INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

The 2019 Inaugural Fearless Steps Challenge: A Giant Leap for Naturalistic Audio

John H.L. Hansen, Aditya Joglekar, Meena Chandra Shekhar, Vinay Kothapally, Chengzhu Yu, Lakshmish Kaushik, Abhijeet Sangwan

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

{john.hansen, aditya.joglekar, meena.chandrashekhar, vinay.kothapally, chengzhu.yu, lakshmish.kaushik, abhijeet.sangwan}@utdallas.edu

Abstract

The 2019 FEARLESS STEPS (FS-1) Challenge is an initial step to motivate a streamlined and collaborative effort from the speech and language community towards addressing massive naturalistic audio, the first of its kind. The Fearless Steps Corpus is a collection of 19,000 hours of multi-channel recordings of spontaneous speech from over 450 speakers under multiple noise conditions. A majority of the Apollo Missions original analog data is unlabeled and has thus far motivated the development of both unsupervised and semi-supervised strategies. This edition of the challenge encourages the development of core speech and language technology systems for data with limited ground-truth/low resource availability and is intended to serve as the "First Step" towards extracting high-level information from such massive unlabeled corpora. In conjunction with the Challenge, 11,000 hours of synchronized 30-channel Apollo-11 audio data has also been released to the public by CRSS-UTDallas. We describe in this paper the Fearless Steps Corpus, Challenge Tasks, their associated baseline systems, and results. In conclusion, we also provide insights gained by the CRSS-UTDallas team during the inaugural Fearless Steps Challenge.

Index Terms: NASA Apollo 11 mission, corpus, speech activity detection, speaker diarization, speaker identification, speech recognition, sentiment detection, pipeline diarization transcripts.

1. Introduction

The Fearless Steps (FS) initiative started in 2013 with the digitization of the Apollo-11 Mission analog tapes based on an NSF CISE project. The last six years have seen the development of the Corpus consisting of over 19,000 hours of audio data from the Apollo 1, 11, 13 and the Gemini 8 missions. The unique nature of this data posed a serious challenge for analysis using conventional speech technologies [1]. This challenge motivated the development of multiple solutions from CRSS catered to the nature and complexity of the Apollo data [2, 3, 4, 5, 6, 7]. The development of algorithms in speech activity detection, speaker diarization, speaker identification, speech recognition and sentiment detection domains were complemented with the development of a 100 hour human annotated subset of the corpus. While complete diarized speaker and ASR text content was provided for all 19,000 hours, the 100 hour human annotated data serves as an effective ground truth for this challenge. This data, referred to as the Fearless Steps (FS) Challenge Corpus is comprised of three mission critical stages from the Apollo-11 mission, viz., **Lift Off**, **Lunar Landing**, and **Lunar Walking**. This data is now being used to hold the inaugural challenge [8]. This paper serves as an overview of the corpora, Challenge Tasks, their associated baseline systems and results.

2. Fearless Steps Corpus

Traditionally, most speech and language technology (SLT) domain applications concentrate on analysis of a single audio stream or channel with one or more speakers involved. The audio in FS Corpus represents 30 individual synchronized analog communications channels with multiple speakers in different locations working real-time to accomplish NASA's Apollo missions [9]. For Apollo-11, this means each channel reflects a single communications loop (channel) that can contain anywhere from 3-65 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 to the mentioned conversational aspects, most of the audio channels suffer from a wide range of issues like high channel noise, system noise, attenuated signal bandwidth, transmission noise, cosmic noise, analog tape static noise, noise from tape aging, etc., with noise levels varying within each channel across time. Several instances of speech are also degraded by the presence of crosstalk and channel feedback, causing echo effects and several cases of overlap due to background speakers [2, 3, 10]. The above mentioned characteristics are some of the many peculiarities that highlight the challenges presented through this corpus. An efficient methodology for improvement of SLT systems on this data relies on data selection, accurate transcription, domain adaptive processing, and benchmarking against state-of-the-art systems. The Challenge development phase elaborates on the first aspect of this methodology.

