Meena M. Chandra Shekar and John H.L. Hansen

Historical Audio Search and Preservation: Finding Waldo Within the Fearless Steps Apollo 11 Naturalistic Audio Corpus

Apollo 11 was the first manned space mission to successfully bring astronauts to the Moon and return them safely. As part of NASA's goal in assessing team and mission success, all voice communications within mission control, astronauts, and support staff were captured using a multichannel analog system, which until recently had never been made available. More than 400 personnel served as mission specialists/support who communicated across 30 audio loops, resulting in 9,000+ h of data. It is essential to identify each speaker's role during Apollo and analyze group communication to achieve a common goal. Manual annotation is costly, so this makes it necessary to determine robust speaker identification and tracking methods. In this study, a subset of 100 h derived from the collective 9,000 h of the Fearless Steps (FSteps) Apollo 11 audio data were investigated, corresponding to three critical mission phases: liftoff, lunar landing, and lunar walk. A speaker recognition assessment is performed on 140 speakers from a collective set of 183 NASA mission specialists who participated, based on sufficient training data obtained from 5 (out of 30) mission channels. We observe that SincNet performs the best in terms of accuracy and F score achieving 78.6% accuracy. Speaker models trained on specific phases are also compared with each other to determine if stress, g-force/atmospheric pressure, acoustic environments, etc., impact the robustness of the models. Higher performance was obtained using i-vector and x-vector systems for phases with limited data, such as liftoff and lunar walk. When provided with a sufficient amount of data (lunar landing phase), SincNet was shown to perform the best. This represents one of the first investigations on speaker recognition for massively large team-based communications involving naturalistic communication data. In addition, we use the concept of "Where's Waldo?" to identify key speakers of interest (SOIs) and track them over the complete FSteps audio corpus. This additional task provides an opportunity for the research community to transition the FSteps collection as an educational resource while also serving as a tribute to the "heroes behind the heroes of Apollo."

Digital Object Identifier 10.1109/MSP.2023.3237001
Date of current version: 1 May 2023

Introduction

Speech technology has evolved dramatically in recent decades with voice communication and voice-enabled devices becoming ubiquitous in the daily lives of consumers. Many research advancements in the speech and language community have been possible through advanced machine learning algorithms and models. However, machine learning algorithms require extensive and diverse audio data to develop effective models. Most existing datasets rely primarily on simulated/recorded speech over limited time periods (e.g., one to maybe several hours). To develop next-generation technologies, there is a requirement for audio materials to be collected in the presence of highly variable background noise and channel conditions, pose significant real-world challenges, be real and not simulated, and include speaker variability (age, dialect, task stress, emotions, etc.). Today, both education and industry rely more on collaborative team-based problem solving. However, there is a lack of resources available to understand and model the dynamics of how individuals with different skill sets blend their expertise to address a common task. Unfortunately, corporations with speech/audio data are reluctant to share data with the community due in part to privacy/legal reasons. Hence, there is a significant need by the speech community for access to "big data" consisting of natural speech that is freely available. Fortunately, a massive audio dataset that is naturalistic, real-world, multispeaker, task directed, and consisting of fully diarized, synchronized audio has recently been made freely available to the community: the FSteps Apollo 11 audio corpus (thanks to the Center for Robust Speech Systems, the University of Texas at Dallas [CRSS-UTDallas] [1]). With 400+ personnel, more than 9,000 h of audio data, a full diarized speaker, and Automatic Speech Recognition transcripts, significant research potential exists through analysis of these data. It is essential to analyze groups of teams that communicate to learn, engage, and solve complex problems. It is not possible to annotate every

speaker manually in this massive corpus, nor is it possible for any individual human being to decipher the interactions taking place among 400+ speakers, making it necessary to employ automatic methods to transcribe and track speakers. In addition, we use the concept of "Where's Waldo?" to identify key SOIs and track them across the complete FSteps audio corpus. This provides an opportunity for the research community to leverage this collection as an educational resource while also serving as a tribute to the heroes behind the heroes of Apollo.

Speech technology and challenges

Speech technology and voice communications have evolved to contribute to smart homes, voice dialing, smart classrooms, and voice-enabled devices. Voice communications have become prominent in the daily lives of consumers, with digital assistants such as Apple's Siri, Amazon Alexa, Google Assistant, JD Ding Dong, and Microsoft Cortana used for completing complex tasks using voice. Such research advancements have been possible because of using advanced machine learning techniques. However, machine learning models are data hungry, and there is an increasing need for freely available large audio datasets to create effective models for voice technologies. Industry giants such as Apple, Amazon, IBM, YouTube, Google, and Microsoft are constrained at some level to share such data with the community due to privacy/legal reasons. Other datasets that do exist rely primarily on simulated or artificial voice problems over a staged limited time period. There is a significant need from the speech and language community to access big-data audio that is natural, depicts real-life scenarios, is devoid of privacy issues, is multispeaker, and is freely available, to develop next-generation technologies [2].

