5 s collars applied. The figure shows the resulting four possibilities (TN, FN, TP, FP) considered in the scoring. The gray areas preceding and trailing the annotated speech are the 0.5 s collar regions.

Scoring Procedure: Information for downloading the scoring software will be available at the OpenSAT website. The four system output possibilities mentioned above determine the probability of a false positive (P_{FP}) and the probability of a false negative (P_{FN}). Developers are responsible for determining a hypothetical optimum setting (θ) for their system that minimizes the DCF value.

P_{FP} = detecting speech where there is no speech, also called a "false alarm";
P_{FN} = missed detection of speech, i.e., not detecting speech where there is speech, also called a "miss";

$$P_{FP} = \frac{total\ FP\ time}{annotated\ total\ nonspeech\ time}$$
$$P_{FN} = \frac{total\ FN\ time}{annotated\ total\ speech\ time}$$

DCF(θ) is the detection cost function value for a system at a given system decision-threshold setting

$$DCF(\theta) = 0.75 \times P_{FN}(\theta) + 0.25 \times P_{FP}(\theta)$$

P_{FN} and P_{FP} are weighted by 0.75 and 0.25, respectively, θ denotes a given system decision-threshold setting.

SD

Diarization error rate (DER), introduced for the NIST Rich Transcription Spring 2003 Evaluation (RT-03S)[3], 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:

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

where

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

[3] https://catalog.ldc.upenn.edu/LDC2007S10

- **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 segments

The recommended open source scoring tool is maintained as a github repo[4].
To score a set of system output RTTMs *dev_001.rttm* against corresponding reference RTTMs *ref_001.rttm* using the un-partitioned evaluation map (UEM) *dev_001.uem*, the command line would be:

```
python score.py -u dev_001.uem -r ref_dev_001.rttm  -s dev_001.rttm
```

The per-file results for DER will be considered for evaluation. For additional details about scoring tool usage, please consult the documentation given in the github repository.

Labeling Criterion:

- Speech segments for which human annotators were unable to identify the the Speaker have been labeled "UNK" and will not be scored
- Overlap instances will not be considered for Diarization
- All the scoring regions will be provided in the UEM files, (refer to Section B.1). For the dev set, the UEM will be available for download.
- Consecutive segments of the same speaker with a non-speech segment of ≤ 1 second come together and are considered as a single segment.
- A forgiveness collar of 0.25 seconds, before and after each reference boundary, will be considered in order to take into account both inconsistent human annotations and the uncertainty about when a speaker begins or ends.

SID

The SID Task will be evaluated for Accuracy of the Top-3 system predictions for a given input file.

[4] https://github.com/nryant/dscore

$$Accuracy = \frac{\sum_{i \in S} N_{sys}(i)}{\sum_{i=1}^{M} N_{ref}(i)}, \qquad S = \{k \in [1, M] : N_{ref}(k) \subseteq N_{sys}(k)\}$$

where,
$N_{ref}(i)$ = speaker labels from ground truth for i^{th} segment,
$N_{sys}(i)$ = system predicted speaker labels for i^{th} segment,
M = total number of segments

Labeling Criterion:

- Speech segments for which human annotators were unable to identify the Speaker have been labeled "UNK" and will not be provided as instances in Dev the Eval sets.
- Speech segments which have an associated speaker sabel may contain background noise and presence of speech from background conversations.
- Consecutive segments of the same speaker with a non-speech segment of ≤ 2 seconds come together and are considered as a single segment.

ASR

Manual transcription for the Dev and Eval sets has been carried out using hand annotated SAD labels as a starting point.

System performance metric computation

- An overall Word Error Rate (WER) will be computed as the fraction of token recognition errors per maximum number of reference tokens (scorable and optionally deletable tokens):

$$WER = \frac{(N_{Del} + N_{Ins} + N_{subst})}{N_{Ref}}$$

where,

N_{Del} = number of unmapped reference tokens (tokens missed, not detected, by the system)
N_{Ins} = number of unmapped system outputs tokens (tokens that are not in the reference)
N_{Subst} = number of system output tokens mapped to reference tokens but non-matching to the reference spelling
N_{Ref} = the maximum number of reference tokens (includes scorable and optionally deletable reference tokens)

The tool for WER calculation is provided by Kaldi [5]

Labeling Criterion:

- Segments or parts of speech unintelligible to the annotators are marked as [unk] and will not be scored during evaluation
- Overlap instances with SNR $\leq$ 0 dB are marked as [unk] and will not be scored during evaluation
- Consecutive segments of the same speaker with a silent of less that 2 seconds come together and are considered as a single segment.