2.1. Challenge Development

The strategic selection of data from the lift off, lunar landing, and lunar walking mission stages helped refine data selection options from the available 11,000 hours. The data was further focused to 5 of the possible 30 channels from those mission stages. Following preliminary analysis on all the channels, the loops with rich information and speech parameter variability (speech density and noise levels) selected were Flight Director (**FD**), Mission Operations Control Room (**MOCR**), Guidance Navigation and Control (**GNC**), Network Controller (**NTWK**), and Electrical Environmental and Consumables Manager (**EECOM**). The 100 hour Challenge Dataset is comprised of roughly 20 hours of each of these channels. The five core tasks which would attempt to solve the most challenging aspects of the data were selected for the Challenge.

1. Speech Activity Detection (**SAD**)
2. Speaker Diarization (**SD**)

http://dx.doi.org/10.21437/Interspeech.2019-2301

3. Speaker Identification (**SID**)
4. Automatic Speech Recognition (**ASR**)
5. Sentiment Detection (**SENTIMENT**)

Baseline System outputs for each task were used as a preliminary step to generate manual annotations. These labels were improved by annotators and accepted as final labels after two stages of inter-rater reliability. This was an extremely time intensive task due to the intricate complexities of the data viz. lack of consistent information on backroom staff speakers, and identifying technical terminologies and abbreviations. Moreover, spontaneous speech with rapid conversational turns and multiple overlap scenarios increased the difficulty of transcription significantly. Domain knowledge gathered from official NASA sources and flight journals and related documents were used to identify these terminologies and transcribe the audio. For the severely degraded channels and speech segments, multichannel information was leveraged to identify recurring speakers. Several speakers and parts of speech were still marked as 'UNK' and '[unk]' for segments which could not be identified with any recognizable speaker or intelligible speech.

2.2. Challenge Dataset

To ensure an equitable distribution of data into training, evaluation, 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 for some channels, and based on importance of the mission, the speech activity density of the corpus varies throughout the mission [1]. A total of 80 hours of audio are provided for task system development. For these 80 hours, a sub-set of 20 hours of human verified ground truth labels and transcripts are provided. An additional set of 20 hours is provided for open test evaluation. In addition, baseline ASR and sentiment detection system outputs are provided for the Training set.

Table 1: *Dataset Release Format for all five Tasks*

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

The evaluation set files are selected to have a higher degree of complexity than the development set, but with a similar distribution of speech. Audio sets for the speaker identification task were processed separately, generating one file per dominant speaker. Segment duration per speaker utterance is restricted to a minimum of 5 seconds. The Table 1 shows the application of given data sets to tasks.

2.2.1. Development Set

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

2.2.2. Training Set

Approximately 60 hours of audio data in addition to their associated baseline system generated sentiment labels and transcripts are provided. This set has no associated ground truth, and has been made available for researchers to use this unlabeled data to leverage their systems.

2.2.3. Evaluation Set

Only the audio files are provided for evaluation set. The Evaluation set consists of roughly similar amounts of clean and noisy channel audio, comprising of 20 hours in total. The helpful statistics about the Evaluation set are given in Table 2.

2.3. General Statistics

Due to the communication characteristics observed for the audio data, there is a presence of background conversation speech in good portion of the audio in addition to the other noise sources mentioned. The Table 2. provides a general analysis of the 100 hours, aiming to shed some light on the properties of the data. The average number of speakers and the speaker duration mean and standard deviation per 30 minute segments of every channel are provided. In addition, the mean and variation of SNR within each channel segment are also displayed [11]. The channel information for the entire data will be released after the Challenge concludes.

Table 2: *Channel/Mission Specialist information: Signal to Noise Ratio Statistics (dB SNR) per channel per 30-min segment for Dev and Eval Sets, average speaker talk duration per channel*

Mission	**SNR**		**Average**	**Spkr Duration**	
Specialist	**mean**	**std**	**# Spkrs**	**mean**	**std**
EECOM	13.3	7.4	16	23.04	6.72
FD	14.7	10.5	11	28.74	6.08
GNC	14.9	11.9	21	25.18	5.58
MOCR	5.1	12.6	13	22.36	5.65
NTWK	10.7	11.2	24	17.12	4.97

3. Challenge Tasks

All the Challenge Tasks were modelled after previously held challenges like the NIST OpenSAT, NIST SRE, DIHARD, and Chime-5 Challenge [12, 13, 14, 15]. Accepted standard metrics for each task are used for evaluation of systems in this study.

