Fearless Steps APOLLO: Advanced Naturalistic Corpora Development

John H.L. Hansen, Aditya Joglekar, Szu-Jui Chen, Meena Chandra-Shekar, Chelzy Belitz

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, szujui.chen, meena.chandrashekar, chelzy.belitz}@utdallas.edu

Abstract

In this study, we present the Fearless Steps APOLLO Community Resource, a collection of audio and corresponding meta-data diarized from the NASA Apollo Missions. Massive naturalistic speech data which is time-synchronized, without any human subject privacy constraints is very rare and difficult to organize, collect, and deploy. The Apollo Missions Audio is the largest collection of multi-speaker multi-channel data, where over 600 personnel are communicating over multiple missions to achieve strategic space exploration goals. A total of 12 manned missions over a six-year period produced extensive 30-track 1-inch analog tapes containing over 150,000 hours of audio. This presents the wider research community a unique opportunity to extract multi-modal knowledge in speech science, team cohesion and group dynamics, and historical archive preservation. We aim to make this entire resource and supporting speech technology meta-data creation publicly available as a Community Resource for the development of speech and behavioral science. Here we present the development of this community resource, our outreach efforts, and technological developments resulting from this data. We finally discuss the planned future directions for this community resource.

Keywords: Apollo Missions, Fearless Steps, Pipeline Diarization, LanguageARC, Explore Apollo, Finding Waldo

1. Introduction

Naturalistic Speech corpora have enabled the development of state-of-the-art Deep Learning Models, which are known to benefit from scale and complexity in the data (Carletta, 2007), (Barker et al., 2018), (Harper, 2015), (Ryant et al., 2018). New deep learning research methodologies including graph neural networks, representation and self-supervised learning, have accelerated the need for massive speech resources, typically on the order of 1000's of hours (Hinton et al., 1999), (Bengio et al., 2013), (Scarselli et al., 2008). Most resources of such scale are either private, or are simulated data. CRSS-UTDallas over the past 7 years has made significant strides in developing a massively naturalistic resource which has made 19,000 hours publicly available, and aims to make over 150,000 hours of speech conversations and corresponding meta-data globally available. We refer to this CRSS-UTDallas driven project as the Fearless Steps (FS) APOLLO Community Resource. The core element of FS-APOLLO is to develop a corpora phase for each digitized Apollo Mission along with a sub-corpus for Speech and Language Technology (SLT) research. We refer to this collection as the FS-APOLLO corpora. Here, we illustrate several novel aspects of the corpora through general data statistics. We will detail the ExploreApollo.org and LanguageARC portals developed for Outreach and Education using this data. A subset of 125 hours of manually annotated audio released as a Challenge Corpus has proven to be an asset to SLT development, with multiple state-of-the-art developed by researchers globally for all core SLT tasks. We will briefly describe this Challenge series, and the pipeline diarization updates.

2. Fearless Steps APOLLO Resource

The Fearless Steps (FS) APOLLO Resource includes the development and deployment of the Apollo Missions audio, it's associated meta-data, and SLT systems to generate automatic labels for the massively unlabeled and expanding corpus collection. Our collaboration with the Linguistic Data Consortium (LDC) is aimed at enabling free distribution of the audio and meta-data for all 12 manned Apollo Missions. Since the initial FS-APOLLO public releases, more than 500 organizations have utilized (Hansen et al., 2018), (Hansen et al., 2019), the 19,000 hours of automatic labelled, and 125 hours of human annotated audio for research on tasks including but not limited to the FS Challenge. In this section, we will elaborate on the development of these corpora.

2.1. Data Collection & Deployment

Digitization process for FS-APOLLO started with Apollo-11. The Soundscriber device displayed in Fig. 1 was used with a CRSS-developed 30-track read-head digitizing solution to convert analog tapes into 44.1Khz lossless digitized audio. The Inter-Range Instrumentation Group (IRIG) timecodes encoded on channel 1 were used to save time-synchronized audio. This process initially yielded 11,000 hours of Apollo-11, 8,000 hours of Apollo-13, Apollo-1 and Gemini-8 recordings. After receiving approval from NASA export control, CRSS-UTDallas started distributing the data online, through workshops and SLT challenges (Joglekar et al., 2020).

