

Visualizing Audio for Anomaly Detection

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

University of Illinois

FODAVA Annual Review, December 8, 2011



Outline

- 1 Research Accomplishments
 - Testbeds
 - Feature Transformations
- 2 Proposed Research
 - Audio Class Discovery
 - Web-Based Multimedia Analytics: Audio Attribute Extraction
- 3 Conclusions

Outline

- 1 Research Accomplishments
 - Testbeds
 - Feature Transformations
- 2 Proposed Research
 - Audio Class Discovery
 - Web-Based Multimedia Analytics: Audio Attribute Extraction
- 3 Conclusions

Testbeds: Milliphone

Goals

- **key metaphor:** 1000 microphones = 1 milliphone
- **research question:** map anomaly to color
- **real-world goal:** analysis of large surveillance installations

Press [7]

The screenshot shows a web browser displaying an article on the Futurity.org website. The browser's address bar shows the URL: www.futurity.org/science-technology/see-the-sounds-audio-as-visual-image/. The Futurity logo is prominently displayed at the top of the page. Below the logo, there are navigation tabs for 'EARTH & ENVIRONMENT', 'HEALTH & MEDICINE', 'SCIENCE & TECHNOLOGY', and 'SOCIETY & CULTURE'. The article title is 'See the sounds: Audio as visual image', and it is categorized under 'SCIENCE & TECHNOLOGY'. The article text begins with 'U. ILLINOIS (US) — New technology lets analysts "see" large amounts of audio data by turning sounds into a visual picture.' Below the text, there are social media sharing buttons for Facebook (12 shares) and Twitter (2 tweets). A large image shows a person interacting with a large, multi-colored data visualization on a wall. The visualization consists of a complex network of lines and nodes in various colors (red, blue, green, yellow). To the right of the article, there are sections for 'DAILY E-NEWS' with an email sign-up form, 'BROWSE BY SCHOOL' with a 'Select School' dropdown, 'FOLLOW FUTURITY' with social media icons for RSS, Facebook, and Twitter, and 'YouTube VIDEOS' with a video player thumbnail.

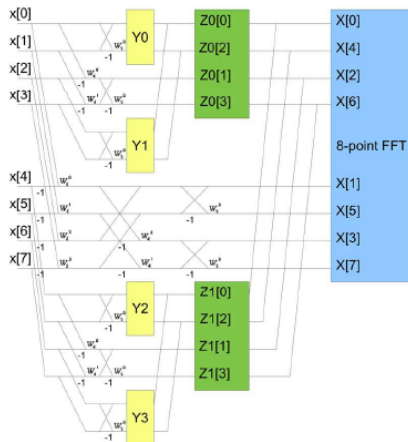
Features: Multiscale Spectrograms

Goals

$$X_n[k] = \sum_{m=0}^{N-1} x[n+m]w[m]e^{-j\frac{2\pi km}{N}}$$

... in less than $\mathcal{O}\{N \log N\}$ per overlapping window [1].

Reusing Butterflies



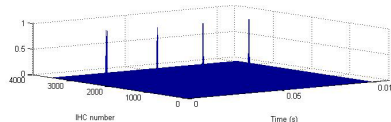
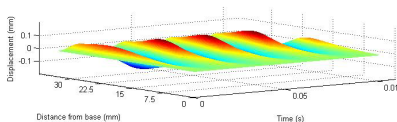
Features: Inner Ear Model

Nonlinear Wave

Propagation in the inner ear is modeled as a nonlinear wave (stiffness varying by place) [5]:

$$K(x) \frac{\partial^2 y}{\partial t^2} = \mu(x) \frac{\partial y}{\partial x} + \epsilon(x) \frac{\partial^2 y}{\partial x^2}$$