3.1. Speech Activity Detection (SAD)

The noise levels as low as -10 dB have been observed for regions of degraded channels. As seen in Section.2, multiple noise types present in the data degrade the quality of speech. Moreover, traditional speech systems fail to converge for drastically varying speech densities seen across the data [5, 16, 17]. These factors are essential considerations in the design of speech systems for this Challenge. The Detection Cost Function (DCF) is a NIST defined function used as the evaluation metric for this task. The goal for system developers will be to determine and select their system detection threshold, θ, that minimizes the overall DCF value. DCF(θ) is the detection cost function value for a system at a given system decision-threshold setting

$$DCF(\theta) = 0.75 \times P_{FN}(\theta) + 0.25 \times P_{FP}(\theta) \quad (1)$$

where, P_{FP} = false alarm, and P_{FN} = missed detection of speech; and P_{FN} and P_{FP} are weighted by 0.75 and 0.25, respectively, θ denotes a given system decision-threshold setting.

3.2. 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. This challenge is focused on Diarization from scratch. The evaluation metric for this task, diarization error rate (DER), introduced for the NIST Rich Transcription Spring 2003 Evaluation 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 speakers (RS) and system speakers (SS) [18, 19]. More concretely, DER is defined as:

$$DER(\%) = \left(\frac{FA + Miss + Error}{Total}\right) \times 100 \qquad (2)$$

where, *Total* = total duration of RS, *FA* = total SS time not attributed to RS, *Miss* = total RS time not attributed to SS, and *Error* = total RS time attributed to wrong speaker segments.

3.3. Speaker Identification (SID)

In addition to the issues faced by diarization systems, Speaker Identification system performance also relies on speech content per segment. The main focus of this challenge task is to identify speakers with drastically varying speech. 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 183 speakers with at least 10 seconds of total speech content, that are distributed in the Development and Evaluation Sets, as shown in Table 3

Table 3: *General statistics for the SID task. The mean, median, and duration/utterance values are expressed in seconds.*

Data	# Spkr	Spkr. duration (s)		Spkr Utterances	
		Mean	Median	Dur/utt	Total
Dev	183	247.7	50.38	5.35	8394
Eval	183	105.3	21.42	4.69	3600

The primary focus of this challenge is be in-set identification of speakers with varying duration of speech and noise levels. A simple Top-5 accuracy metric to gauge system performance. The SID Task will be evaluated for Accuracy of the Top-5 system predictions for a given input file.

$$Accuracy = \frac{\sum_{i \in S} N_{sys}(i)}{\sum_{i=1}^{M} N_{ref}(i)}, \quad \text{for} \quad S = k \in [1, M] : N_{ref}(k) \subseteq N_{sys}(k) \qquad (3)$$

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

3.4. Automatic Speech Recognition (ASR)

Rapid conversational turns during tensed or exited moments are often observed in the FS data. One or two word conversational cues are used as standard protocol for channel calls. Such conversations recorded over noisy channels pose a challenge for ASR systems. The goal of the ASR challenge task is to produce a verbatim, case-insensitive transcript of all words spoken in the noisy spontaneous speech. The Kaldi WER scoring toolkit was used to evaluate system performance [20]. Sections of audio unidentifiable to manual annotators were ignored during scoring.

The overall System WER was computed as the average WER(%) value computed separately for each 30-min audio segment (given by the following equation):

$$WER(\%) = \left(\frac{N_{Del} + N_{Ins} + N_{subst}}{N_{Ref}}\right) \times 100 \qquad (4)$$

3.5. Sentiment Detection (SENTIMENT)

Detecting audio sentiment in natural and spontaneous speaker settings and various speaker interactive scenarios (i.e., 1-way, 2-way, public speech etc.) is challenging. The sentiment system is expected to generate 3 sentiment outcome polarities, namely (1) positive, (2) negative and (3) neutral. A simple Accuracy per 10 millisecond frame is used as the evaluation metric to measure system performance [21, 22].

4. Baseline Systems and Results

Established systems which are either state-of-the-art of have been developed for naturalistic audio have been used to benchmark optimum system performance for this challenge.

4.1. Speech Activity Detection (SAD)

The Combo-SAD system was developed for the spontaneous speech in a highly noisy environment for the RATS corpus [17].

