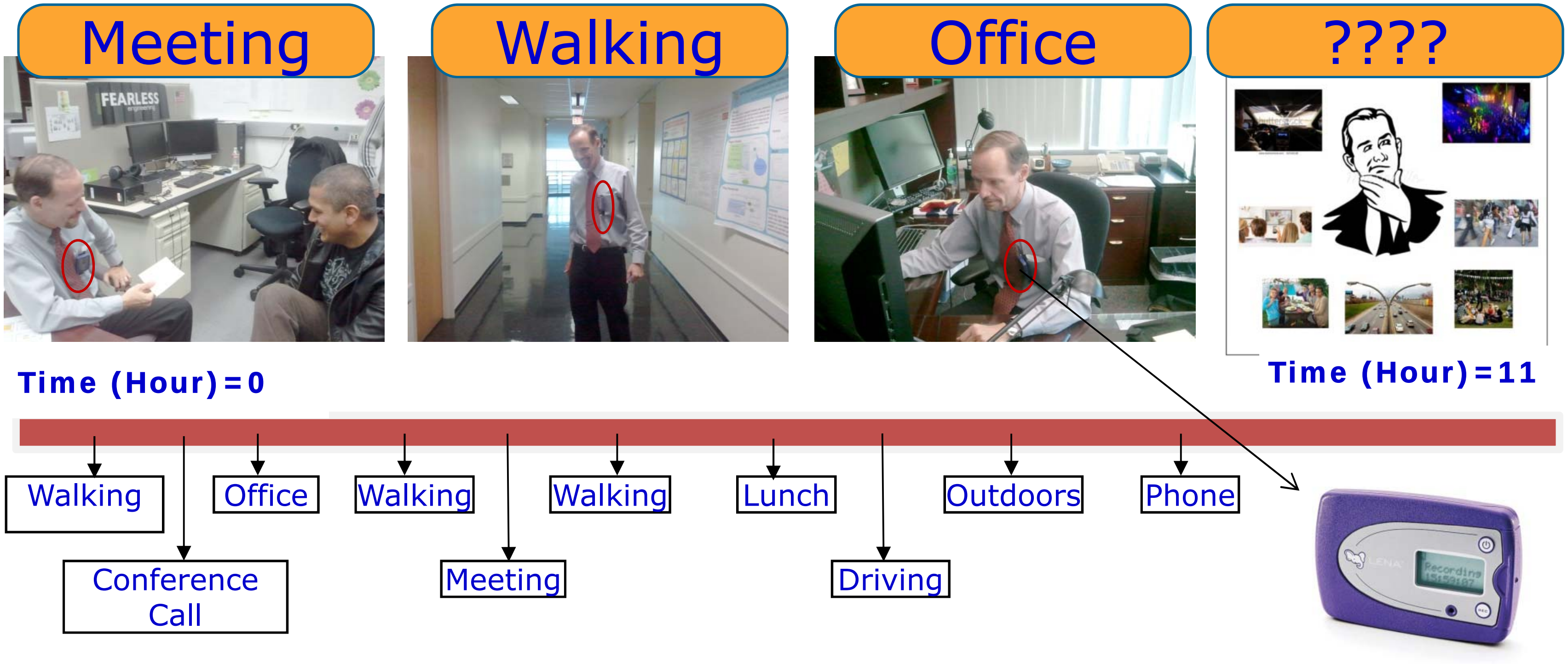
Outline

- ◈ 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

Time (Hour)=11

Walking

Conference Call

Office

Walking

Meeting

Walking

Lunch

Driving

Outdoors

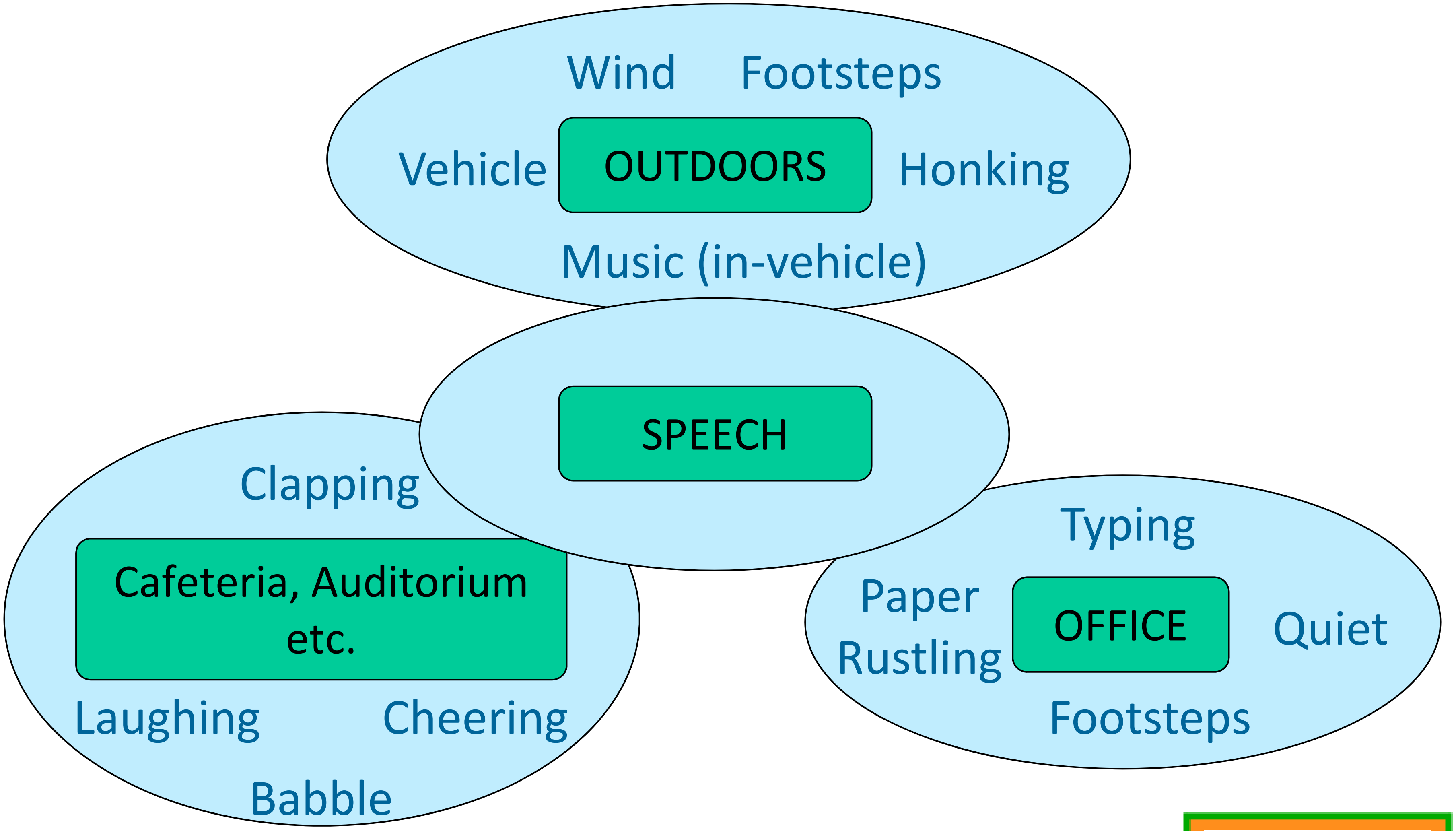
Phone

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

