Pattern Recognition
Applied to Music Signals

1 **Music Content Analysis**

2 **Classification and Features**

3 **Statistical Pattern Recognition**

4 **Gaussian Mixtures and Neural Nets**

**5** **Singing Detection**

Dan Ellis <dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/muscontent/

Laboratory for Recognition and Organization of Speech and Audio
Columbia University, New York
July 1st, 2003

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Music Content Analysis

1
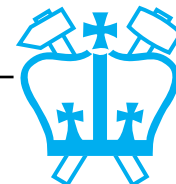
- **Music contains information at many levels**
  - what is it?

- **We'd like to get this information out automatically**
  - fine-level transcription of events
  - broad-level classification of pieces

- **Information extraction can be framed as: pattern classification / recognition or machine learning**
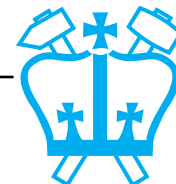 - build systems based on (labeled) training data
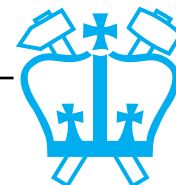
Music analysis

- **What information can we get from music?**
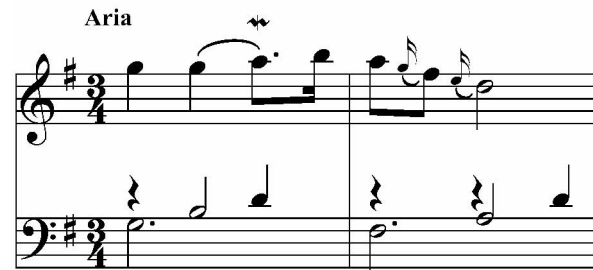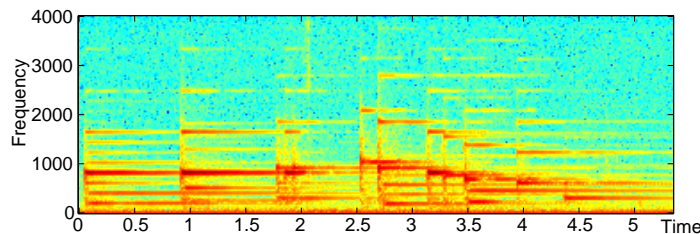
- **Score recovery**
  - extract the 'performance'

- **Instrument identification**

- **Ensemble performance**
  - 'gestalts': chords, tone colors

- **Broader timescales**
  - phrasing & musical structure
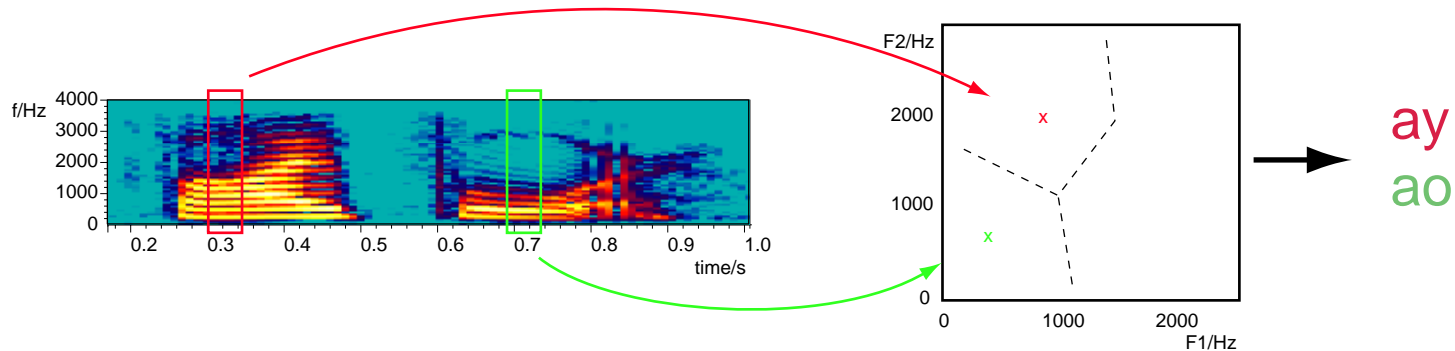 - artist / genre clustering and classification