CONV

Both tracks in the Conversational Analysis Task will be evaluated for Accuracy of the Top-3 system predictions for a given input file.

$$Accuracy = \frac{\sum_{i \in S} N_{sys}(i)}{\sum_{i=1}^{M} N_{ref}(i)}, \qquad S = \{k \in [1, M] : N_{ref}(k) \subseteq N_{sys}(k)\}$$

where,
$N_{ref}(i)$ = speaker labels from ground truth for i^{th} segment,
$N_{sys}(i)$ = system predicted speaker labels for i^{th} segment,
M = total number of segments

Labeling Criterion:

- Segments or parts of speech unintelligible to the annotators are marked as [unk] and will not be scored during evaluation
- Overlap instances with SNR $\leq$ 0 dB are marked as [unk] and will not be scored during evaluation
- Consecutive segments of the same speaker with a silent of less that 2 seconds come together and are considered as a single segment.

[5] http://kaldi-asr.org/doc/tools.html

6 Evaluation Conditions

In the event that researchers have access to the entire Fearless Steps Corpus, they may not use the audio data corresponding to the following tapes for any task: 868, 869, 870, 883, 884, 885, 886[6]. Apart from the aforementioned data, Researchers may use any other data of their choice for system training and development. All the challenge data to be released (audio and transcriptions) are given in Table. 5.

Dataset Release	**Tasks**
Data → Tracks → Train	SID, ASR, CONV
Data → Tracks → Dev	SAD, SD, ASR, CONV
Eval → Data → Tracks	SAD, SD, ASR, CONV
Data → Speakers → Train	SID
Eval → Data → Speakers	SID

Table 5: Task-Data assignment description

7 Evaluation Rules

1. Site registration will be required in order to participate
2. Researchers who register but do not submit a system to the Challenge are considered withdrawn from the Challenge
3. Researchers may use any audio and transcriptions to build their systems, with the exception of the data mentioned in Section 6.
4. Only the audio for the blind eval set (20 hours) will be released. Researchers are expected to run their systems on the blind eval set.
5. Investigation of the evaluation data prior to submission of all systems outputs is not allowed. Human probing is prohibited.

All Challenge participants are required to submit a conference paper(s) describing their systems (and reporting performance on Dev and Eval sets) to the "FEARLESS STEPS CHALLENGE PHASE-3" Special Sessions section at ISCA INTERSPEECH-2021.

[6]Tape Identification Numbers labeled by NASA

8 Evaluation Protocol

- The entire Fearless Steps Corpus (consisting of over 11,000 hours of audio from the Apollo-11 Mission) including the 100 hours is publicly available and requires no additional licence to use the data.
- There is no cost to participate in the Fearless Steps evaluation. Development data and evaluation data will be freely made available to registered participants.
- At least one participant from each team must register on the Fearless Steps Challenge 2021.
- System output submissions will be sent to the official Fearless Steps correspondence email-id. See Appendix. D for system output submission packaging.
- Participants can submit at most 2 system submissions per day.
- Results of submitted systems will be mailed to the registered email-id within a week of the submission.
- It is required that participants agree to process the data in accordance with the following rules.

9 System Input and Output

9.1 Audio Files

All audio files provided will be in the standard single channel (mono) 16-bit PCM 'WAV' format, sampled at 8000 Hz.

For SAD, SD, ASR and SENTIMENT Tasks: Each file in the Train, Dev, and Eval set have a duration between 30 and 33 minutes. There are no other exceptions.

For SID Task: The Dev and Eval sets for this task have segmented Files with a single speaker information. The associated speaker label for the Dev set will be provided in the file name. These files may have a duration ranging from 2.2 to 15 seconds (inclusive).

9.2 Speech Activity Detection (SAD)

For both Input and Output file format descriptions, see Appendix A.3

9.3 Speaker Diarization (SD)

For the non-scoring region Input File format description, see Appendix B.1

For both Input and Output file format descriptions, see Appendix A.1

9.4 Speaker Identification (SID)

For Output file format description, see Appendix A.4

9.5 Automatic Speech Recognition (ASR)

For both Input and Output file format descriptions, see Appendix A.2

9.6 Conversational Analysis (CONV)

For both Input and Output file format descriptions, see Appendix A.2

10 Updates

During registration, researchers are advised to provide an email address they are most active on. Any other updates and changes will be displayed on the website Fearless Steps Challenge Phase-3 and sent through a mailing list to all registered participants. Contact Fearless Steps for questions relevant to FS-3 2021 not covered in this evaluation plan.