FSteps corpus

Establishing the corpus

Apollo 11 was the first manned space mission that landed on the Moon. Virtually all logistics were accomplished through audio, with Apollo missions spanning 7- to 10-days, representing coordinated efforts of hundreds of individuals within NASA Mission Control. Well over 100,000 h of synchronized analog data were produced for the entire program. The Apollo missions [24] represent unique data since they are perhaps some of the few events, where all possible communications were recorded using multiple synchronized channel recorders of these real-world task-driven teams, all of which produced multidimension/location data sources with the potential to be made freely available to the public. For example, recent historical events, such as the U.S. Hurricane Katrina disaster [25], the 9/11 U.S. terrorist attacks [26], or Japan's Fukushima Daiichi nuclear reactor meltdown [27], bear resemblance to the Apollo missions in terms of the need for effective team communications, time-sensitive tasks, and number of team-focused personnel involved. These events consist of critical task operation, complexity of human undertaking, and the degree and timing of intercommunications required. However, access to such data sources for research and scientific study may be difficult, if not impossible, due to both privacy and any coordinated and synchronized recording infrastructure when the event took place [3].

Under U.S. NSF support, CRSS-UTDallas spent six years to recover Apollo audio to establish the FSteps corpus, consisting of digitizing all analog audio with full diarization metadata production (who spoke what and when). The corpus was recovered from 30-track analog tapes, resulting in a corpus containing 30 channels of time-synchronized data, including flight director (FD) loop, air-to-ground capsule communicator (CAPCOM) communication, back-room loops, multiple mission logistics loops, etc. Thus far, CRSS-UTDallas has digitized and recovered as well as developed an advanced diarization (e.g., who said what and when) pipeline and processed 19,000 h of Apollo audio consisting of naturalistic, multichannel conversational speech spanning over 30 time-synchronized channels (i.e., all of Apollo 11 and most of Apollo 13). Significant research potential exists through analysis of this dataset since it is the largest collection of task-specific naturalistic time-synchronized speech data freely available worldwide [1], [4].

FSteps corpus in the news

The CRSS-UTDallas and FSteps Audio corpus have been featured in over 40 television, radio, newspaper, and online news stories from NBC, CBS, BBC-UK, NPR, NSF, ASEE, Discover, NHK-Japan, *National Geographic*, the *Dallas Morning News*, Texas Country Reporter, *Community Impact*, NSF, and others [5], [6], [7], [8], [9]. The most significant recognition was the contribution to the News Network CNN documentary movie on Apollo 11, where CRSS-UTDallas provided all recovered audio including complete diarized transcript speaker/text content that allowed convolutional neural network (CNN) to "stitch" the recovered voice to hundreds of hours of NASA silent 70-mm mission control room video footage (i.e., CRSS-UTDallas was recognized in the film credits). The FSteps corpus poses a unique opportunity for the development of semi-supervised systems for massive data with limited ground truth annotations.

Challenges of the Apollo corpus

The sheer volume and complexity of the NASA Apollo data and the underlying operation provide many research opportunities for audio, speech, and language processing. In the context of Apollo, this is a difficult problem given that audio contains several artifacts, such as 1) variable additive noise, 2) variable channel characteristics, 3) reverberation, and 4) variable bandwidth accompanied by speech production issues (such as stress) and speech capture issues (such as astronaut speech captured while walking on the Moon in space suits). A number of studies have considered detection of speech under stress [10], [11], [12] or recognition of speech under stress [13], [14]. An interesting study of the Apollo astronauts' voice characteristics was conducted over three different acoustic environments as well [15]. For such time and mission-critical naturalistic data, there is extensive and diverse speaker

variability. The diversity and variability of speaker state for astronauts over a 6- to 11-day mission offers a unique opportunity in monitoring individuals through voice communications. The mission-specific aspects can provide further insights regarding speech content, conversational turns, speech duration, and other conversational engagement factors that vary depending on mission phases.

The UTDallas FSteps Apollo data are composed of 19,000 h (9,000 for Apollo 11) possessing unique and multiple challenges over 30 subteam-based channels. For our study, we have selected a subset of 100 h [1] from five speech active channel loops manually transcribed by professional annotators for speaker labels. The 100 h are obtained from three mission critical events: 1) lift off (25 h), 2) lunar landing (50 h), and 3) lunar walking (25 h).

The five channels are

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 100 h are divided into training (60 h), development (20 h), and test (20 h) sets. For the 183 speakers in this 100 h set, we considered a total of 140 speakers who produced at least 15 s of total speaker duration with three or more speech utterances for each speaker. Each speaker had a minimum duration of 1+ s of audio speech. Of the 140 speakers, three speakers are astronauts who are present only in the lunar landing phase. Figure 1 shows the speech content distribution over the five primary channels (FD, MOCR, GNC, NTWK, and EECOM) in three different phases. Although this corpus has 100 h of audio data, the actual speech content consists of about 17 h. Figure 1 shows that there is a nonuniform distribution across most speakers, and some speakers are present in only one of three phases. Very few speakers are present in all three mission phases (note that this is constrained only for this subset). To understand why a speaker may not be present in all three phases, it is necessary to understand how NASA specialists communicate with each other in the MOCR. The next section highlights the MOCR communications protocol.

