# Visual Analytics for Audio

*Mark Hasegawa-Johnson, Camille Goudeseune, Kai-Hsiang Lin, David Cohen, Xi Zhou, Xiaodan Zhuang, Kyung-Tae Kim, Hank Kaczmarski and Tom Huang*


ILLINOIS

*NIPS Workshop on Visual Analytics*

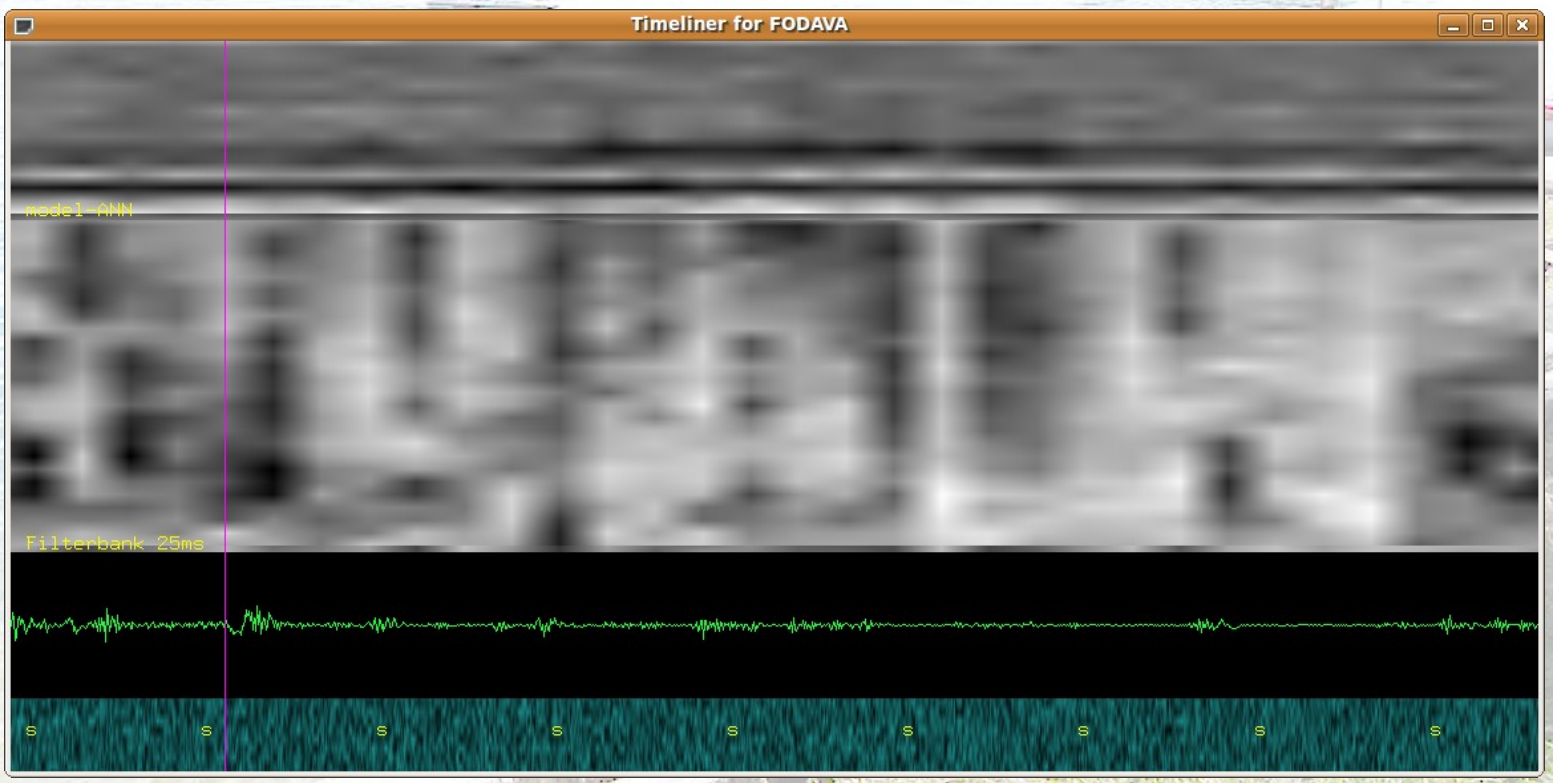
*December 11, 2009*

# Visual Analytics for Audio

- **Visualizing Large Audio Testbeds**
  - **Multi-Day Timeline Audio**
  - **Milliphone = 1000 Microphones**
- **Signal Features**
  - **Multiscale FFT/ Multiscale Filterbanks**
  - **Auditory Salience**
- **Likelihood Features: Acoustic Event Detection**
  - **Minimum Bayes Risk Feature Selection**
  - **Tandem ANN/HMM Event Detection**
  - **UBM/MAP Event Verification**
- **Summary**

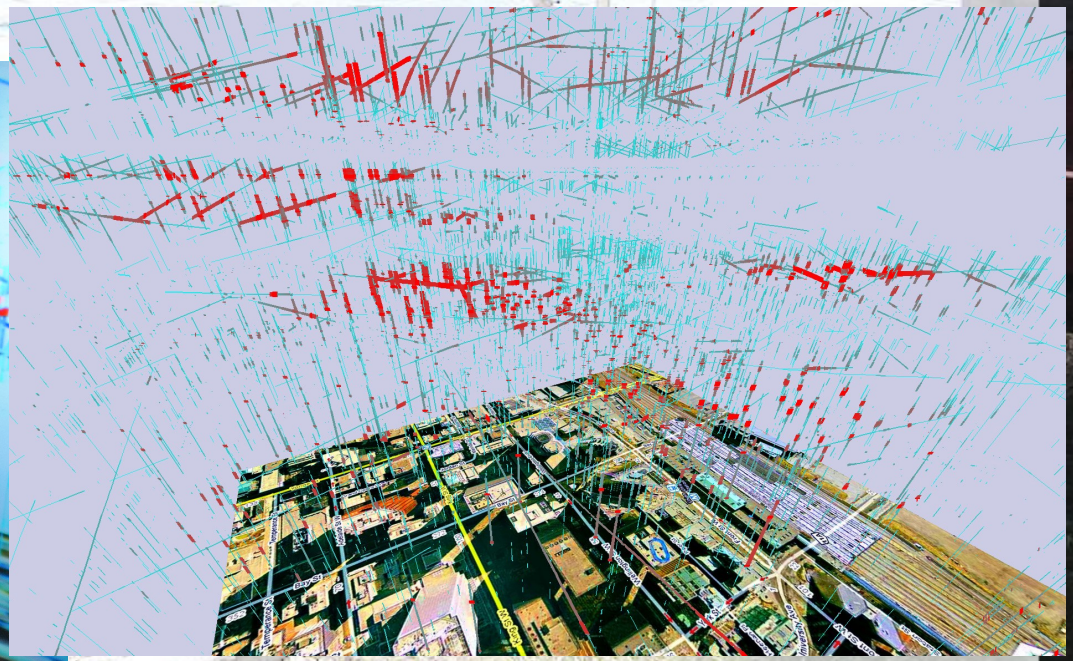
# Testbed #1: Portable Multi-Day Audio Timeliner

- **Dramatis Personae:** Emergency first responders (EFRs)
- **Analysis Object:** One microphone, one month
- **Act 1, Scene 1:** EFRs arrive on scene, download surveillance audio to a handheld
- **Objective:** Event diagnosis, prognosis, & management



