Interspeech 2018
2-6 September 2018, Hyderabad

Fearless Steps: Apollo-11 Corpus Advancements for Speech Technologies from Earth to the Moon

John H. L. Hansen, Abhijeet Sangwan, Aditya Joglekar, Ahmet E. Bulut, Lakshmish Kaushik, Chengzhu Yu

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

{john.hansen, abhijeet.sangwan, aditya.joglekar, ahmet.bulut, lakshmish.kaushik, chengzhu.yu}@utdallas.edu

Abstract

The Apollo Program is one of the most significant benchmarks for technology and innovation in human history. The previously introduced UTD-CRSS Fearless Steps initiative resulted in the digitization of the original analog audio tapes recorded during the Apollo Space Missions. The entire speech data for the Apollo 11 Mission is now being made publicly available with the release of the Fearless Steps Corpus. This corpus consists of a cumulative 19,000 hours of conversational speech spanning over thirty time-synchronized channels. With over six hundred speakers, the corpus has a rich collection of information which can be beneficial for research and advancement in the speech and language community. Recent efforts on this data have led to the generation of pipeline diarization transcripts for the entire speech corpus. Research has also been done to address speech and natural language tasks such as speech activity detection, speech recognition, and sentiment analysis. This paper provides an overview of the Fearless Steps Corpus and highlights the factors that make the processing of this data a challenging problem. To promote further development of algorithms for naturalistic data, five challenge tasks are also organized. We also describe the challenge tasks with details on a fully transcribed subset of the corpus, and initial results achieved by our systems.

Index Terms: Apollo 11 mission, corpus, pipeline diarization transcripts, speech activity detection, speech recognition, speaker variability, conversational analysis, sentiment, topic

1. Introduction

The NASA Apollo Space Program was the first to successfully take astronauts to the moon and bring them back safely. Over the span of four years, the success of the lunar missions was made possible by the dedicated work of over 400,000 personnel supporting the program in various capacities [1, 17, 18, 26]. As a massive collaborative effort, with astronauts flying the missions in outer space, most people who participated in each mission never left the ground. The success of these missions heavily relied on the highly trained team of dedicated scientists, engineers, and specialists working seamlessly together in a cohesive manner. Of the six missions that completed the moon landings and brought the astronauts back to Earth, Apollo 11 was the first mission to carry out this operation, and holds special significance in many aspects, including from a technical and operational perspective. During the eight day mission, vast quantities of data such as audio, video, pictures and telemetry were collected [2, 3]. Amongst these sources, analysis of audio and text serves as a special point of interest since the audio and text records are a comprehensive collection of complex interactions between all personnel including astronauts, mission control, and backroom staff. These interactions were recorded on analog tape recorders, both to support subsequent engineering analysis and for historical purposes [2, 27]. Through the UTD-CRSS Fearless Steps initiative, we were able to get access to these analog tapes. We envision that the extraction of meaningful information from these records will yield applications in multiple domains, including high-level group dynamic analysis to understand the intricate communication characteristics involved in success parameters for large-scale time-critical and mission-critical operations. To make these analyses possible and motivate further research on real-world multi-user-tasks, all the audio data, transcripts, and associated meta-data is being made publicly available. This paper serves an overview of the full corpus release, corpus development process, its merits, challenges, the research progress made through this data, and the potential developments that can be made through the initiation of the Fearless Steps Challenge Tasks.

2. Data Collection

The Apollo 11 mission lasted 8 days 3 hours 18 minutes and 35 seconds. The entire communications between astronauts, flight controllers, and their backroom support teams inside NASA Mission Control Center (MCC) were continuously recorded using two 30-track analog reel-to-reel recording machines, namely Historical Recorder 1 (HR1) and 2 (HR2) [13, 28]. By alternately changing the tapes, continuity was ensured without any loss of data [7, 8, 13]. 29 of the 30 channels/tracks on the analog tape were used to record speech data with one channel recording the Mission Elapsed Time (MET) in an encoded IRIG-B format [5].