Outline

1   **Music Content Analysis**

2   **Classification and Features**

- classification
- spectrograms
- cepstra

3   **Statistical Pattern Recognition**

4   **Gaussian Mixtures and Neural Nets**

5   **Singing Detection**

Lab ROSA

Laboratory for the Recognition and
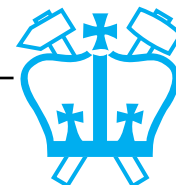Organization of Speech and Audio
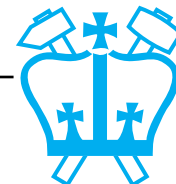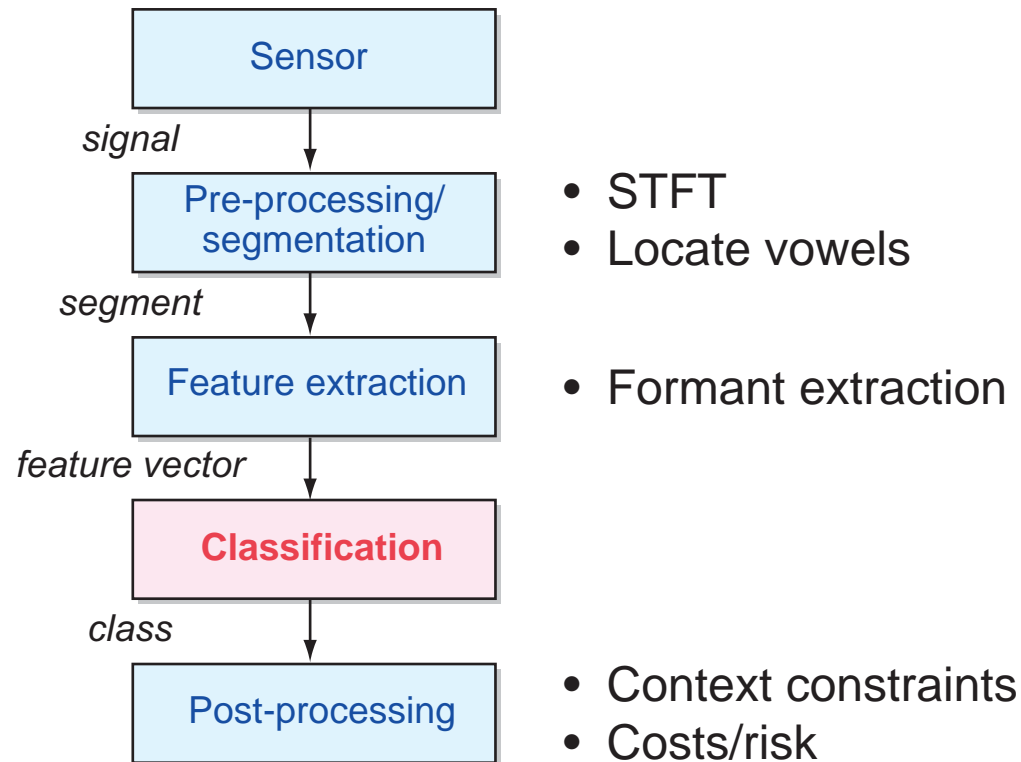
Classification and Features

- **Classification means:**
  **finding categorical (*discrete*) labels
 for real-world (*continuous*) observations**

- **Problems**
  - parameter tuning
  - feature overlap

Classification system parts

Sensor

signal

Pre-processing/ segmentation
- STFT
- Locate vowels

segment

Feature extraction
- Formant extraction

feature vector

Classification

class

Post-processing
- Context constraints
- Costs/risk

- **Right features are critical**
  - place upper bound on classifier
  - should make important aspects visible
  - invariance under irrelevant modifications