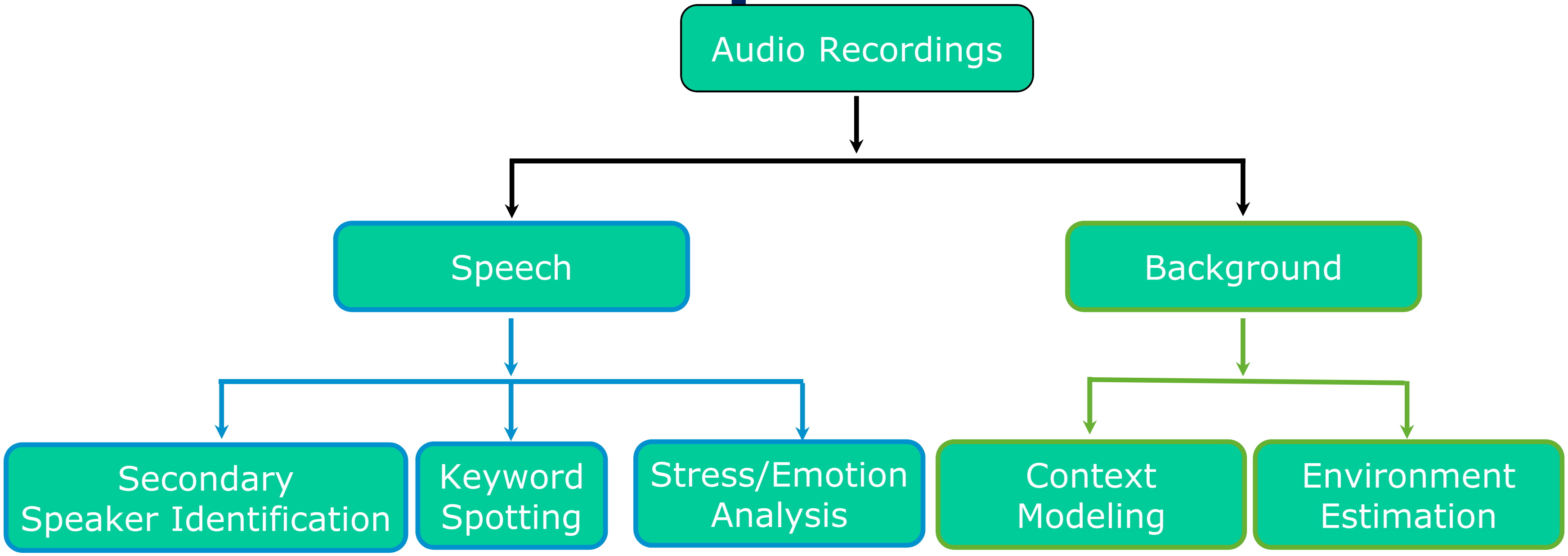
Data Collection and annotation

- 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



Scope and range of experiments



Audio Recordings

Speech

Background

Secondary Speaker Identification

Keyword Spotting

Stress/Emotion Analysis

