SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones

Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, Andrew T. Campbell
Dept. of Computer Science, Dartmouth College
Mobile Phones are the perfect platform for large scale sensor networks.
Sounds captured by the microphone allow us to sense people and the environment.
SoundSense

Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, Andrew T. Campbell
design
Design Issues

• Scalable Sound Classification
• Privacy
• Resource limitations
• Phone context
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Admission Control

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Admission Control

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Admission Control

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Admission Control

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
Admission control lowers the resources consumed during continuous sensing

- Filter out silence and non-informative frames

- Criteria:
  - Energy & Spectral Entropy
SoundSense Pipeline

- Sound Waveform
- Framing
- Feature Extraction
- Coarse Category Classification
- Intra-Category Classification

Admission Control

Acoustic Features

- Decision Tree Classifier
- Markov Model Recognizer

Voice Analysis
Music Analysis
Ambient Sound Learning
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Acoustic Features

Admission Control

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
Discriminative Acoustic Features

- Frequency Spectrum based features:
  - Robust to distance/muffling
  - Computationally very cheap after initial FFT
Discriminative Acoustic Features

Spectral Rolloff for Music and Voice
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Sound Waveform

Admission Control

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
Coarse Category Classification

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Music

Voice

Ambient Sound
Coarse Category Classification

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Music

Voice

Ambient Sound
Coarse Category Classification

[Decision Tree Classifier]

[Acoustic Features]

[Markov Model Classifier]

Music, Music, **Voice**, Music, **Voice**, Music, **Voice**

Music

Voice

Ambient Sound
Coarse Category Classification

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer


Music

Voice

Ambient Sound
Coarse Category Classification

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer


Music, Music, Music, Music, Music, Music, Music, Music

Music

Voice

Ambient Sound
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Admission Control

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
SoundSense Pipeline

Sound Waveform

Framing

Feature Extraction

Coarse Category Classification

Intra-Category Classification

Admission Control

Acoustic Features

Decision Tree Classifier

Markov Model Recognizer

Voice Analysis

Music Analysis

Ambient Sound Learning
Significant sounds in the life of a user

• Unique to the individual

• Not practical to build supervised classifiers for all the potential significant sounds

• Examples:
  ▸ Mystery Sound 1
  ▸ Mystery Sound 2
Significant sounds in the life of a user

• Unique to the individual

• Not practical to build supervised classifiers for all the potential significant sounds

• Examples:
  ▶ Mystery Sound 1
  ▶ Mystery Sound 2
Significant sounds in the life of a user

• Unique to the individual

• Not practical to build supervised classifiers for all the potential significant sounds

• Examples:
  ▶ Mystery Sound 1
  ▶ Mystery Sound 2
Significant sounds in the life of a user

• Unique to the individual

• Not practical to build supervised classifiers for all the potential significant sounds

• Examples:
  
  ▶ Car (turning signal)

  ▶ Mystery Sound 2
Significant sounds in the life of a user

• Unique to the individual

• Not practical to build supervised classifiers for all the potential significant sounds

• Examples:

  ▶ Car (turning signal)
  ▶ Mystery Sound 2
Significant sounds in the life of a user

- Unique to the individual
- Not practical to build supervised classifiers for all the potential significant sounds
- Examples:
  - Car (turning signal)
  - Vacuum Cleaner
Our Approach

1. Incrementally build models of encountered sounds
2. Rank them by order of "importance"
3. Human-in-the-loop labeling
Sounds modeled as multivariate Gaussians

MFCCs Features

\[ \mu_1, \Sigma_1 \]
\[ \mu_2, \Sigma_2 \]
\[ \ldots \]
\[ \text{Bin}_{n-1} \]
\[ \mu_n, \Sigma_n \]
Sounds modeled as multivariate Gaussians

\[
\text{Prob}(p_i = b) = \begin{cases} 
1 & \text{if } i = \arg \max_b \mathcal{N}(m_i; \mu_b, \Sigma_b) \\
0 & \text{and } \mathcal{N}(m_i; \mu_b, \Sigma_b) > \epsilon_{\text{THRESHOLD}} \\
0 & \text{otherwise}
\end{cases}
\]
Sounds modeled as multivariate Gaussians