2.1.1. Naming Convention

Fig. 1 illustrates the file naming convention used to efficiently deploy audio content across all Apollo Missions. The files have been named to create unique ID's for all channels and missions. The file ID's are able to map

Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People @LREC2022, pages 14–19
Marseille, 25 June 2022.

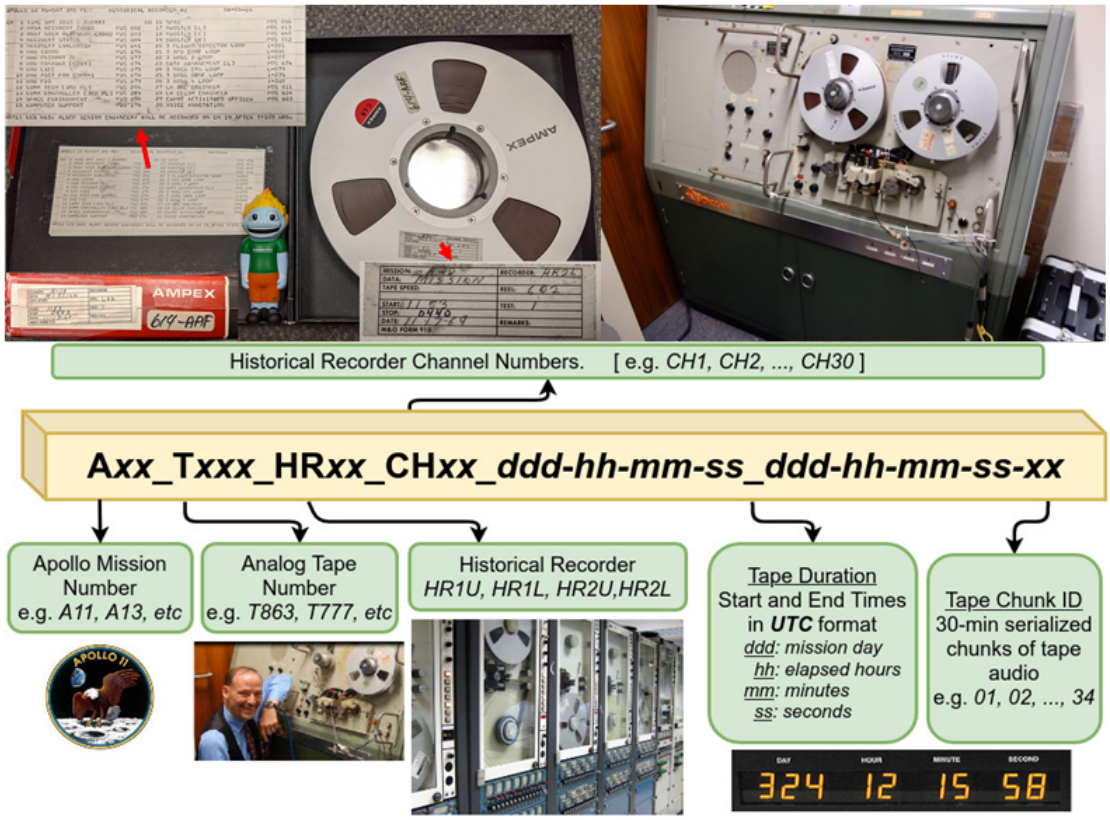

Figure 1: ***(Top)***: CRSS-UTDallas audio capture solution for Apollo analog tapes and file naming convention. Soundscriber playback system modified by CRSS to allow synchronous 30-channel digitization; ***(Bottom)***: File Naming convention; ***(from left to right)***: Apollo Mission, Analog Tape Number, Historical Recorder, Tape Duration IRIG time-code, and Tape chunks. Illustrations of Soundscriber recording system and IRIG time-code also shown.

uniquely to each recording through tape information and associated IRIG tape start and stop time-stamps. Fig. 1 also shows images of the Soundscriber used to record and digitize the Analog Apollo Tapes.

