Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for Unseen Channel and Mission Data across NASA Apollo Audio

Aditya Joglekar[1], *Seyed Omid Sadjadi*[2], *Meena Chandra-Shekar*[1], *Christopher Cieri*[3], *John H.L. Hansen*[1]

[1]Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering, The University of Texas at Dallas (UTD), Richardson, Texas, USA
[2]NIST ITL/IAD/MIG, Gaithersburg MD, USA
[3]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, USA

{aditya.joglekar, meena.chandrashekar, john.hansen}@utdallas.edu[1], omid.sadjadi@nist.gov[2], ccieri@ldc.upenn.edu[3]

Abstract

The Fearless Steps Challenge (FSC) initiative was designed to host a series of progressively complex tasks to promote advanced speech research across naturalistic "Big Data" corpora. The Center for Robuts Speech Systems at UT-Dallas in collaboration with the National Institute of Standards and Technology (NIST) and Linguistic Data Consortium (LDC) conducted Phase-3 of the FSC series (FSC P3), with a focus on motivating speech and language technology (SLT) system generalizability across channel and mission diversity under the same training conditions as in Phase-2. The FSC P3 introduced 10 hours of a previously unseen channel audio from Apollo-11 and 5 hours of novel audio from Apollo-13 to be evaluated over previously established and newly introduced SLT tasks with streamlined tracks. This paper presents an overview of the newly introduced conversational analysis tracks, Apollo-13 data, and analysis of system performance for matched and mismatched challenge conditions. We also discuss the Phase-3 challenge results and evolution of system performance across the three Phases, and next steps in the Challenge Series.

Index Terms: NASA Apollo, speech activity detection, speaker diarization, speaker identification, speech recognition, conversational analysis.

1. Introduction

The importance of naturalistic datasets in the advent of a technology revolution led by deep neural networks has become paramount. With deep learning system performance linearly scaling with the amount of data provided, naturalistic "Big Data" corpora have become a benchmark in determining a competitive edge in the domain of artificial intelligence (AI). Developing good quality naturalistic datasets is challenging [1, 2, 3, 4].

There are additional challenges in developing data with rich information content in multiple SLT domains [5, 6]. The NASA Apollo missions data is a collection of 150,000+ hours of audio with unprompted multi-party conversations recorded over 30 channels. The preserved data contains over 450 personnel in constant air-to-air, air-to-ground, and ground-to-ground communication working collaboratively to solve time-critical challenges. Speech and natural language systems can significantly benefit from this preserved data and vice versa. The Fearless Steps Challenge (FSC) series has led the efforts in promoting such system development by annotating a small portion of this Apollo corpus (115 hours) and establishing challenge tasks to benchmark SLT systems [6, 7, 8].

The goal of FSC Phase-3 (FSC P3) is to assess the ability of speech and language systems in providing consistent performance across channel, noise, speech, speaker, and semantic variabilities. The FSC P3 evaluation set provides such a test platform, sourcing its data from multiple channels from Apollo-11 and Apollo-13 missions [9, 10, 11].

FSC P3 was conducted in collaboration with the National Institute of Standards and Technology (NIST) and Linguistic Data Consortium (LDC) from February through March of 2021. Nine participating organizations, with 14 task-specific teams contributed 235 system submissions. Phase-3 evaluation reported state-of-the-art (SOTA) results on 3 of the 8 challenge tracks. The current edition of the FSC was run entirely online using the NIST OpenSAT evaluation platform (`https://sat.nist.gov/fsc3`). The web platform supported a variety of services including evaluation registration, data distribution, system output submission, submission validation, scoring, and system description/presentation uploads [5, 12, 13].

2. Challenge Tasks

The NASA Apollo Missions audio data were recorded on 30-channel analog tapes ranging from 14 to 17 hours in duration. Large data streams of this nature are often a hindrance in effective development of SLT systems. In an effort to streamline such development, 30-minute audio chunks were annotated from the most high-impact events in Apollo 11 and Apollo 13 missions. These chunks have been presented in the FSC corpus as "audio streams". These streams of undiarized audio are supplemented by annotation files with diarized labels. Essentially, audio streams for Training (*Train*) and Development (*Dev*) sets provided to challenge participants contain a single markup-styled annotation file per stream. Each annotation file contains a ground-truth label or transcription per single-speaker utterance. Participants are expected to diarize the Evaluation (*Eval*) set audio streams in addition to the primary task. This requires using multiple systems to process downstream tasks, often propagating error at each functional block. The FSC P2 established the necessity for developing separate speaker diarization (SD) and automatic speech recognition (ASR) tracks to streamline the development of effective clustering and acoustic models/language models respectively [3, 4, 14].