Communications in the MOCR

A total of 38 astronauts made up the 15 mission crews between 1968 and 1975. Of those, 24 flew to the Moon on nine missions, with 12 being Moon walkers. Two 30-track audio historical recorders were employed to capture all team loops of the mission control center (MCC) resulting in more than 100,000 h of Apollo analog audio. The MCC was organized hierarchically: one FD, one CAPCOM, more than 10–15 MOCR flight controllers, and a corresponding set of backrooms with specialists that support each flight controller. One channel loop connected the FD with the flight controllers, and each backroom had a separate loop to

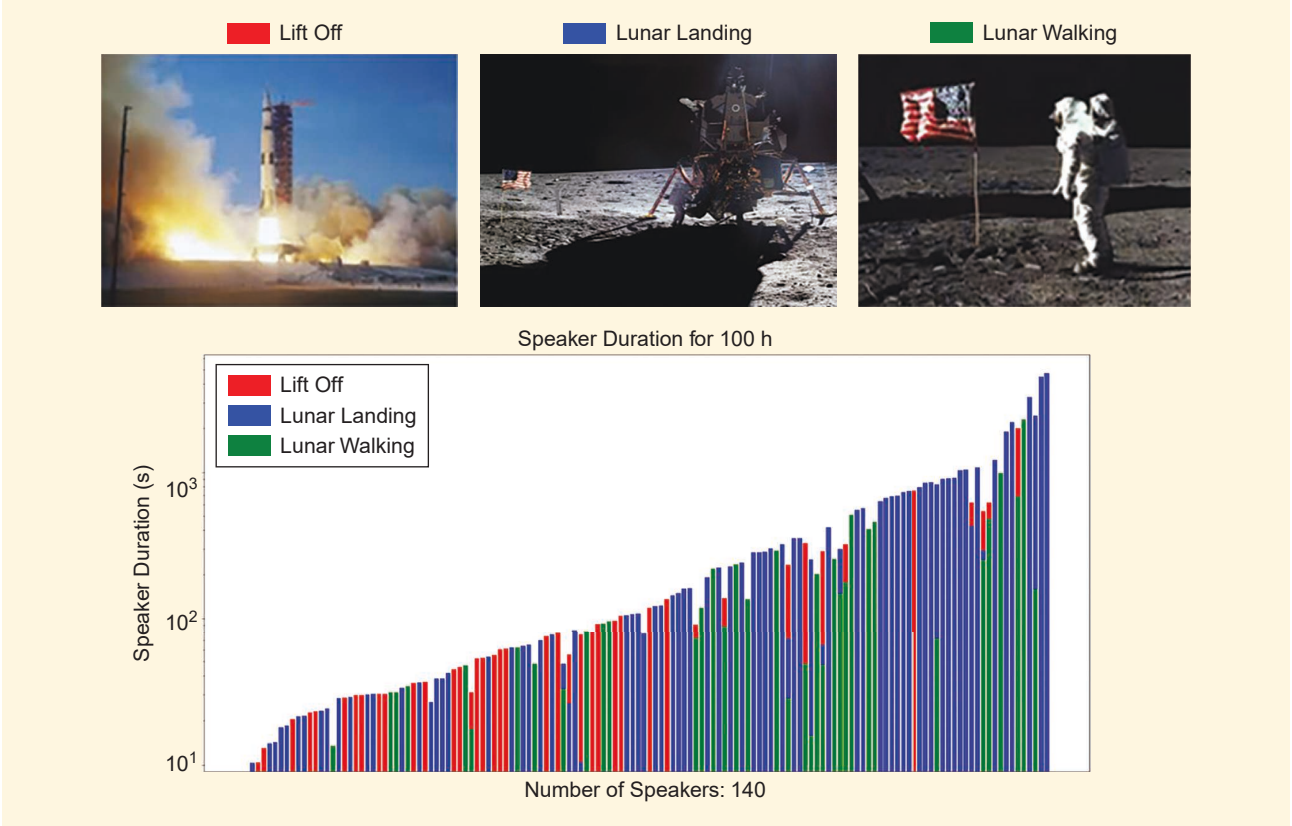


FIGURE 1. Varying speaker duration throughout FS Apollo 11 audio.

connect them with the flight controller who they supported. Two special loops were also recorded, one between the spacecraft and the MCC (CAPCOM) and a second for the news media that included those communications along with public affairs commentary [16]. Only the CAPCOM was able to talk directly with the astronauts.

NASA mission specialists used close-talking microphones and at times phone headsets. Because of the Earth-to-Moon trajectory, communication with the spacecraft was possible for about 90% of this time. Also, audio transmission from Earth to/from the Apollo 11 capsule was achieved through S-band communication with multiple relay stations across Earth back to NASA in Houston, TX, USA (e.g., Goldstone, CA, USA; Madrid, Spain; Honeysuckle, Australia; and Canary Islands). These recordings exhibit highly variable channel characteristics due to the diversity in communication signal paths [3]. Many complex multiparty activities are coordinated using speech, including air traffic control, military command centers, and human spaceflight operations. It is not possible for one person to listen/uncover every event happening or to precisely transcribe all interactions taking place. This represents motivation for an algorithm-based solution to identify, tag, and track all speakers in the Apollo audio. Since this is a massive audio corpus, it requires an effective and robust solution for speaker identification.

Finding Waldo

NASA's Apollo program stands as one of the most significant contributions to humankind. Given the 9,000+ h of Apollo 11 audio, with 400+ speakers over an 8-day mission, it is necessary to tag speaker exchanges with many being short-duration turns. Due to strict NASA communication protocols in such time-critical missions, most personnel employed a compact speaking style, with information turn-taking over 3- to 5-s windows. This poses a unique and challenging research problem for speaker tagging since most speaker recognition systems need 10 s to 5 min for the highest accuracy of finding "needles in a haystack" from a speaker tagging perspective. For example, Figure 2 shows five A11 channels at a time instance during the liftoff. Communication between mission specialists takes place only when there is a specific technical or mission need. To illustrate the rarity of communication turns, we consider a 30-min segment across five channels as shown. Red segments highlight silence, while blue

FIGURE 2. Speech activity detection for three mission critical phases: (a) liftoff, (b) lunar landing, and (c) lunar walk.

segments highlight speech produced by speakers. It is difficult to tag <3-s speech utterances (utterances such as "uh huh," "yes," etc.), as well as the need to assess the possibility of silence between speaker turns. Hence, we divide each 30-min segment into a series of 20-s analysis blocks. If a 20-s block contains greater duration of speech versus silence, this 20-s block is highlighted in blue, whereas a greater duration of silence is highlighted in red. We see there is significant silence compared with speech. Similar speech/silence multichannel plots have been demonstrated for other phases of the mission.