2.2. Fully Annotated Sub-corpus

With currently available technology, a massive naturalistic unlabeled corpus with distinct acoustic and language characteristics is of limited value. A small portion of this corpus audio sampled from mission critical stages can however significantly open the scope of engagement with the larger corpus. CRSS sampled 109hrs of Apollo-11, 10hrs of Apollo-13, & 6hrs of Apollo-8 to generate manual speech, speaker, transcripts, topic, & sentiment annotations. These annotations have been included in packages with audio data for release in four Phases of Challenge Tasks. These challenge tasks included Speech Activity Detection (SAD), Speaker Diarization (SD), Speaker Clustering, Speaker Identification (SID), Speaker Verification (SV), Automatic Speech Recognition (ASR), Sentiment Detection, Topic Identification (TID), & Topic Tracking (Joglekar et al., 2021). Analysis on the challenge corpus has provided key insights in speaker, speech, & noise characteristics.

3. Outreach

Active outreach efforts were performed by CRSS to receive feedback from the wider community on development of supporting meta-data and technologies.

3.1. Workshops

Initial efforts for FS-APOLLO Resource focused on gathering information from three distinct communities while simultaneously digitizing Apollo tapes. This was done to maximize the potential corpus impact for the wider public. Three distinct communities include: **(i)** Speech and Language Technology (SLT), **(ii)** Historical Archives and STEM Education, and **(iii)** Speech and Behavioral Sciences were all approached to provide their expertise on how the data could impact their fields. Salient responses were chosen to construct a research and annotation plan. The community feedback highlighted a need for CRSS to develop speech tools enabling automatic transcription of the entire Apollo-11 and Apollo-13 Corpora. Additional steps like assigning semantic tags to conversations of significance were also identified as essential to drive the desired impact across all communities (Joglekar et al., 2020).

3.2. Pipeline Diarization Baseline

To produce supplemental automatic meta-data, a small 10 hr subset of Apollo-11 was manually annotated for SAD, SD, and ASR tasks. We simultaneously used established corpora to train Deep Neural Network (DNN) based acoustic models (Cieri et al., 2004), and scraped all openly available technical documents pertaining to

NASA, training an N-gram language model based on 4.2-billion words. Using the human annotations to tune our system, pipeline diarization transcripts were created for the entire Apollo-11 and Apollo-13 corpora. These transcripts were used to roll out a second round of human annotations on a respectable-sized corpus of 125 hours, with data sourced from Apollo-11, Apollo-13, and Apollo-8 missions. Improved speech and speaker labels were used to further develop sentiment and conversational topic labels.

3.3. ExploreApollo.org

In an effort to motivate k-12 STEM education, CRSS-UTDallas developed an interactive website to share Apollo data and insights. The website[1] is maintained by CRSS, with UTDallas students contributing through Senior Design project collaborations. Senior design projects organized and managed by CRSS members and staff involve active enhancement of features to increase K-12 student engagement. Fig. 2 shows the improved landing page for the web app. This page provides users with the option to listen to fully transcribed and time-stamped Apollo Missions audio with a visualization panel showing utterance-wise transcripts, speaker information, and images of additional meta-data associated to that timeline. As an illustration, the audio segments with speech from Neil Armstrong taking the first steps on the moon are supplemented with transcripts, astronaut photos, and the news releases of the Apollo landing.

