

FEARLESS STEPS Challenge (FS-2): Supervised Learning with Massive Naturalistic Apollo Data

Aditya Joglekar, John H.L. Hansen, Meena Chandra Shekar, Abhijeet Sangwan

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
The University of Texas at Dallas (UTD), Richardson, Texas, USA

{aditya.joglekar, john.hansen, meena.chandrashekar, abhijeet.sangwan}@utdallas.edu

Abstract

The Fearless Steps Initiative by UTDallas-CRSS led to the digitization, recovery, and diarization of 19,000 hours of original analog audio data, as well as the development of algorithms to extract meaningful information from this multi-channel naturalistic data resource. The 2020 FEARLESS STEPS (FS-2) Challenge is the second annual challenge held for the Speech and Language Technology community to motivate supervised learning algorithm development for multi-party and multi-stream naturalistic audio. In this paper, we present an overview of the challenge sub-tasks, data, performance metrics, and lessons learned from Phase-2 of the Fearless Steps Challenge (FS-2). We present advancements made in FS-2 through extensive community outreach and feedback. We describe innovations in the challenge corpus development, and present revised baseline results. We finally discuss the challenge outcome and general trends in system development across both phases (Phase FS-1 Unsupervised, and Phase FS-2 Supervised) of the challenge, and its continuation into multi-channel challenge tasks for the upcoming Fearless Steps Challenge Phase-3.

Index Terms: NASA Apollo 11 mission, corpus, speech activity detection, speaker diarization, speaker identification, speech recognition, multi-channel audio streams, diarized segments.

1. Introduction

Recent decades have seen tremendous improvements to Speech and Language Technology (SLT) systems. This has only been possible due to thoroughly curated speech and language corpora that have been made publicly available [1, 2, 3, 4, 5]. The ability for systems to adapt to, and extract meaningful information from unlabeled data using limited ground-truth knowledge is a challenge in machine learning and AI [6, 7, 8]. Unfortunately, there is an unlimited amount of unstructured and unsupervised data compared to high quality human annotated data. To effectively address this reality, development of solutions will require consistent improvements to SLT systems. The initially digitized 19,000 hours from the NASA Apollo-11 and Apollo-13 missions [9, 10] represent the largest naturalistic time synchronized multi-channel data. This corpus will be supplemented in continuing efforts with an additional 150,000 hours, enabling research on the largest publicly available corpus till date. Structuring this data through pipeline diarization transcripts, automatic speaker/sentiment tagging, etc., will enable preservation and archiving of historical data. These efforts will massively increase research opportunities, and be of significant benefit to the STEM community. As an initial step to motivate this stream-lined and collaborative effort from the SLT community, UTDallas-CRSS has been hosting a series of progressively complex tasks to promote advanced research on naturalistic “Big Data” corpora. This began with the “Inaugural FEARLESS

STEPS Challenge: Massive Naturalistic Audio (FS-1)”. The first edition of this challenge encouraged the development of core unsupervised/semi-supervised speech and language systems for single-channel data with low resource availability, serving as the “First Step” towards extracting high-level information from such massive unlabeled corpora [11, 12, 13, 14, 15]. As a natural progression following the successful inaugural FS-1 challenge, the FEARLESS STEPS Challenge Phase-2 (FS-2) focuses on the development of single-channel supervised learning strategies. FS-2 Challenge provides 80 hours of ground-truth data through training (Train) and development (Dev) sets, with an additional 20 hours of blind-set evaluation (Eval) data. Based on feedback from the Fearless Steps participants, additional tracks for streamlined speech recognition and speaker diarization have been included in the FS-2. To encourage diversified research interests, participants were also encouraged to utilize the FS-2 corpus to explore additional problems dealing with naturalistic data. The results for this challenge will be presented at the ISCA INTERSPEECH-2020 Special Session.

2. Community Outreach & Feedback

The NASA Apollo Mission Control recordings are rich source of time-critical team based communications. Complex communication characteristics in this corpus can be explored through multiple avenues, and require vast resource utilization [16, 17].