Lab ROSA
Laboratory for the Recognition and
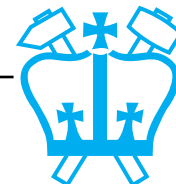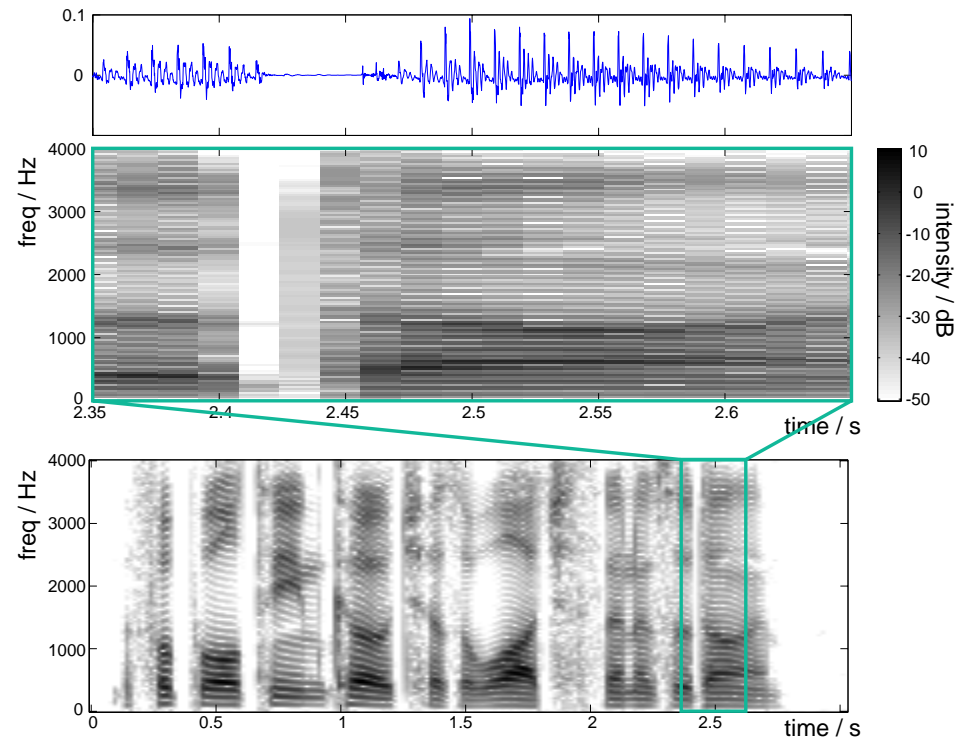Organization of Speech and Audio
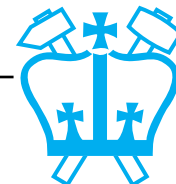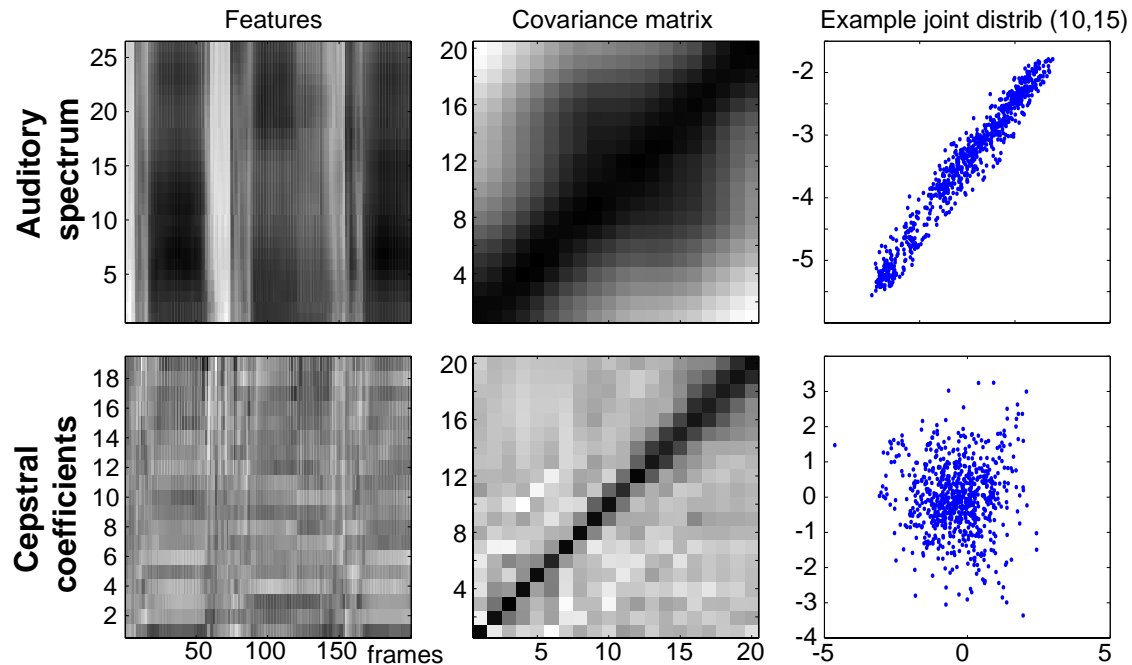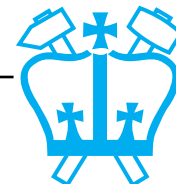
The Spectrogram

- **Short-time Fourier transform:**

$$X[k, m] = \sum_{n=0}^{N-1} x[n] \cdot w[n - mL] \cdot \exp-j\left(\frac{2\pi k(n - mL)}{N}\right)$$