3.4. LanguageARC

LanguageARC was developed by the Linguistic Data Consortium at Univ. of Pennsylvania based upon work supported by NSF. This is a crowd-sourcing platform which helps users to contribute to resources that are then shared for research, education and technology development purposes. There exists paid options that can provide services similar to Amazon's Mechanical Turk, but LanguageArc is popular among the speech and language community with millions of users. Users can freely answer questions about specific data in the form of short tasks. Considering that Apollo data is a largely unlabeled audio dataset, this platform provides an opportunity to provide meta-data for not just Apollo-11 but also other Apollo missions. Currently, users can begin working on three different tasks for Apollo-8: Determine Audio Quality, Transcribe speech, and create speaker count info. per clip. Each audio clip consists of 10 sec. snippets across six specific channels listed: Flight Director (FD), Public Affairs Officer (PAO), Network Controller (NTWK), Mission Operations Control Room (MOCR), Electrical, Environmental, and Consumables Manager (EECOM), and Guidance, Navigation, and Control systems engineer (GNC). Meta-data for the listed tasks are being produced by helpful volunteers. Our goal is to add more missions in the near future and also include more tasks for the current mission.

3.5. Finding Waldo

The Apollo missions represent unique data since all communications were recorded using multiple synchronized channel recorders of real-world task-driven teams. Two 30-track audio historical recorders were employed to capture all team loops of the Mission Control Center (MCC). The MCC was organized hierarchically: one Flight Director (FD), one Capsule Communicator (CAPCOM), more than 15 chief MOCR personnel, and a corresponding set of backrooms with specialists that support multiple specialist teams were time sequenced over 6-12 day missions. The primary speakers operating these five channels are command/owners of these channels. Each mission specialist is designated a speaker role and since the mission spans multiple days, these roles are fulfilled by 3-4 mission specialists. Effective communication is required for teams to work collaboratively to learn, engage, and solve complex problems. To track and tag individual speakers across our Fearless Steps audio corpora, we use the concept of 'where's Waldo' to identify all instances of our speakers-of-interest (SOI) across a cluster of other speakers. We select five SOI: Astronauts Neil Armstrong, Buzz Aldrin, & Michael Collins, with Gene Kranz serving as FD, and Charlie Duke as CAPCOM. Fig. 3 shows each speaker's speech duration in a "Donut" plot. This plot summarizes conversational turn-taking for speakers over an extended time set, providing a global perspective of the speaker interaction between each SOI vs other speakers across audio clips. Identifying these personnel can help pay tribute and yield personal recognition to the hundreds of notable engineers and scientists who made this mission possible. This collection also opens new research options for recognizing team communication, group dynamics, and human engagement/psychology for future deep space missions (Shekar and Hansen, 2021).

3.6. The Soundscriber Playback System

For many years, a majority of the Apollo audio existed on analog tapes stored at the NASA NARA archive[2]. The setup used in recording mission audio was based on two recorders, known as Historical Recorder 1 (HR1) and 2 (HR2), each with an upper and lower tape deck. Both HR1 and HR2 ran continuously, switching between decks as each tape neared the end of its recording limit. The original audio was recorded on 29 of the 30 channels per 17 hour tape. The Soundscriber has been instrumental to the preservation and digitization of the Apollo mission audio. This unit was specifically manufactured for NASA by Soundscriber Corp. (Hansen et al., 2018). Novelty of the NASA Soundscriber system used to record MCC/MOCR communications proved to be a hurdle for digitizing historic mission audio. The only means to recover this audio was using a separate Soundscriber playback system, which allowed someone to listen to only one selected audio channel. Prior to data