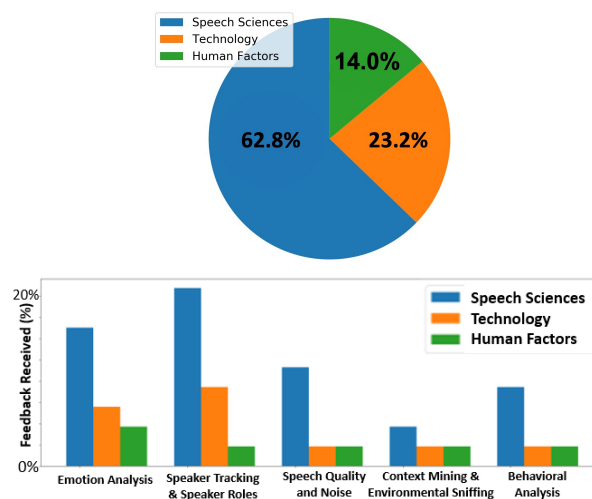


Figure 1: Analysis of community feedback. (top): Participant Breakdown (bottom): Most requested areas of interest

To ensure optimum long-term benefits of exploring this corpus, feedback from researchers in multiple intersecting disciplines is crucial. An essential component of corpora development following the completion of FS-1 was a focus on community

outreach and feedback. Multiple community engagement sessions were conducted with an aim in gathering essential future directions for the evolving FS Apollo corpus. The three communities directly benefiting from this corpus research and development include: (i) Speech Processing Technology (SpchTech), (ii) Communication Science and History (CommSciHist), and (iii) Education/STEM, Preservation/Archives, and Community-use (EducArch), who were consulted through workshop engagements. Community engagements are illustrated in Figure 1.

2.1. Fearless Steps Workshops

User feedback was primarily collected through 6 workshops at STEM and archival events (including IS-19, JSALT-19, ASA-19, ASRU-19). Online surveys of researchers downloading the FS corpus enabled access to feedback globally. A sample of the feedback from the above mentioned communities is illustrated in Figure 1. The salient responses across all communities focused on availability of more labeled data for system development, linking unstructured audio data with relevant meta-data through robust semi-supervised SLT systems, convenient data access, and retrieval through pipeline diarization transcripts.

2.2. Inaugural Fearless Steps (FS-1) Challenge

Over 170,000 hours of synchronized audio data were collected by NASA during the Apollo missions. Digitizing this audio with synchronized SLT pipeline processing would enable streamlined information access and retrieval to all communities. Due to resource limitations on developing manual annotations, speech and language systems capable of extracting meaningful information using limited ground-truth resources are necessary. FS-1 was designed with this premise, providing 20 hours of development set ground-truth, and 20 hours of evaluation set for five tasks: Speech Activity Detection (SAD), Speaker Diarization (SD), Speaker Identification (SID), Speech Recognition (ASR), and Sentiment Detection. These 40 hours of data was selected from channels with comparatively lower levels of degradation. A lexicon and language model based on 4.2 billion NASA mission text content was also freely provided [18, 19]. Semi-supervised and unsupervised systems optimized for the Apollo data were used as baseline systems [20, 21, 22, 23, 24]. These systems have been used as benchmarks for evaluating the variability introduced in FS-2 by an additional 60 hours of audio from highly degraded channels.

Table 1: Comparison of baseline results for FS-1 and FS-2 evaluation sets. Evaluation Metrics for FS-1 and FS-2: SAD: DCF (%), SID: Top-5-Acc (%), SD: DER (%), ASR: WER (%), with Relative degradation in performance for same systems (%)

Fearless Steps System(s) Performance on Eval Set			
Task	FS-1 (%)	FS-2 (%)	Rel. Degradation (%)
SAD	11.70	13.60	16.20
SD	68.23	88.27	29.37
SID	47.00	41.70	11.27
ASR	88.42	84.05	- 4.90

With a goal to maintain competitiveness in FS-2, higher content of degraded audio was selected to form the Eval set in FS-2 to offset the advantage of Train set ground-truth availability. This is detailed in Section 4.1. Table-1 provides a comparison of baseline system performance for all tasks over the Eval sets of FS-1 and FS-2. Significant degradation in system performance in three out of four tasks is observed. The evaluation metrics used for tasks SAD, SD, SID, and ASR were detection cost function (DCF), diarization error rate (DER), top-5 accuracy (%), and

word error rate (WER) respectively [7, 25, 26].