- **Plot STFT $X[k, m]$ as a grayscale image:**

Cepstra

- **Spectrograms are good for visualization; Cepstrum is preferred for classification**

  - dct of STFT: $c_k = \mathrm{idft}(\log|X[k,m]\|)$

- **Cepstra capture coarse information in fewer dimensions with less correlation:**

Features   Covariance matrix   Example joint distrib (10,15)

Auditory spectrum

Cepstral coefficients

frames

Outline

1 **Music Content Analysis**

2 **Classification and Features**

3 **Statistical Pattern Recognition**
- Priors and posteriors
- Bayesian classifier

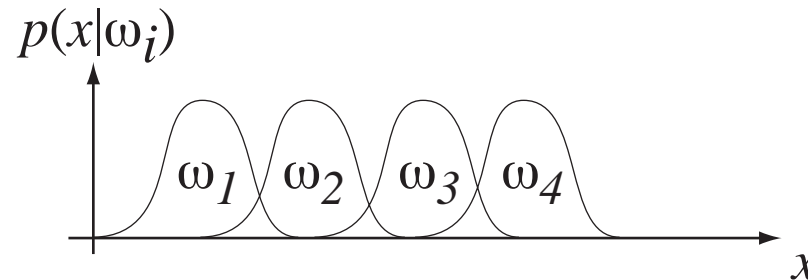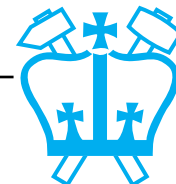**4** **Gaussian Mixtures and Neural Nets**

5 **Singing Detection**

Lab ROSA
Laboratory for the Recognition and
Organization of Speech and Audio
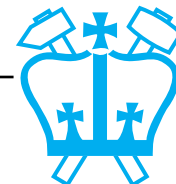
Statistical Pattern Recognition

- **Observations are random variables whose distribution depends on the class:**

Class   Observation

ω_i x

⬤ ⟶ ○ ⟶ $\Pr(\omega_i | x)$

(hidden) $p(x|\omega_i)$

discrete   *continuous*
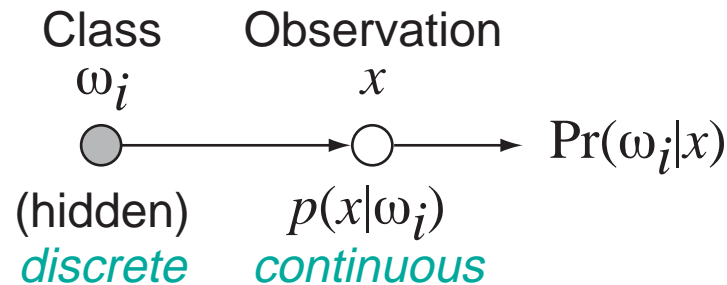
- **Source distributions** $p(x|\omega_i)$

  - reflect variability in feature

  - reflect noise in observation

 - generally have to be estimated from data (rather than known in advance)

  $p(x|\omega_i)$

  ω_1 ω_2 ω_3 ω_4

  x

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Priors and posteriors

- **Bayesian inference can be interpreted as updating prior beliefs with new information, x:**

Bayes' Rule: Likelihood

$$Pr(\omega_i) \cdot \frac{p(x|\omega_i)}{\sum_j p(x|\omega_j) \cdot Pr(\omega_j)} = Pr(\omega_i|x)$$

Prior
probability   Posterior
probability

'Evidence' = p(x)

- **Posterior is prior scaled by likelihood & normalized by evidence (so Σ(posteriors) = 1)**

- **Objection: priors are often unknown**
  - but omitting them amounts to assuming they are all equal
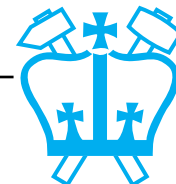
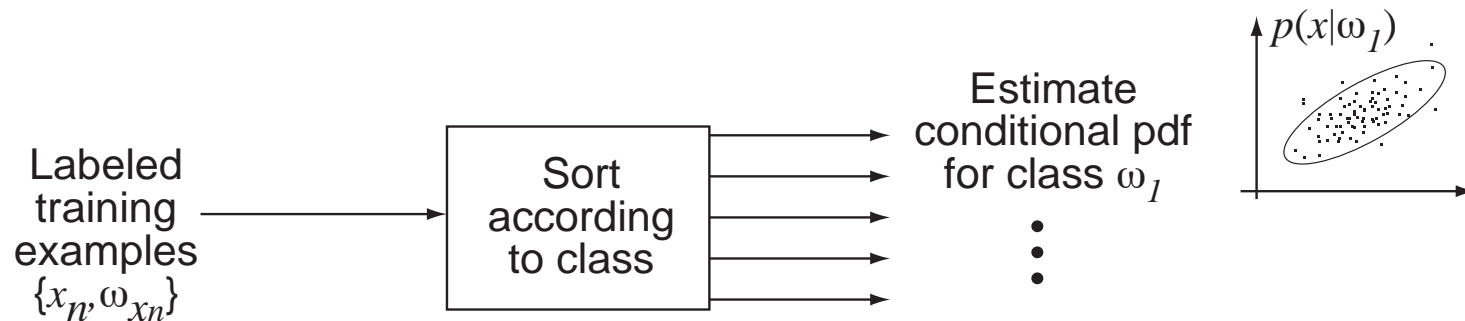# Bayesian (MAP) classifier