This format has been extended for the FSC P3, which provides separate tracks for diarized "audio segments" in the speaker diarization (SD), automatic speech recognition (ASR), and conversational analysis (*CONV*) tasks. The entire list of tasks and tracks available for the FSC P3 are presented below. The five original channels (*FD*, *MOCR*, *EECOM*, *GNC*, *NTWK*) used in the FSC P1 and P2 have been preserved in Train and Dev sets, with no additions to the data-sets [6, 7, 8].

- *TASK 1*: Speech Activity Detection ***(SAD)***
- *TASK 2*: Speaker Identification ***(SID)***
- *TASK 3*: Speaker Diarization
  - (3.a.) *Track 1*: using system SAD ***(SD_track1)***
  - (3.b.) *Track 2*: using reference SAD ***(SD_track2)***
- *TASK 4*: Automatic Speech Recognition
  - (4.a.) Track 1: using system SAD ***(ASR_track1)***
  - (4.b.) Track 2: using diarized audio ***(ASR_track2)***
- *TASK 5*: **Conversational Analysis**
  - (5.a.) Track 1: using system SAD ***(CONV_track1)***
  - (5.b.) Track 2: using diarized audio ***(CONV_track2)***

Tasks established in previous challenges are still core speech tasks, and do not directly benefit spoken language understanding (SLU) of the multi-party conversations. With SOTA word error rates (WER) as high as 24%, SLU systems cannot be expected to effectively extract meaningful information regarding topic, sentiment, emotion, or semantic context. This effect was observed in the FSC P1 sentiment detection task. Accordingly, a new task with separate tracks was created with an aim to identify key conversational moments in the data [9, 15, 16, 17].

2.1. Conversational Analysis

Methodologies to identify salient events in continuous audio streams can significantly reduce the cost of information retrieval [18, 19, 20, 21]. The significance of such methodologies is more pronounced for the Apollo Missions, which can have intermittent time-critical events, followed by large periods of inactivity or normal conversations. Identification of such "Hotspots" in over 150,000 hours of audio data can help both STEM and non-STEM researchers in efficient retrieval and analysis of high value content. These hotspot events are essentially an extractive summarization of conversations between Mission Control personnel. A total of 25 conversational cues critical to successful deployment of the Apollo Missions were identified as conversational "Hotspots" and presented as diarized segments for a classification task termed "Hotspot Detection" (CONV_track2). The task of finding salient events and providing accurate extractive summarization in continuous audio streams is presented as a separate track for the FSC P3 (CONV_track1).

2.2. Challenge Deployment

The NIST OpenSAT web platform (`https://sat.nist.gov`) was used to conduct the FSC P3. Participants were allowed to download Train, Dev, and Eval sets after agreeing to the terms and conditions of the Challenge. The scoring toolkit developed for the FSC P2 was used in the validation and scoring back-end for the platform. Participants were provided with basic analytics for their submissions to assist in their system development efforts. Every team was allocated a single submission slot per task/track with multiple submissions (up to 3 per day) allowed to update the system performance.