2.1. Digitization of Analog Tapes

The records stored by the United States National Archives and Records Administration (NARA) were used to digitize the original analog tapes, by designing a new read-head for the SoundScriber player (as shown in Fig. 1) [2, 4]. The read-head was developed specifically with the aim of digitizing the 30 channels simultaneously, thus preserving the synchronicity of the data, enabling individual channels from each HR1 and HR2 to be indexed and stored separately. Synchronous multichannel reconstruction of the entire mission is made possible by using

10.21437/Interspeech.2018-1942

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

Figure 1: *(left): The SoundScriber device used to decode 30 track analog tapes, and (right): The UTD-CRSS designed read-head decoder retrofitted to the SoundScriber [13]*

2.2. Text and Other Data Sources

In addition to the audio collection, understanding the mission specific aspects through text sources can provide insights useful for further analysis of the audio data [1, 3]. Fig. 2 illustrates a perspective for the eight stages the entire mission can be classified into. The amount of speech content, conversation, speech durations, and other such factors vary depending on the stage of the mission.

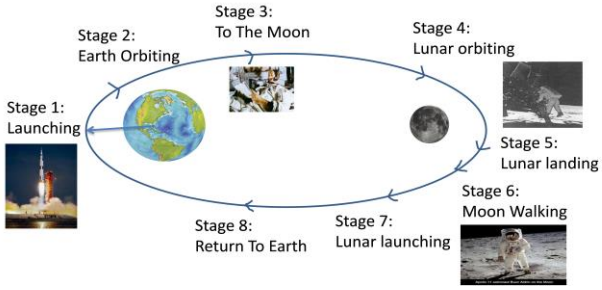

Figure 2: *Overview of the timeline of Apollo 11 mission [3]*

Document linking of text available sources with the speech data using timing information has helped create meaningful illustrations and visualizations as can be seen by the application developed by UTDesign and UTD-CRSS team [29].

3. Corpus Development

The Apollo data is unique and poses multiple challenges. Comprised of 19,000 hours of naturalistic multi-channel data, it is characterized by multiple classes of noise and degradation and several overlap instances over the 29 channels. Most audio channels are degraded due to high channel noise, system noise, attenuated signal bandwidth, transmission noise, cosmic noise, analog tape static noise, noise due to tape ageing, etc. [9, 10, 28]. Moreover, the noise conditions and signal-to-noise ratio (SNR) levels change over time and channels, varying from 0 dB to 20 dB. Some channels even have the presence of babble noise depending on the location of the personnel in the MCC. For the Apollo missions, head-mounted Plantronics microphones were used, but some in-spacecraft recordings were made using fixed far-field microphones which also picked up the presence of environmental noise (e.g., glycol cooling pumps and thruster firings) that varied over time [13]. These conditions severely degrade the performance of conventional speech activity detection algorithms [10, 13, 16]. Due to the time-critical nature of the mission, multiple instances with rapid switching of speakers exist. In many cases, extremely short duration responses are common. As an example, the average duration for speakers during mission status updates are close to 0.5 seconds with as many as 15 speakers occurring in turns in a span of 10 seconds. This poses a serious challenge for speaker recognition and diarization systems [6, 20, 28]. The speech density varies dramatically over time, depending on the channel, the stage of the mission, and issues encountered during the mission. Some instances show more than 20 active speakers at a time, carrying conversations for 15 minutes at a stretch, and some other instances show extended periods of silence, usually for hours. For astronauts, their vocal tract characteristics have been observed to have considerable changes through different stages of the mission [3]. These factors render thresholding mechanisms and diarization system performances degraded. The conversational content in the missions was specific to the standards maintained by NASA for efficient air-to-air, air-to-ground, and ground communications. Using standard language models for this application can lead to misclassification of words detected, posing a significant challenge for speech recognition. Hence, there is a need for incorporating NASA specific vocabulary to existing language models [9, 19]. Analyzing unprompted speech has its own challenges [13, 28]. All the speech recorded in the corpus is unprompted, and hence subject to significant variations in speech characteristics for every speaker. These are some challenges have shown to significantly degrade the quality of generalized SAD, ASR, SID, and Diarization models. Apollo specific application developments have shown to drastically improve system accuracy on these models [3, 9, 14, 16, 28].