Sentiment Detection task from FS-1 provided participants with rudimentary labels of ‘positive’, ‘neutral’, and ‘negative’. However, all communities expressed interest in descriptive labels for emotion and behavioral analysis, as seen in Figure 1. Hence, sentiment detection task was removed from FS-2, and will be reintroduced in FS-3 as emotion detection task with 100 hours of improved labels.

3. FS-2 Challenge Tasks

The consensus from the community on requirement of increased transcribed data, and incremental task-targeted labeling prompted focused efforts on providing more variety in core-speech tasks. Hence, for FS-2, two separate challenge tracks were introduced for diarization and speech recognition. The speaker diarization track SD_track2 focuses on developing robust speaker embedding and clustering algorithms, while SD_track1 caters to the more challenging task of diarization from scratch. Equivalently, the Speech Recognition track ASR_track2 focuses on transcribing diarized speech segments (each segment contains noisy speech from a single speaker), while ASR_track1 incorporates the broader scope of transcribing noisy overlapped multi-speaker continuous streams. All challenge tasks for FS-2 are given in the following list:

- 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)

The evaluation metrics for all tasks are consistent with the previous challenge, and described in Section 2.2 [26, 27, 28]. A scoring toolkit¹ was made publicly available for this challenge.

4. Corpus Re-Deployment (FS-2)

The five selected channels Flight Director (FD), Mission Operations Control Room (MOCR), Guidance Navigation and Control (GNC), Network Controller (NTWK), and Electrical Environmental and Consumables Manager (EECOM) from FS-1 were preserved with improved labeling for FS-2. The high degree of variability in speech and noise characteristics across these five channels has been explored previously [1, 2, 19, 29]. In FS-2, we introduce 60 hours of additional speech transcriptions and speaker labels from these channels to the existing 40 hours to provide sufficient data for supervised system training.

4.1. Data Set Selection

The Dev, and Eval sets provided through FS-1 were developed using 70% audio streams selected from clean channels, and 30% selected from degraded channels. The Train, Dev, and Eval sets for FS-2 were categorized with scope to introduce multi-channel tasks in future challenges, while maintaining progressive difficulty in verification sets. The intention behind this data set