2.3. Performance Metrics

The performance metrics and conditions for the FSC P3 were largely the same as the conditions for the FSC P2. A NIST defined detection cost function (DCF) measure was used for scoring the speech activity detection (SAD) task, with a forgiveness collar of *0.25* seconds, which was reduced from the collar duration of *0.5* seconds for the FSC P2. Both tracks for SD and ASR were evaluated using diarization error rate (DER) and word error rate (WER) for the same testing conditions as the FSC P2. Speaker Identification (SID) task performance metric was updated from Top-5 % Accuracy (Top-5 Acc.) to Top-3 % Accuracy (Top-3 Acc.). The newly introduced task tracks CONV track-1 and track-2 were evaluated using separate metrics. Track-2 using diarized segments was evaluated using Top-3 % Accuracy, and track-1 using audio streams was tested using the measure Jaccard error rate (JER) [22, 23, 24]. The Jaccard index has been traditionally used in image segmentation and more recently in speaker diarization in the DIHARD Challenge series [22, 25]. JER as an initial measure provides a good representation of diarizing and identifying conversational labels for the diarized segments. For each reference speaker *ref* the speaker-specific JER_{ref} is computed as:

$$\text{JER}_{\text{ref}}(\%) = \left(\frac{FA + Miss}{Total}\right) \times 100, \tag{1}$$

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,
- *Miss* is the total reference speaker time not attributed to a system speaker.

The JER metric for CONV track1 could be replaced with a different metric for future Phases depending on the evolution of the data and labels developed for this task.

3. Data

The amount of raw unlabelled data generated in the world has been exponentially increasing compared to the available ground-truth annotated data. A key feature of robust SLT systems is their ability to adapt to such real-world data with varying acoustic characteristics, in low-resource training conditions.

3.1. Unseen Channel & Mission

The operations and propulsion (OPS&PRO) channel from Apollo-11 was annotated for testing system performance over unseen noise characteristics. It provides a good basis for testing multiple tasks, since the content of speech due to the technical nature of the channel will be different from the other Apollo-11 channels which have more general mission status related conversations. A total of 5 hours of Apollo-13 data was selected from three separate channels, providing additional variability in channel noise, air-to-ground communication noise, different speakers, and speech conversations of a very different nature compared to the conversations seen in Apollo-11.

3.2. General Statistics

Table 3 summarizes the overall statistics for the audio streams data. The key feature we notice is the variability in the Eval set even for basic analysis parameters. Tables 1 and 2 present

Table 1: *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 [7]*

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	264.3	38.2	(3.19 , 5834.46)	4.09	(1.8 , 16.22)	14077

Table 2: *General statistics for the CONV_track2 task. The mean, median, minimum, and maximum values for cumulative label durations, and individual labels are all expressed in seconds.*

Data set	# Hotspots	Lab. Duration (s)			Lab. Utterances (s)		
		mean	median	(min , max)	mean	(min , max)	total
Train	25	1546.8	796.4	(193.26 , 6274.85)	2.4	(0.5 , 30.4)	16059
Dev	25	464.8	233.7	(45.11 , 2626.05)	2.5	(0.5 , 33.2)	4662
Eval	25	976.3	435.9	(59.05 , 8036.36)	2.6	(0.5 , 29.96)	9360

Table 3: *Overall Statistics of audio streams for the FSC P3. The mean, min, and max values are expressed in seconds.*

	Train	Dev	Eval			
			A-11	A-13	Unseen	**Total**
# Streams	125	30	40	10	18	68
Dur. (hrs)	63.5	15.3	20.1	5	9.1	34.4
% Speech	29.4	32.5	34.4	33.4	37.3	35
#Spkrs/Stream	19	24	20	13	25	20

Table 4: *Duration Statistics of audio segments for ASR_track2. The mean, min, and max values are expressed in seconds [7]*

Data set	Segments	Utterance Duration (s)		
		mean	min	max
Train	35,473	2.85	0.10	70.37
Dev	9,203	2.97	0.12	67.39
Eval	21,846	2.98	0.10	162.75

statistics on the diarized segments for the SID and CONV_track2 tasks, while Table 4 provides insight into the distribution of segments in the Train, Dev and Eval sets.

3.3. Evaluation Set Variability

The unseen channel and mission data present in the evaluation set are set up in a blind format. Participants were not provided with mission or channel labels during the evaluation phase, thereby making it possible to evaluate systems for their generalizability to unseen data conditions.