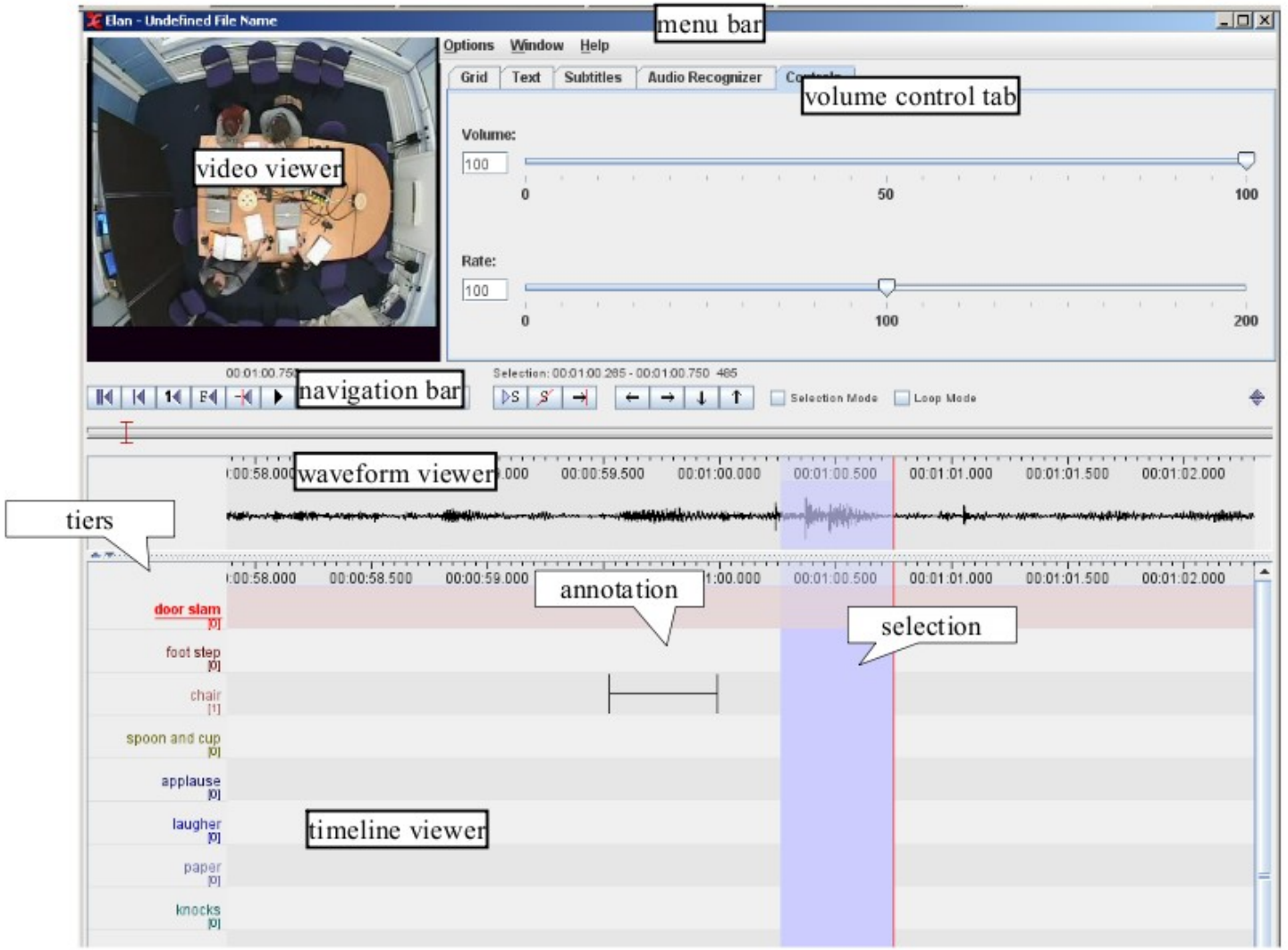
# Testbed #2: 1000 Microphones = One Milliphone

- **Dramatis Personae:** Command center data analysts
- **Analysis Object:** 1000 microphones, 24 hours
- **Act 2, Scene 1:** Analyst in a Virtual Reality Theater (the Beckman CUBE) seeks anomalies in a large dataset
- **Objective:** Find the anomalies



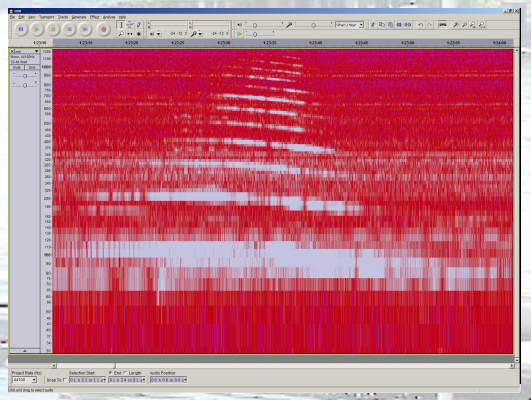
# Dataset #1: Meeting Room Audio

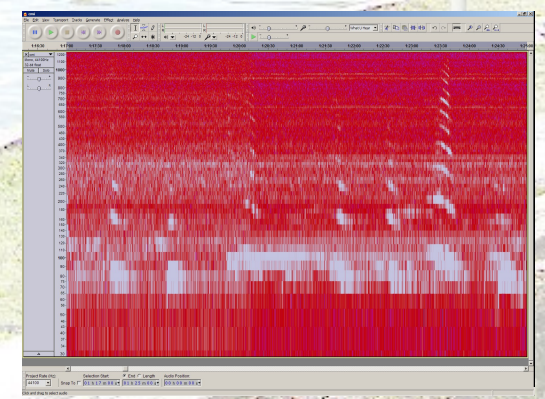
30 annotators, 24 hours of data, 14 acoustic events

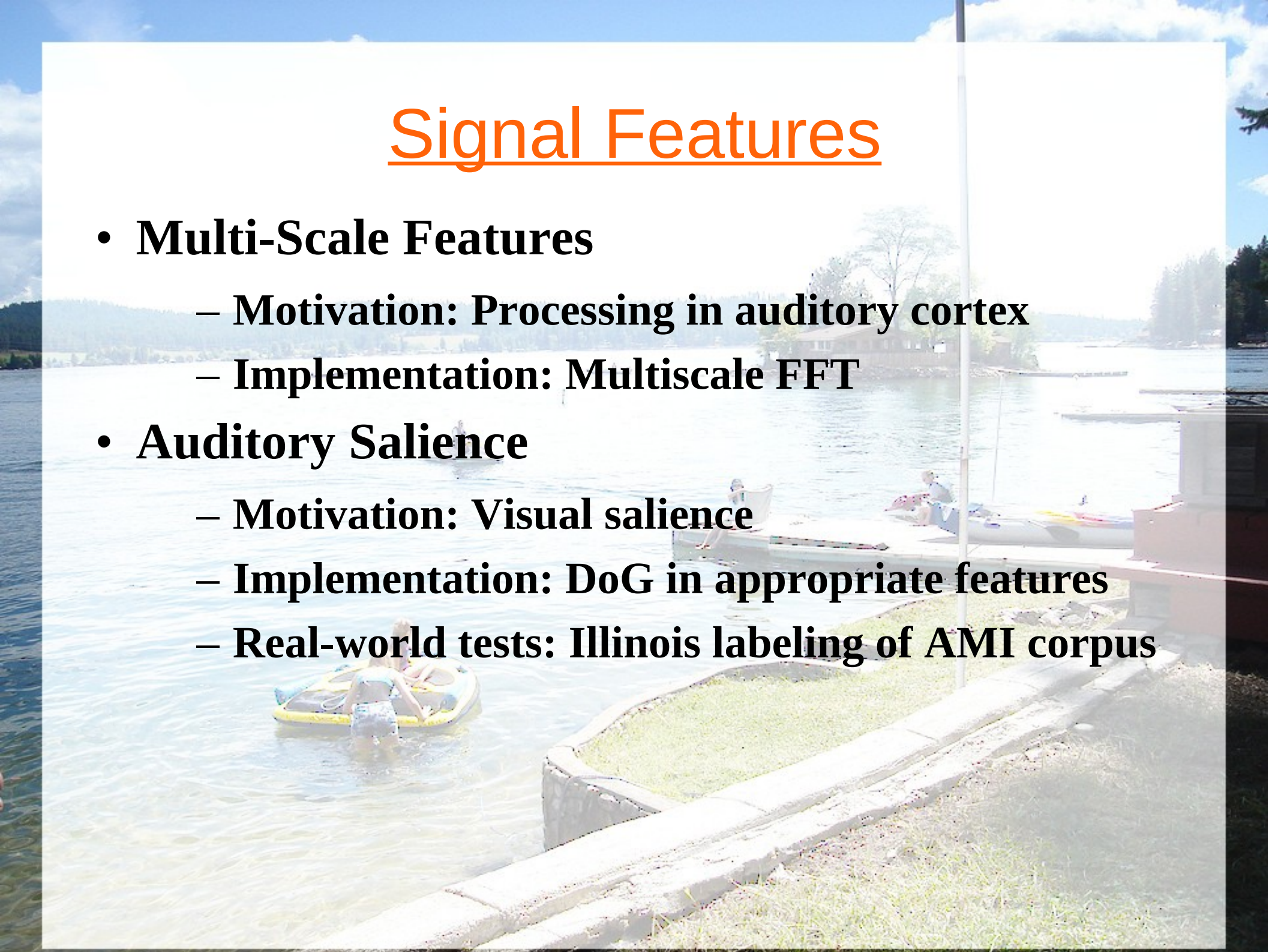


# Dataset #2: Willard Airport

24 hours of audio, `labeled' by commercial airplane takeoff & landing records (inadequate!)