¹https://github.com/aditya-joglekar/FS02_Scoring_Toolkit

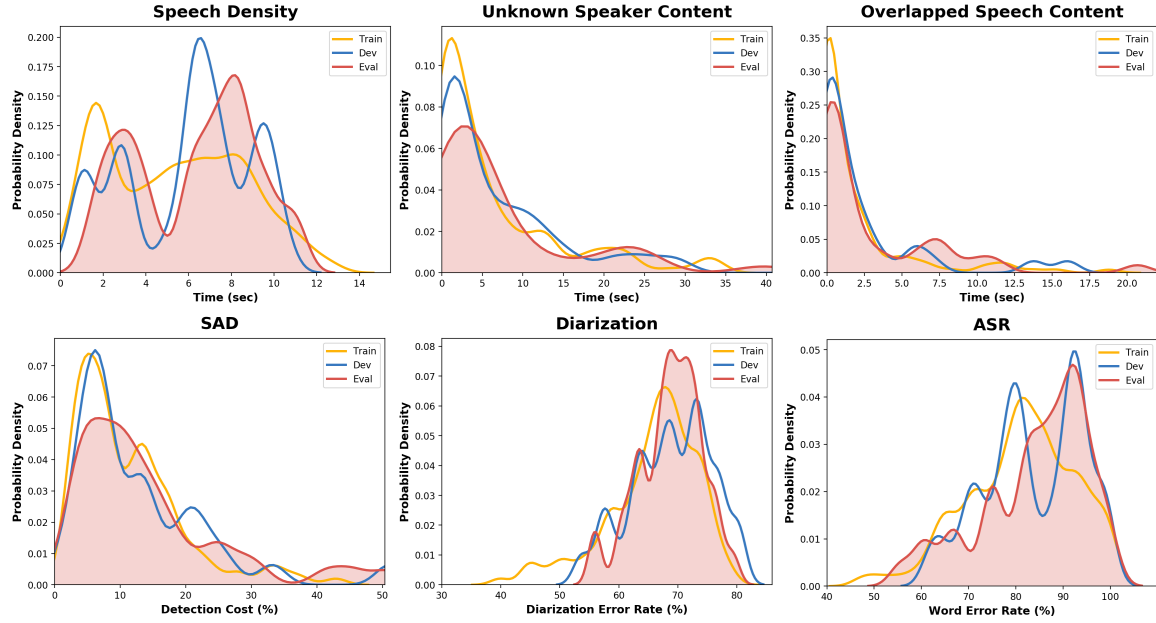


Figure 2: Probability Distributions of decision parameters for Train, Dev, and Eval sets.

Table 2: 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

design was to replicate naturalistic system development processes [5, 6, 30]. The FS-2 Challenge Corpus audio is divided into (i) audio streams, and (ii) audio segments. Audio streams reflect unaltered digitized audio from the Apollo missions. Audio segments are short duration speech sections diarized from the audio streams. Each segment contains a continuous speech utterance from a single speaker. Section 4.2 describes the process of splitting 100 hours into Train, Dev, and Eval sets. Section 4.3 provides more insight into development of segment based tasks SID and ASR_track2.

4.2. Audio Streams

Performance of SLT systems is dependent on factors like overlap content present in the data, amount of unintelligible speech, speech density variation, amount of data with unknown speakers, etc. In addition to this, the unsupervised baseline systems are useful in providing a measure of degradation in a given audio stream. We use the term 'decision parameters' to cumulatively describe the above measures. Using this methodology, it is possible to provide sets with progressive levels of difficulty across multi-channel audio streams in spite of inter-channel variations found in the Apollo data [1]. We perform this process by calculating all decision parameters for 100 hours of audio streams individually. These parameters are then normalized to generate degradation scores across the 100 hours. These scores are time-aligned across 5 channels and averaged to provide a single degradation score per 30-minute time chunk. These scores are finally categorized into three sets by progressive order of degradation. 5 channel segments with a cumulative highest degradation across all deci-

sion parameters are thus included in the Eval set, followed by Dev set. The streams with the least performance degradation are selected into the Train set. Trends observed from Figure 2 explain that even when the overall degradation across multiple channels is large, due to the variances in channel characteristics, the distributions for Train, Dev, and Eval sets have similar means, but differing distributions. Such varying distributions across decision parameters can aid in assessing the robustness of systems and their ability to generalize to data with a high degree of cross-channel variability.

Table 3: Duration Statistics of audio segments for ASR_track2. The mean, min, and max values are expressed in seconds.

Data set	Segments	Utterance Duration (s)		
		mean	min	max
Train	35,474	2.85	0.10	70.37
Dev	9,203	2.97	0.12	67.39
Eval	13,714	2.78	0.10	53.04

4.3. Audio Segments

SID task in FS-1 challenge provided 183 speakers a minimum of 10 seconds of training data. FS-2 SID task extends this set by adding over 30,000 additional utterances for 218 speakers. With shorter utterance durations and larger variations in speaker durations as seen in Table-2, FS-2 provides a more challenging task over FS-1. This data also encapsulates the challenges faced in speaker tagging for Apollo corpora. While a few personnel had major speaking roles, most backroom staff in the mission control

audio recordings had limited but integral speaking roles, making unbalanced and low resource speaker identification essential for a real-world scenario. Table-3 illustrates the general duration statistics of audio segments provided for the ASR_track2 task. While this task has the advantage of having fully diarized segments, the single word utterance durations shorter than 0.2 secs pose a challenge to ASR systems.

5. Baseline Systems

The SAD, ASR, and speaker diarization baseline systems from the first challenge were retrained and optimized for usage in this challenge [2]. Both tracks for SD and ASR tasks were evaluated using the same system, with differing configurations. Baseline results for all tasks are provided in Table-4.