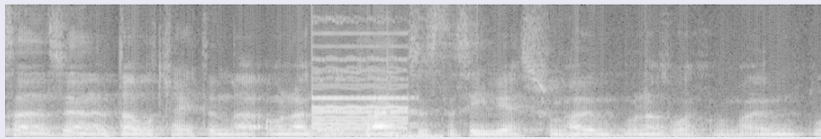
Sparsified Features



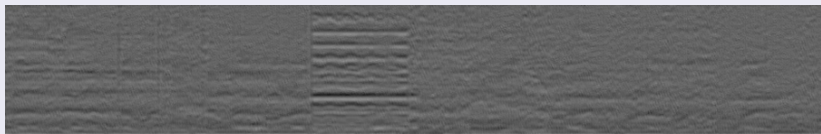
Auditory Brainstem Model

- Spherical Bushy Cells:
 $B_{SB}(f, t) = W \sum_i A(i, f, t) + VB_{SB}(f, t - 1)$, where W is a diagonal input weight matrix and V is a diagonal matrix of forgetting factors.
- Stellate Cells: $B_S(f, t) = W \sum_i A(i, f, t) + VB_S(f, t - 1)$, where W is a matrix whose rows integrate over inputs with similar frequencies, and V is a diagonal matrix of forgetting factors.
- Globular Bushy Cells: $B_{GB}(i, t) = W \sum_f A(i, f, t) + VB_{GB}(i, t - 1)$, where W is a diagonal input weight matrix and V is a diagonal matrix of forgetting factors.
- Multipolar Cells: $B_M(i, t) = W \sum_f A(i, f, t) + vB_M(i, t - 1)$, where W is a matrix whose rows integrate over inputs with similar intensities, and V is a diagonal matrix of forgetting factors.
- Octopus Cells: $B_O(f, t) = W \sum_i A(i, f, t)$, where W is a matrix with all diagonal values equal to 1 and all off diagonal values less than 1.

Experiments Recently Completed: Spectrogram, or . . .



Saliency-Maximizing Features?



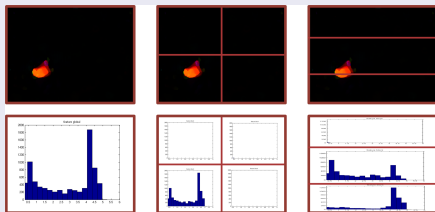
$f(X)$ chosen to maximize the information conveyed ($I(\phi, Y)$) in its highly-salient first visible glimpse ($\phi(f(X))$) [6].

$$f^* = \operatorname{argmax}_f E_{X,Y} \{I(\phi(f(X)), Y)\}$$

Log Likelihood Audiovisual Features

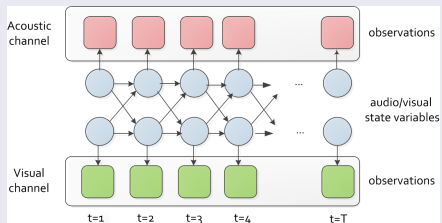
Acoustic event detection system works well [10], was recently extended to audiovisual:

Visible Features: Movement



Optical flow based overlapping spatial pyramid histograms for a footstep event.

Fusion: CHMM



Dynamic Bayesian network representation of a coupled hidden Markov model (significantly reduces error [4]).

Outline

- 1 Research Accomplishments
 - Testbeds
 - Feature Transformations
- 2 Proposed Research
 - Audio Class Discovery
 - Web-Based Multimedia Analytics: Audio Attribute Extraction
- 3 Conclusions

1. Class Discovery

Problem Statement

- The labeled dataset \mathcal{D}_L contains examples drawn from only one of the two classes, say, $y_i = 1$ for $1 \leq i \leq l$
- The unlabeled dataset \mathcal{D}_U contains examples drawn from two classes, say, $y_i \in \{0, 1\}$ for $l + 1 \leq i \leq l + u$
- Both of the two classes have piece-wise-compact distributions, e.g., GMMs

Likelihood Model

$$P(x, y, L) = \sum_{k=1}^K \mathcal{N}(x | \mu_k, \Sigma_k) P(y | k) P(L | y)$$

$$L \in \{\text{labeled, unlabeled}\}$$

Semi-Supervised Learning for Class Discovery

Parameter Set

$$\theta = \{\mu_k, \Sigma_k, P(y|k), P(L|y)\}$$

For class $y = 0$, we set $P(L = \text{labeled} | y = 0) = 0$.