Context Modeling

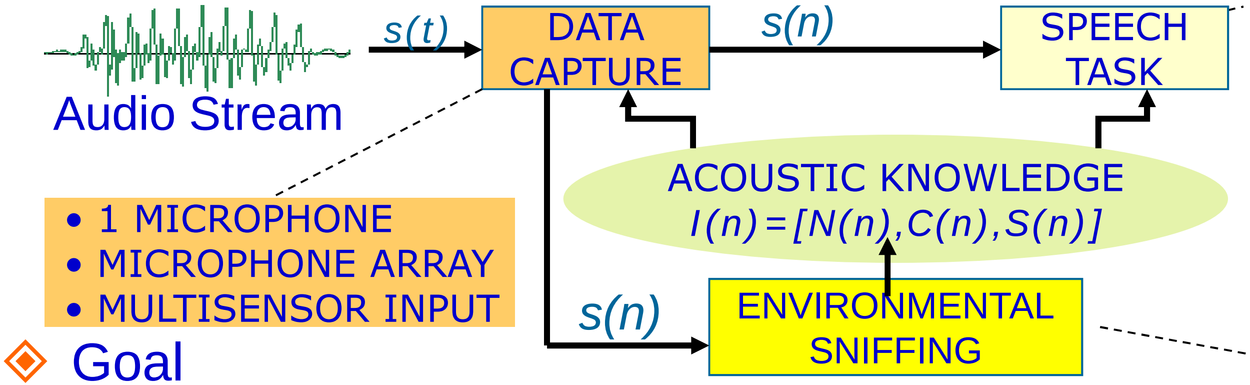
Environment Estimation

- 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



- 1 MICROPHONE
- MICROPHONE ARRAY
- MULTISENSOR INPUT

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

- 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.