[1] app.exploreapollo.org

[2] https://www.archives.gov/space

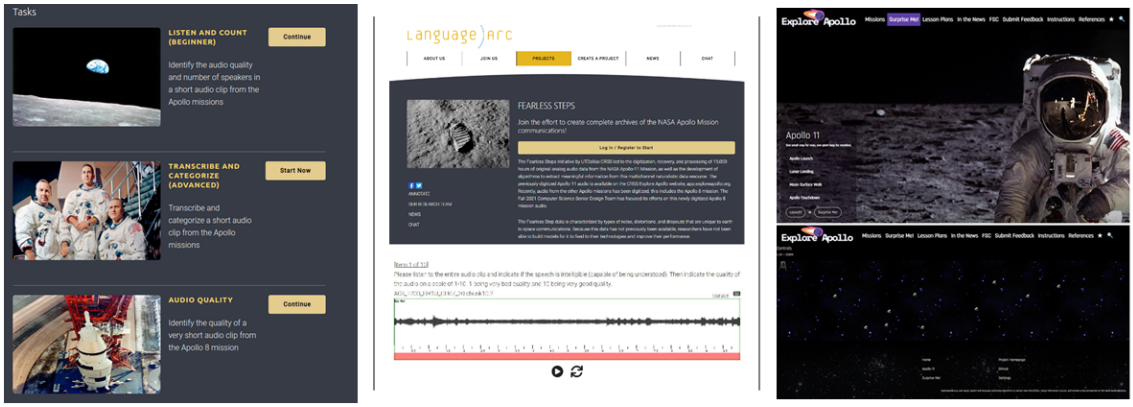

Figure 2: ***(left)***: Established tasks on LanguageARC for Apollo-8; ***(center)***: Fearless Steps Project on LanguageARC. ***(right)***: Explore Apollo Website. ***(right top)***: Landing page for the website provide options to browse to the audio playback section, games section, or the challenge tasks, ***(right bottom)***: An illustration of a single player web-app game on the website where the user has to move up or down to escape the incoming asteroids.

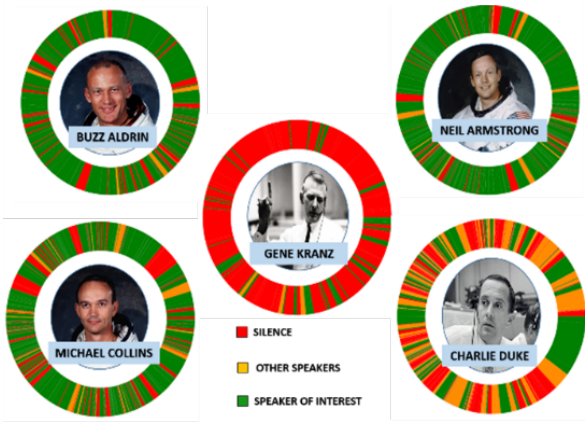

Figure 3: Speaker Duration for Speakers of Interest vs Other MCC Personnel

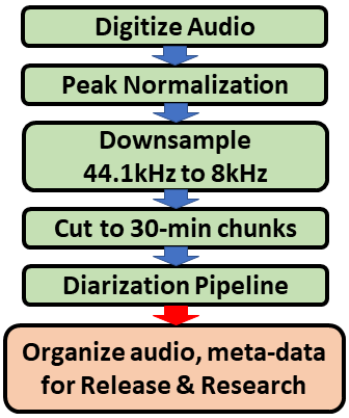

Figure 4: An overview of the steps followed to prepare for distribution of Apollo mission audio data.

recovery efforts, no multi-channel Soundscriber playback units existed to play these tapes. Based on this fact, CRSS estimates that less than 2% of all available audio has ever been heard/recovered since initial recording in the 1960's/70's. A long collaboration between CRSS-UTDallas and NASA engineers/technicians identified one playback system (a second had been dismantled but, eventually, used for parts in the restoration of the other). The original system was modified by CRSS-UTDallas, allowing simultaneous 30-track digitization (Sangwan et al., 2013). The Soundscriber playback system, along with its modifications, can be seen in Fig. 1. This development reduced digitization time by a factor of 30. Digitizing channels simultaneously allowed time synchronization while supporting tape preservation (greatly reducing the stress placed on aging tapes), providing a great resource for both researchers and historians alike.