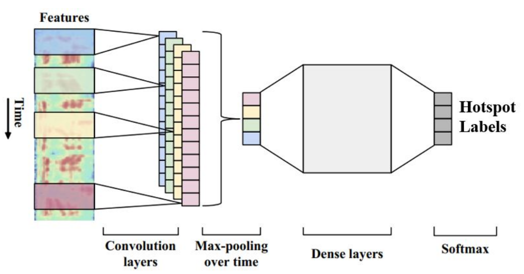

Figure 1: *CNN-Saliency network illustration, used as baseline system for both tracks of Conversational Analysis [26]*

4. Results

In this section, we analyze baseline performance for all tasks, and evaluate the system generalizability factor for submissions from all challenge participants.

4.1. Baseline Results

All the original FSC2 baseline systems were used without modifications for evaluating the FSC3 Eval set [27, 28, 29] with the exception of the extractive summarization tasks (CONV track1 and track2). The baseline results for all tasks are reported in Table 5. We notice a degradation in performance on the FSC3 Eval sets for all tasks/tracks, caused by the addition of the unseen channel and mission data. Since most of the baseline systems are unsupervised and rely on core acoustic features [30, 31], their degraded performance indicates added acoustic complexity in the Eval set due to the unseen mission and channel inclusion.

The baseline system used for both tracks of the conversational analysis task was originally developed for emotion detection[26]. The convolutional network illustrated in Figure 1 is designed to aggregate high level context over time (in this case, log-Mel-Filterbank feature frames) to generate a single summarization, which fits well with the structure of this task.

4.2. Best Systems Comparison

Table 6 provides a comparative illustration of the improvements in SOTA for Phase-2 and Phase-3[32, 33, 34, 35]. We report significant improvements in FSC3 Top system performance over FSC2 for SID task, ASR_track2 and both SD tracks. However, the SOTA for SAD and ASR_track1 tasks are maintained from FSC2. Comparisons to the FSC2 systems have been conducted

Table 5: *Baseline Results for FSC P3 Development and Evaluation Sets [7]*

Fearless Steps Phase-3 Baseline Results			
Task	**Metric**	**Dev (%)**	**Eval (%)**
SAD	DCF	12.84	15.16
SD_track1	DER	79.72	88.27
SD_track2	DER	68.68	77.91
SID	Top-3 Acc.	75.20	72.46
ASR_track1	WER	83.80	92.3
ASR_track2	WER	80.50	86.4
CONV_track1	JER	58.6	71.5
CONV_track2	Top-3 Acc	67.1	54.2

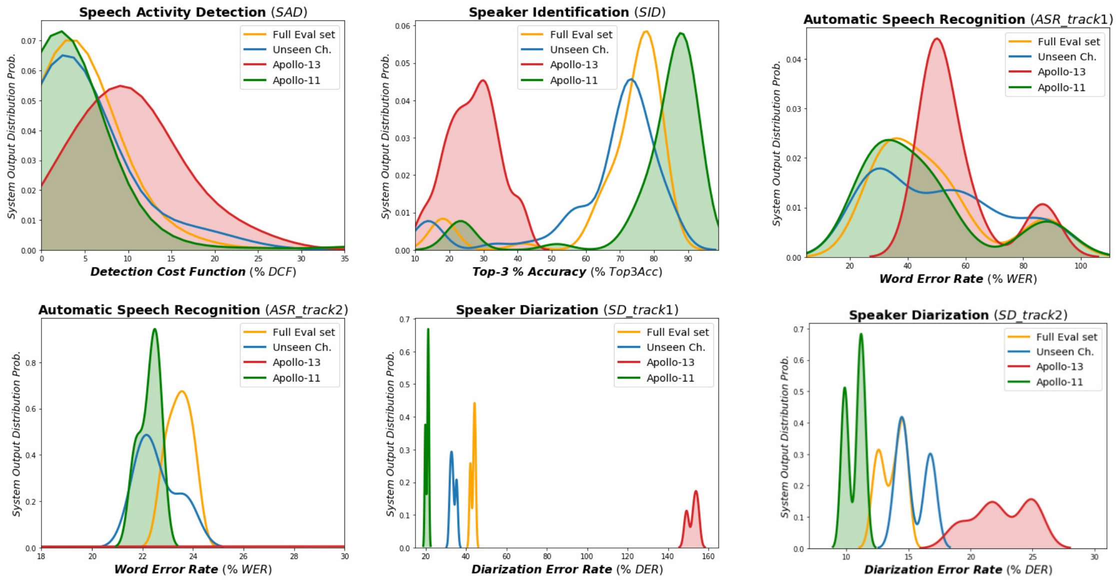

Figure 2: *Distribution of system submission results between the Apollo-11, Unseen Channel, Apollo-13 segments/streams of the Eval set.*