[1] M. Akbacak, J.H.L. Hansen, "Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 465-477, Feb. 2007.
[2] M. Akbacak, J.H.L. Hansen, "Advances in Acoustic Noise Sniffing for Robust In-Vehicle Systems," Chapter 10 in *Advances for In-Vehicle and Mobile Systems: An International Perspective*, Springer-Verlag Publishers, 2006.
[3] M. Akbacak, J.H.L. Hansen, "ENVIRONMENTAL SNIFFING: Robust Digit Recognition for an In-Vehicle Environment," INTERSPEECH-2003, pp.2177-2180, Geneva, Switzerland, Sept. 2003.

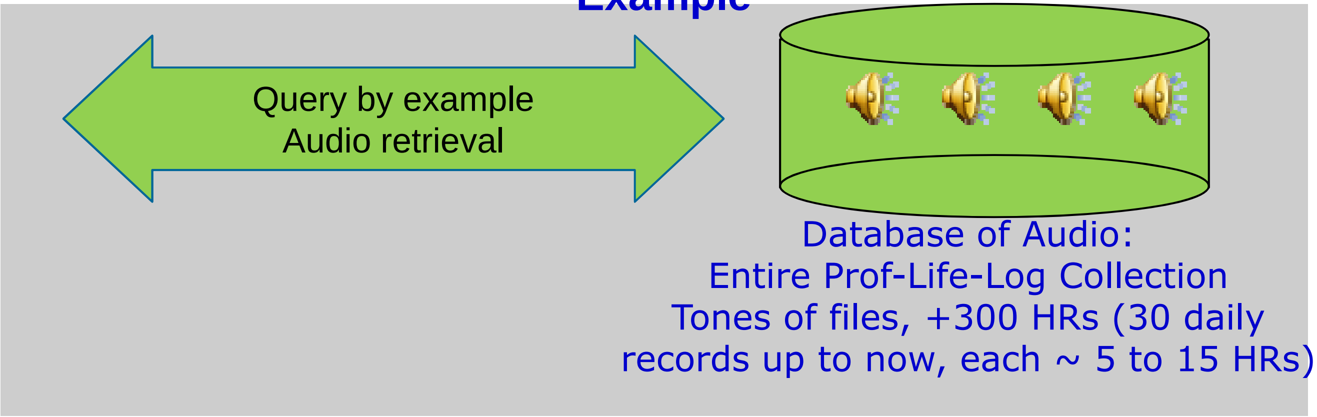
ASV system

- Acoustic Signature Vector (ASV) system structure

Input Query

Typically small duration
10s to 1min

System 1 : Query by Example

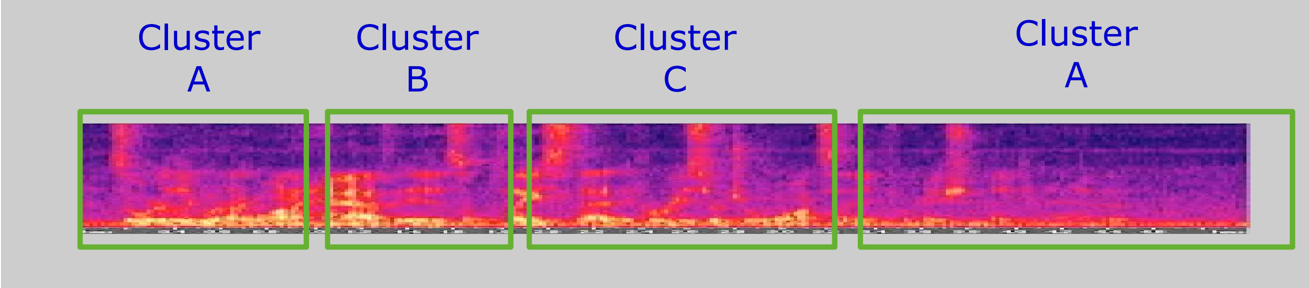


Query by example
Audio retrieval

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

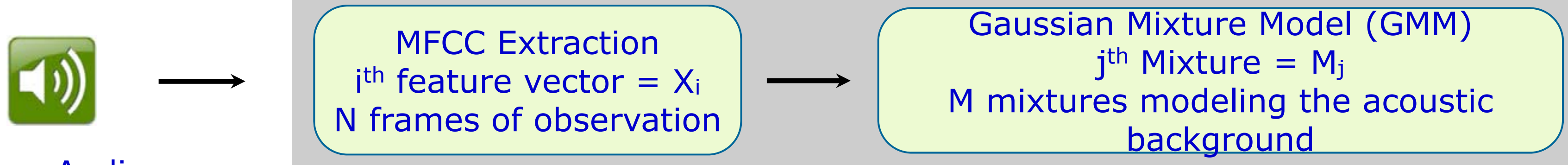
Long Duration
1 day or more

System 2 : Automatic Clustering of Homogenous Audio



Acoustic signature vector (ASV)

Acoustic Signature Vector Computation



Audio Sample

MFCC Extraction
i^{th} feature vector = X_i
N frames of observation

