Towards Better Real-world Acoustic Event Detection

Mark Hasegawa-Johnson, Xiaodan Zhuang, Xi Zhou, Camille Goudeseune, Hao Tang, Kai-Hsiang Lin, Mohamed Omar, and Thomas Huang

University of Illinois

Presentation at Seoul National University, May 30, 2012
Outline

1. Introduction: Task Definitions, Fundamentals
2. Discriminative Feature Selection for Acoustic Event Detection
3. Symplectic Maps: Discriminative Feature Transform for Phone Classification
4. Supervectors: Compensating for Unknown Sources of Variability
5. K-12 Outreach: Beckman Open House
6. Conclusions
Outline

1. Introduction: Task Definitions, Fundamentals
2. Discriminative Feature Selection for Acoustic Event Detection
3. Symplectic Maps: Discriminative Feature Transform for Phone Classification
4. Supervectors: Compensating for Unknown Sources of Variability
5. K-12 Outreach: Beckman Open House
6. Conclusions
Task Definitions

Task #1: Acoustic Event Detection
- Detect non-speech acoustic events (door slam, chair movement, paper shuffle) in a meeting room
- What happened when?

Task #2: Speech Phone Classification
- Given an acoustic spectrum $x_i$, specify the phone label $y_i$
- A heavily-studied problem, therefore the baselines are well understood
Task #1: Non-Speech Acoustic Event Detection

Motivation

“Activity detection and description is a key functionality of perceptually aware interfaces working in collaborative human communication environments... detection and classification of acoustic events may help to detect and describe human activity...” (CLEAR-AED Task Brief)

Difficulties

- Negative SNR (speech is “background noise”)
- Unknown spectral structure
- Different spectral structure for each event type
Difficulty #1: Negative SNR
Difficulty #2: Unknown Spectral Structure

Key Jingle

Footsteps

Speech
Fundamentals of Machine Learning

- **What is Random:**
  - **The Test Data:** \((x, y)\) are drawn from unknown distribution \(P(x, y)\), where \(x \in \mathcal{X}\) is observation, \(y \in \mathcal{Y}\) is the unknown correct label
  - **The Training Data:** \(D = \{(x_1, y_1), \ldots, (x_n, y_n)\}\) are drawn i.i.d. from \(P(x, y)\)

- **What is Learned:** The hypothesis function \(h(x)\) is selected from a function space \(\mathcal{H} : \mathcal{X} \rightarrow \mathcal{Y}\) with covering number \(N_{\mathcal{H}}\) (covering number = number of meaningfully distinct hypotheses).

- **How it is Evaluated:** Selecting \(h(x)\), if the true label is \(y\), incurs cost \(\ell(h(x), y) \in [0, 1]\).

The Goal of Machine Learning

Choose \(h(x) \in \mathcal{H}\) in order to minimize \(R(h) = E_P[\ell(h(x), y)]\)
Probably Approximately Correct (PAC) Learning

- **Risk:**
  \[ R(h) = E_P \left[ \ell(h(x), y) \right] \]

- **Empirical Risk Estimate:**
  \[ \hat{R}(h, D) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) \]

- **PAC Bound (Haussler, 1983):** with probability at least \(1 - \delta\),
  \[ \max_{h \in \mathcal{H}} \left| R(h) - \hat{R}(h, D) \right| \leq \sqrt{\frac{\ln 2N_{\mathcal{H}} - \ln \delta}{2n}} \]

- **Methods for Controlling Complexity:** the function-class entropy, \( \ln N_{\mathcal{H}} \), can be controlled by reducing the size of the feature space, or simplifying the classifier structure.
Outline

1. Introduction: Task Definitions, Fundamentals
2. Discriminative Feature Selection for Acoustic Event Detection
3. Symplectic Maps: Discriminative Feature Transform for Phone Classification
4. Supervectors: Compensating for Unknown Sources of Variability
5. K-12 Outreach: Beckman Open House
6. Conclusions
Discriminative Feature Selection for AED