- **Optimal classifier is** $\hat{\omega} = \underset{\omega_i}{\mathrm{argmax}}\; Pr(\omega_i | x)$

  but we don't know $Pr(\omega_i | x)$

- **Can model conditional distributions**
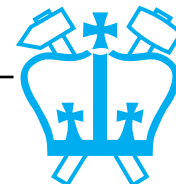 $p(x | \omega_i)$ **then use Bayes' rule to find MAP class**

Labeled training examples $\{x_n, \omega_{xn}\}$ → Sort according to class → Estimate conditional pdf for class ω_1 → $p(x|\omega_1)$

- **Or, can model directly e.g. train a neural net to map from inputs x to a set of outputs $Pr(\omega_i)$**

  - discriminative model
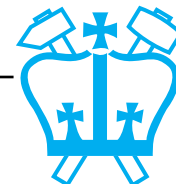
Outline

1 **Music Content Analysis**

2 **Classification and Features**

3 **Statistical Pattern Recognition**

4 **Gaussian Mixtures and Neural Nets**
- Gaussians
- Gaussian mixtures
- Multi-layer perceptrons (MLPs)
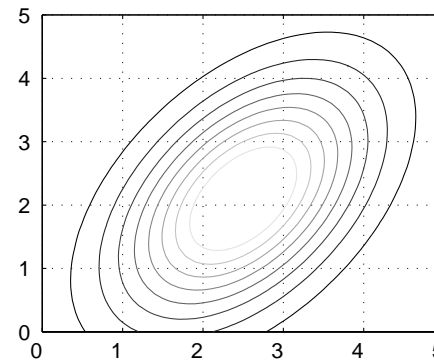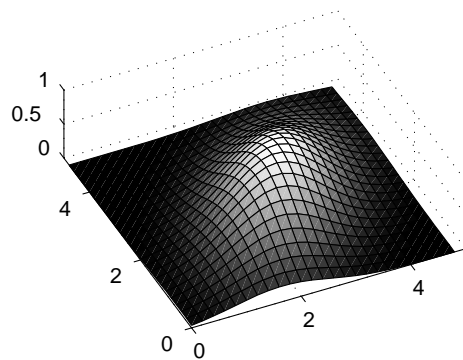- Training and test data

5 **Singing Detection**

LabROSA
Laboratory for the Recognition and
Organization of Speech and Audio

- **Gaussians** as parametric distribution models:

$$p(\mathbf{x} | \omega_i) = \frac{1}{(\sqrt{2\pi})^d |\Sigma_i|^{1/2}} \cdot \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x} - \mathbf{\mu}_i) \right]$$

- **Described by** d **dimensional mean vector** μ_i **and** d x d **covariance matrix** Σ_i

- **Classify by maximizing log likelihood i.e.**

$$\underset{\omega_i}{\mathrm{argmax}} \left[-\frac{1}{2}(\mathbf{x} - \mathbf{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x} - \mu_i) - \frac{1}{2}\log|\Sigma_i| + \log Pr(\omega_i) \right]$$