Gaussian Mixture Model (GMM)
j^{th} Mixture = M_j
M mixtures modeling the acoustic background

Sample is converted into an ASV

Likelihood Matrix

$$\begin{bmatrix} L_{11} & L_{12} & \cdots & L_{1N} \\ L_{21} & \ddots & & \\ \vdots & & & \\ L_{M1} & & & L_{MN} \end{bmatrix}$$

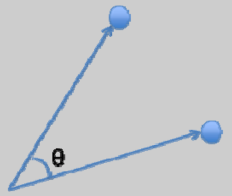
Acoustic Signature Vector (ASV) is obtained by summing the likelihoods for each mixture

$$\begin{bmatrix} \Sigma L_{1k} \\ \Sigma L_{2k} \\ \\ \Sigma L_{Mk} \end{bmatrix}$$

Cosine Distance to measure similarity

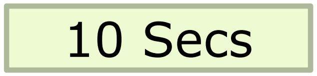
A and B are ASVs of 2 audio samples

$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

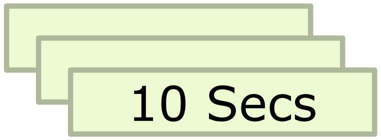


Experiments

- 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



  - 3-Query



  - “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

ASV System (EER%)

System	EER%
1-Query,Pure	24.93
1-Query, S-R	23.09
3-Query,Pure	21.76
3-Query, S-R	**19.06**

GMM_UBM System (EER%)

System	EER%
1-Query,Pure	29.08
1-Query, S-R	30.15
3-Query,Pure	27.64
3-Query, S-R	**27.16**