Zhuang et al., ICASSP 2008

- **Problem:** what acoustic features are relevant for detecting non-speech acoustic events?
- **Input:** \((x_i \in \mathbb{R}^D)\) includes many acoustic features invented for speech processing (MFCC, PLP, energy, ZCR)
- **Output:** \((f_i \in \mathbb{R}^d)\) selects the most useful features: 
  \[ f_i = Wx_i \]
  where \(W^T = [w_1, \ldots, w_K]\), and \(w_k\) is an indicator vector (only one non-zero element)
- **Hidden Markov Modeling:** the label sequence \(Y^* = [y_1^*, \ldots, y_N^*]\), \(y_i \in \{\text{keyjingle}, \text{footstep}, \ldots\}\) is chosen by a hidden Markov model observing \(F = [f_1, \ldots, f_N]\):
  \[ Y^* = \arg \max p(F|Y)p(Y) \]
Bayes Error Rate

Zhuang et al., ICASSP 2008

Bayes Error Rate

Let $w_k$ be an indicator vector (all zeros except for one element). The Bayes-optimal error rate of a classifier observing feature $w_k^T x$ is

$$P(\text{error}) = \int \int P\left( y \neq \arg \max p(w_k^T x, y) \right) \, dy \, dx$$

Bayes Error Rate Approximated on a Database

$$\mathcal{F}(w_k) = \frac{1}{N} \sum_{i=1}^{N} \delta \left( y_i \neq \arg \max p(w_k^T x_i, y_i) \right)$$
Feature Selection Algorithms

### Hard-Bayes-Error Feature Selection

For $k = 1, \ldots, K$, choose the indicator vector $w_k$ ($w_k$ is all zeros except for one nonzero element) to minimize

$$
F(w_k) = \frac{1}{N} \sum_{i=1}^{N} \delta \left( y_i \neq \arg \max p(w_k^T x_i, y_i) \right)
$$

### Soft-Bayes-Error Feature Selection

For $k = 1, \ldots, K$, choose the indicator vector $w_k$ ($w_k$ is all zeros except for one nonzero element) to minimize

$$
F_S(w_k) = \frac{1}{N} \sum_{i=1}^{N} \text{rank} \left( y_i \left| w_k^T x_i \right. \right)
$$
Acoustic Event Detection Accuracy (Percent)

Zhuang et al., ICASSP 2008

MFCC26DAZ = 26 Mel-frequency cepstral coefficients + deltas + acceleration

DERIVE26DAZ = 26 Derived features + deltas + acceleration

DERIVE78 = 78 Derived features

All of these have accuracy below 32%!! Need some more ideas, which are coming in the next two sections.
<table>
<thead>
<tr>
<th></th>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction: Task Definitions, Fundamentals</td>
</tr>
<tr>
<td>2</td>
<td>Discriminative Feature Selection for Acoustic Event Detection</td>
</tr>
<tr>
<td>3</td>
<td>Symplectic Maps: Discriminative Feature Transform for Phone Classification</td>
</tr>
<tr>
<td>4</td>
<td>Supervectors: Compensating for Unknown Sources of Variability</td>
</tr>
<tr>
<td>5</td>
<td>K-12 Outreach: Beckman Open House</td>
</tr>
<tr>
<td>6</td>
<td>Conclusions</td>
</tr>
</tbody>
</table>
Discriminative Feature Transform

Work in progress...

- **Problem:** what projection of the acoustic spectrogram is relevant for recognizing non-speech acoustic events?
- **Output:** \((f_i \in \mathbb{R}^d)\) selects the most useful features:

\[
f_i = \sum_{k=1}^{K} c_k \sigma(w_k^T x_i)
\]

where \(c_k \in \mathbb{R}^d\) and \(w_k \in \mathbb{R}^D\) are arbitrary real-valued weight vectors, and \(\sigma(z) = 1/(1 + e^{-z})\).