Many characteristics of this data can also be observed in situations not involving space missions. Therefore, analysis on this corpus can also be applicable to any high impact multi-party speaker situations. The corpus addresses general and specific problems in both speech and natural language domains. The development of improved algorithms relies on application-specific domain knowledge, which we will discuss in the coming sections.

3.1. Communication Protocols

Specific protocols followed during the mission were imperative for ensuring successful communication.

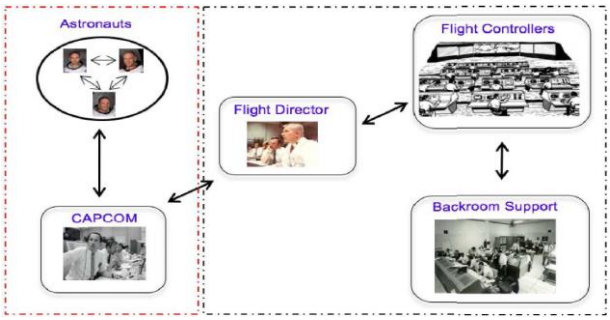

Figure 3: *Apollo-11 communication overview. The black dotted parts are ground communications between hundreds of flight controllers and their 'backroom' support staff. The red dotted parts are space-to-ground communications. [28]*

Knowledge of these communication characteristics can be leveraged to achieve better inferences through informed analysis. Fig. 3 shows a basic structure of the communication protocols. Only the Capsule Communicator (CAPCOM) could directly communicate with the astronauts, with the Flight Director (FLIGHT) in control of accessing all other channel loops. With multiple speakers joining in on these loops at various points in time, all personnel would address the channel owner by their assigned channel names [16, 17, 25]. In fact, all backroom staff are present on multiple channel loops. Audio markers such as 'Quindar Tones' were used to infer communication with the astronauts.

Another useful feature for analysis of astronauts' interactional dynamics is the chord diagram representation of their conversation content as shown in Fig. 4. [9, 23] This chord diagram is a visual map of the density of conversations each astronaut had with each other and the CAPCOM.

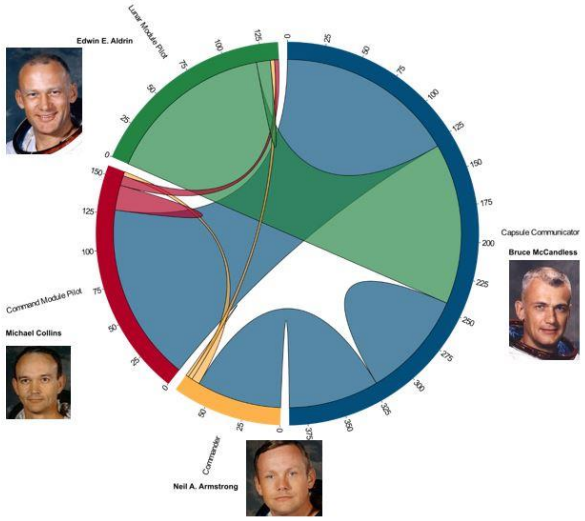

Figure 4: *Chord Diagram of Astronauts' Conversations [23]*

3.2. Mission Elapsed Time Decoding

Time codes are necessary to link events to audio instances across all channels. However, time-varying noise elements affected the channel. Existing tools for decoding the MET timecodes were not tuned to compensate for the channel artifacts, giving a decoding error rate of more than 10%. To overcome the time, frequency and amplitude distortions present in the recorded channel, a dynamic thresholding system was developed, achieving decoding accuracy of 99.2%. The decoded timing information format describes the 'Year, Month, Day, Hour, and Seconds' per data frame in that order.