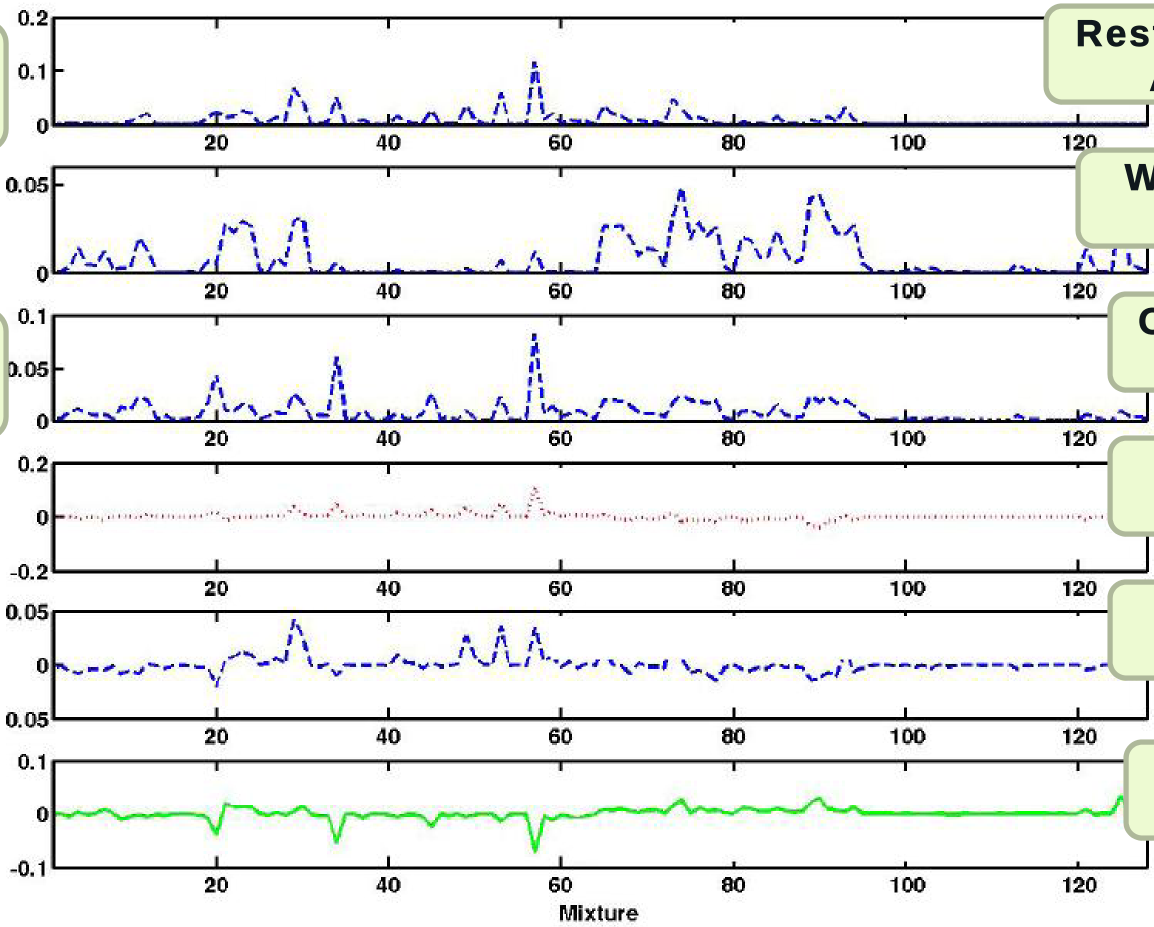
Environment Detection performance here shows promise, but the actual EER needs to be in the 5-10% range to be useful for a practical system.

Acoustic Signature Vector example

Restaurant

Walking

Outdoor



Restaurant ASV

Walking ASV

Outdoor ASV

R-W ASV

R-O ASV

W-O ASV

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

Experiments

ASV System based clustering (F-Measure)

System	F-Measure % Cosine distance	F-Measure % Euclidean distance
1-Query,Pure	61.78	60.74
1-Query, S-R	63.37	63.23
3-Query,Pure	75.09	74.54
3-Query, S-R	**79.82**	**77.47**