\[ \mu'_b = \frac{\sum_{i=1}^t \text{Prob}(p_i = b)w_i}{\sum_{i=1}^t \text{Prob}(p_i = b)} = \frac{1}{t} (\mu_b(t - 1) + w_{j, \text{Prob}(p_j = b) = 1}) \]

\[ \Sigma'_b = \frac{\sum_{i=t-P}^t \text{Prob}(p_i = b)(w_i - \mu'_b)(w_i - \mu'_b)^T}{\sum_{i=1}^m \text{Prob}(p_i = b)} \]
Sounds modeled as multivariate Gaussians
Sound Rank estimates the “importance” a sound

Sound Rank based on:

- Duration of the event
- Frequency of the encounter
Sound Rank estimates the “importance” a sound

More Important

Less Important
Human-in-the-loop to identify and validate sounds

MFCCs Features

SR = 50
SR = 30
... SR = 10
SR = 3
Human-in-the-loop to identify and validate sounds

MFCCs Features

SR = 50
SR = 30
SR = 10
SR = 3

Human-in-the-loop to identify and validate sounds
Human-in-the-loop to identify and validate sounds

MFCCs Features

\[
\begin{align*}
\text{SR} &= 50 \\
\text{SR} &= 30 \\
\text{ignore} \\
\text{SR} &= 3
\end{align*}
\]
Human-in-the-loop to identify and validate sounds
Human-in-the-loop to identify and validate sounds

MFCCs Features

SR = 50
hide

... ignore

SR = 3

![Human-in-the-loop](image)
Human-in-the-loop to identify and validate sounds

MFCCs Features

SR = 50
hide
ignore
SR = 3
Human-in-the-loop to identify and validate sounds

MFCCs Features

- label
- hide
- ignore
- SR
- 3

[Image of person]

[Sound wave]

ignore
MFCCs Features

- label
- hide
- ignore

HMM Smoother
Output event
evaluation
iPhone implementation

- Jail broken iPhone
- C, C++, Objective C
- 300 KB application footprint
SoundSense implementation has low levels of resource consumption on the iPhone

<table>
<thead>
<tr>
<th></th>
<th>cpu usage range</th>
<th>memory usage range</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUI only</td>
<td>&lt; 1%</td>
<td>2.79 MB</td>
</tr>
<tr>
<td>silence</td>
<td>1 - 5%</td>
<td>3.75 MB</td>
</tr>
<tr>
<td>music/speech</td>
<td>8% - 20%</td>
<td>3.80 MB</td>
</tr>
<tr>
<td>ambient sound</td>
<td>11% - 22%</td>
<td>3.80 MB ~ 5 MB</td>
</tr>
</tbody>
</table>
SoundSense implementation has low levels of resource consumption on the iPhone.

<table>
<thead>
<tr>
<th></th>
<th>cpu usage range</th>
<th>memory usage range</th>
</tr>
</thead>
<tbody>
<tr>
<td>GUI only</td>
<td>&lt; 1%</td>
<td>2.79 MB</td>
</tr>
<tr>
<td>silence</td>
<td>1 - 5%</td>
<td>3.75 MB</td>
</tr>
<tr>
<td>music/speech</td>
<td>8% - 20%</td>
<td>3.80 MB</td>
</tr>
<tr>
<td>ambient sound</td>
<td>11% - 22%</td>
<td>3.80 MB ~ 5 MB</td>
</tr>
</tbody>
</table>
### Coarse Classification using Decision Tree only

<table>
<thead>
<tr>
<th></th>
<th>ambient sound</th>
<th>music</th>
<th>voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>ambient sound</td>
<td>0.91</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>music</td>
<td>0.13</td>
<td>0.81</td>
<td>0.05</td>
</tr>
<tr>
<td>voice</td>
<td>0.06</td>
<td>0.14</td>
<td>0.78</td>
</tr>
</tbody>
</table>
Coarse Classification using Decision Tree only

<table>
<thead>
<tr>
<th></th>
<th>ambient sound</th>
<th>music</th>
<th>voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>ambient sound</td>
<td><strong>0.91</strong></td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>music</td>
<td>0.13</td>
<td>0.81</td>
<td>0.05</td>
</tr>
<tr>
<td>voice</td>
<td>0.06</td>
<td>0.14</td>
<td>0.78</td>
</tr>
</tbody>
</table>
Coarse Classification using Decision Tree only