Table 6: *Comparison of the best systems developed for all FSC P2 and P3 challenge tasks.* FS-3 *(in bold) represents system performance over the entire blind set provided for participants. Relative improvement of top-ranked system per task in the FSC P3 seen channel data* (FSC3-A11) *(underlined) over FSC P2 evaluation set is illustrated.* FSC3-A13 *and* FSC3-UnkCh *represent the top system performance for the Unseen Mission and Unseen channel data respectively.*

Comparison of Best System Submissions on FSC3 Eval set (and sub-sets)						
Task	**FSC2 (%)**	**FSC3 (%)**	**FSC3-A11 (%)**	**FSC3-A13 (%)**	**FSC3-UnkCh (%)**	**Rel. Imp. (%)**
SAD	1.07	**1.47**	1.16	2.37	1.67	*-7.57 %*
SID	92.39	**83.27**	93.26	23.77	85.16	**12.9 %**
SD_track1	28.85	**42.20**	19.92	149.1	32.33	**30.95 %**
SD_track2	26.55	**12.32**	9.82	19.05	14.15	**53.59 %**
ASR_track1	24.01	**29.96**	26.87	47.94	26.82	*-11.9 %*
ASR_track2	24.26	**22.85**	21.69	100	21.82	**10.59 %**

over the same data evaluated by participants in FSC2. We also observe that the top systems for every task/track were able to adapt efficiently to the unseen channel data.

4.3. System Generalizability

Figure 2 displays a comparative performance of all system submissions across the Apollo-11, Unseen Channel, and Apollo-13 sub-sets of the Eval set. We observe that while the top systems were able to generalize well to the unseen mission and channel data, a majority of the systems had degraded performance. The unseen channel performance for all tasks was seen to be closely related to the seen channel data, and in some cases even showed improved performance. Based on this analysis, we hypothesize that system generalizability is a key component in developing systems for the remaining Apollo Missions.

5. Discussion

We notice that system generalizability was positively correlated with a better overall performance. We report that the systems competing in the FSC P3 showed improved performance for channel variabilities, but significantly lacked the ability to generalize to the unseen mission data. It should be noted that the imbalance in the Eval sub-sets could have caused some bias in the overall submission results. We plan to create and provide equally weighted channel and mission data for the next challenge.

6. Conclusions

Through the Phase-3 of the FSC, we introduced a new challenge task that aims to extract high level context from conversations. We tested system capability to generalize for previously unseen channel and mission variability. In the next Phase of the Challenge we plan to extend the Training and Development datasets to include more data from the Apollo missions 8 and 10. In conclusion, we assert the need for further development in SLT systems for naturalistic data. Future efforts in the Fearless Steps Challenge series will increasingly involve the goal of developing systems that are highly adaptable and robust to out-of-domain data.

7. 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 (CRSS Transcription Team) for leading the ground-truth development efforts on the FSC P3 Challenge Corpus.

8. References

[1] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, "Two decades of speaker recognition evaluation at the National Institute of Standards and Technology," *Computer Speech & Language*, vol. 60, p. 101032, 2020.

[2] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[3] M. Harper, "The automatic speech recognition in reverberant environments (ASpIRE) challenge," in *Proc. IEEE ASRU Workshop*, 2015, pp. 547–554.

[4] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. INTERSPEECH*, 2018, pp. 1561–1565.

[5] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014, pp. 224–230.

[6] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proc. INTERSPEECH*, 2018, pp. 2758–2762.

[7] A. Joglekar, J. H. Hansen, M. C. Shekar, and A. Sangwan, "Fearless steps challenge (FS-2): Supervised learning with massive naturalistic Apollo data," in *Proc. INTERSPEECH*, 2020, pp. 2617–2621.

[8] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio," in *Proc. INTERSPEECH*, 2019, pp. 1851–1855.