- **Hidden Markov Modeling:** the label sequence \(Y^* = [y_1^*, \ldots, y_N^*]\), \(y_i \in \{\text{keyjingle, footstep, \ldots}\}\) is chosen by a hidden Markov model observing \(F = [f_1, \ldots, f_N]\):

\[
Y^* = \arg \max p(F|Y)p(Y)
\]
The Baum-Welch Algorithm

Hidden Markov model parameters are trained to maximize the expected log likelihood, with expectation over the unknown state sequence $Q = [q_1, \ldots, q_N]$

$$\mathcal{J} = E_Q \{ \log p(F, Q) \}$$

$$\mathcal{J} = -\frac{1}{2} \sum_{i=1}^{N} \sum_{q} p(q_i = q|F, Y)(f_i - \mu_q)^T \Sigma_q^{-1}(f_i - \mu_q) - \ldots$$
Baum-Welch Back-Propagation

The neural network can be trained, using standard gradient descent methods, in order to minimize $\mathcal{J}$. For example,

$$f_i = \sum_{k=1}^{K} c_k \sigma(w_k^T x_i)$$

\[
\frac{\partial \mathcal{J}}{\partial c_k} = \sum_{i=1}^{N} \sum_{q} p(q_i = q|F) \left( \frac{\partial \mathcal{J}}{\partial f_i | q_i = q} \right) \left( \frac{\partial f_i}{\partial c_k} \right) \\
= \sum_{i=1}^{N} \sum_{q} p(q_i = q|F) \Sigma_q^{-1}(\mu_q - f_i)\sigma(w_k^T x_i)
\]
The Problem of Spurious Maxima

- It is always possible to train a mixture Gaussian so that $\mathcal{J} = \infty$
  - Solution: Give one of the Gaussians a zero variance ($\Sigma_q = 0$)
  - This is called “over-training”
- In Baum-Welch Back-Propagation, the same result is obtained for $\|c_k\| \to 0$
  - Solution: require $\|c_k\| = 1$, or more generally, $\|\frac{\partial f_i}{\partial x_i}\| = 1$
Methods for Avoiding Spurious Maxima

- Constrained optimization: maximize

\[ \mathcal{L} = J + \sum_{k} \lambda_k (\|c_k\| - 1) \]

with Lagrange multipliers \( \lambda_k \) chosen so that \( \|c_k\| = 1 \)

- Symplectic Maximum Likelihood Transform (SMLT, Omar and Hasegawa-Johnson, 2004): replace the neural network with one that computes a *volume preserving* transform:

\[ |J_f(x)| = 1 \]

where \( J_f(x) \) is the Jacobian of the transform
SMLT + GMM for Phone Classification
Omar and Hasegawa-Johnson, 2004

- Compute phone label $y_i$ given MFCC cepstrum $x_i$
- Symplectic maximum likelihood transform (SMLT) computes $f_i(x_i)$
- Maximum likelihood linear transform (MLLT) computes $f_i = Wx_i$
- Gaussian mixture model (GMM) computes $p(f_i|y_i)$
- Database: TIMIT

<table>
<thead>
<tr>
<th>Features</th>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>GMM</td>
<td>73.7%</td>
</tr>
<tr>
<td>MLLT</td>
<td>GMM</td>
<td>74.6%</td>
</tr>
<tr>
<td>SMLT</td>
<td>GMM</td>
<td>75.6%</td>
</tr>
</tbody>
</table>
Outline

1. Introduction: Task Definitions, Fundamentals
2. Discriminative Feature Selection for Acoustic Event Detection
3. Symplectic Maps: Discriminative Feature Transform for Phone Classification
4. Supervectors: Compensating for Unknown Sources of Variability
5. K-12 Outreach: Beckman Open House
6. Conclusions
Compensating for Unknown Sources of Variability

(Hatch and Stolcke, 2006)

**Surprising idea:** 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?
Estimating the Instance PDF: MAP Adaptation

(Gauvain and Lee, 1994)

**Mixture Gaussian Model**

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

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

**MAP Adaptation to the p’th instance**

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

$$\gamma_k(t) = \frac{w_{ck} \mathcal{N}(\bar{x}_t; \bar{\mu}_k, \Sigma_k)}{\sum_j w_{cj} \mathcal{N}(\bar{x}_t; \bar{\mu}_j, \Sigma_j)}$$