EM Algorithm

E-Step:

$$Q(\theta, \theta^{(i-1)}) = E \left[\log p_{\theta}(\mathcal{Y}_L, \mathcal{X}_L, \mathcal{X}_U) \mid \mathcal{X}_L, \mathcal{X}_U, \theta^{(i-1)} \right]$$

M-Step:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} Q(\theta, \theta^{(i-1)})$$

Experimental Test: Proposed: Audio Events

- Every high-energy event is an audio event
- Labeled data contain a number of known event types. Call these \mathcal{D}_L .
- Unlabeled data are permitted to have events from the known classes, or from one class for which no trained model exists

Experimental Test: Completed: Speech Prosody

- Every syllable is either $y_i = 1$ (prosodic phrase final) or $y_i = 0$ (nonfinal)
- Syllables that are known to be word-nonfinal are therefore also phrase-nonfinal. Call these \mathcal{D}_L .
- Syllables that are word-final may be either phrase-final or phrase-nonfinal. Call these \mathcal{D}_U .
- $x_i = 25$ acoustic features based on pitch, duration, and energy of the syllable

Classification Results: Speech Prosody [3]

- “Class Discovery” case: labeled examples of the phrase-nonfinal class, but no labeled examples of the phrase-final class.
- “Semi-Supervised” case: All syllables followed by silence are automatically considered to be phrase-final, thus providing \mathcal{D}_L with examples from both classes.

	nonbreak	break w/o silence	minor break, silence	major break, silence	total
Chance	100%	0%	0%	0%	79%
Class Discovery	70	83	87	87	73
Semi-Supervised	84	66	77	91	82
Supervised					89

2. Web-Based Multimedia Analytics: Attribute Extraction

Visual analysis of text often leverages parametric semantic spaces, computed using methods such as latent semantic analysis.

Typically a parametric semantic space is computed by creating a feature vector for each document, then transforming the document vector using a transform matrix:

$$\vec{x}_m = W\vec{d}_m. \quad (1)$$

Proposed Audio Attributes

- 1 Spoken Term Detection (STD): Based on our multilingual spoken term detector [9].
- 2 Acoustic Event Detection (AED): Based on our meeting-room AED system [10].
- 3 Gaussian Mixture Model Supervectors (GMMSV): A histogram-like summary of the mapping from audio cepstograms to classes [8].

Research Questions

- Web data are **much** harder than other data we've seen. If STD, AED, GMMSV accuracy suffers, is visualization still useful?
- STD, AED, GMMSV will result in dramatically different projections of any particular audio database.

Outline

- 1 Research Accomplishments
 - Testbeds
 - Feature Transformations
- 2 Proposed Research
 - Audio Class Discovery
 - Web-Based Multimedia Analytics: Audio Attribute Extraction
- 3 Conclusions

Conclusions

- 1 Testbeds
 - 1 Timeliner publicly released
 - 2 Milliphone working in demonstrations
- 2 Features
 - 1 Physiological features in development
 - 2 Saliency-maximizing features in review
 - 3 Likelihood features published and extended
- 3 Proposals
 - 1 Class discovery
 - 2 Audio document summaries

References



David Cohen, Camille Goudeseune, and Mark Hasegawa-Johnson.
Efficient simultaneous multi-scale computation of FFTs.
Technical Report GT-FODAVA-09-01, Georgia Institute of Technology, 2009.



M Hasegawa-Johnson, C Goudeseune, J Cole, H Kaczmarski, H Kim, S King, T Mahrt, JT Huang, X Zhuang, KH Lin, HV Sharma, Z Li, and TS Huang.
Multimodal speech and audio user interfaces for k-12 outreach.
In *Proc. APSIPA*, 2011.



Jui-Ting Huang and Lin-Shan Lee.
Detection of prosodic words in Mandarin Chinese.
In *Proc. SpeechProsody, Dresden*, 2006.



Po-Sen Huang, Xiaodan Zhuang, and Mark Hasegawa-Johnson.
Improving acoustic event detection using generalizable visual features and multi-modality modeling.
In *Proc. ICASSP*, 2011.



Sarah King.
Auditory brainstem model for acoustic landmark detection.
Talk delivered at Prosody/ASR meeting, Oct 2011.



Kai-Hsiang Lin, Xiaodan Zhuang, Camille Goudeseune, Sarah King, Mark Hasegawa-Johnson, and Thomas S. Huang.
Improving faster-than-real-time human acoustic event detection by saliency-maximized audio visualization.
in review.



Steve McGaughey.
See the sounds: Audio as visual image.
futurity.org, 10.1016, 2011.



Xi Zhou, Xiaodan Zhuang, Hao Tang, Mark A. Hasegawa-Johnson, and Thomas S. Huang.
Novel Gaussianized vector representation for improved natural scene categorization.