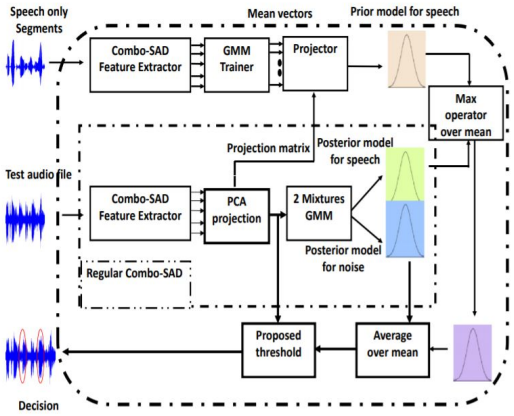

Figure 1: *Speech Activity Detection Baseline System Description*

An extension of this algorithm to include detection for audio segments with a high degree of speech density variation was developed for the Apollo data. This improved system, referred to as TO-ComboSAD, is used as the Baseline system for the challenge [5] and is shown in Figure 1.

4.2. Speaker Diariazation (SD)

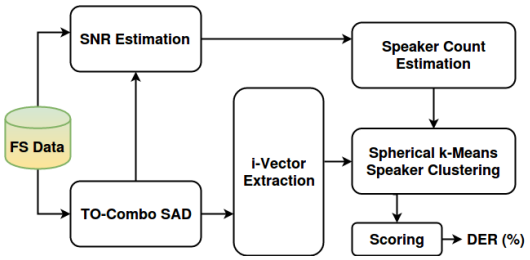

Figure 2: *Speaker Diarization Baseline System Description*

Since this task is designed for diarization from scratch, the best available SAD system used as the SAD baseline system is

used as the initial block of the diarization process. i-Vectors for the speech segments are extracted using a pre-trained i-vector extractor and UBM on SRE (04-08) data. This is followed by speaker clustering using spherical k-means method. Clustering is optimized using mixtures of von Mises-Fisher distributions, which have been found to be fruitful for naturalistic audio streams similar to the data observed in this corpus [23]. Figure 2 shows the diarization system block diagram.

4.3. Speaker Identification (SID)

The main baseline system was modeled using the state-of-the-art i-vector PLDA. We used Kaldi to obtain our speaker models [20], and trained the Universal Background Model (UBM) and TV-matrix using all SRE (04-08), and Switchboard (SWB II-p02,p03 and Cellular-p01,p02) data. Due to the low-resource availability, the PLDA model is trained with the SRE-04 to SRE-08 data, and subsequently adapted with i-vectors extracted from the unlabeled Train set [24, 25, 26]. Figure 3. describes the adapted-PLDA front-end and back-end system used as baseline for this task. The development set is then used as enrollment and the evaluation set is used as the test set. The top 5 likelihood scores for every test utterance are then evaluated to get the system performance.

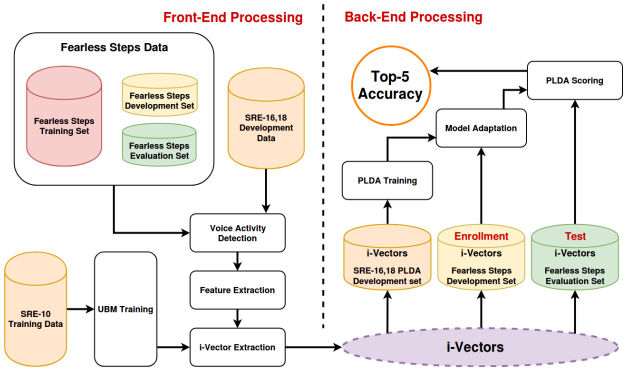

Figure 3: *Speaker Identification Baseline System Description*

4.4. Automatic Speech Recognition (ASR)

An Apollo mission-specific speech recognition system was developed using Kaldi ASR toolkit [20]. The acoustic models are trained using a mix-style approach, where acoustic data from multiple corpora are used. Individual language models are developed from these different sources and are fused to develop a comprehensive language model. NASA uses lot of scientific terms and abbreviations. All scientific terms were collated from the above said text sources which amounted to more than 4 billion words. A new pronunciation dictionary was generated for words and abbreviations that are not in the standard dictionary [4, 27]. This system was also used to generate the pipeline diarization transcripts for the entire 19,000 hours of the Corpus. The lexicon and language model used to generate these pipeline diarization transcripts have also been released to the public.