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

$$\bar{\mu}^{(p)}_k = \frac{\sum_{t \in p} \gamma_k(t) \bar{x}_t + \nu \bar{\mu}_k}{\sum_{t \in p} \gamma_k(t) + \nu}$$
A UBM is a Gaussian mixture model (GMM) whose parameters, $\theta$, are trained on all one-centisecond-frames in the whole database, $\mathcal{D}$, in order to maximize

$$
\mathcal{L}(\mathcal{D}, \theta) = \prod_{j, t \in \mathcal{D}} p_\theta(x_{tj}), \quad p_\theta(x_{tj}) = \sum_{k=1}^{K} w_k \mathcal{N}(x_{tj}; \mu_k, \Sigma_k)
$$

The UBM can be used to perform soft clustering of the data:

$$
\gamma_{tj}(k) = p(q_{tj} = k|x_{tj}) = \frac{w_k \mathcal{N}(x_{tj}; \mu_k, \Sigma_k)}{p(x_{tj})}
$$

Summing the UBM over a whole audio clip produces a cluster histogram:

$$
h_j[k] = \sum_{t=1}^{T_j} \gamma_{tj}(k)
$$
Beyond the Histogram: The Cluster Shift Vector

- Histogram computes zero\(^{th}\)-order statistics (the number of frames corresponding to each cluster)
- The Cluster-Shift vector computes first-order statistics (the amount by which the frames in cluster \(k\) are shifted away from the cluster mean):

\[
m_{jk} = \frac{\tau \mu_k + \sum_{t=1}^{T_j} \gamma_{tj}(k)x_{tj}}{\tau + \sum_{t=1}^{T_j} \gamma_{tj}(k)}
\]

\[
s_j = \begin{bmatrix}
    w_1^{1/2} \Sigma_1^{-1/2}(m_{j1} - \mu_1) \\
    \vdots \\
    w_K^{1/2} \Sigma_K^{-1/2}(m_{jK} - \mu_K)
\end{bmatrix}
\]

where \(\tau\) is a hyper-parameter, typically chosen in the range \(0 \leq \tau \leq 10\).
## Attributes of the Cluster-Shift Vector

- **Relatively sparse:** if an audio clip contains fewer than $\tau$ keypoints from the $k^{th}$ cluster, then the corresponding cluster-shift subvector $m_{jk} \approx 0$

- **Audio clips as distributions:** The cluster-shift vectors $m_{jk}$ can be interpreted as mean vectors of an adapted distribution:

  \[
  p_j(x) = \sum_{k=1}^{K} w_k \mathcal{N}(x; m_{jk}, \Sigma_k)
  \]

- **Cluster-Shift Euclidean distance $\approx$ KL Divergence:** The symmetrized Euclidean distance between two distributions, $D(p_i, p_j)$, is well approximated as

  \[
  D(p_i, p_j) \approx \|s_i - s_j\|^2_2
  \]
The Gaussianized Vector: Steps 1 and 2...

With a small change to the cluster-shift vector representation, it is possible to ensure that the vectors $s_j$ are exactly Gaussian (Zhou et al., CVPR 2008):

1. **Cluster Membership Vector:** Define binary vector $\eta$ of dimension $K$, such that

   $$p(\eta[k] = 1) = w_k, \quad p(\eta[k] = 1|x_{tj}) = \gamma_{tj}(k)$$

2. **Monte Carlo Cluster Assignment:** Randomly generate binary vector instances $\eta_{tj}$ of random vector $\eta$ corresponding to each of the frames, i.e., using $\gamma_{tj}$ as the multinomial parameters.
The Gaussianized Vector: ... Steps 3 and 4

3. **Hard-Decision Cluster-Shift:** Redefine $m_{jk}$ as

$$m_{jk} = \frac{1}{n_k} \sum_{t=1}^{T_j} \eta_{tj}[k] x_{tj}$$

where $n_k = \sum_t \eta_{tj}[k]$ is the number of vectors for which $\eta_{tj}[k] = 1$.