To track and tag individual speakers across our FSteps audio dataset, we use the concept of Where's Waldo? to identify all instances of our SOIs across a cluster of other speakers. The resulting diarization of Apollo 11 audio and text material captures the complex interaction between astronauts, mission control, scientists, engineers, and others, creating numerous possibilities for task/content linking. Figure 3 shows a t-distributed stochastic embedding (T-SNE) representation of each SOI x-vector embeddings versus non-SOI x-vector embeddings. The speaker embeddings form a separate cluster for each speaker model, making it possible for us to extract a particular speaker from a cluster of speakers. In this example, we select five SOIs: Astronaut Neil Armstrong, Astronaut Buzz Aldrin, Astronaut Michael Collins, FD Gene Kranz, and CAPCOM Charlie Duke. Figure 4 shows each speaker's speech duration in what is referred to as a "donut plot" for the speaker/other speaker plus silence plot. Figure 4 provides an intriguing global perspective of the speaker interaction between each SOI versus other speakers and silence across the audio clips. We see for CAPCOM, there is significant speaker turn-taking activity compared with non-SOI speakers where CAPCOM normally is the prime speaker with the astronauts. Analyzing the handful of speakers present in the small audio dataset of 100 h can be extended to the complete Apollo 11 mission (with 9,000 h of data) as well as future efforts for the complete Apollo program (with 150,000 h of audio) since many speakers are common throughout the Apollo missions. This big-data community-based audio resource will support the team members of the Apollo program and their families. Identifying these personnel can help pay tribute and yield personal recognition to the hundreds of notable engineers and scientists who made this feat possible. This collection opens new research options for recognizing team communication, group dynamics, and human engagement/psychology for future deep space missions.

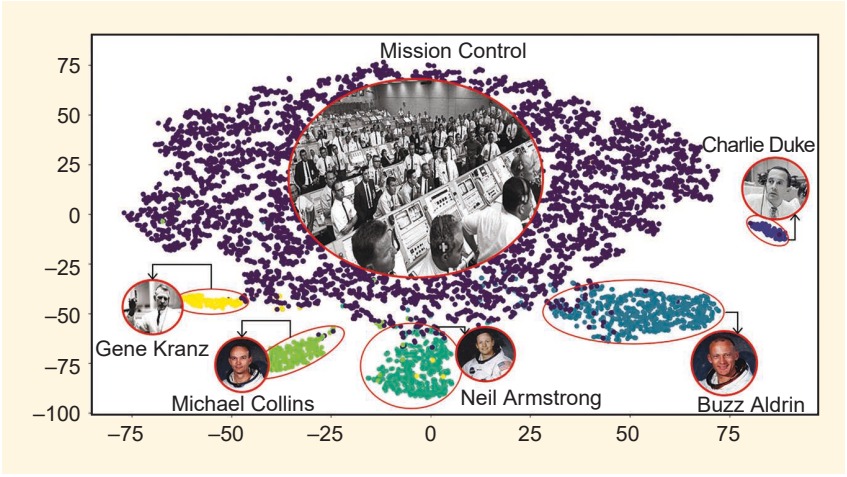


FIGURE 3. T-distributed stochastic embedding (T-SNE) representation of speaker utterances on FS Apollo 11 for mission control and SOIs.

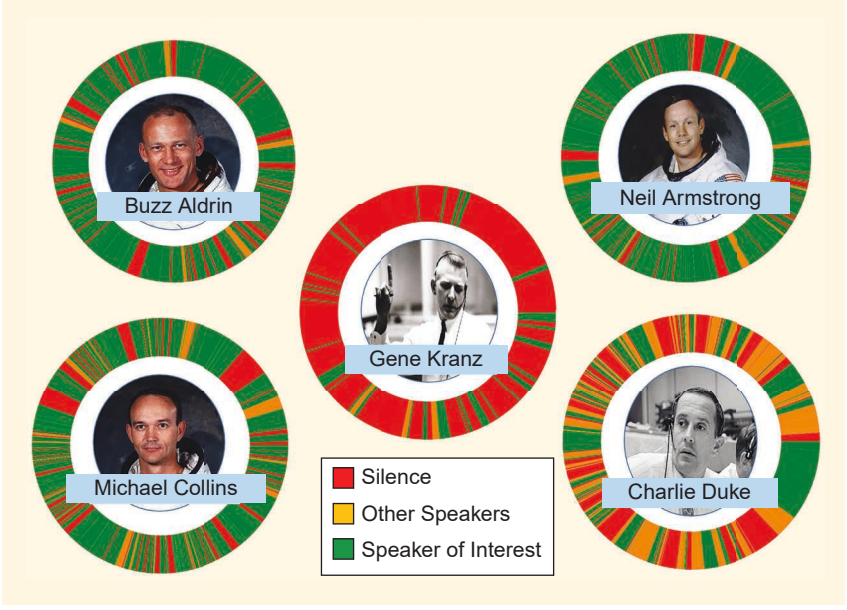


FIGURE 4. Speaker duration of particular speakers versus other speakers.