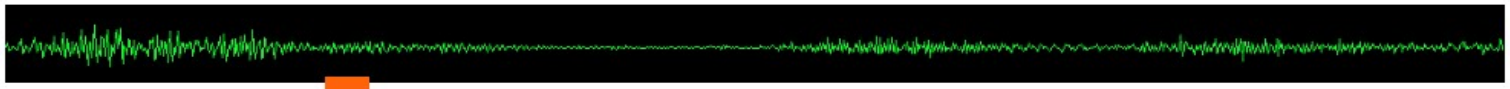


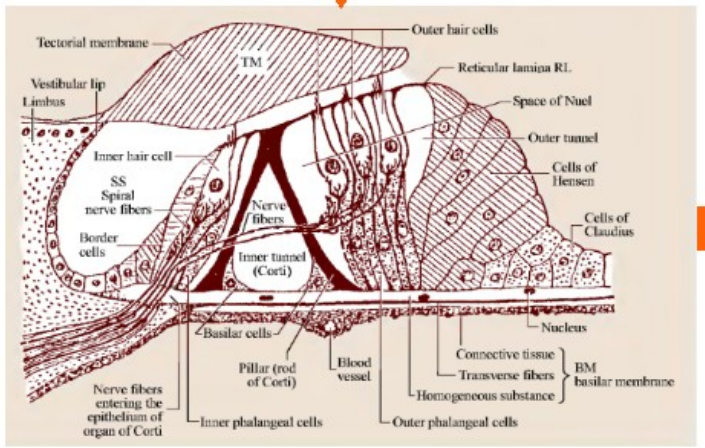
# Signal Features

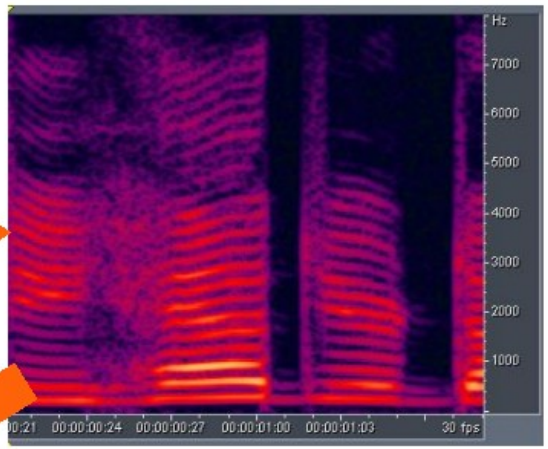
- **Multi-Scale Features**
  - **Motivation: Processing in auditory cortex**
  - **Implementation: Multiscale FFT**
- **Auditory Salience**
  - **Motivation: Visual salience**
  - **Implementation: DoG in appropriate features**
  - **Real-world tests: Illinois labeling of AMI corpus**

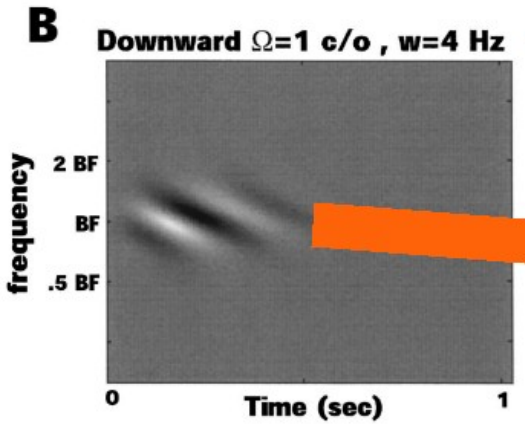
# Motivation: Rate-Scale Features

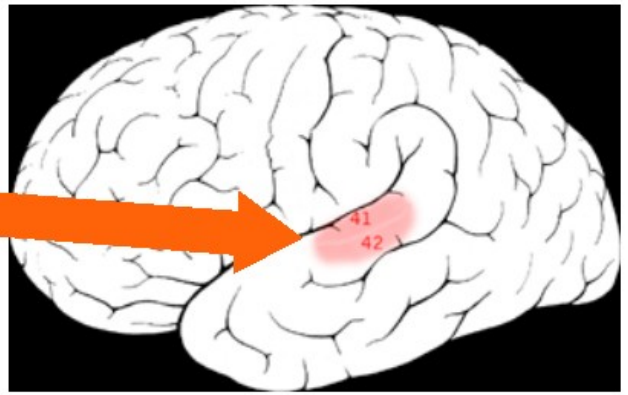
(Carlyon & Shamma, 2003)





# Multiscale FFT: Formulas

(Cohen et al., FODAVA 09-01)

$$X[2k] = F_{N/2}\{x_0[n] + x_1[n]\},\ \ 0 \le k < N/2$$

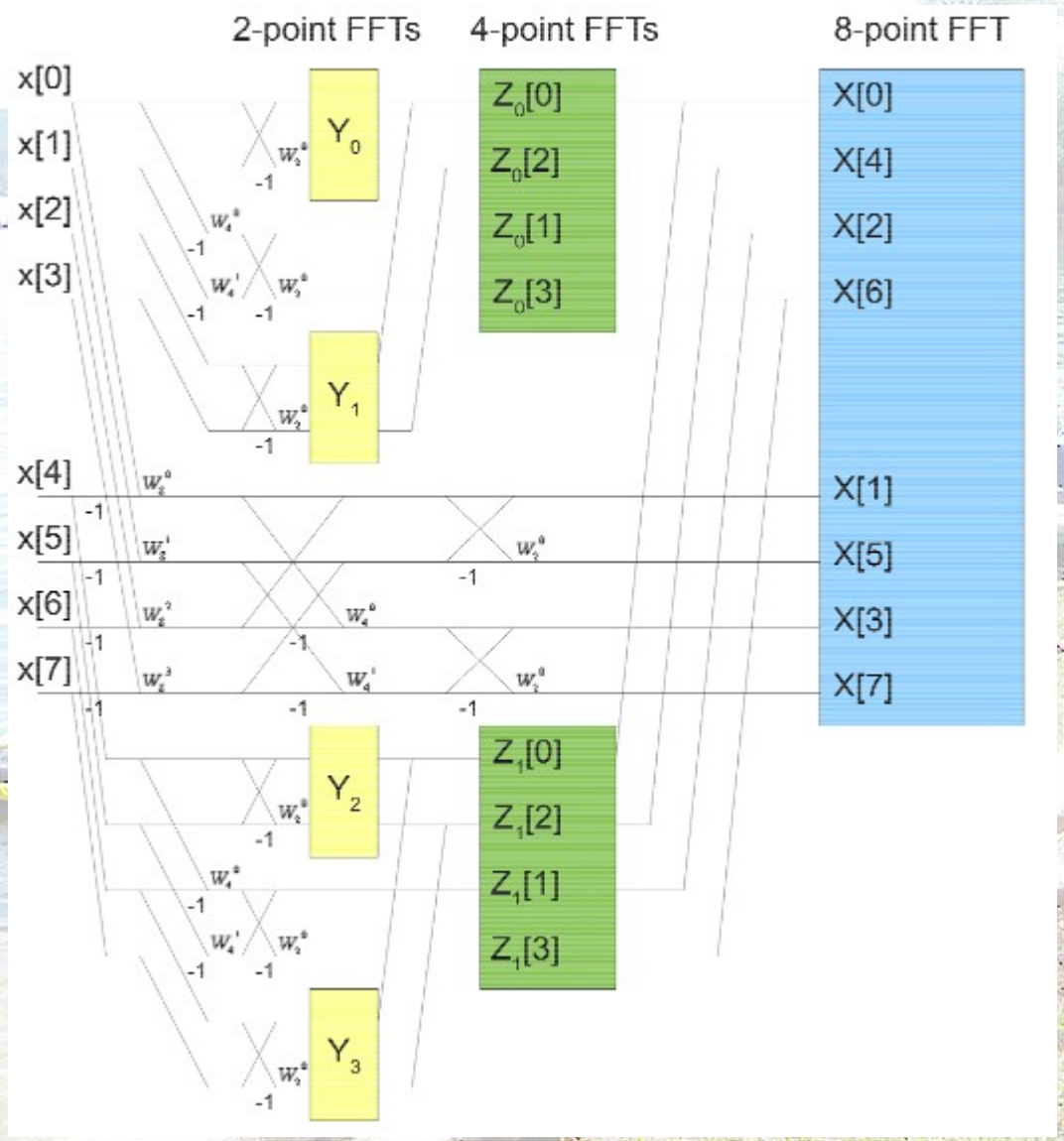
$$X[2k+1] = F_{N/2}\{e^{-j2\pi n/N}(x_0[n] - x_1[n])\},\ \ 0 \le k < N/2$$