4. **Count-Dependent Supervector:** Redefine $s_j$ as

$$s_j = \left[ (\Sigma_1/n_1)^{-1/2}(m_{j1} - \mu_1) \\ \\
(\Sigma_2/n_2)^{-1/2}(m_{j2} - \mu_2) \\ \\
\vdots \\ \\
(\Sigma_K/n_K)^{-1/2}(m_{jK} - \mu_K) \right]$$
## Attributes of The Gaussianized Vector

- Every element is conditionally Gaussian:

\[
p(x_{tj} | \eta_{tj}[k] = 1) = \frac{p(\eta_{tj}[k] = 1 | x_{tj}) p(x_{tj})}{p(\eta_{tj}[k] = 1)} = \frac{\gamma_{tj}(k) p(x_{tj})}{w_k}
\]

but

\[
\gamma_{tj}(k) = \frac{w_k \mathcal{N}(x_{tj}; \mu_k, \Sigma_k)}{p(x_{tj})}
\]

so

\[
p(x_{tj} | \eta_{tj}[k] = 1) = \mathcal{N}(x_{tj}; \mu_k, \Sigma_k)
\]

- The mean vector is the average of \( n_k \) elements, so

\[
m_{jk} \sim \mathcal{N}(m_{jk}; \mu_k, \Sigma_k / n_k)
\]

- The supervector is normalized by \((\Sigma_k / n_k)^{-1/2}\), so

\[
s_j \sim \mathcal{N}(0, 1)
\]
Normalizing the Instance PDF: WCCN

Parameterize the $p^{th}$ instance

1. Instance PDF is parameterized by a supervector, $s_p$.

$$s_p = \begin{bmatrix} w_1^{1/2} \Sigma_1^{-1/2} (m_{1p} - \bar{\mu}_1) \\ \vdots \\ w_K^{1/2} \Sigma_K^{-1/2} (m_{Kp} - \bar{\mu}_K) \end{bmatrix}$$

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

Normalize the $p^{th}$ instance

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

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
(Zhuang et al., Pattern Recognition Letters, 2010)

MAP Decoding Using Tandem NN-GMM-HMM

Rescored Using WCCN Supervectors

MAP Decoding
For segmentation & classification

Hypothesized Boundaries and event labels
(one best or lattice)

Confidence rescoring / event classification using new method

Refining output result According to AED metric

Improved Detection Result
## Acoustic Event Detection Results

### Without Supervectors

<table>
<thead>
<tr>
<th>Inst.</th>
<th>AED Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIT</td>
<td>4.4</td>
</tr>
<tr>
<td>ITC</td>
<td>23.4</td>
</tr>
<tr>
<td>TUT</td>
<td>14.7</td>
</tr>
<tr>
<td>UIUC</td>
<td>36</td>
</tr>
<tr>
<td>STI2R</td>
<td>22.9</td>
</tr>
<tr>
<td>UPC</td>
<td>23</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>21.9</td>
<td>39.6</td>
<td>49.8</td>
<td>63.0</td>
<td>29.6</td>
<td>15.4</td>
<td>36.8</td>
<td>54.7</td>
<td>41.7</td>
<td>26.0</td>
<td>38.3</td>
<td>38.6</td>
</tr>
</tbody>
</table>

MFCC = mel-frequency cepstral coefficients, FB = filterbank, T = Tandem, S = Supervector
Outline

1. Introduction: Task Definitions, Fundamentals
2. Discriminative Feature Selection for Acoustic Event Detection
3. Symplectic Maps: Discriminative Feature Transform for Phone Classification
4. Supervectors: Compensating for Unknown Sources of Variability
5. K-12 Outreach: Beckman Open House
6. Conclusions
Engineering Open House

- First held in 1907, when it attracted 1600 people
- Annual since 1952
- Two-day event
- Primary audience: K-12 field trips and families

Beckman Open House

The Beckman Institute Open House is a biennial event held in conjunction with the University of Illinois College of Engineering Open House. It features numerous exhibits showcasing fun and interesting research topics, including brain-computer interfaces, robots learning grammar, and the process of photosynthesis, as rendered through 3-D computer simulations. Things to try at the Open House include donning goggles to simulate alcohol intoxication and trying to shoot bullets, spelling out sentences using only your brainwaves, seeing what it’s like to get an MRI, and taking the helm of an electron scanning microscope.

A new entry joining the 2011 Open House is Bert, the iCub humanoid robot from Steve Levinson’s Language Acquisition and Robotics group. The rare (only one in the Western Hemisphere) advanced robot is sure to wow you at this year’s Beckman Institute Open House!