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

Task	Metric (%)	Dev	Eval
Speech Activity Detecion	DCF	8.6	11.7
Speaker Diarization	DER	65.2	68.2
Speaker Identification	Top-5 Acc.	58.1	47.0
Speech Recognition	WER	71.2	88.4
Sentiment Detection	Acc.	46.2	49.7

4.5. Sentiment Detection (SENTIMENT)

Due to the neutral nature of speech all Apollo communications loop personnel were trained to speak in, this task encouraged the use of both speech and text to develop multi-modal systems to extract sentiment information. Hence, the above defined ASR model serves as the front-end model for the Sentiment baseline system. This baseline is generated using keywords extracted from part-of-speech (POS) tagging, followed with iterative Maximum Entropy (ME) optimization [21, 22].

4.6. Baseline Results and System Submissions

The Baseline results for both development and evaluation sets are provided in Table 4. The results highlight the level of complexity for the data, for which systems are either state-of-the-art or developed specifically for this data. These results emphasize further the importance of collaborative efforts to develop domain-specific strategies. While deep learning strategies comprised of 85% of all competing systems, less than half achieved better results than the baseline, showing no significant correlation between performance and types of statistical models used. However, more than 90% of systems with domain-aware strategies outperformed the baseline systems, irrespective of the network complexity or depth of layers. The best systems for their respective challenge tasks achieved absolute improvements over the baseline results by 8.4% for SAD, 42% for SID, 25% for ASR, and 24% for SENTIMENT. The final rankings of system submissions for each task will be released on July 20, 2019 exactly at 9:56pm CST (50th Anniversary of the First Moon Walk!).

5. Discussion

The first edition of the Fearless Steps Challenge has been met with significant interest. At the time of writing, we received 150 registrations from 75 organizations. A total of 16 organizations participated in the Challenge, submitting 116 competing systems across all five tasks. Multiple systems were able to incorporate knowledge specific to the challenging scenarios presented through this corpus with state-of-the-art machine learning technologies to achieve superior performance over the baseline systems. The keen interest shown by the research community worldwide as well as the feedback from participants highlight the importance of having a publicly available naturalistic corpus of involving unscripted teams solving real-world problems of historical significance. We hope to promote further growth in the SLT domains through such community engagements.

6. Conclusions

Challenging datasets have laid the foundation towards achieving tremendous progress seen in the SLT domain over the past three decades. The Fearless Steps initiative strives to continue such efforts to benefit the community. Having seen several algorithms developed for this challenge, this is just the first step towards exploring such corpora. The next steps for this Challenge will address supervised and multi-channel approaches though the second and third editions of the Fearless Steps Challenge.

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.

8. 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] 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 *INTERSPEECH*, 2013, pp. 1135–1139.

[3] C. Yu, J. H. Hansen, and D. W. Oard, "Houston, we have a solution': A case study of the analysis of astronaut speech during nasa apollo 11 for long-term speaker modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[4] L. N. Kaushik, "Conversational speech understanding in highly naturalistic audio streams," Ph.D. dissertation, University of Texas at Dallas, 2018.

[5] 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.

[6] C. Yu and J. H. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.

[7] "Crss-utdallas explore apollo app," https://app.exploreapollo.org/, accessed: 2019-03-29.

[8] "Fearless steps challenge 2019 website," http://fearlesssteps.exploreapollo.org/, accessed: 2019-03-29.

[9] "National archives," www.archives.gov, accessed: 2018-10-24.

[10] C. Yu and J. H. Hansen, "A study of voice production characteristics of astronuat speech during apollo 11 for speaker modeling in space," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 1605–1614, 2017.

[11] T. A. Allen, "Nist speech signal to noise ratio measurements," 2016.

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

[13] 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

[14] "Nist opensat 2017," https://www.nist.gov/itl/iad/mig/opensat, accessed: 2019-03-01.

[15] 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.

[16] V. Kothapally and J. H. Hansen, "Speech detection and enhancement using single microphone for distant speech applications in reverberant environments." in *INTERSPEECH*, 2017, pp. 1948–1952.

[17] 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.

[18] "Diarization scoring toolkit," https://github.com/nryant/dscore, accessed: 2019-03-29.

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

[20] 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.

[21] L. Kaushik, A. Sangwan, and J. H. Hansen, "Automatic audio sentiment extraction using keyword spotting," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[22] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Sentiment extraction from natural audio streams," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 8485–8489.

[23] 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

[24] 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.

[25] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[26] 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.

[27] A. Stolcke, "Srilm-an extensible language modeling toolkit," in *Seventh international conference on spoken language processing*, 2002.