3.3. Pipeline Diarization Transcript Generation

Transcribing the Apollo data manually is extremely resource intensive. Furthermore, due to the vast nature of the audio, an alternate route to manual transcription must be found. Generation of transcripts for the entire data has been of utmost importance to be able to make further progress on language tasks [9]. An Apollo-centric diarization scheme that was developed to refine the ASR system output greatly improved the quality of generated transcripts [13]. Leveraging multichannel information to improve the quality of the transcripts is one of the many ways to solve the problem of speech recognition using multiple modes of information.

The diarization transcripts and timecode information is being provided for the entire data with the release of the corpus.

4. 100-Hour Challenge Corpus

To create challenge tasks and test the performance of developed systems, we had to select a subset of the Fearless Steps Corpus which could be transcribed manually by professional annotators, and would be representative of the challenges encountered in the comprehensive corpus, while being a rich source of information.

4.1. Challenge Corpus Selection

The Stages '1', '5' and '6' (see Fig. 1) which were high impact mission-critical events were found to be ideal for the development of the 100-hour Challenge Corpus. With the quality of speech data varying between 0 and 20 dB SNR in this challenge corpus [10, 23, 24, 28], the channel variability and complex interactions across all five channels of interest discussed in the previous section are mostly encapsulated in these 100 hours. These three major events the multichannel data is chosen from are:

1. Lift Off (25 hours)
2. Lunar Landing (50 hours)
3. Lunar Walking (25 hours)

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

1. Flight Director (FD)
2. Mission Operations Control Room (MOCR)
3. Guidance Navigation and Control (GNC)
4. Network Controller (NTWK)
5. Electrical, Environmental and Consumables Manager (EECOM)

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

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

Table 1: *Total speech duration (in hours) of the released dataset for each channel type.*

The distribution of total speech content in each of the channels for every event has been given in Table 1, with the total speech content for each event provided in the final column, and over each channel provided in the final row. Total speech content in the challenge corpus amounts to approximately 36 hours.

4.2. Transcription

Manual Transcription efforts were conducted over a portion of the Apollo Data using the LDC Transcriber Tool (see Fig. 5) [22]. The transcripts in this format include the Speech, Speaker

and Diarization information, which is being released along with the annotations specific to each task [24, 25, 28].

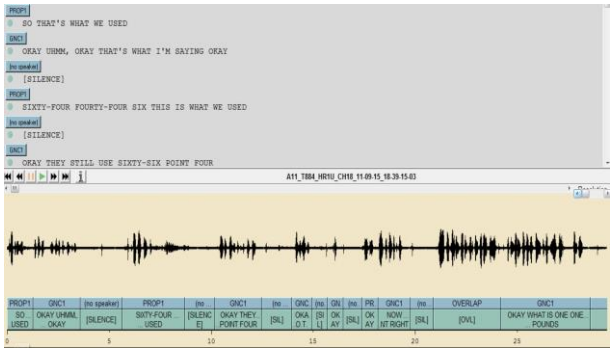

Figure 5: *Illustration of the Apollo Transcripts using the Transcriber tool*

4.3. Data Distribution for Challenge Tasks

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

	Speech Duration (h)	**Silence Duration (h)**	**Average no. of Speakers/hour**
Train	25.9	31.3	35
Dev	3.5	6.6	29
Test	5.5	10.2	31

Table 2: *Total duration of speech and silence (in hours) of the dataset, and average number of speakers per hourly segments.*

To address and balance all the challenges [1, 2, 3, 11, 12, 14, 15] of the corpus, we try to achieve a similar difficulty distribution for the train, development, and test dataset. The general statistics about the dataset are described in Table 2. For the preparation of the subset, we included speech segments over the five channels that have similar speech activity, which could be used for multichannel analysis, and have a consistent number of speakers between train, development, and test datasets.

