Voice enabled domains: Diarization & ASR

Diarization: Who spoke What, When, Where, and How?
- Single Channel+Speaker vs. Multi Channels+Speakers

Examples of Monitoring Communications
- Prof-Life-Log/PLTL: monitoring daily human interactions
- Fearless Steps Apollo-11 Corpus: Earth to the Moon - Multi-Stream Channels & Speakers

Diarization can be MUCH Richer for Knowledge Extraction

Speech Recognition – challenges in spontaneous conversational speech (not “prompted”– like Apple Siri, etc): coarticulation issues

Can also leverage multiple streams simultaneously....

Past multi-stream efforts focused mostly on “the Meeting task”

Limited work on independent audio streams, non face-to-face scenarios
Unscripted ~16hr daily speech collection in natural environments

DIARIZATION advancements: SAD; Environment Sniffing; ASR/KWS

Unrestricted topics, vocabulary, language, environments, etc.

Recent Advancements: Combo-SAD, Word-Count Estimation

Sample Personal Audio Recording Analysis

Speech Activity Detection
- Speech: 88%
- Background: 12%

Environmental Sniffing
- Student Union: 11%
- Office: 44%
- Outdoor: 6%
- Restaurant: 35%
- Car: 3%

Speaker Diarization
- Primary Spkr: 27%
- Background: 12%
- Spkr A: 11%
- Spkr D: 13%
- Spkr B: 20%
- Spkr C: 17%

Primary Speaker
- Speaker D
- Speaker C
- Speaker B
- Speaker A

Background
- Student Union
- Outdoor
- Restaurant
- Car
- Office

Speech Activity Detection

Environmental Sniffing

Speaker Diarization
With Word Counts, possible to analyze differences in conversation engagement between days, people, environments, etc.

Correlation: Similar days: A & I ($\rho = 0.87$)  Diverse days: B & F ($\rho = 0.27$)
‘Houston We Have a Solution’: Novel Speech Processing Advancements for Analysis of Large Asynchronous Multi-Channel Audio Corpora

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

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
School of Brain & Behavioral Sciences (Speech & Hearing)
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

NSF CISE Sponsored (2012-2017)
2.0 Apollo: Audio Interaction Analysis

30 Channel Audio

NASA SoundScriber Audio Capture

Backroom Audio Loops

Seat Position: Operator ID tracking (3 person’s in 8-9hr shifts)
2.1 Apollo: Communication Flow

- Hierarchy of information exchange

30 Channel Audio

NASA SoundScriber Audio Capture

Astronauts

Capsule Communicators

Flight Directors

Control Room and Backrooms

Surgeon (Medical)
Control (Navigation)
EECOM (Electrical)
Booster (Propulsion)

GNC (Navigation)
Network (Telemetry)
FDO (Flight Data)
TELMU (EVA)
Obtained permission to digitize audio: Apollo 11, 13, 9, 1 and Gemini-8 +19,000 hours of mission data available for research and analysis Unique opportunity to learn how backroom teams work & collaborated to accomplish complex technical tasks CRSS-UTDallas developed: (i) digitizing solution for 30-track tapes; (ii) organized multi-channel data/missions; (iii) advanced state-of-the-art diarization pipeline {SAD, ASR w/ NASA specific Lexicon, Language Model – based on +4.5M words}; (iv) processed all 19,000hrs; (v) obtained NASA export control release of data

Soundscriber: Apollo Audio 30 Track Tape Player
(i) Orig. 1-track Head
(ii) UTDallas designed 30 Track Head
2.2 Apollo-11: Audio Analysis

- Corpus: air-to-ground from Apollo-11 mission
- Mission Duration: **8 days, 3 hours 18 minutes, 35 seconds.**
- Apollo-11: separated into 8 stages

The timeline of Apollo 11 mission:

1. **Stage 1:** Launching
2. **Stage 2:** Earth Orbiting
3. **Stage 3:** To The Moon
4. **Stage 4:** Lunar orbiting
5. **Stage 5:** Lunar landing
6. **Stage 6:** Moon Walking
7. **Stage 7:** Lunar launching
8. **Stage 8:** Return To Earth
Deep Neural Networks based Acoustic models developed using SWB + Fisher audio databases

