Outline

- Introduction
  - Prof-Life-Log corpus
    - Collection paradigm
    - Data collection and annotation
    - Commonly encountered environments
    - Scope and range of experiments

- Acoustic Signature Vector (ASV) system
  - Acoustic Signature Vector (ASV) system structure
  - Acoustic Signature Vector Computation

- Experiments and results
- Summary and future works
Collection paradigm

- Meeting
- Walking
- Office
- ???

Time (Hour) = 0
- Walking
- Office
- Walking
- Lunch
- Outdoors
- Phone
- Conference Call
- Meeting
- Driving

- Unscripted speech collection in natural environments
- Unrestricted topics, vocabulary and language use
- Analysis of daily acoustics and voice communications:
  - An acoustic signature per subject
Data Collection and annotation

- Data collected in daily sessions
  - Data recorded on mobile digital LENA unit
  - Each session can last from 8-16 hours (full work-day)

- 45+ sessions collected so far and corpus is growing

- Rich diversity of acoustic environments
  - 50+ environments annotated so far (e.g. office, restaurant, clapping, wind, car, babble, computer-use etc.)

- Rich diversity in topics and speaking style and material

- A small subset is focused on collecting various commonly encountered environments (pure environment with no speech)

- 7+ hours of data annotated so far (over 20 sessions).
Commonly encountered environments

OUTDOORS

Wind
Footsteps
Honking

Music (in-vehicle)

SPEECH

Clapping

Cafeteria, Auditorium etc.
Laughing
Cheering
Babble

OFFICE

Typing
Paper Rustling

Quiet
Footsteps
Scope and range of experiments

Audio Recordings

Speech
  - Secondary Speaker Identification
  - Keyword Spotting
  - Stress/Emotion Analysis

Background
  - Context Modeling
  - Environment Estimation

- Automatic Speech Recognition (ASR)
- Speaker Diarization
- Speaker Identification
- Environmental Sniffing

- Keyword Spotting
- Sentiment/Opinion Estimation
- Speaker Context Modeling
- Speech Background Separation
Environmental Sniffing

General system architecture

- Audio Stream
  - 1 MICROPHONE
  - MICROPHONE ARRAY
  - MULTISENSOR INPUT

DATA CAPTURE

- s(t)
- s(n)

SPEECH TASK

- ASR
- SPEECH CODING
- SPEAKER ID
- SPEECH ENHANCEMENT
- LANGUAGE ID
- NOISE TRANSCRIPTION
- INFORMATION RETRIEVAL

ACOUSTIC KNOWLEDGE

- \( I(n) = [N(n), C(n), S(n)] \)

ENVIRONMENTAL SNIFFING

- PSD ESTIMATE
- IMPULSIVE
- STATIONARITY
- PERIODICITY
- NARROWBAND/TONE
- BROADBAND

Goal

- Detect, classify and track acoustic conditions, extract acoustic knowledge.
- PASSIVE: Provide the acoustic knowledge.
- ACTIVE: Give smart decisions, direct subsequent speech systems.


ASV system

Acoustic Signature Vector (ASV) system structure

System 1: Query by Example

Input Query

Typically small duration
10s to 1min

Query by example
Audio retrieval

Database of Audio:
Entire Prof-Life-Log Collection
Tones of files, +300 HRs (30 daily records up to now, each ~ 5 to 15 HRs)

System 2: Automatic Clustering of Homogenous Audio

Long Duration
1 day or more

Cluster A
Cluster B
Cluster C
Cluster A

Database of Audio:
Entire Prof-Life-Log Collection
Tones of files, +300 HRs (30 daily records up to now, each ~ 5 to 15 HRs)
Acoustic signature vector (ASV)

Acoustic Signature Vector Computation

MFCC Extraction

\[ j^{th} \text{ feature vector } = X_i \]

N frames of observation

Gaussian Mixture Model (GMM)

\[ j^{th} \text{ Mixture } = M_j \]

