FODAVA-Partner: Visualizing Audio for Anomaly Detection

Mark Hasegawa-Johnson, Camille Goudeseune, Hank Kaczmarski and Thomas Huang

University of Illinois

December 12, 2012
A trained data analyst can detect anomalies at a glance when data are appropriately transformed.
Outline

1. Research Summary
   - Publications
   - Outreach
   - Topics of Active Research

2. Key Results
   - Saliency-enhanced features halve analyst errors
   - Audio visualization octoplies anomaly detection speed

3. Example Result: Generative-to-Discriminative Mapping Reduces AED Error 20%

4. Conclusions: Results of this Research
Outline

1. Research Summary
   - Publications
   - Outreach
   - Topics of Active Research

2. Key Results
   - Saliency-enhanced features halve analyst errors
   - Audio visualization octoplies anomaly detection speed

3. Example Result: Generative-to-Discriminative Mapping Reduces AED Error 20%

4. Conclusions: Results of this Research
Publications: Journal Articles & Technical Reports


Publications: Conference & Workshop Papers


Results: Public Dissemination and K-12 Outreach

**Dissemination & Outreach**

- **Beckman Open House Exhibits** in 2009, 2011
- **Beckman Cube Tour**
  - Groups: K-12 and international visitors, ~350 groups/year
- **Press Release** on futurity.org

**Millipphone in the Beckman Cube**

- [See the sounds: Audio as visual image](https://www.futurity.org/science-technology/see-the-sounds-audio-as-visual-image/)
Topics of Active Research

- **Data Transformations**
  - Biology: **Auditory Modeling Features**
    King & Hasegawa-Johnson, CoLing 2012
  - Psychology: **Salience-Maximizing Features**
    Lin et al., ICASSP 2012
  - Statistics: **Log Likelihood Features**
    Zhuang et al., PRL 2010
  - DSP: **Multiscale Spectrograms**
    Cohen, Goudeseune & Hasegawa-Johnson, GT-FODAVA-09-01

- **Software Testbeds**
  - Multiscale Zooming: **Timeliner**
    Goudeseune, ACM WIMMPD 2012
  - Geospatial VA: **Milliphone**
    McGaughey, Futurity, November 2011

- **Data Mining & Learning Theory**
  - Unknown Class Discovery
    Huang & Hasegawa-Johnson, 2008
  - Web-Based Multimedia Analytics
Outline

1. Research Summary
   - Publications
   - Outreach
   - Topics of Active Research

2. Key Results
   - Saliency-enhanced features halve analyst errors
   - Audio visualization octoplies anomaly detection speed

3. Example Result: Generative-to-Discriminative Mapping Reduces AED Error 20%

4. Conclusions: Results of this Research
Key Results of this Research

1. Saliency-enhanced features halve the error rate of human analysts
2. Audio visualization permits anomaly detection at 8X real-time
3. Generative-to-discriminative modeling reduces acoustic event detection errors by 20%
Saliency-enhanced features halve the error rate of human analysts

In our 2012 ICASSP paper, we demonstrated that human analysts tasked with detecting anomalies in a large audio file can halve their error rates (F-score increases from 0.3 to 0.6) by the use of a visualization tool in which visual saliency of the spectrogram is a monotonic function of estimated probability of an audio anomaly.

Fig. 5. (a) F-score of human AED using different audio visualization; (b) Three-way ANOVA of the F-score
Audio visualization permits anomaly detection at 8X real-time

In our 2011 APSIPA paper we showed that the use of zoomable audio visualization tools allows some users to find audio "easter eggs" (anomalies, e.g., motorcycles, cuckoo clocks, and spaceships added in to a background composed of eight hours of orchestral music) at a rate eight times faster than they would achieve by simply listening to the audio.
Outline

1. Research Summary
   - Publications
   - Outreach
   - Topics of Active Research

2. Key Results
   - Saliency-enhanced features halve analyst errors
   - Audio visualization octoplies anomaly detection speed

3. Example Result: Generative-to-Discriminative Mapping Reduces AED Error 20%

4. Conclusions: Results of this Research
Bayesian Modeling: Instead of saying that the class PDF generates instances,
- Say that the class PDF generates instance PDFs, and each instance PDF generates exactly one instance.

