INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Multi-Channel Apollo Mission Speech Transcripts Calibration

Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen

Center for Robust Speech Systems (CRSS), Eric Jonsson School of Engineering,
The University of Texas at Dallas (UTD), Richardson, Texas, U.S.A.
{lakshmish.kaushik, abhijeet.sangwan, john.hansen}@utdallas.edu

Abstract

NASA's Apollo program is a great achievement of mankind in the 20th century. Previously we had introduced UTD-CRSS Apollo data digitization initiative where we proposed to digitize Apollo mission speech data (~100,000 hours) and develop Spoken Language Technology based algorithms to analyze and understand various aspects of conversational speech[1]. A new 30 track analog audio decoder is designed to decode 30 track Apollo analog tapes and is mounted on to the NASA Soundscriber analog audio decoder (in place of single channel decoder). Using the new decoder all 30 channels of data can be decoded simultaneously thereby reducing the digitization time significantly. We have digitized 19,000 hours of data from Apollo missions (including entire Apollo-11, most of Apollo-13, Apollo-1, and Gemini-8 missions). Each audio track corresponds to a specific personnel/position in NASA mission control room or astronauts in space. Since many of the planned Apollo related spoken language technology approaches need transcripts we have developed an Apollo mission specific custom Deep Neural Networks (DNN) based Automatic Speech Recognition (ASR) system. Apollo specific language models are developed. 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 aging, etc,. In this paper we propose a novel method to improve the transcript quality by using Signal-to-Noise ratio of channels and N-Gram sentence similarity metrics across data channels. The proposed method shows significant improvement in transcript quality of noisy channels. The Word Error Rate (WER) analysis of transcripts across channels shows significant reduction.

Index Terms: Speech recognition, Apollo mission, Multi-Track speech, Transcript enhancement

1. Introduction

NASA's Apollo program is a great achievement of the 20th century. During 9 lunar missions that were launched, 12 astronauts walked on the surface of the moon. Countless scientists, engineers, technicians worked behind the scenes to ensure that both the solutions and the execution of the detailed flight plan was a success ($> 400K$). Furthermore, all mission operations were meticulously recorded, documented, annotated and discussed (e.g.interviews, oral histories, press conferences etc.). For example, more than 100,000 hours of audio was recorded pertaining to conversation of Astronauts in space and Mission Control center with (20 positions in control room; each focusing on different aspects of mission) and hundreds of scientists assisting the mission from the backrooms. More detailed description is given in [1]. Till date a very minute fraction of the data is digitized, annotated and available in public domain. All the audio data is stored in more than seven hundred, 30 track analog tapes. Each tape can record 14 hours of data. The speech data flow from all the sources on ground (Mission Control Room, Back rooms) and space (Lunar Module) was recorded in real-time over the entire mission. Analog tapes degrade over time. Hence digitizing the entire data would preserve the entire operational aspects of Apollo project work-flow forever, but it will be a great resource to develop spoken language technology based systems like [12, 13, 14, 15].

The data in analog tapes can be decoded (analog-to-digital conversion) using a machine called Soundscriber shown in Figure 1. Until recently NASA's audio department had only one Sound Scriber machine with single read head. Only one track could be decoded at a time. In order to decode and digitize the entire apollo mission it would around 11 years to decode all the data; round the clock. This is not a practical solution. In this paper, we present a newly designed 30 track head (now mounted on the NASA Soundscriber) which can decode all 30 tracks simultaneously reducing the decoding time enormously as explained in Section 2.

Once data is digitized, next step is to transcribe the data. NASA export control has to review and approve that there is no sensitive data for the public release of audio data. They review the transcribed text and approve/disapprove the release of a data portions. Better the transcripts the release process gets expedited. We have developed a very robust Apollo Specific Automatic Speech Recognition (ASR) system. But the transcripts generated are still very noisy because most of the audio channels suffer from various noisy conditions and degradations (one or more). This results in a high Word Error Rate (WER). Our ASR system is DNN based and performing acoustic model adaptation is a challenge and with very limited availability of transcribed data makes it difficult to pursue acoustic domain methods. In Apollo data, the data being recorded from the personnel occur across tracks/channels which is explained in Section 2.3. Hence we have developed a system that exploits the SNR and N-Gram sentence similarity metrics of transcripts across channels to improve the transcript quality and is presented in Section 4. The results and analysis is discussed in Section 5