5. Challenge Tasks

As an effort to motivate an initial research direction, UTD-CRSS will also be hosting five challenge tasks for which guidelines following NIST standards will be provided. These following Tasks are designed to advance research efforts not only in Speech Processing and Machine Learning, but also in Natural Language Understanding:

1. Speech Activity Detection (SAD)
2. Speaker Diarization
3. Speaker Identification (SID)
4. Automatic Speech Recognition (ASR)
5. Keyword Spotting for Joint Topic and Sentiment Detection

These five tasks include selected speech segments which would have the highest impact in terms of speech, text and language analysis. Consolidated transcripts with ground truth labels for the above-mentioned tasks will be made available for this challenge. Guidelines for the tasks will be provided at the release of the challenge. The next section will provide an overview on the results achieved by our systems developed specifically for the Apollo data and in general the motivation behind holding these challenges tasks.

6. Experimental Results

Research on the corpus has seen the development of multiple unsupervised and supervised strategies. It should be noted that since these experiments were performed at different stages of the corpus development and manual transcription efforts, the results given below have been evaluated over specific subsets of the data annotated for testing performance of the algorithms.

For the task of speech activity detection, the unsupervised strategy To-Combo-SAD achieved an equal error rate (EER) of 10% [10]. A supervised curriculum learning based approach yielded an 11.2% EER over highly degraded channels (~5dB) [9, 21]. For speaker diarization, active learning based clustering techniques yielded a diarization error rate (DER) of 30% [14, 28]. A baseline SID system using an i-vector-PLDA framework gave a classification accuracy of 59.1% [6].

To adapt to the highly technical nature of the conversational content in the recordings, an application specific language model adapted to NASA and other such organizational standards had to be created. The language model trained with over four million relevant words together with a DNN based acoustic model improved the accuracy of the ASR system significantly [9, 13]. The word error rate (WER) varied drastically with the noise conditions, achieving 18% for clean speech and as much as 120% for heavily degraded channels.

7. Conclusions

The Fearless Steps Corpus poses multiple challenges in the speech and language domain. Due to the nature of this data, solving these challenges can lead to the development of generalized algorithms with significant applications in research and industry. In addition to this, with the advent of highly optimized algorithms and large-scale computing power, the speech and language processing technologies have the potential to play an important role in future deep-space missions. The challenge tasks have been created to encourage directed collaborative effort towards advancements in speech technologies. As such, there is still significant research potential, since the corpus stands as the largest collection of unprompted speech. Having merely scratched the surface initiating the challenge tasks, there are still various novel problems that can be elegantly solved using this data. Researchers who wish to work on this corpus will be provided with the corpus and associated meta-data, and are encouraged to use this data for any novel topics of their interest.

8. Acknowledgements

This project was funded by AFRL under contract FA8750-15-1-0205, NSF 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. UTD-CRSS lab would like to thank Gregory H. Wiseman and Karen L. Walsemann of National Aeronautics and Space Administration (NASA) for their never-ending support throughout the digitization process. We would like to thank Tuan Nguyen for helping out tirelessly in digitization process.

9. References

[1] A. Sangwan, L. Kaushik, C. Yu, J. H. L. Hansen and Douglas W. Oard. "Houston, we have a solution: using NASA Apollo program to advance speech and language processing technology." INTERSPEECH. 2013.

[2] Douglas W. Oard, J. H. L. Hansen, A. Sangwan, B. Toth, L. Kaushik, and C. Yu. "Toward Access to Multi-Perspective Archival Spoken Word Content." In Digital Libraries: Knowledge, Information, and Data in an Open Access Society, 10075:77–82. Cham: Springer International Publishing, 2016.

[3] C. Yu, J. H. L. Hansen, and Douglas 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," INTERSPEECH. 2014.

[4] Pear Jr, B. Charles, "Accessories and auxiliary equipment." Magnetic Recording in Science and Industry (1967): 420.

[5] J. B. Houston, "Converting Time Signals from BCD to IRIG-B." (1982).

[6] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel,, J. Silovsky, (2011). "The Kaldi speech recognition toolkit". In IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society.