11 Interspeech 2021 Special Session

The results of the challenge will be presented at a special session at Interspeech 2021, held from August 30^{th} - September 3^{rd}, 2021 in Brno, Czech Republic. Researchers wishing to submit papers should do so through the Interspeech submission portal. Additional instructions will be provided once the Interspeech submission portal opens. Researchers will be notified privately of their final system ranking, on 1^{st} April (deadline for submission of the paper update deadline). System rankings will be announced publicly on the 50^{th} Anniversary of the first moon walk.

12 Acknowledgements

This project was supported in part by AFRL under contract FA8750-15-1-0205, NSF-CISE Project 1219130, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H. L. Hansen. We would also like to thank Tatiana Korelsky and the National Science Foundation (NSF) for their support on this scientific and historical project. A special Thanks to Katelyn Foxworth for leading the ground-truth development efforts for the FS-02 Challenge Corpus.

13 Licensing Details

All the conversations between Astronauts and and Mission Control Personnel during the Apollo-11 Mission were recorded by NASA. The tireless efforts of CRSS-UTD transcribers and researchers contributed to the shaping of this enormous amounts of data into a well-defined corpus to address various speech and language tasks for naturalistic audio, a portion of which is now made publicly available to the speech community through this Challenge via a creative commons license.

Note: The Creative Commons License is restricted to the efforts made by CRSS-UTD, which involves 100 hours of Challenge Corpus (audio) data sampled from 8Khz, along with its meta-data generated separately. The license also covers all the scripts which were used in the preparation of the corpus and systems built to support the tasks in this Challenge, along with the webpages developed to host the Challenge.

The Entire Fearless Steps Corpus consisting of over 11,000 hours of audio from the Apollo-11 Mission is publicly available under the 'NASA Media Usage Guidelines'.

— NASA Media Usage Guidelines —

Still Images, Audio Recordings, Video, and Related Computer Files for Non-Commercial Use:

NASA content - images, audio, video, and computer files used in the rendition of 3-dimensional models, such as texture maps and polygon data in any format - generally are not copyrighted. You may use this material for educational or informational purposes, including photo collections, textbooks, public exhibits, computer graphical simulations and Internet Web pages. This general permission extends to personal Web pages.

News outlets, schools, and text-book authors may use NASA content without needing explicit permission. NASA content used in a factual manner that does not imply endorsement may be used without needing explicit permission. NASA should be acknowledged as the source of the material. NASA occasionally uses copyrighted material by permission on its website. Those images will be marked copyright with the name of the copyright holder. NASA's use does not convey any rights to others to use the same material. Those wishing to use copyrighted material must contact the copyright holder directly.

For Additional Information regarding Commercial and Non-Commercial Use: Please visit: NASA Multimedia Guidelines Webpage.

Appendix A Label File Format Specification

A.1 RTTM File Format

Systems should output their diarizations as Rich Transcription Time Marked (RTTM) files [7]. A NIST defined File Format, the RTTM files are text files containing one turn per line, each line containing nine space-delimited fields:

- **Type** segment type; should always by "SPEAKER"
- **File ID** file name; basename of the recording minus extension (e.g., "FS_P01_eval_023")
- **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*** (*Optional*) system confidence (probability) that information is correct; should always be <NA>

For instance:

```
SPEAKER FS_P01_dev_001 1 256.04 2.35 <NA> <NA> EECOM1 <NA>
SPEAKER FS_P01_dev_001 1 358.08 3.06 <NA> <NA> FD1 <NA>
SPEAKER FS_P01_dev_001 1 368.97 2.22 <NA> <NA> GNC1 <NA>
```

[7] https://web.archive.org/web/20170119114252/http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf

A.2 JSON File Format

The transcriptions are provided in JSON format for each file as <file_ID>.json [8]. The JSON file includes the following pieces of information for each utterance:

- **Speaker ID** Token: "speakerID"
- **Transcription** Token: "words"
- **Conversational Label** Token: "conv"
- **Start Time** Token: "startTime"
- **End Time** Token: "endTime"

For instance:

```
{
  "speakerID":  "NEIL",
  "words":      "It's one small step for man,",
  "conv":       "ASTRO_TALK"
  "startTime":  "1325.203"
  "endTime":    "1327.501"
},
{
  "speakerID":  "NEIL",
  "words":      "One, Giant leap for mankind.",
  "conv":       "ASTRO_TALK"
  "startTime":  "1330.162"
  "endTime":    "1332.89"
}, ....
```