Lab
ROSA
Laboratory for the Recognition and
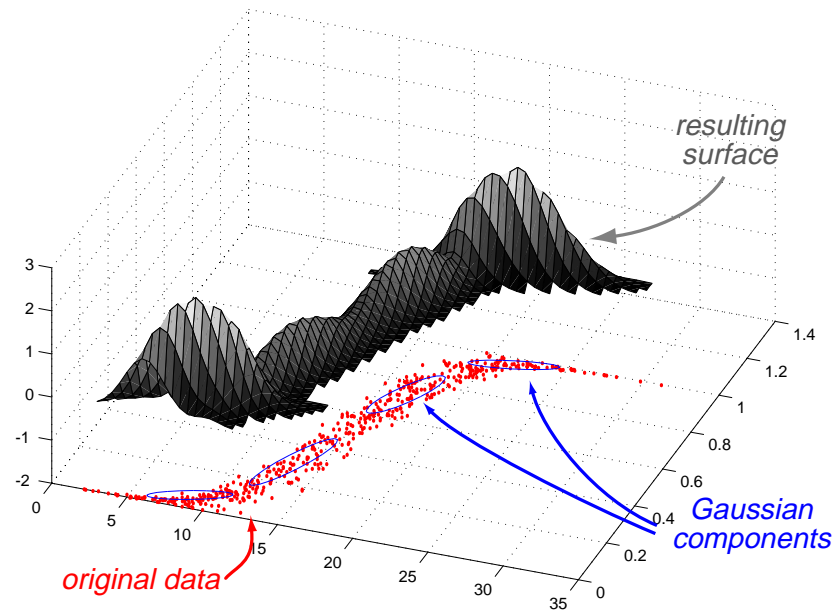Organization of Speech and Audio
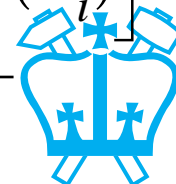
Gaussian Mixture models (GMMs)

- **Weighted sum of Gaussians can fit any PDF:**
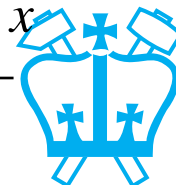
  i.e. $$p(x) \approx \sum_k c_k p(x|m_k)$$

  weights c_k

  Gaussians $p(x|m_k)$
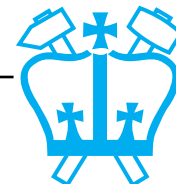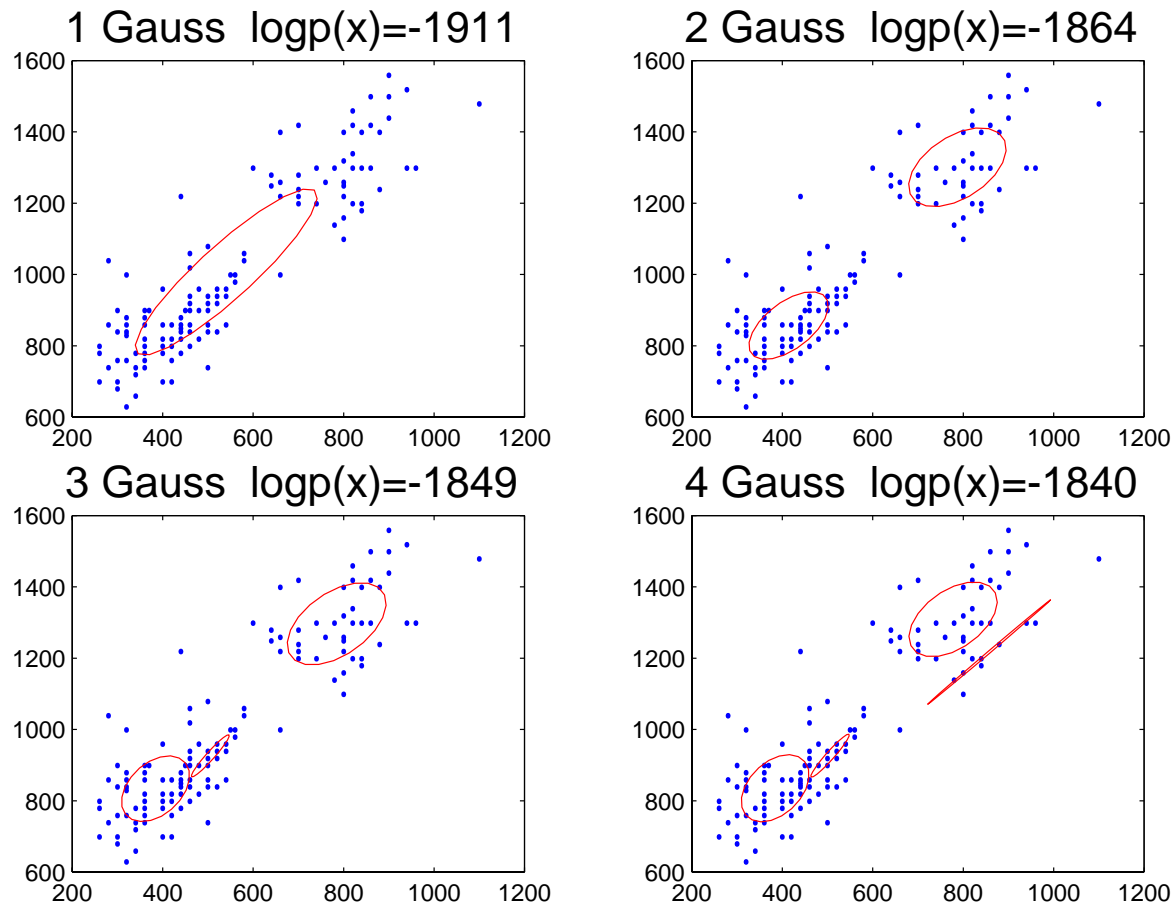
  - each observation from random single Gaussian?