3.7. The Data Preparation Pipeline

Prior to diarization, digitized audio needs to undergo preprocessing steps. These steps maximize the corpus utility for communities interested in SLT research, historical preservation and team-communication study. The specific steps, described in Fig. 4, were selected to prepare the raw digitized data for diarization process pipeline. Original 44.1 kHz data was preserved separately, and a copy of the data was used for preprocessing to account for future pipeline optimizations. The initial data triage pipeline included moving 17hr digitized channel audio to functional 30min audio chinks with proper filename conventions. New code was developed automatically identify and remove spikes caused due to tapr start and stop. From there, peak normalization was applied and the audio was downsampled to 8kHz (maintaining all relevant information), and cut into uniform 30-minute audio streams, synchronous across all 30 channels on a given tape. These streams are then named as described in Fig. 1, as well as transcribing information gathered from tape heat sheets.

4. FS Challenge Research Corpus

The the FS-APOLLO Corpora is a collection of digitized and largely unlabeled audio data. The fully labelled, multi-functional subset extracted from mission critical phases in the Corpora is referred to as the Fearless Steps Challenge (FSC) Corpus (Joglekar et al., 2020), (Joglekar et al., 2021) (Joglekar et al., 2022).

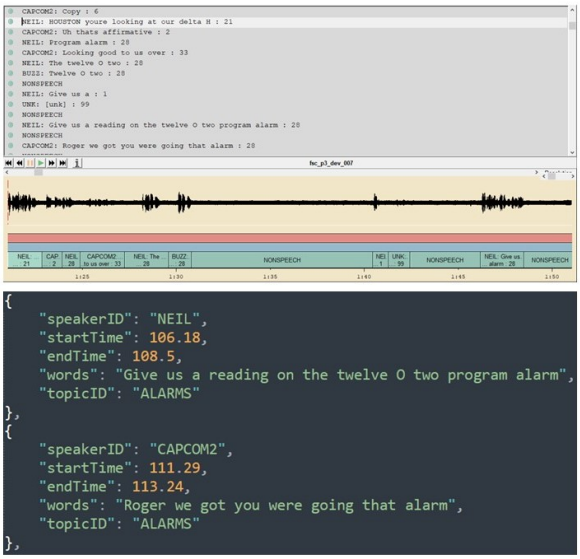

Figure 5: ***(Top)***: Illustration of multi-domain labels in transcriber tool `.trs` format. ***(Bottom)***: Annotations converted to `.json` format for FS challenge phases.

4.1. Fearless Steps Corpora Development

The years 2020 and first half of 2021 were marked by slowly moving digitization efforts due to COVID-19 restrictions. Even with these restrictions, CRSS-UTDallas was able to digitize an additional 50,000 hours of audio. This audio is recorded at 44.1Khz at NASA, JSC which houses the only existing system that can play the Apollo analog tapes. Entire Apollo 8, 9, and 10 were digitized, providing valuable information on MCC speakers. Since the core MCC team remained unchanged over the course of 10 years of the Apollo program, we have a collection of aging-based naturalistic speech corpus which will be developed soon. An illustration of the generated transcriptions are displayed in Fig. 5. The `.trs` files generated by the annotators using the LDC transcriber tool (Cieri and Liberman, 2006) were processed to generate the `.json` files. the Json format was provided to researchers trying to perform speech tasks on continuous audio streams.

4.2. Pipeline Diarization Advancements

The initial system developed in 2017 is a simple DNN with a N-gram language model, with word-error-rate (WER) around 80% on FSC Phase-2 development set. Recently, we further advanced a new baseline system using the advanced hybrid architecture in the Kaldi speech recognition toolkit (Povey et al., 2011). A scenario representation trained in self-supervised manor is incorporated with conventional MFCC and i-vector features to boost the performance on WER (Chen et al., 2021). The results are shown in Table 1.

5. Future Community Resource Direction

CRSS-UTDallas strives towards making continual progress to advance SLT and improve the three FS-APOLLO community resources. Our immediate goals include promoting self-supervised learning, and releasing over 50,000 hours of the already digitized data to be used for training general representations. We also aim to employ our speaker tracking system 'Finding Waldo' across missions to analyse changes in the speaker traits during the entire duration of the Apollo Program (around 10 years).