[8] https://www.json.org/

A.3 SAD File Format

Systems should output their SAD as text (txt) files [9] A NIST defined File Format, the text files are text files containing one turn per line, each line containing nine tab-delimited fields:

- **Test** Test Definition File Name (Value: X)
- **TestSet ID** contents of the id attribute TestSet tag (Value: X)
- **Test ID** contents of the id attribute of the TEST tag (Value: X)
- **Task** SAD <== a literal text string, without quotations (Value: SAD)
- **File ID** contents of the id attribute of the File tag (Value: X)
- **Interval start** an offset, in seconds from the start of the audio file for the **start** of the speech/non-speech interval (Value: floating number)
- **Interval end** an offset, in seconds from the start of the audio file for the **end** of the speech/non-speech interval (Value: floating number)
- **Type** In system output: speech/non-speech without quotation marks (Value: speech/non-speech)
  In the reference: S/NS for speech/non-speech
- ***Confidence Score*** (*Optional*) A value in the range 0 thorugh 1.0, with higher values indicating greater confidence about the presence/absence of speech

For instance: SAD system output file

```
X   X   X   SAD   X   0.00   5.77    speech       0.500000
X   X   X   SAD   X   5.77   6.37    non-speech   0.500000
X   X   X   SAD   X   6.37   11.22   speech       0.500000
```

Interval overlapping will be disallowed and will fail in validating your files. Example of overlapping:

X	X	X	SAD	X	0.00	**5.77**	speech	0.500000
X	X	X	SAD	X	**5.13**	6.37	non-speech	0.500000

[9] http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf

A.4 SID Output File Format

The SID output file should be a text file containing one test-segment per line, each line containing five space-delimited fields

- **Test** Test Definition File Name
- **Prediction 1** Top System SpeakerID Prediction
- **Prediction 2** 2nd Most Likely System SpeakerID Prediction
- **Prediction 3** 3rd Most Likely System SpeakerID Prediction
- **Prediction 4** 4th Most Likely System SpeakerID Prediction
- **Prediction 5** 5th Most Likely System SpeakerID Prediction

For instance: SID system output file

```
FS_P01_dev_FD1_001    FD1   GNC1  INCO      NEIL  BUZZ
FS_P01_dev_NEIL_025   FD1   AGC   FIDO2     NEIL  BUZZ
FS_P01_eval_001       FD1   GNC2  INCO      NEIL  BUZZ
FS_P01_eval_005       FIDO  GNC1  GUIDANCE  NEIL  BUZZ
```

File Naming conventions: No additional text file is provided for the speaker identification labels. Every segment consists of one speaker to evaluate. Researchers are expected to retrieve the Speaker label for each segment from the file name.

Development Set:

```
FS_P01_dev_<Speaker ID>_<Dev Utterance ID>
```

Evaluation Set:

```
FS_P01_eval_<Eval Utterance ID>
```

Appendix B Supplementary File Format Specification

The Supplementary Files are any files in addition to the audio and (system and ground-truth) label files that are necessary to evaluate the performance of a system accurately for all given Tasks.

B.1 UEM File Format

The scoring region for each audio file is provided separately through the NIST defined UEM format[10]. These regions will be provided to the scoring tool via UEM files. Un-partitioned evaluation map (UEM)[11] 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., "FS_P01_dev_001")
- **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:

```
FS_P01_dev_001   1  234.32  737.54
FS_P01_eval_001  1  832.35  1800.00
FS_P01_eval_007  1  124.46  624.23
```

[10] https://web.archive.org/web/20170119114252/
[11] https://catalog.ldc.upenn.edu/docs/LDC2004S11/readme.html#INPUT_UEM

Appendix C Data Resources for Training

This appendix identifies a (non-exhaustive) list of publicly available corpora the researchers may use for system training.[12] [13]

Corpora containing meeting speech

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

Conversational telephone speech (CTS) corpora

- 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)

[12] https://www.ldc.upenn.edu/
[13] https://coml.lscp.ens.fr/dihard/2018/data.html

Other 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)
- 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:

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

Section 1: Task

The Challenge Task for which the system is to be evaluated. Researchers can submit multiple systems for one or all of the tasks. However, each Submission should include a single system evaluation for a single task. If the same team of researchers wishes to submit multiple systems for a separate task, that submission should be done separately.

Section 2: Abstract

A 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 other publicly available corpora 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 re-implement it. 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.

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 30 minute File must be reported.

For SID: system execution time for entire Eval Set must be reported