2. Apollo Data Corpus Development

2.1. 30 Track Apollo Analog Tape

Each analog tape has 30 channels/tracks onto which the speech data was written on 29 channels/tracks. Mission Elapsed Time (MET) is written on one the first track in IRIG format [2, 3]. During Apollo mission there were two analog recorders namely Historical Recorder 1 (HR1) and Historical Recorder 2 (HR2) which are not functional presently. First a blank tape would be mounted on HR1 and data was recorded on all 30 channels.

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-1778

Before the 14 hour tape on HR1 runs out of recording tape, a new tape was mounted on HR2 around 1/2 hour before the HR1 runs out of recoding tape. By alternatively changing the tapes, continuity was ensured without any loss of data.

2.2. UTD-CRSS 30 Track Analog Decoder

The data was recorded over 50 years ago and presently there is only one operational Soundscriber device shown in Figure 1 located at NASA, Houston. U.S.A., that can decode (cannot encode) the analog tapes.

Figure 1: *Sound Scriber device to decode the 30 track analog tapes.*

Moreover it decodes only one track at a time as it has a single head decoder. The analog decoding rate being real-time it would take more than 100K hours to digitize the data. This is not a practical solution if the plan is to digitize huge quantities of data ($>$100K hours). Hence, we have developed a 30 track head that can decode all the 30 tracks simultaneously. Figure 2 shows the newly designed 30 track decoder that was mounted on to the Sound Scriber. The decoded data from this newly designed 30 track head decoder is passed on to 30 channel analog to digital converter mixers. The real-time 30 tracks of output data from the mixer is stored in wav format.

Figure 2: *UTD-CRSS designed 30 track analog tape decoder.*

We have successfully digitized *19,000 hours* of speech data from various Apollo missions (including entire Apollo-11, most of Apollo-13, Apollo-1, and Gemini-8 missions) in 3 months. This includes head installation time and testing of the new decoder mounts. There are more planned session to digitize the remaining data. This enormous speech corpus of 19,000 hours is used to develop various experiments in the field of Spoken Language Technology (SLT). This enormous audio archive is the first of its kind, of organized missions based speech data.

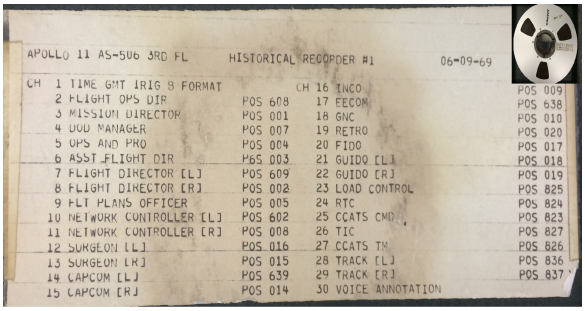

Figure 3: *Various designated channel information in 30 track Apollo mission analog tape*

2.3. Noisy Transcript Problem Definition

Given any Apollo Analog tape there are 30 analog channels and each has a designated purpose. Figure 3 gives an example track sheet showing the mapping of each channel to a particular personnel, position, backroom loop. For example, channel 7 is assigned to Flight Director audio only (Leads the flight control team and has overall operational responsibility), channel 15 is assigned to CAPCOM (capsule communicator) that links the lunar module to earth MCC. Some channels are dedicated to one personnel and some channels has audio from multiple sources. For example on Flight director loop audio from lunar module and MCC personnel is recorded. In GUIDO channel (Guidance officer) there is audio from Flight director, MCC, backroom assisting scientists and sometime from other channels. Audio loops is an audio track containing data from sub groups of astronauts and main scientists in Machine Control Center and backrooms (group of scientists assisting the MCC personnel) that communicate to solve specific issues that may arise by creating personalized groups. The participant's speech data appears in the loops and also in their designated channels/tracks if present. There can be different sets of channels (across channels) with same subset of data in a given time frame; (for eg, channels 5, 7, 13 might have same data over short/long durations. Similarly, channels 10, 21, 24 might have same data over short/long durations.). This can be viewed as two groups of channels occurring in same time frame which share common data across channels occurring only in their groups. The main purpose of the above explanation is to convey that, during the conversation to accomplish a task, the audio information from one channel would get recorded on multiple channels and they are all synchronized in time (wrt to MET). There are 2-3 department of defense channels (DOD) that are not processed for any experiment.