$$X[2k] = X_0[k] + X_1[k],\ \ 0 \le k < N/2$$

$$X[2k+1] = X_0[k + ½] - X_1[k + ½],\ \ 0 \le k < N/2$$

# Multiscale FFT: Butterflies

(Cohen et al., FODAVA 09-01)



# Digression: How to Display Audio Events?

**Channel Model: Information in the Visual Channel**

$$\psi(\vec{r}) = h(\vec{r}) * \phi(\vec{r}) + v(\vec{r})$$

$$C = \int \left(1 + \frac{|H(\vec{\vec{\Omega}})|^2 S_\phi(\vec{\Omega})}{S_v(\vec{\Omega})}\right) d\vec{\Omega}$$

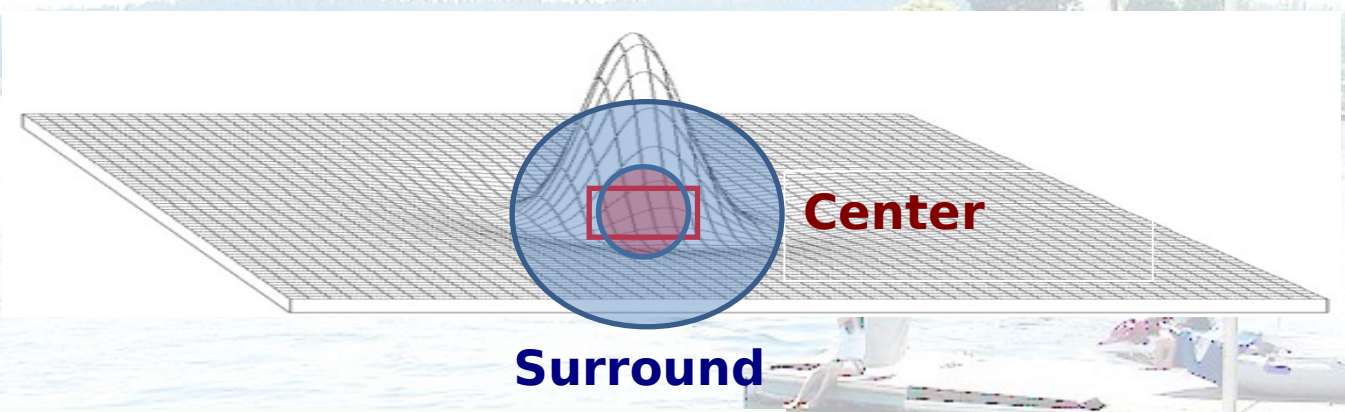
**To Maximize Information Transfer, Choose**

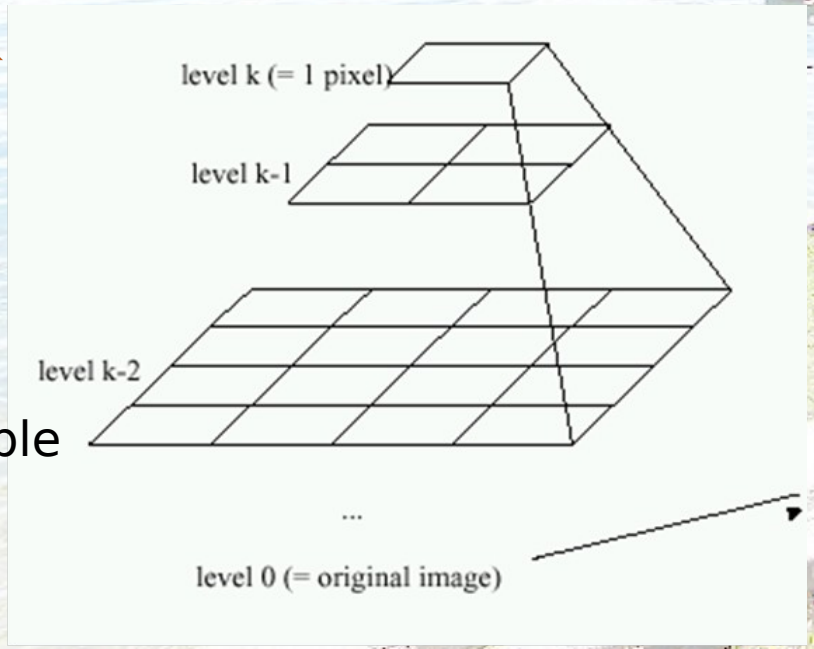
$$\begin{aligned} \Delta\vec{r} &= \mathrm{BW}\left(H(\vec{\Omega})\right) \\ &\approx 2\mathrm{pixels} \\ \Delta\phi(\vec{\Omega}) &\propto \sqrt{\frac{|H(\vec{\Omega})|^2}{S_v(\vec{\Omega})}} \\ &\approx \frac{1}{128}\max\phi(\vec{r}) \end{aligned}$$