Table 4: *Baseline Results for Development and Evaluation Sets*

Fearless Steps Phase-02 Baseline Results			
Task	Metric	Dev (%)	Eval (%)
SAD	DCF	12.50	13.60
SD_track1	DER	79.72	88.27
SD_track2	DER	68.68	67.91
SID	Top-5 Acc.	75.20	72.46
ASR_track1	WER	83.80	84.05
ASR_track2	WER	80.50	82.23

5.1. Speaker Identification

The SID baseline system developed for FS-1 used i-Vectors for front-end processing [23]. This system was more suited to the FS-1 SID data since it had at least 10 seconds of speech content per speaker. Due to the challenging nature of the current FS-2 SID data (≤ 4 utterances per speaker on average), this system was rendered inadequate. Moreover, for speakers in the Apollo data, x-Vector and i-Vector embeddings have low separability, forming separate clusters for same speaker utterances from different channels. This is illustrated with a t-SNE plot of i-Vector and x-Vector embeddings for 140 speakers in Figure 3 [31, 32, 33].

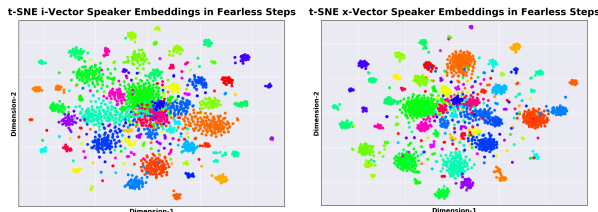


Figure 3: *Reduced dimensional i-Vector embedding (left), and x-Vector embedding (right) t-SNE plots for 140 speakers [33]*

To provide an alternate baseline system more suited to the revised SID data, SincNet system was used [34]. Input data was normalized and preprocessed to provide speech frames using the rVAD system (which ranked 4th in the FS-1 SAD task) [35]. rVAD system threshold was optimized to provide strict speech boundaries. The SincNet was trained for 360 epochs. This system (shown in Figure 4) provided a Top-5 Accuracy of 72.46%, which was a 30% absolute improvement over the FS-1 SID baseline system.

6. Discussion

FS-2 Challenge concluded with 111 system submissions across all tasks. While this was similar to the 116 system submissions received for FS-1 challenge, participation for both tracks of SD and ASR tasks was noticeably higher. The systems developed for

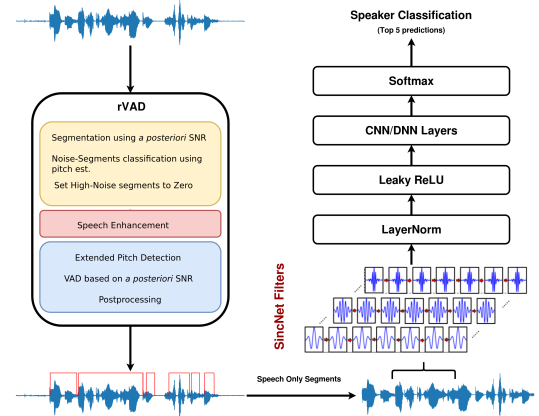


Figure 4: *rVad-SincNet based SID baseline system [34, 35]*

FS-2 also exhibited vast improvements in performance compared to the best systems developed for FS-1 challenge [2, 11, 12, 13, 15], as seen in Table-5. We observed relative improvements of 67%, 57%, and 62% for SAD, Speaker Diarization from scratch, and Speech Recognition from audio streams tasks respectively. These top ranked systems from the community will be used to develop baselines for the next phase of the challenge, FS-3.

Table 5: *Comparison of the best systems developed for all FS-1 and FS-2 challenge tasks. Relative improvement of top-ranked system per task in FS-2 over FS-1 is illustrated.*

Comparison of Best System Submissions			
Task	FS-1 (%)	FS-2 (%)	Rel. Imp. (%)
SAD	3.31	1.07	67.67 %
SID	89.94	92.39	2.72 %
SD_track1	68.23	28.85	57.71 %
SD_track2	N/A	26.55	N/A
ASR_track1	63.97	24.01	62.46 %
ASR_track2	N/A	24.26	N/A

7. Conclusions

The FEARLESS STEPS Challenge Phases are aimed at developing robust speech and language systems for multi-party naturalistic audio. FS-2 enabled the development of new state-of-the-art supervised systems for core-speech tasks on Apollo data through its Challenge Corpus. Train, Dev, and Eval sets compatible for multi-channel challenges were also developed. Final Phase (FS-3) of the Fearless Steps initiative will include single and multi-channel core-speech tasks on the available 100 hours, and 20 hours of yet unrevealed Apollo-13 multi-channel audio ("Houston, we've had a problem"!)). System advancements through FS-2 have also accelerated the development of conversational analysis and natural language understanding tasks for FS-3 like hot-spot detection, topic summarization, and emotion detection.