Most of the audio channels suffer from different issues (one or many) like high channel noise, system noise, attenuated signal bandwidth, transmission noise, cosmic noise, analog tape static noise, noise from tape aging, etc,. Some channels are more noisier than others.

One of the basic operation performed on the digitized audio data is transcript generation using ASR system. We have developed an Apollo mission focused custom ASR system as explained in Section 3. But, transcript quality is not consistent (Word Error Rate (WER) varies across channels).

There is audio data overlap from one channel to other. Due to different channel conditions the transcript quality is different.

A new methodology that analyzes the noise levels across channels and employs text similarity to improve the overall transcript quality is proposed in the Sections 4.

3. Speech Recognition System

Apollo mission focused custom automatic speech recognition system (ASR) is developed using KALDI speech recognition system [4]. The acoustic models are trained using a mix-style approach, where acoustic data from multiple corpora are used. We have used Conversational Telephone Corpora consisting of Switchboard [5] and Fisher corpora [6] to train our acoustic models. The above corpora constitutes around 2000 hours of speech data. Part of the training data is corrupted with additive noise at various SNRs (signal to noise ratio) to create a mixture of clean and noisy conditions (to mimic various kinds of noises like office, fan, cafeteria, car road). The acoustic models use Mel frequency cepstral coefficients (MFCCs) along with deep neural networks (DNNs).

For the purpose of developing robust and comprehensive language models we collected text data from various sources relating to Apollo Missions, Gemini, and International Space Station (ISS). It contains materials from various sources like, Books, articles, mission-reports [8, 9, 10], research-reports. We collected over 100 books, 300 technical reports, all NASA mission reports from Apollo, Gemini, ISS and various other sources. We also have collected interview transcripts of many astronauts and mission directors. Also to model day-to-day conversation databases like, (i) Switchboard, (ii) Fisher (iii) UW191 [7] are also used to develop language model. Using this huge textual database collection, we have developed a set of efficient and comprehensive language models. Individual language models are developed from these different sources and are fused to develop a very comprehensive language model. NASA uses lot of scientific terms and abbreviations. All scientific terms were collated from the above said text sources, new pronunciation dictionary was generated for words and abbreviations that are not in the standard dictionary. After careful review these words were augmented to the standard pronunciation dictionary. The final ASR vocabulary size is about 120K words.

The ASR uses a 2-pass decoding strategy, where an fMLLR transform is estimated in the first pass and used to normalize the MFCCs. In the second pass, the normalized MFCCs are used with DNN models to obtain the output recognition lattice. Subsequently, one-best transcripts are obtained from the lattice. A new method of Speech Activity Detection have been proposed to solve Apollo data specific issues [11].

4. SNR and N-gram text similarity based Transcript Calibration

From Sections 2.3 it is known that same data would be present across different channels at the same time but the channel quality is different. Some are more degraded than others. Having the knowledge that data chunks occur across channels, we propose method to use the transcript of cleaner/less noisier channels to improve the transcript quality of noisier tracks given the constraint that we identify effectively where does similar data occur across channels. To achieve this we propose a method to improve the dirty transcript as shown in Figure 4.