# Salience

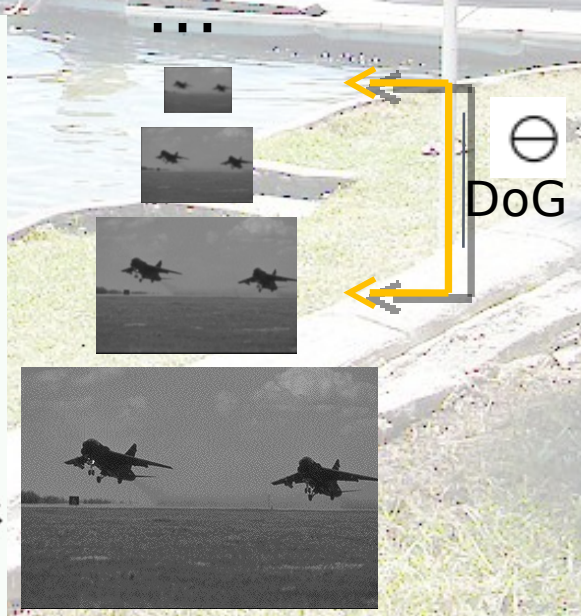
Salient objects/events **POP OUT**

Computer vision: Difference of Gaussians



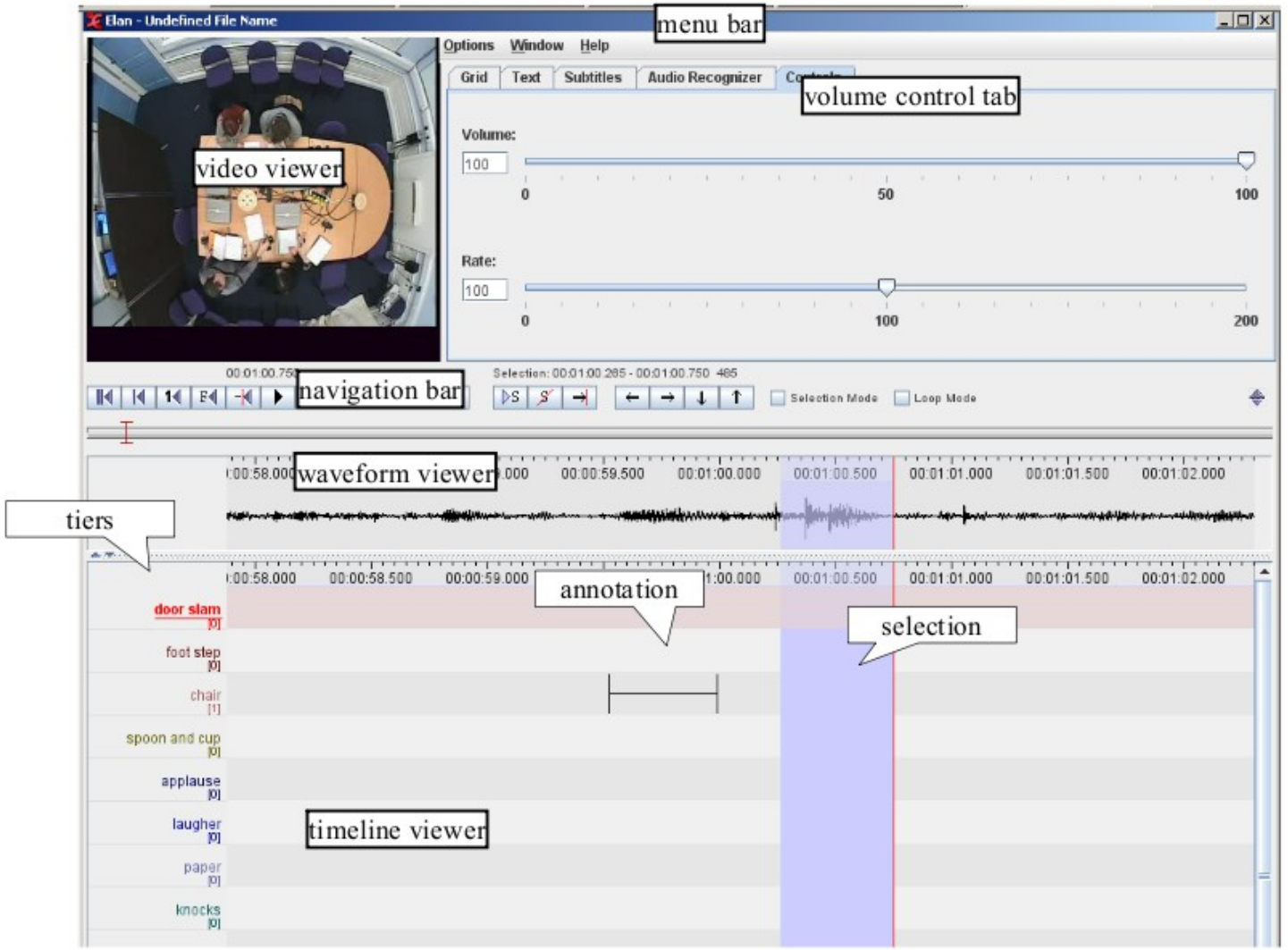

Smooth
then
down-sample



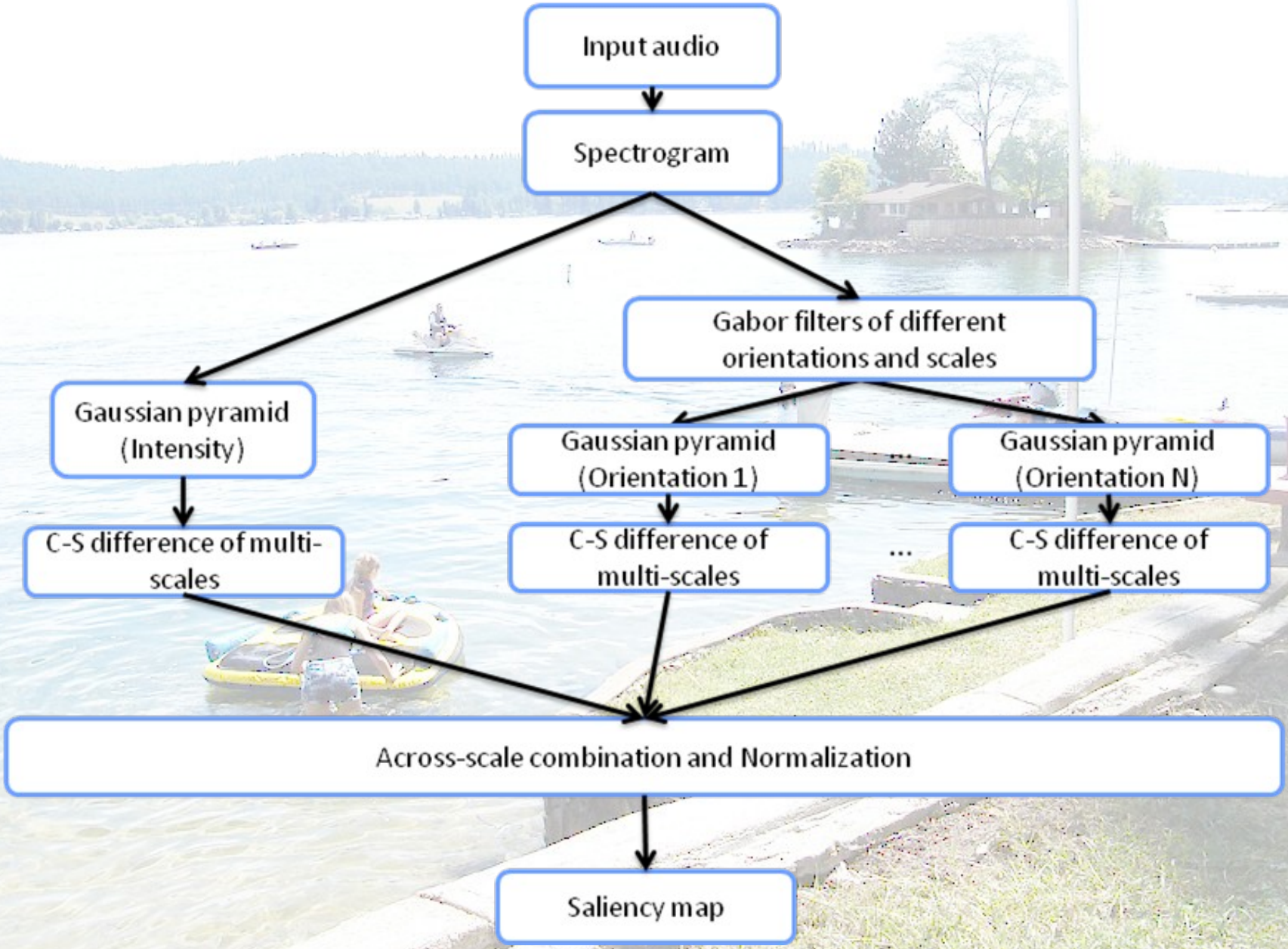
# Audio Salience: Data Annotation

First pass: subjects label “salient” vs. “non-salient”