Speaker recognition systems

In the last decade, state-of-the-art speaker recognition systems have evolved from Gaussian mixture models (GMMs) to using deep neural networks (DNN) to train speaker models. The combination of an i-vector and a distance measure has become the dominant approach for text-independent speaker recognition. For our study, we choose three baseline systems which are GMM-based i-vector systems [17], DNN-based x-vector systems [18], and CNN-based SincNet systems [19]. For the i-vector and x-vector systems, input features of 12 Mel-frequency cepstral coefficients with a frame length of 25 ms. Delta and double-delta features are appended to create a 39-dimensional

feature vector. An energy-based voice activity detector selects features corresponding to speech frames. To extract i vectors for each speaker utterance, the Universal Background Model (UBM) was trained on the National Institute of Standards and Technology-Speaker Recognition Evaluation 16 [20] corpora to create a 2,048-component full-covariance GMM. A 600-dimensional i-vector extractor was developed and extracted. To extract x vectors, a feed-forward DNN that computes speaker embeddings from variable-length acoustic segments was used. The DNN was trained to classify the N speakers in the training data. DNN embeddings are trained on the SRE16 dataset, and extracted x vectors are 512-dimensional vectors. The Kaldi speech recognition tool kit was used to train both i vectors and x vectors. For the CNN-based SincNet architecture, each raw speech waveform is split into chunks of 200 ms with 10 ms overlap. The first convolutional layer uses sinc functions with 80 filters of length L = 251 samples followed by two standard CNN layers, both using 60 filters of length 5, and finally, three fully connected layers composed of 2,048 neurons with batch normalization and layer normalization. For our study, we have combined the speaker recognition system with several dimensionality reduction and scoring methods such as principal component analysis (PCA), linear discriminant analysis (LDA), cosine distance scoring (CDS), random forest (RF), and XGBoost.

As seen from Figure 1, speaker duration is not uniform across all speakers, suggesting a data imbalance. Therefore, we show that several evaluation metrics to determine where performance of each baseline system could fail (e.g., micro-average, macro-average, accuracy). F score is defined as the harmonic mean between precision and recall. The microaverage calculates the contributions of all speaker models to compute the average metric, whereas the macroaverage computes the metric independently and then computes the average. The microaverage can help in reflecting any class imbalance in the dataset. The results show that for all three baseline systems, the microaverage is greater than the macroaverage, indicating that these systems are classifying speaker models with smaller sample sets inaccurately. The SincNet solution performs the best in terms of both accuracy and f scores. This evaluation suggests there are viable solutions for tagging moderately short speaker turns in this Apollo collection.

Speaker recognition from Earth to the Moon

Naturalistic and long-duration continuous audio recordings are very interesting and challenging in terms of speech activity detection, speaker recognition, and speech analysis. The performance of speaker recognition systems on the Apollo 11 audio dataset can be impacted because of various acoustic environments (such as Earth, deep space, or the surface of the Moon). This new Apollo dataset over the span of 8 days, 3 h, 18 min, and 35 s, or a total of approximately 196 h, provides new opportunities for speech technology and team dynamics analysis. Other factors that can impact the performance of speaker recognition systems include the mismatch between the training and the test environments. Previous studies have addressed this issue by using an acoustic modeling framework (GMM-UBM) that is trained on specific noisy environments [21]. Another study explores speaker modeling methods for speaker verification in noisy environments by focusing on building hybrid classifiers and using utterance partitioning [22]. However, this study deals with environments where

Table 1. Performance of baseline speaker recognition systems.

System	Scoring/Classification System	F Score (%)		Accuracy (%)
		Microaverage	Macroaverage	
i-vector	CDS	48	42	47
	RF + PCA	40	13	39.74
	RF + LDA	54	20	49.5
	XGBoost + PCA	55	23	54.72
	XGBoost + LDA	62	28	62.25
x-vector	CDS	47	38	46.8
	RF + PCA	58	25	57.87
	RF + LDA	67	31	67.05
	XGBoost + PCA	70	39	70.27
	XGBoost + LDA	73	41	73.4
SincNet	–	79	53	78.6