resulting surface

Gaussian components

original data

- **Find c_k and m_k parameters via EM**

  - easy if we knew *which m_k generated each x*

GMM examples

- **Vowel data fit with different mixture counts:**

1 Gauss logp(x)=-1911

2 Gauss logp(x)=-1864

3 Gauss logp(x)=-1849

4 Gauss logp(x)=-1840

Neural networks

- **Don't model distributions** $p(x|\omega_i)$ **,**
  instead, model posteriors $Pr(\omega_i|x)$

- **Sums over nonlinear functions of sums**
  → large range of decision surfaces

- **e.g. Multi-layer perceptron (MLP)**
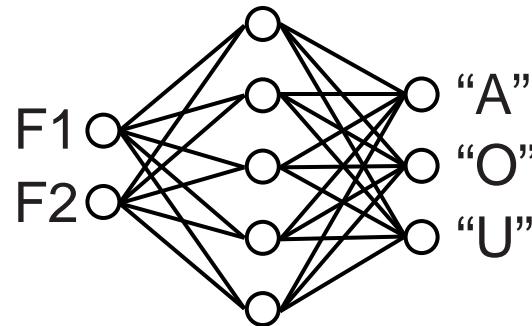 **with 1 hidden layer:**

$$y_k = F[\sum_j w_{jk} \cdot F[\sum_j w_{ij} x_i]]$$

Input layer   Hidden layer   Output layer
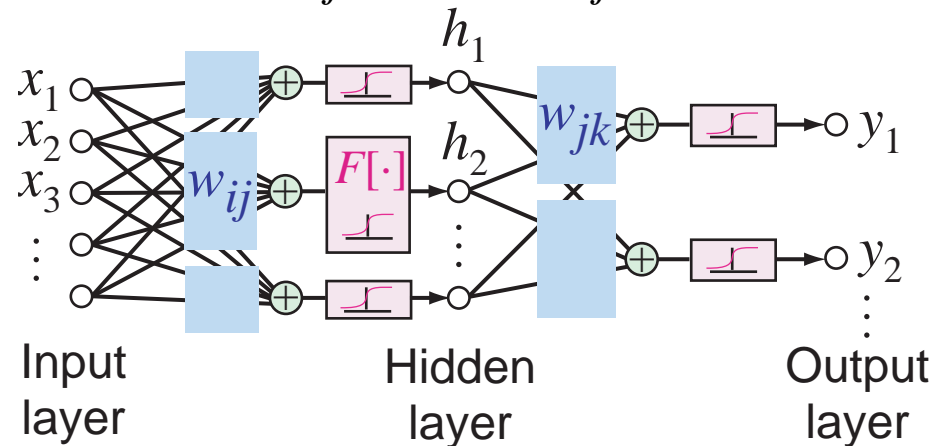
- **Train the weights** w_{ij} **with back-propagation**