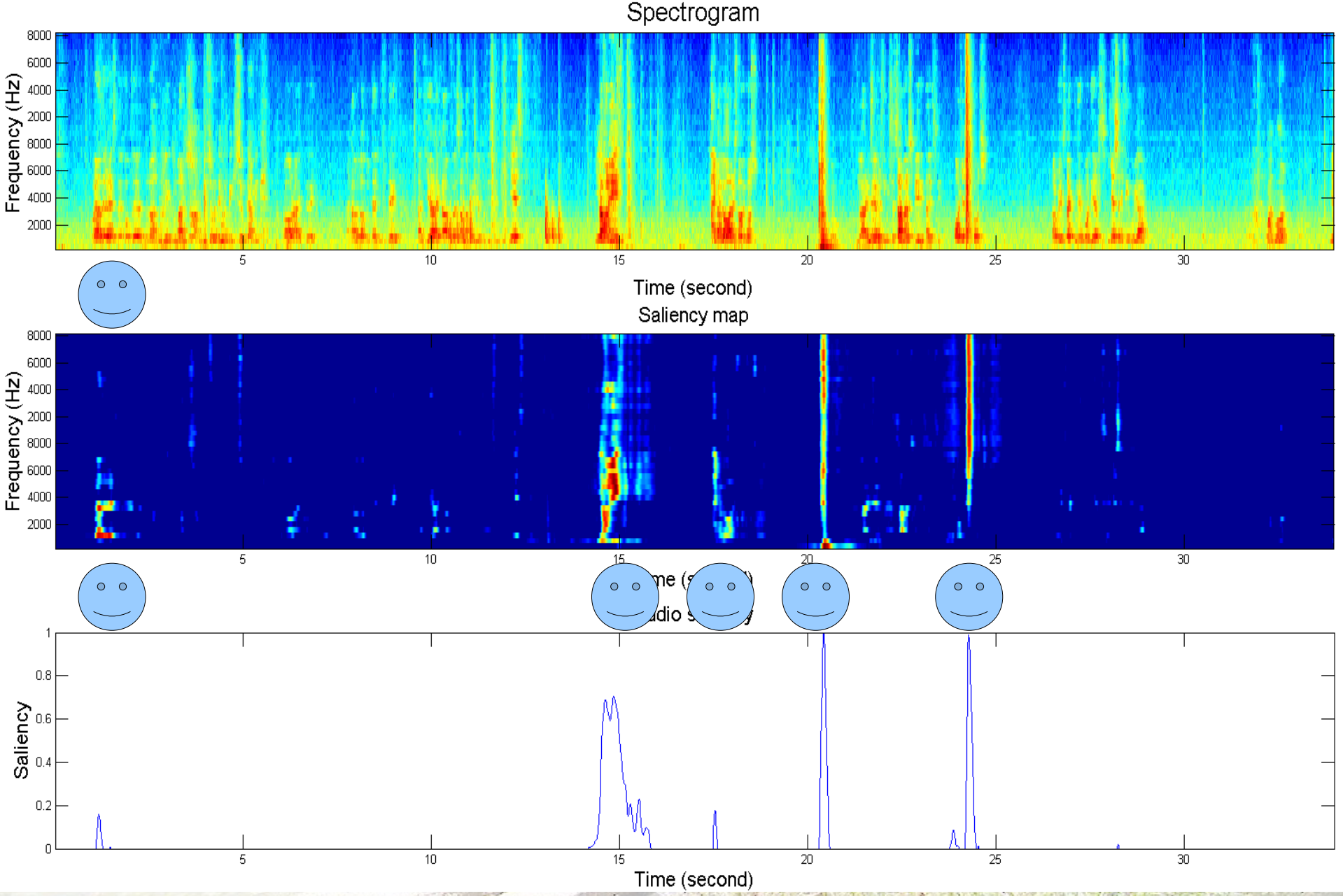
Second pass: subjects identify audio events



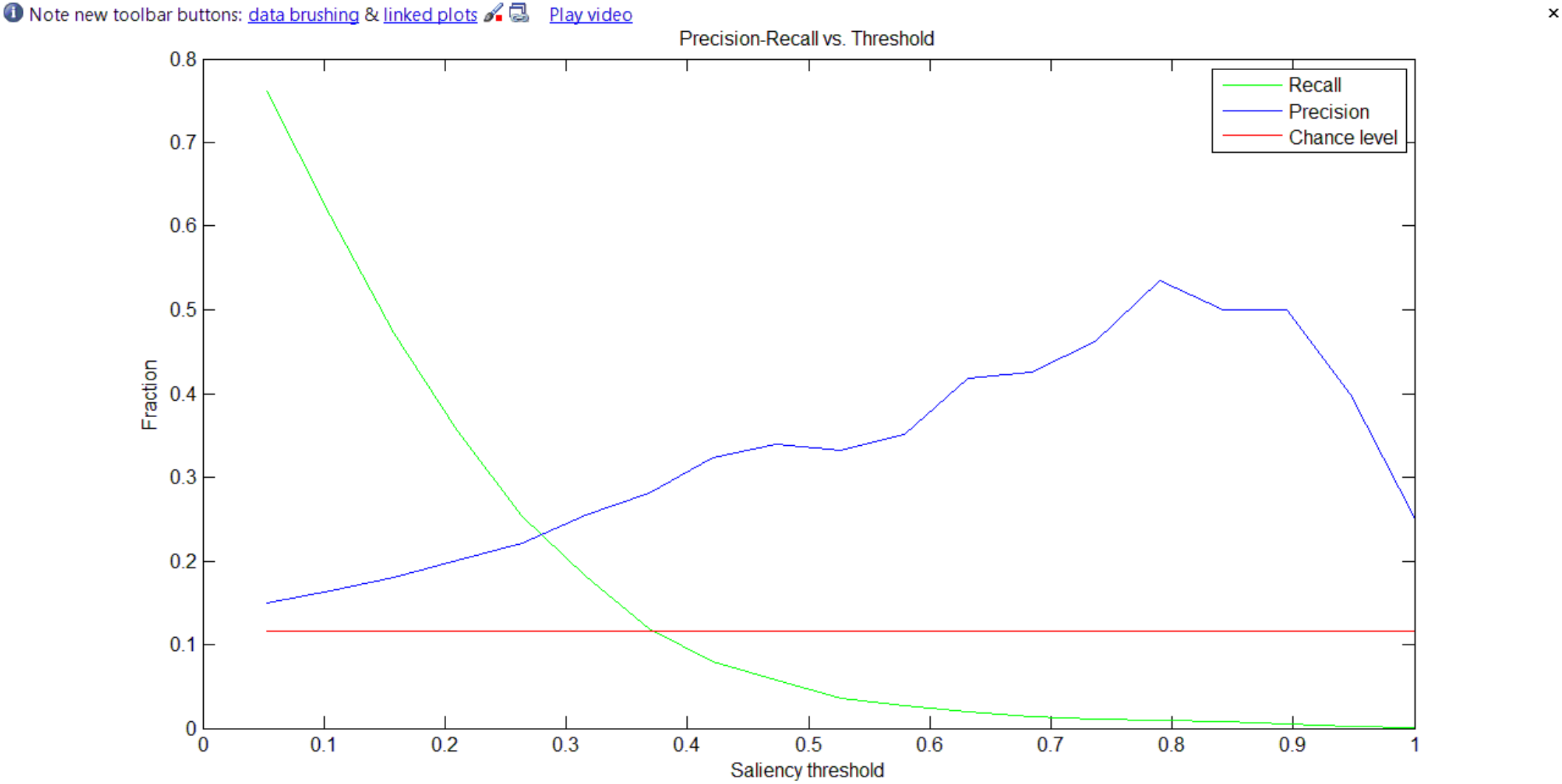
# Audio Salience: Signal Model



# Audio Salience: Sample Results



# Audio Salience EER: ~76%



# Likelihood Features

- **Acoustic Event Detection: Why is it hard?**
  - **Relevant information scattered across multiple scales, multiple (t,f) coordinates**
  - **Low SNR**
- **Minimum Bayes risk feature selection**
- **ANN/HMM hybrid AED**
- **UBM/MAP rescoring**