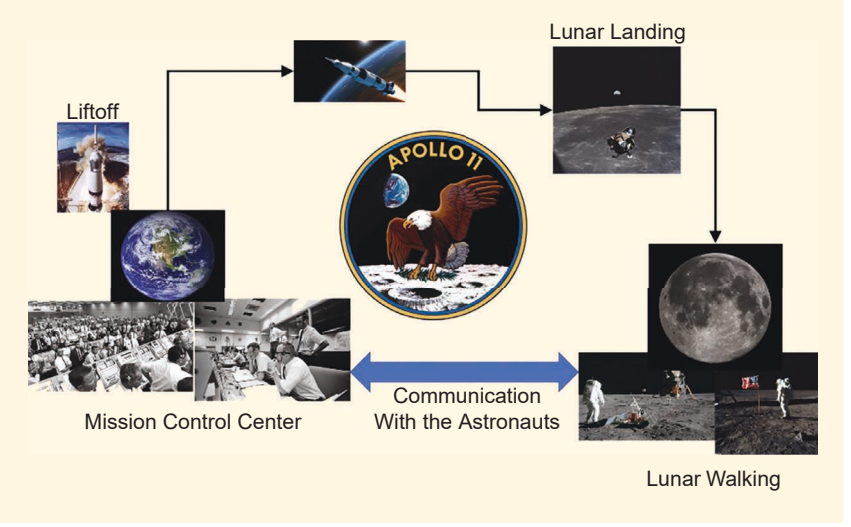
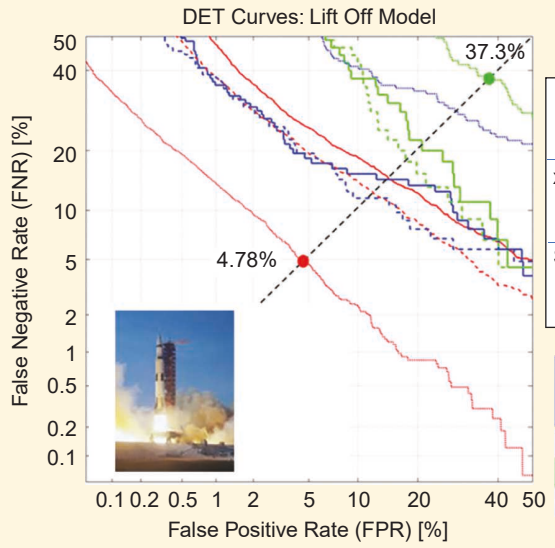


FIGURE 5. Speaker recognition from Earth to the Moon.

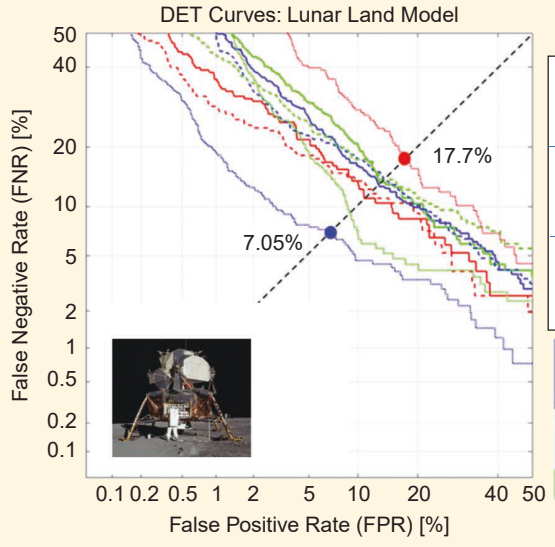


Speaker ID Performance: Lift Off

i-Vector:	Lift Off:	14.634%
	Lunar Land:	14.563%
	Lunar Walk:	20.836%
x-Vector:	Lift Off:	12.248%
	Lunar Land:	11.651%
	Lunar Walk:	17.778%
SincNet:	Lift Off:	4.789%
	Lunar Land:	29.008%
	Lunar Walk:	37.363%

System	Test Model EER (%)		
	Liftoff	Lunar Landing	Lunar Walking
Ivec + PLDA	14.63%	14.56%	20.83%
Xvec + PLDA	12.24%	11.65%	17.77%
SincNet	4.78%	29%	37.36%

(a)

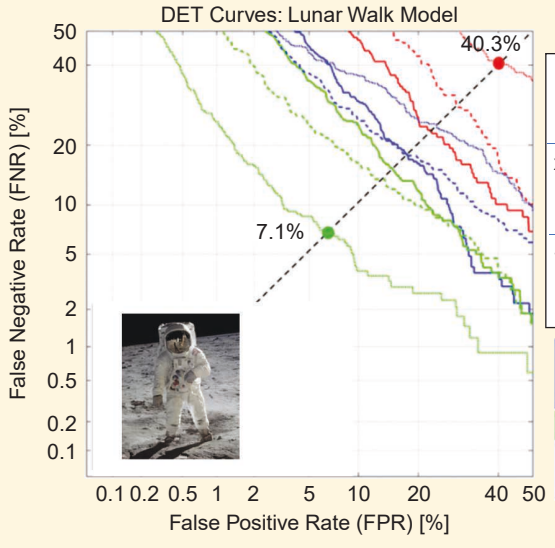


Speaker ID Performance: Lunar Landing

i-Vector:	Lift Off:	11.111%
	Lunar Land:	13.005%
	Lunar Walk:	13.546%
x-Vector:	Lift Off:	11.954%
	Lunar Land:	13.904%
	Lunar Walk:	13.944%
SincNet:	Lift Off:	17.722%
	Lunar Land:	7.052%
	Lunar Walk:	9.126%

System	Test Model EER (%)		
	Liftoff	Lunar Landing	Lunar Walking
Ivec + PLDA	11.11%	13%	13.54%
Xvec + PLDA	11.95%	13.9%	13.94%
SincNet	17.72%	7.05%	9.12%

(b)



Speaker ID Performance: Lunar Walk

i-Vector:	Lift Off:	22.847%
	Lunar Land:	17.910%
	Lunar Walk:	15.313%
x-Vector:	Lift Off:	29.040%
	Lunar Land:	17.940%
	Lunar Walk:	13.750%
SincNet:	Lift Off:	40.316%
	Lunar Land:	23.973%
	Lunar Walk:	7.101%

System	Test Model EER (%)		
	Liftoff	Lunar Landing	Lunar Walking
Ivec + PLDA	22.84%	17.91%	15.31%
Xvec + PLDA	29.04%	17.93%	13.75%
SincNet	40.31%	23.97%	7.1%