8. 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 FS-2 Challenge Corpus.

9. References

- [1] 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*, 2018, pp. 2758–2762. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1942>
- [2] 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*, 2019, pp. 1851–1855. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2019-2301>
- [3] 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.
- [4] M. Harper, "The Automatic Speech Recognition in Reverberant Environments (ASpIRE) challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 547–554.
- [5] 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*, 2018, pp. 1561–1565. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1768>
- [6] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas *et al.*, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.
- [7] F. R. Byers, J. G. Fiscus, S. O. Sadjadi, G. A. Sanders, and M. A. Przybocki, "Open Speech Analytic Technologies Pilot Evaluation OpenSAT Pilot," NIST, Tech. Rep., 2019.
- [8] G. E. Hinton, T. J. Sejnowski, T. A. Poggio *et al.*, *Unsupervised Learning: Foundations of Neural Computation*. MIT press, 1999.
- [9] 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 *INTER-SPEECH*, 2013, pp. 1135–1139.
- [10] "National Archives," www.archives.gov, accessed: 2018-10-24.
- [11] B. Sharma, R. K. Das, and H. Li, "Multi-level Adaptive Speech Activity Detector for Speech in Naturalistic Environments," *Proc. Interspeech 2019*, pp. 2015–2019, 2019.
- [12] 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*. International Speech Communication Association, 2019.
- [13] G. Deshpande, V. S. Viraraghavan, and R. Gavvas, "A Successive Difference Feature for Detecting Emotional Valence from Speech," in *Proc. SMM19, Workshop on Speech, Music and Mind 2019*, 2019, pp. 36–40.
- [14] P. Fallgren, Z. Malisz, and J. Edlund, "How to annotate 100 hours in 45 minutes," *Proc. Interspeech 2019*, pp. 341–345, 2019.
- [15] V. Manohar *et al.*, "Semi-Supervised Training for Automatic Speech Recognition," Ph.D. dissertation, Johns Hopkins University, 2019.
- [16] J. H. Hansen, A. Joglekar, A. Sangwan, and C. Yu, "Fearless Steps: Taking the next step towards advanced speech technology for naturalistic audio," *The Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2956–2956, 2019.
- [17] 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.
- [18] A. Stolcke, "SRILM - an Extensible Language Modeling Toolkit," in *Seventh international conference on spoken language processing*, 2002.
- [19] L. N. Kaushik, "Conversational Speech Understanding in Highly Naturalistic Audio Streams," Ph.D. dissertation, University of Texas at Dallas, 2018.
- [20] 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.
- [21] V. Kothapally and J. H. Hansen, "Speech Detection and Enhancement Using Single Microphone for Distant Speech Applications in Reverberant Environments," in *INTER-SPEECH*, 2017, pp. 1948–1952.
- [22] 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*, 2018, pp. 3603–3607. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-50>
- [23] F. Bahmaninezhad and J. H. L. Hansen, "i-Vector/PLDA speaker Recognition using Support Vectors with Discriminant Analysis," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5410–5414.
- [24] W. Xia, J. Huang, and J. H. Hansen, "Cross-lingual Text-independent Speaker Verification Using Unsupervised Adversarial Discriminative Domain Adaptation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5816–5820.
- [25] "NIST Rich Transcription Spring 2003 Evaluation," <https://catalog.ldc.upenn.edu/LDC2007S10>, accessed: 2019-03-01.
- [26] 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 *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.
- [27] 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.
- [28] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "First DIHARD Challenge Evaluation Plan," 2018.
- [29] A. Ziaei, L. Kaushik, A. Sangwan, J. H. Hansen, and D. W. Oard, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [30] W. Xia and J. H. Hansen, "Speaker Recognition with Nonlinear Distortion: Clipping Analysis and Impact," *Proc. Interspeech 2018*, pp. 746–750, 2018.
- [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] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.
- [33] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
- [34] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.
- [35] 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.