Also back this year are the popular and renowned tools of the Beckman Institute’s Illinois Simulator Laboratory! You will be able to see traveling versions of the CAVe, the Driving Simulator, the Flight Simulator, and the Motion Capture Suite.

www.beckman.illinois.edu/events/boh2011
2011 Open House

- **Background:** Opera and symphony music
- **Random anomalies:** Cuckoo clock; Motorcycle; UFO; Cow; Bird; Pac-Man; Goat

Audio Anomalies

- Visitor given a spectrogram-like visualization, but with anomalies emphasized (unsupervised anomaly detection)
- Goal: find as many as possible in **80 seconds**
Signal representation computes an estimate of the degree of audio anomaly at each time instant, and proportionally enhances the visual salience of the spectrogram.
Competition: The Scoreboard

Score Histogram

Visitors

Anomaly-discovery acceleration ratio
Visualization Accelerates Anomaly-Finding
(Hasegawa-Johnson, Goudeseune et al., APSIPA 2011)

8xRT = 24 Anomalies in 80 Seconds

- 80 seconds of audio contain, on average, 3 anomalies
- If a visitors finds 3 anomalies, that counts as “real-time performance”
- The best-scoring visitors (including one 7-year-old girl!) found up to 26 anomalies (more than 8× real time)
Controlled Study: Subjects Seek Anomalies

In a separate controlled study, we found that matching visual salience to audio anomaly improves the accuracy with which human subjects find audio anomalies planted by the experimenter.

![Graph and Table]

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>1</td>
<td>142.65</td>
<td>0</td>
</tr>
<tr>
<td>Order</td>
<td>1</td>
<td>0</td>
<td>0.9778</td>
</tr>
<tr>
<td>File</td>
<td>3</td>
<td>0.3</td>
<td>0.8234</td>
</tr>
<tr>
<td>Type*order</td>
<td>1</td>
<td>0.4</td>
<td>0.5334</td>
</tr>
<tr>
<td>Type*file</td>
<td>3</td>
<td>1.86</td>
<td>0.154</td>
</tr>
<tr>
<td>Order*file</td>
<td>3</td>
<td>0.29</td>
<td>0.8356</td>
</tr>
</tbody>
</table>
Outline

1. Introduction: Task Definitions, Fundamentals
2. Discriminative Feature Selection for Acoustic Event Detection
3. Symplectic Maps: Discriminative Feature Transform for Phone Classification
4. Supervectors: Compensating for Unknown Sources of Variability
5. K-12 Outreach: Beckman Open House
6. Conclusions
Conclusions: Feature Selection and Transformation

- Non-speech acoustic event spectra $\neq$ speech spectra
Conclusions: Feature Selection and Transformation

- Non-speech acoustic event spectra $\neq$ speech spectra
- Acoustic event detection benefits from discriminative feature selection
Conclusions: Feature Selection and Transformation

- Non-speech acoustic event spectra $\neq$ speech spectra
- Acoustic event detection benefits from discriminative feature selection
- Soft-Bayes-Error selection is better than Hard-Bayes-Error selection
Conclusions: Feature Selection and Transformation

- Non-speech acoustic event spectra \( \neq \) speech spectra
- Acoustic event detection benefits from discriminative feature selection
- Soft-Bayes-Error selection is better than Hard-Bayes-Error selection
- Discriminative feature selection can be generalized to discriminative feature transformation
Conclusions: Feature Selection and Transformation

- Non-speech acoustic event spectra ≠ speech spectra
- Acoustic event detection benefits from discriminative feature selection
- Soft-Bayes-Error selection is better than Hard-Bayes-Error selection
- Discriminative feature selection can be generalized to discriminative feature transformation
- SMLT (a form of discriminative feature transformation) outperforms MFCC and MLLT for phoneme classification in TIMIT
Conclusions: Normalizing for Unknown Sources of Variability

- The class PDF generates instance PDFs; the instance PDFs generate instances.
  - Instance PDF can be estimated using MAP (regularized) learning.
Conclusions: Normalizing for Unknown Sources of Variability

- 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.
Conclusions: Normalizing for Unknown Sources of Variability

- 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.

- The utility of AED can be tested in an interactive framework. Given a display in which the visual salience of a frame is matched to degree of audio anomaly, subjects find anomalies 8× faster, and with 2× the accuracy.
Thank you!