Long 30 track audio data is broken into chunks of 10 minutes each. There will be 30 chunks of audio from 30 channels for every 10 minutes. For the purpose of explanation consider

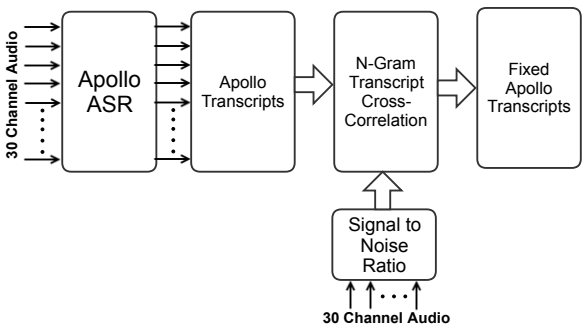

Figure 4: *Steps to improve the transcript quality across channels. Transcripts in noisier channels are fixed using SNRs and N-Gram text similarity metric across channels.*

data streams from a 10 minute time-frame. IRIG channel is neglected as it is time information. All the DOD channels are not processed. As first step, all remaining 26 to 27 audio chunks are passed through the Apollo ASR system one by one, to generate transcripts of all tracks. Once the transcripts are generated some are better than the others. Next, Signal to Noise Ration (SNR) of all the the audio chunks are calculated from the above discussed allowable channels and it is ordered from least noisy to most noisy. We assume that the transcript generated from cleaner track is better than that of the noisier track.

We assume that the ASR transcripts generated from cleaner tracks are better than the noisy tracks. There are 27 tracks of audio data and we do not have prior information regarding which tracks have the replicated data. It may appear on any tracks. Only information that we have is that all the tracks are synced with respect to time. So, in order to fix the issue we propose N-Gram text similarity based method. Each 10 minute transcripts of all 27 channels and is broken into 10 second transcript chunks. Here, N-Gram refers to the 10 second transcript chunk. The length of this chunk is not constant.

In the next step we perform Iterative Transcript Correction (ITC) as explained in Algorithm 1. Consider first 10 second chunks of 27 channels ordered as per the SNR. From Algorithm 1 procedure, iterate over the present master track (L_n), starting from track 2, calculate the text similarity distance of the 10 second transcript chunk of track L_n and track M. Here track M would be one of the subsequent tracks of L_n. text similarity distance is calculated using Levenshtein distance. If the similarity distance between L_n and M is grater or equal to 50%, than replace the transcript of lower SNR track M with the transcript of higher SNR track L_n. Once transcript of a given channel is changed it wont be changed again in succeeding iterations and thereby having the best version of transcript across all channels. In order to ensure all subsets/subgroups of data (explained in Section 2.3) is corrected, the replacement procedure in Algorithm 1 is performed across all channels and comparing it with the subsequent channels and replacing the transcript if the similarity measure is higher than the threshold. This ensures that all the channels (including collection of a sub set of channels) are taken into consideration resulting in better transcripts across all the channels. This proves to be the simple solution to generate high quality transcripts given our unique data situation. By fixing data in such large scale has helped CRSS experiments and NASA to review the transcripts for public consumption faster.

Algorithm 1 Iterative transcript correction for a 29 tracks of 10 second transcript chunk

Variables
  L = List of tracks ordered with highest SNR to lowest SNR
  L_n = Present master track
  M = Present comparison track
  $flag_M$ = Track M transcript replaced flag = 0
end Variables
procedure ITC($L, M, L_m, flag_M$) ▷ Iterative Transcript Correction
  for ($L_n = 2; L_n < 29; L_n++$) **do** ▷ Iterate over tracks wrt List L
    for $M = L + 1; M < 30; M++$ **do** ▷ Iterate over the subsequent tracks in order of SNR wrt List L
      Calculate the sentence similarity distance of track L_n with track M using Levenshtein method
      if similarity between L_n and $M >= 50\%$ **then**
        if $flag_M = 0$ **then**
          Replace transcript chunk of track M with transcript chunk of track L_n
          $flag_M = 1$
        end if
      end if
    end for
  end for
end procedure

5. Results and Analysis

Out of 9,000 hours of automatically transcribed data, a subset of data consisting of 10 minutes across 27 channel data (resulting in 270 minutes of audio data) is considered. It is chosen where there is significant audio activity across many channels. This data is hand transcribed by human transcribers by listening to the audio. This transcribed text is used for the purpose of accuracy evaluation. There is higher speech activity across many channels during major events in a mission, for e.g., take-off, lunar landing, e.t.c. In this data we have considered data during Lunar module landing on the moon. Where they undertake the operation fo ejecting the lunar module and positioning the lunar module in right orbit to land exactly at the right landing site *Mare Tranquilitatis*.