<table>
<thead>
<tr>
<th></th>
<th>ambient sound</th>
<th>music</th>
<th>voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>ambient sound</td>
<td>0.91</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>music</td>
<td>0.13</td>
<td>0.81</td>
<td>0.05</td>
</tr>
<tr>
<td>voice</td>
<td>0.06</td>
<td>0.14</td>
<td>0.78</td>
</tr>
</tbody>
</table>
Coarse Classification using Decision Tree and Markov Model smoothing

<table>
<thead>
<tr>
<th></th>
<th>ambient sound</th>
<th>music</th>
<th>voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>ambient sound</td>
<td>0.94</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>music</td>
<td>0.04</td>
<td>0.92</td>
<td>0.04</td>
</tr>
<tr>
<td>voice</td>
<td>0.03</td>
<td>0.07</td>
<td>0.90</td>
</tr>
</tbody>
</table>
# Unsupervised Adaptive Classifier

## Learning New Sound Classes

<table>
<thead>
<tr>
<th>Sound Rank</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Phone rubbing against clothes</td>
</tr>
<tr>
<td>2nd</td>
<td>Highway driving</td>
</tr>
<tr>
<td>3rd</td>
<td>Walking outdoors</td>
</tr>
<tr>
<td>4th</td>
<td>Driving w/ acceleration</td>
</tr>
<tr>
<td>5th</td>
<td>In-town driving</td>
</tr>
<tr>
<td>6th</td>
<td>Walking indoors</td>
</tr>
<tr>
<td>7th</td>
<td>Taking out trash</td>
</tr>
</tbody>
</table>
Unsupervised Adaptive Classifier
Learning New Sound Classes

<table>
<thead>
<tr>
<th>Sound Rank</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>Phone rubbing against clothes</td>
</tr>
<tr>
<td>2nd</td>
<td><strong>Highway driving</strong></td>
</tr>
<tr>
<td>3rd</td>
<td>Walking outdoors</td>
</tr>
<tr>
<td>4th</td>
<td><strong>Driving w/ acceleration</strong></td>
</tr>
<tr>
<td>5th</td>
<td><strong>In-town driving</strong></td>
</tr>
<tr>
<td>6th</td>
<td>Walking indoors</td>
</tr>
<tr>
<td>7th</td>
<td>Taking out trash</td>
</tr>
</tbody>
</table>
# Unsupervised Adaptive Classifier
## Learning New Sound Classes

<table>
<thead>
<tr>
<th>Sound Rank</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td><strong>Phone rubbing against clothes</strong></td>
</tr>
<tr>
<td>2nd</td>
<td>Highway driving</td>
</tr>
<tr>
<td>3rd</td>
<td>Walking outdoors</td>
</tr>
<tr>
<td>4th</td>
<td>Driving w/ acceleration</td>
</tr>
<tr>
<td>5th</td>
<td>In-town driving</td>
</tr>
<tr>
<td>6th</td>
<td>Walking indoors</td>
</tr>
<tr>
<td>7th</td>
<td>Taking out trash</td>
</tr>
</tbody>
</table>
Sound Sense Diary: Discovering Life Patterns

![Diagrams showing sound events over time](image)
Sound Sense Diary: Discovering Life Patterns
Music Tagger: Building Apps with SoundSense
Related Work

• recognition/audio scene analysis using resource constrained devices

• e.g., Clarkson et al. ’98, Smith et al. ‘06

• specific types of sound analysis:

  • speech recognition, music genre classification, speaker identification

• sensing with mobile phones:

  • UCLA, UIUC, Nokia, Microsoft, Motorola, Intel, UW, Duke
Summary

- SoundSense addresses the needs of sound classification using mobile phones:
  - Privacy, Energy Efficiency & robustness to Phone Context
  - Scalability via unsupervised learning of everyday sounds
- Future Work - Improving Ambient Sound Learning
  - Revising Sound Rank
  - Coping with fragmented learnt classes
Thanks for listening

Questions?