(c)

FIGURE 6. Detection error tradeoff (DET) curves for speaker models: (a) liftoff model, (b) lunar land model, and (c) lunar walk model. EER: equal error rate; PLDA: probabilistic linear discriminant analysis.

speakers are present in the same NASA control rooms on Earth with three astronauts either on Earth, in deep space, or on the Moon. It is therefore essential to analyze robust speaker recognition methods that address mismatch between such environments [2]. An additional challenge for these naturalistic data is the variation of speaker duration according to the mission stage, and each phase of the mission has different speakers. We note that the speakers present in one phase of the mission may not necessarily be present in other phases of the mission (e.g., during Apollo 11 launch at Kennedy Space Center, FL, USA, versus mission operations at Johnson Space Center, TX, USA, 10 min after launch until mission completion). For example, in this 100-h audio dataset, Buzz Aldrin's audio is present only in the lunar landing phase of the mission. Hence, our efforts here will be an open set speaker recognition (Figure 6). To access Apollo audio, visit app.explore.apollo.org [28].

In this evaluation, phase 1 of the Apollo 11 mission is liftoff, consisting of 52 speakers with a total audio duration of 25 h (20 h for training and 5 h for test) from five primary channels (see Figure 2). Figure 6 shows detection error tradeoff curves for the three systems. Speaker models were trained on the liftoff phase of the mission and tested on all three phases of the Apollo 11 mission. The best average equal error rate (EER) was obtained by the x-vector + PLDA system. SincNet performs visibly well in the liftoff phase; however, it has poor performance in other phases. Factors affecting poor performance in this phase could be because other phases do not contain speakers that were present in the liftoff phase. The lunar landing stage contains 50 h (40 h for training and 10 h for test) of audio with a total of 92 speakers from the five primary channels. A similar analysis based on speaker models trained on the lunar landing phase and tested on all other phases shows that the best average EER was obtained by the SincNet system. This phase has twice the amount of data (~50 h) compared with the other two phases (~25 h). The lunar walk stage has 25 h (20 h for training and 5 h for test) of audio with 37 speakers. The best average EER was obtained by the i-vector system. SincNet's performance was heavily degraded because of a lack of adequate data, although the x-vector system's performance is similar to the i-vector system performance [23].

Conclusions

Establishing and assessing speech technology such as speaker recognition over a massive naturalistic corpus with highly variable background and noise conditions represents a challenging goal but is expected to not only help advance robust speaker models for future deep space missions but also allow for exploring engagement analysis for multiparty speaker situations. In this study, we have analyzed the performance of alternative speaker recognition systems to understand the impact of mission task stress, multispeaker common communication channel loops, time-sensitive assessment, and mission decision speaker content. To demonstrate the challenges of the Apollo corpus, we 1) illustrate the rarity of communication turns by plotting speech activity for three mission critical phases across five channels, 2) analyze speaker duration of SOI versus non-SOI and silence, 3) compare various state-of-the-art speaker recognition technologies for this corpus, and 4) train on a specific phase of a 6- to 8-day Apollo mission and test on all phases of mission.

We observe that there is significant silence (~80 h of silence out of a total core 100 h of the FSteps challenge corpus) compared with speech. Further analysis on identifying and tracking instances of our SOI versus non-SOI reveals an intriguing global perspective of speaker interaction between astronauts and NASA mission specialists. Finally, we note that when provided with a sufficient amount of data, SincNet was shown to perform the best in terms of accuracy and F score. The complete Apollo mission program (Apollo 1 through Apollo 17) audio data exceed 150,000+ h, where Apollo 11 and Apollo 13 were recovered through the efforts of CRSS-UTDallas. Therefore, it is not possible to manually annotate the currently available amount of 19,000 h of audio (Apollo 11 and Apollo 13), and hence, this analysis was used to establish best practices for corpus development for improved speaker recognition. Finally, the concept of Where's Waldo? provides an opportunity for the research community to transition the FSteps collection as an educational resource, advancing speech technology, preserving the "words spoken in space," as well as serving as a lasting tribute to the heroes behind the heroes of Apollo.

Acknowledgment

This project was funded by NSF-CISE Award 2016725, and partially by the University of Texas at Dallas Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

Authors

Meena M. Chandra Shekar (meena.chandrashekar@utdallas.edu) received her B.E. degree in electronics and communication from Visvesvaraya Technological University, India, in 2016 and her M.S. degree in signals and systems from the University of Texas at Dallas in 2020. She has been a Ph.D. student in electrical engineering at the University of Texas at Dallas since fall of 2018; she is now working at the Center for Robust Speech Systems, Erik Jonsson School of Engineering and Computer Science, the University of Texas at Dallas, Richardson, TX 75080 USA, under the supervision of Prof. John Hansen. Her research interests include speaker identification, speaker verification, speaker tracking, and speaker diarization.

John H.L. Hansen (john.hansen@utdallas.edu) received his Ph.D. and M.S. degrees from Georgia Institute of Technology and B.S.E.E. degree from Rutgers University. He joined the University of Texas at Dallas, Richardson, TX 75080 USA, in 2005, where he currently serves as Jonsson School associate dean for research, professor of electrical engineering. He established CRSS, which is focused on interdisciplinary research in speech processing, hearing sciences, and language technologies. He has supervised 99 Ph.D./M.S. thesis candidates, and has authored 865

journal/conference papers in the field. He is a recipient of the 2021 IEEE SPS Meritorious Service Award, an ISCA Fellow, and past president of ISCA. He has also served as an organizer/chair of IEEE ICASSP and ISCA INTERSPEECH conferences. He is a Fellow of IEEE.