Out of the 27 channels considered there are two sets of channels that has subset data overlap. First set consists of channels 3, 17, 7, 14 ordered by SNR from highest to lowest. Second set has channels 19, 21, 15 ordered by SNR. Using the Iterative Transcript Calibration (ITC) method explained in 4 the 10 minute data is broken in 10 second chunks and the transcript is modified at local level in each iteration across all chunks resulting in modified transcripts over ten minutes of data. Once we have the modified transcripts across all required channels WER is calculated for the channels in the data sets discussed above. Table 1 shows the WER of these data sets and channels before and after applying ITC method comparing with the transcribed text.

From Table 1, it can be observed that the WER is consistently reduced in noisier channels. WER has reduced significantly across all noisier channels. Even though the channel 14 and channel 15 from sets 1 and 2 respectively are less than 10dB SNR, yet the improvement caused in transcript domain is considerably higher. This method helps to calibrate and generate decent quality transcript for NASA review and release of the data in an expedited manner. By improving the transcript using simple text based methods, we have reduced the computational time to generate accurate transcripts.

Table 1: *Word Error Rate Analysis*

Method	**Channel Set-1 [WER % (SNR dB)]**			
	ch 3	ch 17	ch 7	ch 14
Before ITC	24% (20dB)	26% (16dB)	28% (12dB)	31% (9dB
After ITC	**24%**	**22%**	**23%**	**27%**

Method	**Channel Set-2 [WER % (SNR dB)]**		
	ch 19	ch 21	ch 15
Before ITC	21% (16dB)	27% (14dB)	30% (11dB)
After ITC	**21%**	**22%**	**24%**

Significant effort has been employed to develop comprehensive Apollo mission specific language model and pronunciation dictionary without which the WER would be significantly higher. This has contributed for better transcripts and helped to apply text based calibration methods to improve dirty transcripts.

6. Conclusions

In this paper we have presented a new method adopted to decode 30 track Apollo analog tape data which has accelerated the digitizing time significantly (From around 11 years round the clock to a few months at most). This will not only help in restoring the data from aging tapes but also provide significant data source for the SLT community to study such large mission group data. A new method to improve the transcript quality using the SNR noise levels and N-Gram text similarity based methods. The resulting transcripts are of significantly higher quality across all channels. The method employed is simple yet powerful. The next steps is to develop spoken language technology based solution to model and analyze the interaction of mission personnel and develop methods to model group behavior in hierarchical organizational environments.

7. 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.

8. References

[1] A. Sangwan, et al. "Houston, we have a solution: using NASA apollo program to advance speech and language processing tech-

nology." INTERSPEECH. 2013.

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

[3] J. B. Houston "Converting Time Signals From BCD to IRIG-B." (1982).

[4] 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.

[5] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development.", In ICASSP, pp. 517520, 1992.

[6] C. C. David, D. Miller, and K. Walker. The fisher corpus: a resource for the next generations of speech-to-text. In in Proceedings 4th International Conference on Language Resources and Evaluation, pages 6971, 2004.

[7] A. Stolcke, R. Gadde, M. Graciarena, K. Precoda, A. Venkataraman, D. Vergyri, W. Wang, and J. Zheng, Speech to text research at SRI-ICSI-UW, in NIST RT-03 Workshop. May 2003, DARPA

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

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

[10] "Apollo Mission Reports" https://www.hq.nasa.gov/alsj/alsj-mrs.html

[11] A. Ziaei, L. Kaushik, A. Sangwan, J. H. Hansen, & D. W. Oard, (2014). Speech activity detection for NASA apollo space missions: challenges and solutions. In Interspeech (pp. 1544-1548).

[12] 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.

[13] 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

[14] M. Wollmer, et al. "Youtube movie reviews: Sentiment analysis in an audio-visual context." Intelligent Systems, IEEE (2013): pages 46-53.

[15] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, & D. Zhang, (2005). Automatic analysis of multimodal group actions in meetings. IEEE transactions on pattern analysis and machine intelligence, 27(3), 305-317.