Why it’s useful: Instance PDF is drawn from an arbitrarily high-dimensional space (the space of all possible PDFs).
- It is always possible to find a transformation of that space in which intra-class variability is smaller than inter-class variability.

Obvious limitations:
- How do you estimate a PDF from one instance?
- In which transformation of the “space of all possible PDFs” is intra-class variability smaller than inter-class variability?
Mixture Gaussian Model

$x$ is the signal log spectrum; $c$ is the acoustic event label. The PDF $p(x|c)$ is modeled as a stochastic mixture of Gaussian kernels with means $\mu_k$ and covariances $\Sigma_k$.

$$p(x|c) = \sum_m w_{ck} \mathcal{N}(x; \mu_k, \Sigma_k)$$

MAP Adaptation to the $p$'th instance

$\gamma_k(t)$ is the posterior probability that observation $x_t$, one of the observations from the $p^{th}$ instance, belongs to Gaussian kernel $k$.

$$\gamma_k(t) = \frac{w_{ck} \mathcal{N}(x_t; \mu_k, \Sigma_k)}{\sum_j w_{cj} \mathcal{N}(x_t; \mu_j, \Sigma_j)}$$

The adapted mean vectors, $\mu_k^{(p)}$, describe the $p^{th}$ instance PDF. Their resemblance to the type PDF is controlled by the inertia parameter $\nu$.

$$\mu_k^{(p)} = \frac{\sum_{t \in p} \gamma_k(t) x_t + \nu \mu_k}{\sum_{t \in p} \gamma_k(t) + \nu}$$
Parameterizing and Normalizing the Instance PDF

Parameterize the $p^{th}$ instance

1. Instance PDF is parameterized by a supervector, $\vec{s}_p$.

   \[
   \vec{s}_p = \begin{bmatrix}
   \Sigma_1^{-1/2} (\vec{\mu}_1^{(p)} - \vec{\mu}_1) \\
   \vdots \\
   \Sigma_K^{-1/2} (\vec{\mu}_K^{(p)} - \vec{\mu}_K)
   \end{bmatrix}
   \]

2. Inter-instance variability is parameterized by a within-class covariance matrix, $R = \text{COV}(\vec{s}_p)$.

Normalize the $p^{th}$ instance

3. Supervectors are then normalized using within-class covariance normalization (WCCN): $\tilde{s}_p = R^{-1} \vec{s}_p$.

4. In the WCCN supervector space $\tilde{s}_p$, intra-class variability is less than inter-class variability, therefore any classifier can work well (e.g., nearest-centroid).
Experimental Test: Non-Speech Acoustic Event Detection

Difficulties

- Negative SNR (speech is “background noise”)
- Unknown spectral structure
- Different spectral structure for each event type

Key Jingle

Footsteps

Speech
WCCN Supervectors Rescore Tandem MAP Decoding

MAP Decoding Using Tandem NN-GMM-HMM

1. Contextual window for Tandem
2. Hidden layer
3. Transformation
4. Logarithmic transformation
5. PCA
6. ANN output
7. Current feature frame
8. HMM

Rescoring Using WCCN Supervectors

1. MAP Decoding for segmentation & classification
2. Hypothesized boundaries and event labels (one best or lattice)
3. Confidence rescoring / event classification using new method
4. Refining output result according to AED metric
5. Improved detection result
## Acoustic Event Detection Results

### Without Supervectors

<table>
<thead>
<tr>
<th>CLEAR 2007 AED RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inst.</td>
</tr>
<tr>
<td>AIT</td>
</tr>
<tr>
<td>ITC</td>
</tr>
<tr>
<td>TUT</td>
</tr>
<tr>
<td>UIUC</td>
</tr>
<tr>
<td>STI2R</td>
</tr>
<tr>
<td>UPC</td>
</tr>
</tbody>
</table>