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

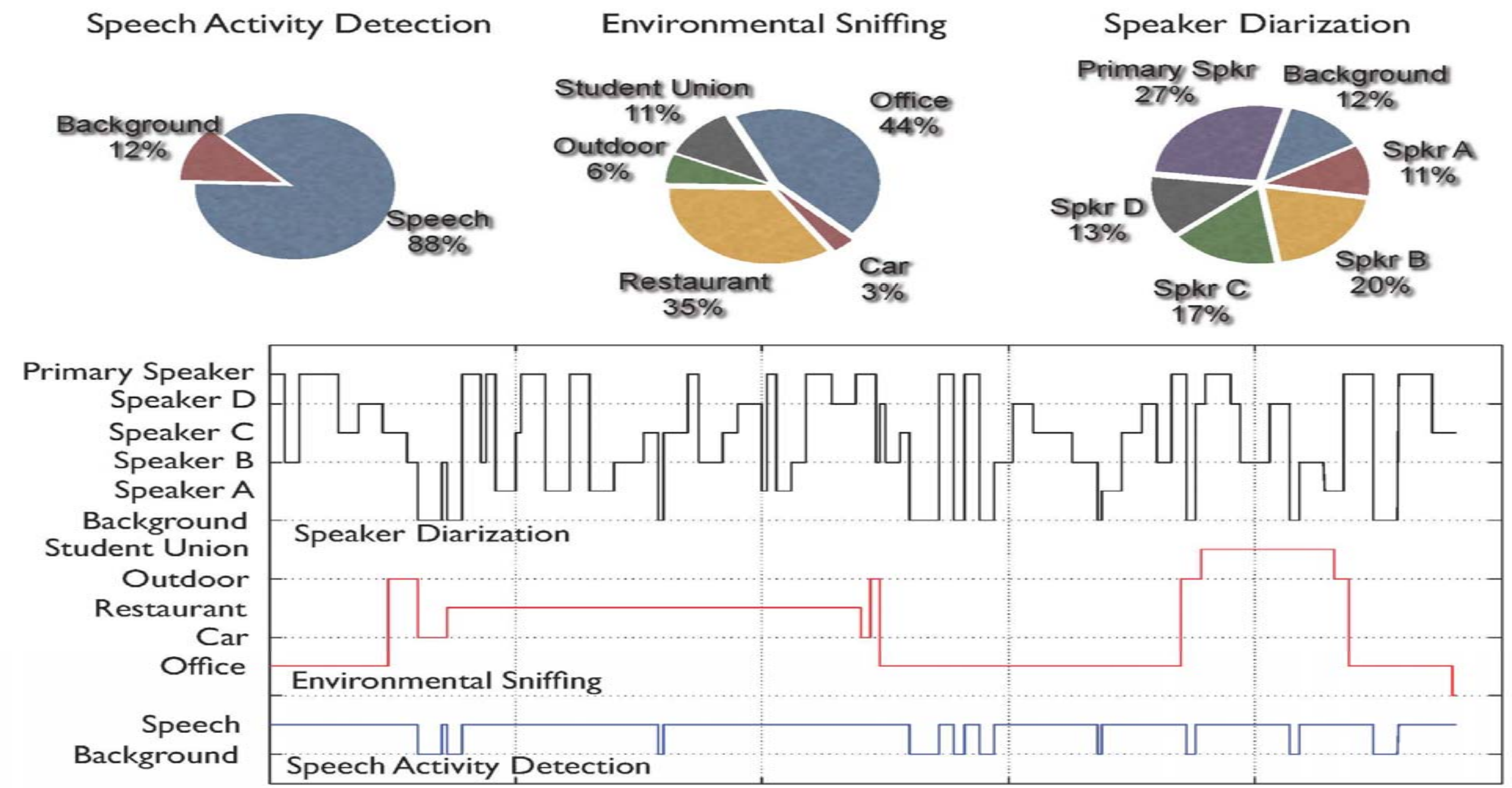
- Environment ID & tracking
- Keyword spotting (KWS)
- Topic ID
- Adult Distribution / Diversity

(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



[1] A. Ziaei, A. Sangwan, J.H.L. Hansen, "*Prof-Life-Log: Personal Interaction Analysis on Naturalistic Audio Streams*." ICASSP'2013, Vancouver, Canada