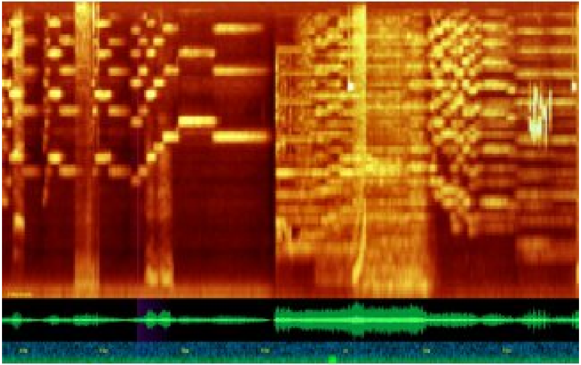
FODAVA-Partner: Visualizing Audio for Anomaly Detection

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

Publications: Journal Articles & Technical Reports

1. Lin, Zhuang, Goudeseune, King, Hasegawa-Johnson, & Huang, "Saliency-maximized Audio Visualization and Efficient Audio-visual Browsing for Faster-than-real-time Human Acoustic Event Detection," in preparation
2. Zhuang, Zhou, Hasegawa-Johnson, & Huang, "Real-world Acoustic Event Detection," *Pattern Recognition Letters* **31**(2):1543-1551
3. Zhou, Zhuang, Tang, Hasegawa-Johnson, & Huang, "Novel Gaussianized Vector Representation for Improved Natural Scene Categorization," *Pattern Recognition Letters* **31**(8):702-708
4. Cohen, Goudeseune & Hasegawa-Johnson, "Efficient Simultaneous Multi-Scale Computation of FFTs," *FODAVA Technical Report* GT-FODAVA-09-01

Publications: Conference & Workshop Papers

1. Goudeseune, "Effective browsing of long audio recordings," *ACM International Workshop on Interactive Multimedia on Mobile and Portable Devices*, 2012
2. King & Hasegawa-Johnson, "Detection of Acoustic-Phonetic Landmarks in Mismatched Conditions Using a Biomimetic Model of Human Auditory Processing," *CoLing* 2012
3. Lin, Zhuang, Goudeseune, King, Hasegawa-Johnson and Huang, "Improving Faster-than-Real-Time Human Acoustic Event Detection by Saliency-Maximized Audio Visualization," *ICASSP* 2012
4. Hasegawa-Johnson, Goudeseune, Cole, Kaczmarski et al., "Multimodal Speech and Audio User Interfaces for K-12 Outreach," *APSIPA* 2011
5. Kim & Mark Hasegawa-Johnson, "Optimal Multi-Microphone Speech Enhancement in Cars," *IEEE DSP in Cars Workshop* 10.1.1.150.8462:1-4, 2009

Publications: Published Abstracts

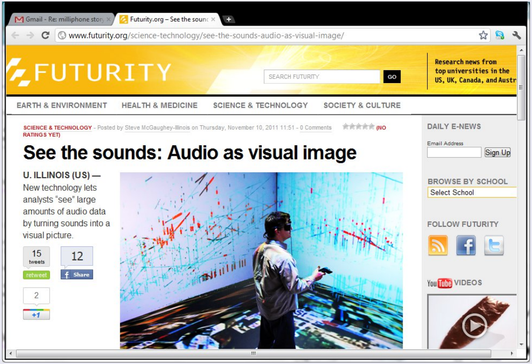
1. Hasegawa-Johnson, Huang, King and Zhou, "Normalized recognition of speech and audio events," *JASA* **130**:2524, 2011
2. Hasegawa-Johnson, Goudeseune, Lin et al., "Visual Analytics for Audio," *NIPS Workshop on Visual Analytics*, 2009
3. Hasegawa-Johnson, "Pattern Recognition in Acoustic Signal Processing," *Machine Learning Summer School* 2009
4. Hasegawa-Johnson, Zhuang, Zhou, Goudeseune & Huang, "Adaptation of tandem HMMs for non-speech audio event detection," *JASA* **125**:2730, 2009
5. Hasegawa-Johnson, "Tutorial: Pattern Recognition in Signal Processing," *JASA* **125**:2698, 2009

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

Milliphone in the Beckman Cube



Topics of Active Research

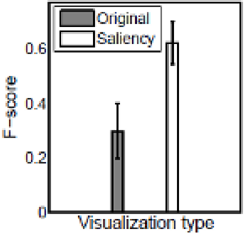
- Data Transformations
  - Biology: **Auditory Modeling Features**
    King & Hasegawa-Johnson, CoLing 2012
  - Psychology: **Salience-Maximizing Features**
    Lin et al., ICASSP 2012
  - Statistics: **Log Likelihood Features**
    Zhuang et al., PRL 2010
  - DSP: **Multiscale Spectrograms**
    Cohen,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



(a)

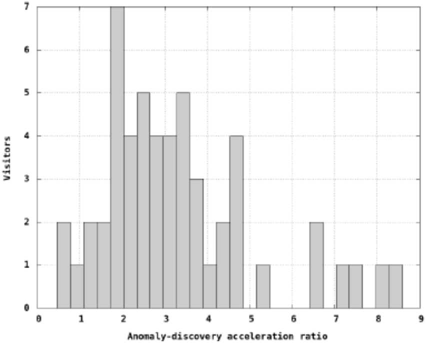
Source	d.f.	F	P
Type	1	142.65	0
Order	1	0	0.9778
File	3	0.3	0.8234
Type*order	1	0.4	0.5334
Type*file	3	1.86	0.154
Order*file	3	0.29	0.8356

(b)

Fig. 5. (a) F-score of human AED using different audio visualization; (b) Three-way ANOVA of the F-score

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.