Table 1: The ASR system is trained on FSC Phase-2 corpus, evaluated on the FSC Phase-4 corpus

Updated Baselines for Fearless Steps Phase-4			
SLT Task	**Metric**	**Dev (%)**	**Eval (%)**
SAD	DCF	4.24	7.57
ASR_track1	WER	28.74	46.3
ASR_track2	WER	24.32	39.4
P2_ASR_track2	WER	26.16	28.9

6. Conclusion

This study has described the data development, label development, and outreach initiatives conducted so far for the Fearless Steps Apollo Community resource. Naturalistic data development is needed for both technology and scientific / society / historical impact. We aim to make this resource an integral part of the systems that will be developed to learn high-level knowledge directly from speech conversations.

7. Acknowledgements

This project was supported by NSF-CISE Community Resource Project 2016725, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen. A special thanks to Katelyn (CRSS-UTDallas Transcription Team) for leading the ground-truth development efforts on the FS Challenge Corpora. Further thanks to the numerous undergraduate senior design students who have contributed to supporting CRSS including Sesank as lead on the preprocessing pipeline and Sapnaben as lead in organizing information from images of digitized tapes.

8. Bibliographical References

Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *Proc. Interspeech 2018*, pages 1561–1565.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Chen, S.-J., Xia, W., and Hansen, J. H. (2021). Scenario aware speech recognition: Advancements for apollo fearless steps & chime-4 corpora. *arXiv preprint arXiv:2109.11086*.

Cieri, C. and Liberman, M. (2006). More data and tools for more languages and research areas: A progress report on ldc activities. In *LREC*, pages 779–782.

Hansen, J. H., Joglekar, A., Shekhar, M. C., Kothapally, V., Yu, C., Kaushik, L., and Sangwan, A. (2019). The 2019 Inaugural Fearless Steps Challenge: A Giant Leap for Naturalistic Audio. In *Proc. Interspeech 2019*, pages 1851–1855.

Harper, M. (2015). The automatic speech recogition in reverberant environments (aspire) challenge. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 547–554. IEEE.

Hinton, G. E., Sejnowski, T. J., Poggio, T. A., et al. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Joglekar, A., Hansen, J. H., Shekar, M. C., and Sangwan, A. (2020). FEARLESS STEPS Challenge (FS-2): Supervised Learning with Massive Naturalistic Apollo Data. In *Proc. Interspeech 2020*, pages 2617–2621.

Joglekar, A., Sadjadi, S. O., Chandra-Shekar, M., Cieri, C., and Hansen, J. H. (2021). Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for Unseen Channel and Mission Data Across NASA Apollo Audio. In *Proc. Interspeech 2021*, pages 986–990.

Joglekar, A., Chen, S.-J., Chandra-Shekar, M., Belitz, C., Yousefi, M., and Hansen, J. H. (2022). Apollo Fearless Steps: Datasets, Challenge Tasks, and SLT system developments for NASA Apollo Missions Audio. In *Manuscript Submitted to Proc. Interspeech 2022*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Sangwan, A., Kaushik, L., Yu, C., Hansen, J. H., and Oard, D. W. (2013). 'houston, we have a solution': using nasa apollo program to advance speech and language processing technology. In *INTERSPEECH*, pages 1135–1139.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2008). The graph neural network model. *IEEE trans. on neural networks*, 20(1):61–80.

Shekar, M. M. C. and Hansen, J. H. (2021). Historical audio search and preservation: "finding waldo" within the fearless steps apollo-11 naturalistic audio corpus. *IEEE Signal Processing Magazine *In Review*.

9. Language Resource References

Carletta, J. (2007). Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190.

Cieri, C., Graff, D., Kimball, O., Miller, D., and Walker, K. (2004). Fisher english training speech part 1 transcripts. *Philadelphia: Linguistic Data Consortium*.

Hansen, J. H., Sangwan, A., Joglekar, A., Bulut, A. E., Kaushik, L., and Yu, C. (2018). Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon. In *Proc. Interspeech 2018*, pages 2758–2762.

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2018). First dihard challenge evaluation plan.