# Non-Speech Audio Events

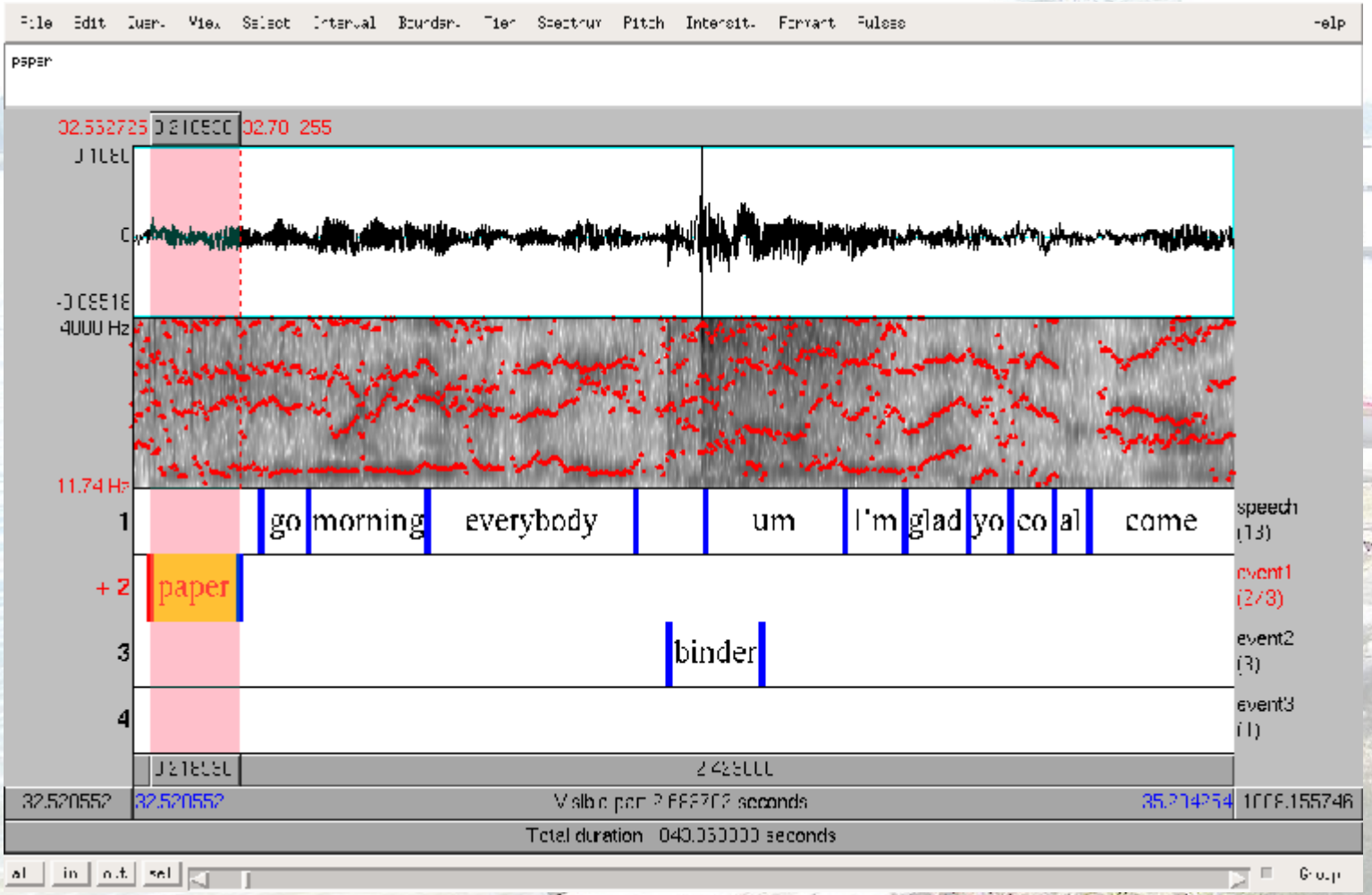
## Motivation

"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)

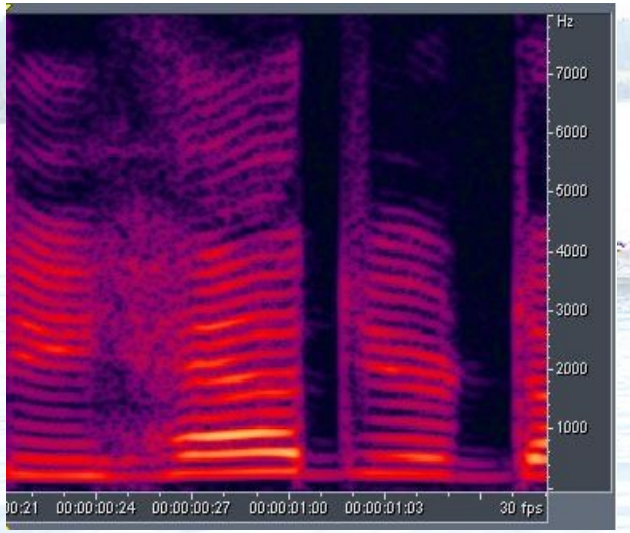
## Difficulties

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

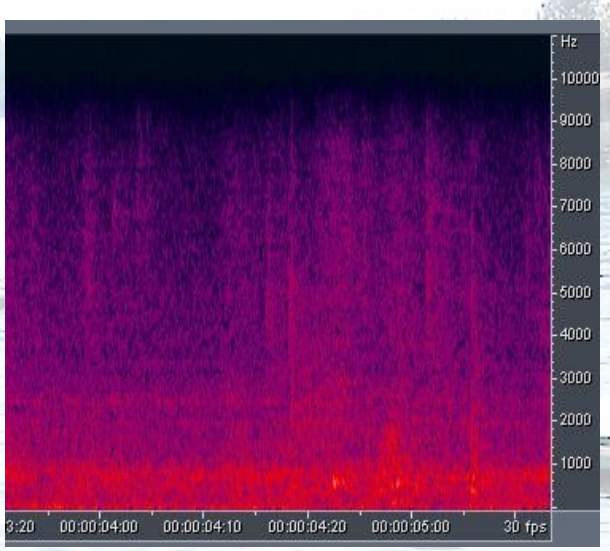
# Difficulty #1: Negative SNR



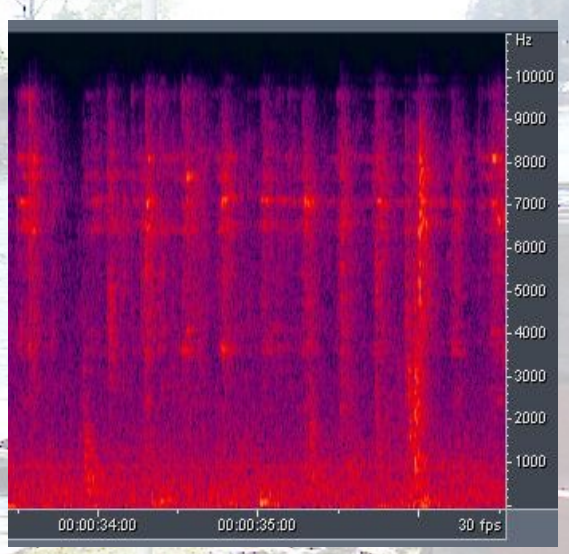
# Difficulty #2: Class-Dependent Scale & Structure




SPEECH




FOOTSTEP




KEY JINGLE

# Discriminative Feature Selection

(Zhuang et al., ICASSP 2008)

- **Problem:** what acoustic features are relevant for detecting non-speech acoustic events?
- **Input:** $(x_i \in \Re^D)$ includes many acoustic features invented for speech processing (MFCC, PLP, energy, ZCR)
- **Output:** $(f_i \in \Re^d)$ 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

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

## 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

(Zhuang et al., ICASSP 2008)

## 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

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

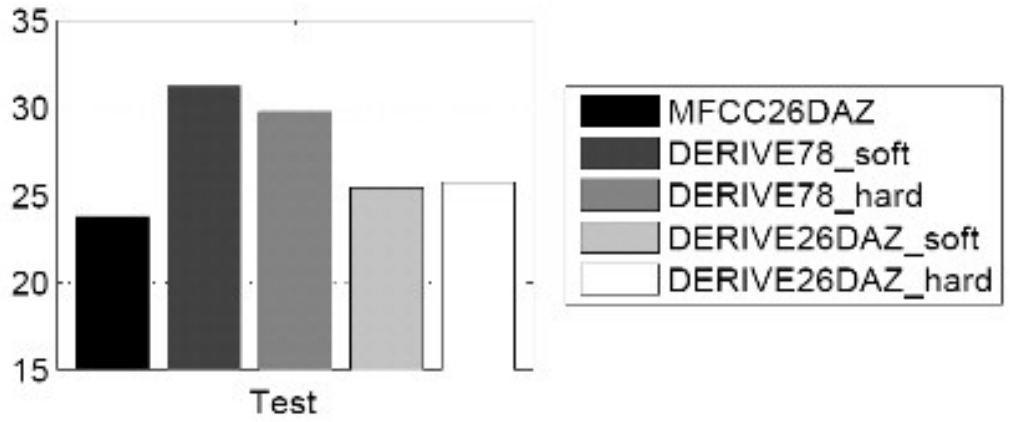
## Soft-Bayes-Error Feature Selection

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

$$\mathcal{F}_S(w_k) = \frac{1}{N} \sum_{i=1}^{N} \text{rank} \left( y_i \left| w_k^T x_i \right. \right)$$