### With WCCN Supervectors

<table>
<thead>
<tr>
<th></th>
<th>ap</th>
<th>cl</th>
<th>cm</th>
<th>co</th>
<th>ds</th>
<th>kj</th>
<th>kn</th>
<th>kt</th>
<th>la</th>
<th>pr</th>
<th>pw</th>
<th>st</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>78.3</td>
<td>26.9</td>
<td>29.5</td>
<td>24.2</td>
<td>56.3</td>
<td>39.9</td>
<td>7.7</td>
<td>0.0</td>
<td>39.0</td>
<td>35.2</td>
<td>14.1</td>
<td>28.7</td>
<td>28.2</td>
</tr>
<tr>
<td>FB</td>
<td>34.5</td>
<td>21.8</td>
<td>25.4</td>
<td>24.9</td>
<td>38.9</td>
<td>27.2</td>
<td>11.7</td>
<td>0.0</td>
<td>49.1</td>
<td>13.8</td>
<td>11.7</td>
<td>28.1</td>
<td>27.8</td>
</tr>
<tr>
<td>Adaboost</td>
<td>44.4</td>
<td>25.5</td>
<td>31.3</td>
<td>31.2</td>
<td>57.3</td>
<td>33.2</td>
<td>13.5</td>
<td>1.9</td>
<td>51.3</td>
<td>36.7</td>
<td>17.6</td>
<td>36.8</td>
<td>34.0</td>
</tr>
<tr>
<td>Adaboost+T</td>
<td>52.6</td>
<td>21.9</td>
<td>37.2</td>
<td>51.3</td>
<td>63.0</td>
<td>29.6</td>
<td>11.5</td>
<td>0.0</td>
<td>54.2</td>
<td>42.7</td>
<td>25.8</td>
<td>34.6</td>
<td>35.3</td>
</tr>
<tr>
<td>Adaboost+S</td>
<td>44.4</td>
<td>25.0</td>
<td>33.7</td>
<td>31.2</td>
<td>56.6</td>
<td>33.2</td>
<td>20.9</td>
<td>35.5</td>
<td>51.3</td>
<td>36.7</td>
<td>19.2</td>
<td>41.3</td>
<td>37.5</td>
</tr>
<tr>
<td>Adaboost+T+S</td>
<td>52.6</td>
<td>21.5</td>
<td>37.4</td>
<td>47.9</td>
<td>63.0</td>
<td>29.6</td>
<td>13.6</td>
<td>44.8</td>
<td>58.6</td>
<td>42.7</td>
<td>26.7</td>
<td>44.4</td>
<td>41.2</td>
</tr>
</tbody>
</table>

MFCC = mel-frequency cepstral coefficients, FB = filterbank, T = Tandem, S = Supervector
Conclusions: Acoustic Event Detection

- The class PDF generates instance PDFs; the instance PDFs generate instances.
  - Instance PDF can be estimated using MAP (regularized) learning.
- The space of all possible PDFs is a very large space indeed; lots of interesting normalization methods are possible.
  - (Simple) Within-class covariance normalization is very effective.
  - After WCCN, (simple) minimum-centroid classification seems to work better (often) than any other classifier.
Outline

1. Research Summary
   - Publications
   - Outreach
   - Topics of Active Research

2. Key Results
   - Saliency-enhanced features halve analyst errors
   - Audio visualization octoplies anomaly detection speed

3. Example Result: Generative-to-Discriminative Mapping Reduces AED Error 20%

4. Conclusions: Results of this Research
Conclusions: Results of this Research

- **Publications:** 9 papers, 5 published abstracts
- **Outreach:** 2 Open Houses, $\sim 1000$ Tour Groups, 1 Press Release
- **Key Results**
  - Saliency-enhanced features halve the error rate of human analysts
  - Audio visualization permits anomaly detection at 8X real-time
  - Generative-to-discriminative modeling reduces acoustic event detection errors by 20%
Thank you!