References

[1] J. H. L. 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*, Sep. 2018, pp. 2758–2762, doi: 10.21437/Interspeech.2018-1942.

[2] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015, doi: 10.1109/MSP.2015.2462851.

[3] A. Sangwan et al., "Houston, we have a solution': Using NASA Apollo program to advance speech and language processing technology," in *Proc. Interspeech*, 2013, pp. 1135–1139.

[4] J. H. L. Hansen et al., "The 2019 inaugural fearless steps challenge: A giant leap for naturalistic audio," in *Proc. Interspeech*, Sep. 2019, pp. 1851–1855, doi: 10.21437/Interspeech.2019-2301.

[5] J. Hansen, "*Explore Apollo*," Audio Archive, Texas Country Reporter. [Online Video]. Available: https://www.youtube.com/watch?v=CTJtRNMac0E&ab_channel=TexasCountryReporter

[6] J. M. Kera, "Listen to unheard audio from NASA's Apollo missions, dusted off by UT Dallas researchers," *Houston Public Media*, Dec. 20, 2017. Accessed: Oct. 22, 2020. [Online]. Available: https://www.houstonpublicmedia.org/articles/news/2017/12/20/257844/listen-to-unheard-audio-from-nasas-apollo-missions-dusted-off-by-ut-dallas-researchers/

[7] C. Meyers, "Hear the backstage story of the Apollo program with newly released audio," *Discover Mag.*, 2018. Accessed: Oct. 22, 2020. [Online]. Available: https://www.discovermagazine.com/the-sciences/hear-the-backstage-story-of-the-apollo-program-with-newly-released-audio

[8] D. Kamp, "The found footage that provides a whole new look at the Apollo 11 moon landing," *Vanity Fair*, 2018. Accessed: Oct. 22, 2020. [Online]. Available: https://www.vanityfair.com/hollywood/2018/12/apollo-11-50th-year-anniversary

[9] E. Ruby, "UT-Dallas celebrates personal ties with NASA 50 years after 'one small step for man,'" *Dallas News*, Jul. 20, 2019. Accessed: Oct. 22, 2020. [Online]. Available: https://www.dallasnews.com/news/education/2019/07/20/ut-dallas-celebrates-personal-ties-with-nasa-50-years-after-one-small-step-for-man/

[10] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, Mar. 2001, doi: 10.1109/89.905995.

[11] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 4, pp. 307–313, Jul. 1996, doi: 10.1109/89.506935.

[12] D. A. Cairns and J. H. L. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *J. Acoust. Soc. Amer.*, vol. 96, no. 6, pp. 3392–3400, Dec. 1994, doi: 10.1121/1.410601.

[13] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Commun.*, vol. 20, nos. 1–2, pp. 151–173, Nov. 1996, doi: 10.1016/S0167-6393(96)00050-7.

[14] J. H. L. Hansen and M. A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 407–415, Sep. 1995, doi: 10.1109/89.466655.

[15] C. Yu and J. H. L. Hansen, "A study of voice production characteristics of astronuat speech during Apollo 11 for speaker modeling in space," *J. Acoust. Soc. Amer.*, vol. 141, no. 3, pp. 1605–1614, Mar. 2017, doi: 10.1121/1.4976048.

[16] D. W. Oard, A. Sangwan, and J. H. L. Hansen, "Reconstruction of Apollo mission control center activity," in *Proc. SIGIR Workshop Explor., Navig. Retrieval Inf. Cultural Heritage*, Jan. 2013, p. 5.

[17] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011, doi: 10.1109/TASL.2010.2064307.

[18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

[19] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Dec. 2018, pp. 1021–1028, doi: 10.1109/SLT.2018.8639585.

[20] S. O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2018 NIST speaker recognition evaluation," in *Proc. Interspeech*, Sep. 2019, pp. 1483–1487, doi: 10.21437/Interspeech.2019-1351.

[21] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1711–1723, Jul. 2007, doi: 10.1109/TASL.2007.899278.

[22] K. S. Rao and S. Sarkar, "Speaker verification in noisy environments using Gaussian mixture models," in *Robust Speaker Recognition in Noisy Environments*. Cham, Switzerland: Springer International Publishing, 2014, pp. 29–47.

[23] M. C. Shekar, "Knowledge based speaker analysis using a massive naturalistic corpus: Fearless steps Apollo-11," ProQuest dissertation, MS-EE Thesis, Univ. Texas, Dallas, TX, USA.

[24] "Apollo program." Wikipedia. Accessed: Jun. 19, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Apollo_program

[25] "U.S. Hurricane Katrina." Wikipedia. Accessed: Jun. 19, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Hurricane_Katrina

[26] "U.S. September 11 terrorist attacks." Wikipedia. Accessed: Jun. 19, 2020. [Online]. Available: https://en.wikipedia.org/wiki/September_11_attacks

[27] "Japan Fukushima Daiichi nuclear reactor meltdown." Wikipedia. Accessed: Jun. 19, 2020. [Online]. Available: https://en.wikipedia.org/wiki/Fukushima_Daiichi_nuclear_disaster

[28] "Explore Apollo.org" Accessed: Jan. 27, 2023. [Online]. Available: https://app.exploreapollo.org/

SP