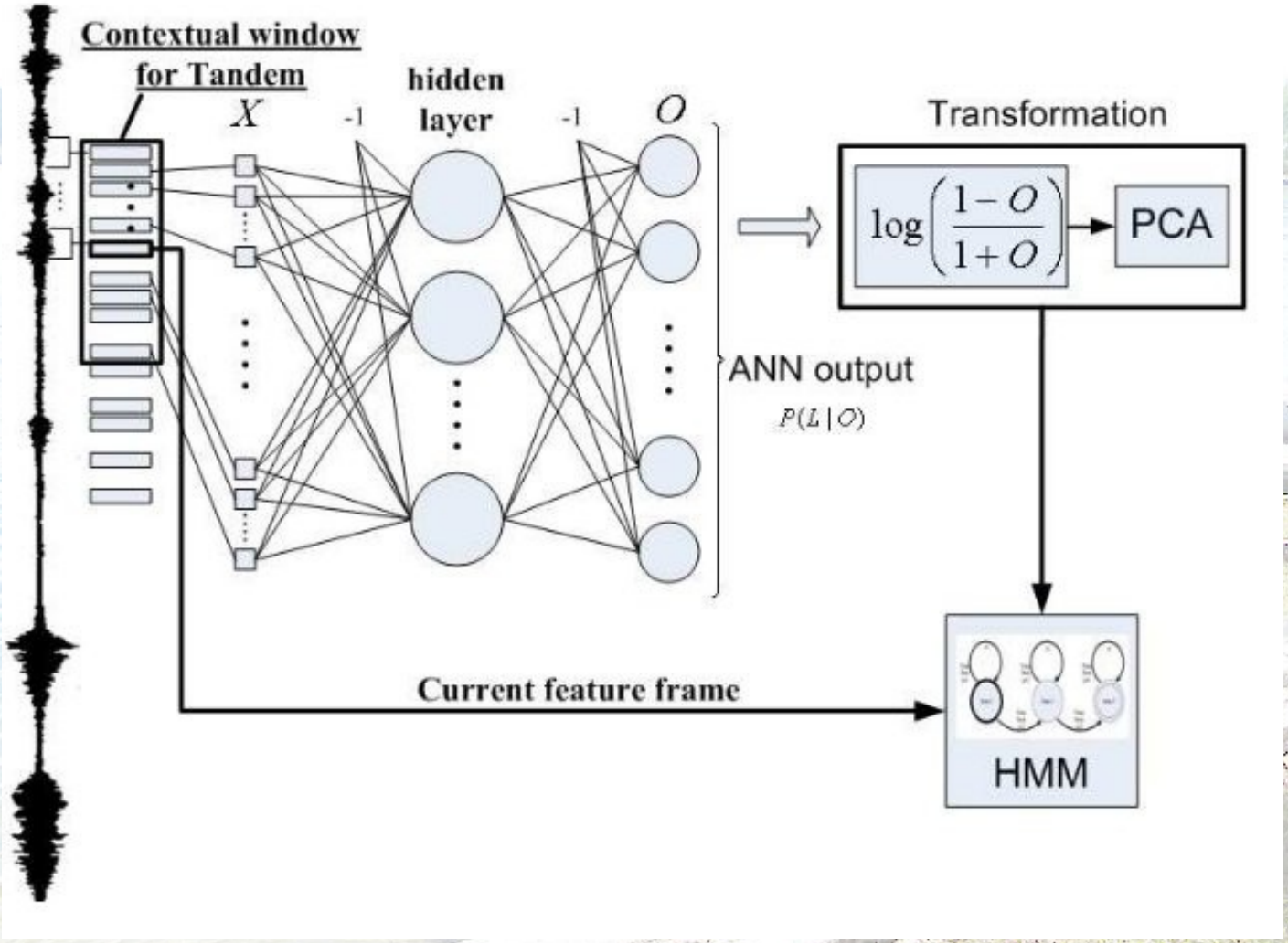
# Acoustic Event Detection Accuracy

(Zhuang et al., ICASSP 2008)




- MFCC26DAZ = 26 Mel-frequency cepstral coefficients + deltas + acceleration
- DERIVE26DAZ = 26 Derived features + deltas + acceleration
- DERIVE78 = 78 Derived features

# Tandem ANN/HMM for AED



# Rescore Using Supervectors

**Mixture Gaussian Model**

$$p(\vec{x}|q) = \sum_k c_{qk} \mathcal{N}(\vec{x}; \vec{\mu}_k, \Sigma_k)$$

**MAP Adaptation to the p'th Detected Segment**

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

**Supervector Representation**

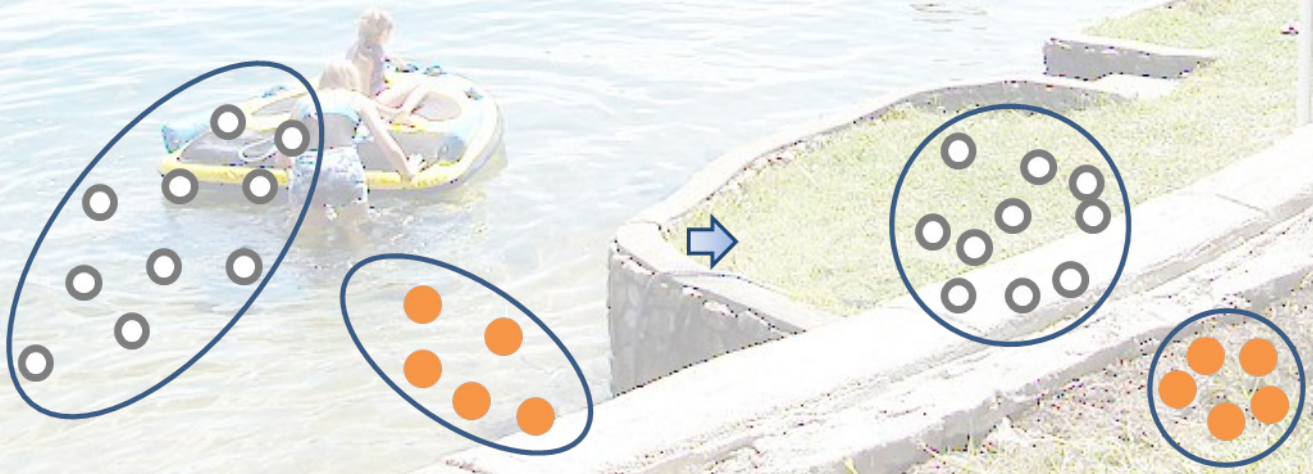
$$\vec{s}_p = \begin{bmatrix} \Sigma_1^{-1/2}(\vec{\mu}_1^{(p)} - \vec{\mu}_1) \\ \vdots \\ \Sigma_K^{-1/2}(\vec{\mu}_K^{(p)} - \vec{\mu}_K) \end{bmatrix}$$

# Within Class Covariance Normalization (WCCN)

- ***Z*-normalize** the supervector to reduce the effect of irrelevant variability using a robust regularized covariance matrix:

$$S=(\gamma S+(1-\gamma)I)$$

- *Z-normalization results in better linear separability*



# AED Accuracy with Tandem & 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 | 21.9 | **39.6** | 49.8 | **63.0** | 29.6 | 15.4 | **36.8** | **54.7** | 41.7 | **26.0** | 38.3 | **38.6** |

Effectiveness of system components

Metric: AED-ACC (%) = weighted multi-class F score

- 'T' = Tandem ANN/HMM recognizer
- 'S' = GMM-Supervector rescoring
- 'Adaboost' = minimum Bayes risk feature selection
- Adaboost+T+S was #1 overall in CLEAR 2007 competition

# Summary & Conclusions

- Audio events occur at scales ranging from 2ms to 20 minutes
  - Efficient computation: multiscale FFT
  - Classifier training: Minimum Bayes risk selection
- Audio events occur at negative SNR
  - Noise compensation: Tandem ANN/HMM
  - Noise compensation: Supervector WCCN
- Visualization is most effective with...
  - Signal features, match the salience of audio & image
  - Log likelihood features; task matching useful but not necessary