[7] "Apollo Flight Journal" https://history.nasa.gov/afj/

[8] "Apollo Lunar Surface Journal" https://history.nasa.gov/alsj/

[9] L. Kaushik. "Conversational Speech Understanding in highly Naturalistic Audio Streams*"* PhD Dissertation, The University of Texas at Dallas, 2017.

[10] A. Ziaei, L. Kaushik, A. Sangwan, J. H.L. Hansen, & D. W. Oard, (2014). "Speech activity detection for NASA Apollo Space Missions: Challenges and Solutions." (pp. 1544-1548) INTERSPEECH. 2013.

[11] L. Kaushik, A. Sangwan, and J. H. L. Hansen. "Sentiment extraction from natural audio streams." In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pp. 8485-8489. IEEE, 2013.

[12] L. Kaushik, A. Sangwan, and J. H. L. Hansen. "Automatic sentiment extraction from YouTube videos." In Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on, pp. 239-244. IEEE, 2013.

[13] L. Kaushik, A. Sangwan, and J. H.L. Hansen. "Multi-Channel Apollo Mission Speech Transcripts Calibration," 2799–2803. IINERSPEECH, 2017.

[14] C. Yu and J. H. L. Hansen, "Active Learning Based Constrained Clustering For Speaker Diarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 11, pp. 2188-2198, Nov. 2017. doi: 10.1109/TASLP.2017.2747097

[15] L. Kaushik, A. Sangwan and J. H. L. Hansen, "Automatic Sentiment Detection in Naturalistic Audio," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 8, pp. 1668-1679, Aug. 2017.

[16] C. Yu, and J. H. L. Hansen. "A study of voice production characteristics of astronaut speech during Apollo 11 for speaker modeling in space." Journal of the Acoustic Society of America (JASA), 2017 Mar: 141(3):1605.

[17] NASA Mission Control Positions: https://www.nasa.gov/centers/langley/news/factsheets/Apollo.html

[18] Apollo 11 Mission Reports: https://www.hq.nasa.gov/alsj/a11/a11mr.html

[19] A. Sangwan, and J. H. L. Hansen. "Keyword recognition with phone confusion networks and phonological features based keyword threshold detection." In Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on, pp. 711-715. IEEE, 2010.

[20] S. Ranjan and J. H. L. Hansen, "Curriculum Learning Based Approaches for Noise Robust Speaker Recognition", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 197-210, 2018.

[21] S. Ranjan, A. Misra, and J. H. L. Hansen, "Curriculum learning based probabilistic linear discriminant analysis for noise robust speaker recognition," in Proc. ISCA INTERSPEECH, 2017, pp. 3717–3721.

[22] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. "Transcriber: A Free Tool for Segmenting, Labeling and Transcribing Speech," (1998).

[23] A. Joglekar, C. Yu, L. Kaushik, A. Sangwan, J. H. L. Hansen, "Fearless Steps Corpus: A Review Of The Audio Corpus For Apollo-11 Space Mission And Associated Challenge Tasks" In NASA Human Research Program Investigators' Workshop (HRP), 2018.

[24] L. Kaushik, A. Sangwan, J. H. L. Hansen, "Apollo Archive Explorer: An Online Tool To Explore And Study Space Missions" In NASA Human Research Program Investigators' Workshop (HRP), 2017.

[25] L. Kaushik, C. Yu, A. Sangwan, J. H. L. Hansen, "The Heroes Behind the Heroes of Apollo-11: Role of STEM" In ASEE Gulf Southwest Annual Regional Conference, 2017.

[26] Apollo 11 Mission Overview: https://www.nasa.gov/mission_pages/apollo/missions/apollo11.html

[27] G. Swanson, "We have liftoff! The story behind the Mercury, Gemini and Apollo air to ground transmissions." Spaceflight 43(2), 74–80 (2001).

[28] C. Yu. "Robust Speaker Modeling In Non-neutral Environments With Application To Large Scale Multi-speaker Audio Streams" PhD Dissertation, The University of Texas at Dallas, 2017.

[29] Explore Apollo Document Linking Application: https://app.exploreapollo.org/