M mixtures modeling the acoustic background

Likelihood Matrix

\[
\begin{bmatrix}
L_{11} & L_{12} & \ldots & L_{1N} \\
L_{21} & \cdot & \cdot & \cdot \\
\vdots & \ddots & \ddots & \vdots \\
L_{M1} & & & L_{MN}
\end{bmatrix}
\]

Acoustic Signature Vector (ASV) is obtained by summing the likelihoods for each mixture

\[
\begin{bmatrix}
\sum L_{1k} \\
\sum L_{2k} \\
\sum L_{Mk}
\end{bmatrix}
\]

Cosine Distance to measure similarity

\[ \text{sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \]

A and B are ASVs of 2 audio samples

Sample is converted into an ASV

Email: {ali.ziae, abhijeet.sangwan, John.Hansen}@utdallas.edu
Experiments

- 36 dimensional MFCCs extracted from a known template (or example) recording of the environment (or three recordings) is used to initiate search (i.e., the label assigned to the segment is known).
- Speech part, (i) preserved or, (ii) removed
- All segments that match this template are retrieved
- Measure EER (equal error rate) to estimate performance in comparison to GMM-UBM system
- F-measure to estimate clustering performance for ASV features

Test scenarios:
- 1-Query 10 Secs
- 3-Query 10 Secs

“Pure” = homogenous environment sounds, 1 sound per block
“S-R” = open audio streams with a mixture of sounds, with speech part removed using a VAD (Voice Activity Detection)
Experiments

<table>
<thead>
<tr>
<th>System</th>
<th>EER%</th>
<th>System</th>
<th>EER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Query, Pure</td>
<td>24.93</td>
<td>1-Query, Pure</td>
<td>29.08</td>
</tr>
<tr>
<td>1-Query, S-R</td>
<td>23.09</td>
<td>1-Query, S-R</td>
<td>30.15</td>
</tr>
<tr>
<td>3-Query, Pure</td>
<td>21.76</td>
<td>3-Query, Pure</td>
<td>27.64</td>
</tr>
<tr>
<td>3-Query, S-R</td>
<td>19.06</td>
<td>3-Query, S-R</td>
<td>27.16</td>
</tr>
</tbody>
</table>

Environment Detection performance here shows promise, but the actual EER needs to be in the 5-10% range to be useful for a practical system.
Acoustic Signature Vector example

Restaurant ASV
Walking ASV
Outdoor ASV
R-W ASV
R-O ASV
W-O ASV

R-W = Difference between Restaurant’s ASV and Walking’s ASV
R-O = Difference between Restaurant’s ASV and Outdoor’s ASV
W-O = Difference between Walking’s ASV and Outdoor’s ASV
Experiments

ASV System based clustering (F-Measure)

<table>
<thead>
<tr>
<th>System</th>
<th>F-Measure % Cosine distance</th>
<th>F-Measure % Euclidean distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-Query, Pure</td>
<td>61.78</td>
<td>60.74</td>
</tr>
<tr>
<td>1-Query, S-R</td>
<td>63.37</td>
<td>63.23</td>
</tr>
<tr>
<td>3-Query, Pure</td>
<td>75.09</td>
<td>74.54</td>
</tr>
<tr>
<td>3-Query, S-R</td>
<td>79.82</td>
<td>77.47</td>
</tr>
</tbody>
</table>

Environment Detection clustering using ASV features show speech removal and multi query strategies help to improve classification between environments.
Conclusion

- Prof-Life-Log corpus presented
  - Collection is naturalistic and contains real-world environments
  - Very useful for many speech tasks
  - Easy to transition for infant/child language assessment scenarios

- Environment ID & tracking
- Keyword spotting (KWS)
- Topic ID
- Adult Distribution / Diversity
  (Male/Female %'s; Age %'s, etc)
Conclusion

- Environment Estimation
  - Detecting mixed-environments is challenging.
  - In all situations, longer test duration/removing speech parts, ASV system outperforms GMM-UBM
Direction

Speech Activity Detection

- Background: 12%
- Speech: 88%

Environmental Sniffing

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

Speaker Diarization

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