Apollo Language Model Development
- Text data from Missions: Apollo, Gemini, and International Space Station (ISS) collected
- Sources for NASA related text: Books, articles, mission-reports, research-reports.
- Text Corpus: 100 + books, 300 technical reports, NASA mission reports from Apollo, Gemini, ISS plus other sources.
- To model Neutral Speech Text: used the text data from (i) Switchboard, (ii) Fisher (iii) UW191
- NASA technical abbreviations collection

New pronunciation dictionary developed for NASA terms

Language model is application specific
2.4 Sentiment Detection:

Apollo-11 Sentiment over complete mission: Air-to-Ground

- Viewing Earth from space
- Mission Accomplished!
- Viewing Moon from space
- Moon Landing
- Technical Problems
- Making difficult maneuver
- Lift Off!
- Journey towards the Moon
- Moon Landing
- Journey back to the Earth
- Return to Earth
## 2.5 Spoken Language Technology (SLT) Results Summary

<table>
<thead>
<tr>
<th>Technology</th>
<th>Accuracy</th>
<th>Inference (CRSS-UTDallas)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASR Engine</td>
<td>12% (EER)</td>
<td>We have developed a state-of-the-art Speech Recognition system that can recognize speech in naturalistic spontaneous noisy conditions</td>
</tr>
<tr>
<td>Word Count</td>
<td>&gt; 95%</td>
<td>Our method can detect word count accurately. This could be a good measure for the intensity of interaction; also can be language independent.</td>
</tr>
<tr>
<td>Sentiment/Opinion</td>
<td>89%</td>
<td>Our method can accurately detect the opinion expressed by the speaker which gives the overall positivity/negativity of a conversation</td>
</tr>
<tr>
<td>Laughter Count</td>
<td>7% (EER)</td>
<td>We can accurately calculate the count of laughter and fillers in a conversation which could indicate the mood of the conversation (laughter) and confidence of the speaker (fillers)</td>
</tr>
<tr>
<td>Speaker Recognition</td>
<td>2% (EER)</td>
<td>Our state of the art I-Vector + PLD system can recognize and track speakers effectively</td>
</tr>
</tbody>
</table>
2.6 Apollo-11: Who’s Talking to Whom?

**Parameter** | **Value**
--- | ---
Conversation Count | 10
Word Count | 1050
# Turns Taken | 60
Topic of discussion | Lift-off

**Parameter** | **Value**
--- | ---
Sentiment | Positive
Emotion | Positive
Stress Levels | High
FEARLESS STEPS CORPUS

Presented by

ROBUST SPEECH TECHNOLOGIES LAB (CRSS)
THE UNIVERSITY OF TEXAS, AT DALLAS

An Audio Corpus of the complete Apollo-11 Mission

To be released at INTERSPEECH 2018

190,000 hours of massive naturalistic multi-channel data

About the Corpus:

Full Corpus: Mission Control, Air-To-Ground, all Back-Room communications, IRIG time for 30 synchronized channels.

A 100 hour sub-challenge Corpus will be released with 5 time-synchronized challenge tasks established.

Pipeline Diarization Transcripts will be provided in addition to the voice activity and ASR transcripts for the Corpus.

The Challenge Corpus will be open source, freely distributed (with a small fee for complete corpus delivery on hard drive).

The Challenge Tasks:
Automatic Speech Recognition (ASR)
Keyword Spotting and Sentiment Detection
Speaker Identification (SID)
Speaker Diarization
Voice Activity Detection (VAD)
Special Thanks, Next Steps, & Questions

✈ NASA – John Stohl, Greg Wiseman
   (Audio Engineering; Houston, TX)
✈ Doug Oard (Univ. of Maryland); Co-PI on NSF project
✈ U.S. National Science Foundation - sponsorship

Next Steps:

✈ Corpus Release to PUBLIC
   (30-chan; ~19k hrs); NASA Export Control
✈ “FEARLESS STEPS” Corpus – 5 Challenge Tasks: (i) SAD, (ii) SID, (iii) Diarization,
   (iv) ASR, (v) KWS/Sentiment
✈ Document Linking – Diarization: ASR/KWS, SID, speaker state/stress

Questions