[9] A. Joglekar and J. H. Hansen, "Fearless steps, NASA's first heroes: Conversational speech analysis of the apollo-11 mission control personnel," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2956–2956, 2019.

[10] J. C. Gorman, P. W. Foltz, P. A. Kiekel, M. J. Martin, and N. J. Cooke, "Evaluation of latent semantic analysis-based measures of team communications content," in *Proceedings of the Human Factors and Ergonomics Society annual meeting*, vol. 47, no. 3. SAGE Publications Sage CA: Los Angeles, CA, 2003, pp. 424–428.

[11] P. W. Foltz, D. Laham, and M. Derr, "Automated speech recognition for modeling team performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 47, no. 4. SAGE Publications Sage CA: Los Angeles, CA, 2003, pp. 673–677.

[12] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 NIST speaker recognition evaluation," in *Proc. INTERSPEECH*, 2017, pp. 1353–1357.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *STIN*, vol. 93, p. 27403, 1993.

[14] A. Sangwan, L. Kaushik, C. Yu, J. H. Hansen, and D. W. Oard, "'houston, we have a solution' : Using NASA Apollo program to advance speech and language processing technology," in *Proc. INTERSPEECH*, 2013, pp. 1135–1139.

[15] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 22, no. 12, pp. 1688–1700, 2014.

[16] F. Valente and A. Vinciarelli, "Language-independent socio-emotional role recognition in the AMI meetings corpus," in *Proc. INTERSPEECH*, 2011.

[17] S. Raaijmakers, K. P. Truong, and T. Wilson, "Multimodal subjectivity analysis of multiparty conversation," in *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 466–474.

[18] S. Somasundaran, J. Ruppenhofer, and J. Wiebe, "Detecting arguing and sentiment in meetings," in *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, 2007, pp. 26–34.

[19] J. Carletta and J. Kilgour, "The NITE XML toolkit meets the ICSI meeting corpus: Import, annotation, and browsing," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2004, pp. 111–121.

[20] A. Dielmann and S. Renals, "DBN based joint dialogue act recognition of multiparty meetings," in *Proc. IEEE ICASSP*, vol. 4, 2007, pp. IV–133.

[21] C. Lai, J. Carletta, S. Renals, K. Evanini, and K. Zechner, "Detecting summarization hot spots in meetings using group level involvement and turn-taking features." in *Proc. INTERSPEECH*, 2013, pp. 2723–2727.

[22] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD challenge evaluation plan," 2018.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU Workshop*, 2011.

[24] "NIST rich transcription spring 2003 evaluation," https://catalog.ldc.upenn.edu/LDC2007S10, accessed: 2019-03-01.

[25] N. Ryant, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, "Third DIHARD challenge evaluation plan," *arXiv preprint arXiv:2006.05815*, 2020.

[26] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *Proc. IEEE ICASSP*, 2017, pp. 2741–2745.

[27] Z.-H. Tan, N. Dehak *et al.*, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech & Language*, vol. 59, pp. 1–21, 2020.

[28] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 1021–1028.

[29] H. Dubey, A. Sangwan, and J. H. Hansen, "Robust speaker clustering using mixtures of von Mises-Fisher distributions for naturalistic audio streams," in *Proc. INTERSPEECH*, 2018, pp. 3603–3607.

[30] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, March 2013.

[31] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[32] B. Sharma, R. K. Das, and H. Li, "Multi-level adaptive speech activity detector for speech in naturalistic environments," *Proc. INTERSPEECH*, pp. 2015–2019, 2019.

[33] A. Vafeiadis, E. Fanioudakis, I. Potamitis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Two-dimensional convolutional recurrent neural networks for speech activity detection," in *Proc. INTERSPEECH*, 2019.

[34] A. Gorin, D. Kulko, S. Grima, and A. Glasman, ""This is Houston. Say again, please". the Behavox system for the Apollo-11 fearless steps challenge (Phase II)," *arXiv preprint arXiv:2008.01504*, 2020.

[35] Q. Lin and M. L. Tingle Li, "The DKU speech activity detection and speaker identification systems for fearless steps challenge Phase-02," *Proc. INTERSPEECH*, pp. 2607–2611, 2020.