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

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?

Estimating the Instance PDF: MAP Adaptation

Mixture Gaussian Model

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

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

MAP Adaptation to the p'th instance

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

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

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

$$\vec{\mu}_k^{(p)} = \frac{\sum_{t\in p}\gamma_k(t)\vec{x}_t + \nu\vec{\mu}_k}{\sum_{t\in p}\gamma_k(t) + \nu}$$

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, $R = \mathrm{COV}(\vec{s}_p)$.

Normalize the p^{th} instance

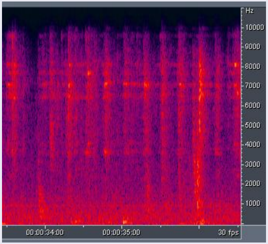
3. Supervectors are then normalized using within-class covariance normalization (WCCN): $\tilde{s}_p = R^{-1}\vec{s}_p$.
4. In the WCCN supervector space $\tilde{s}_p$, intra-class variability is less than inter-class variability, therefore any classifier can work well (e.g., nearest-centroid).

Experimental Test: Non-Speech Acoustic Event Detection

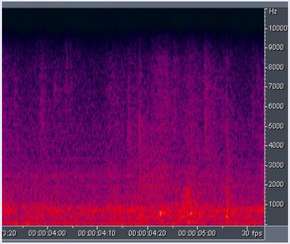
Difficulties

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

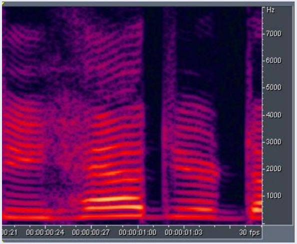
Key Jingle



Footsteps

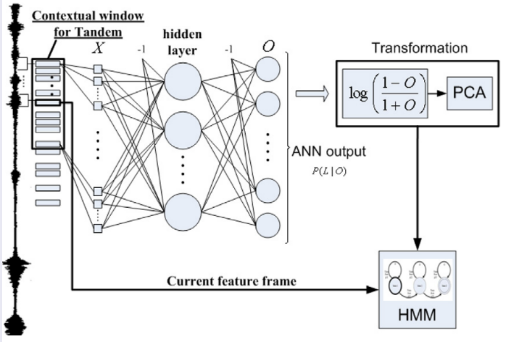


Speech

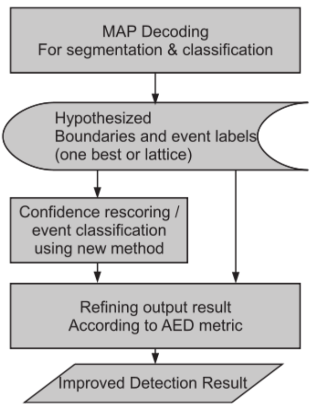


WCCN Supervectors Rescore Tandem MAP Decoding

MAP Decoding Using Tandem NN-GMM-HMM



Rescored Using WCCN Supervectors



Acoustic Event Detection Results

Without Supervectors

CLEAR 2007 AED RESULTS

Inst.	AED Accuracy
AIT	4.4
ITC	23.4
TUT	14.7
UIUC	**36**
STI2R	22.9
UPC	23

With WCCN Supervectors

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

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

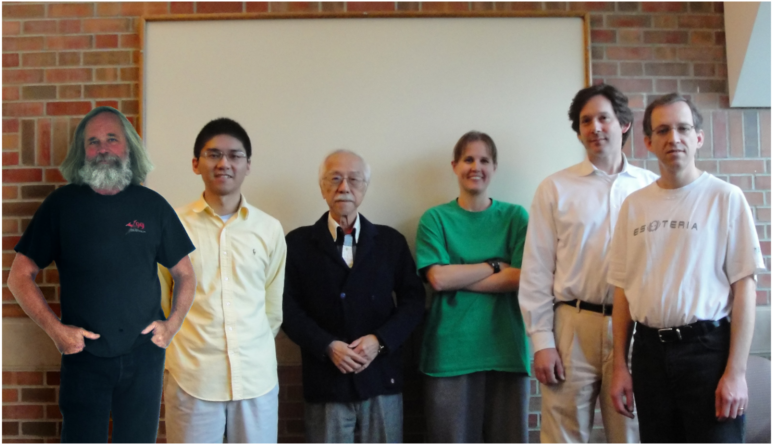
Conclusions: Acoustic Event Detection

- 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

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!