Neural net example

- **2 input units (normalized F1, F2)**
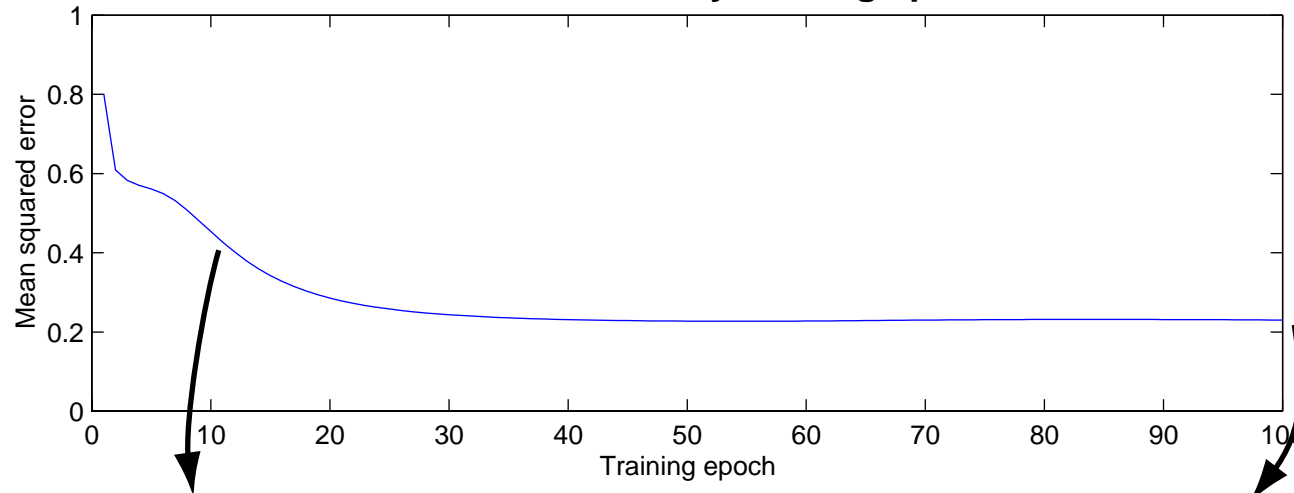
- **5 hidden units, 3 output units ("U", "O", "A")**

- **Sigmoid nonlinearity:**

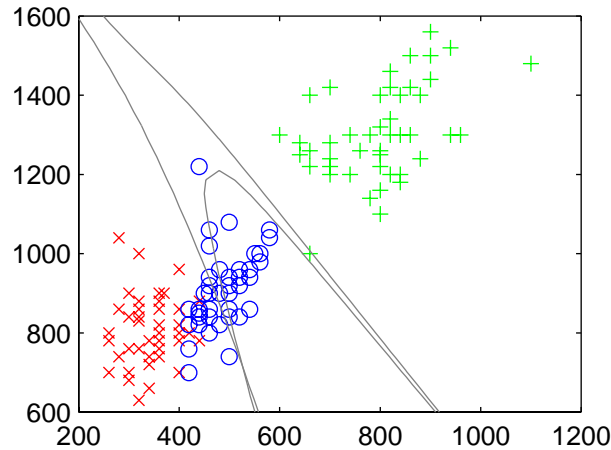$$F[x] = \frac{1}{1 + e^{-x}} \quad \Rightarrow \quad \frac{dF}{dx} = F(1 - F)$$
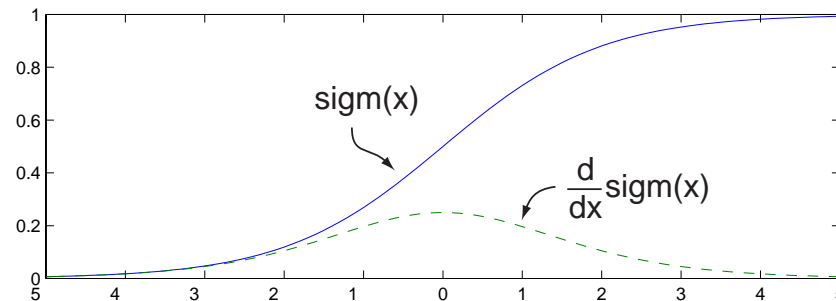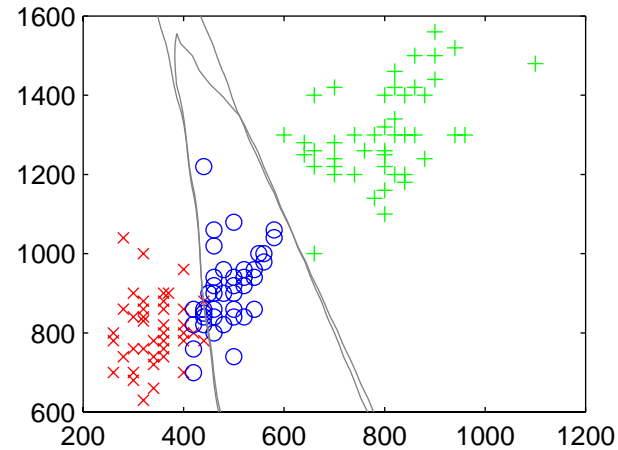
Neural net training

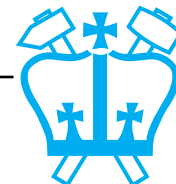## 2:5:3 net: MS error by training epoch

Contours @ 10 iterations
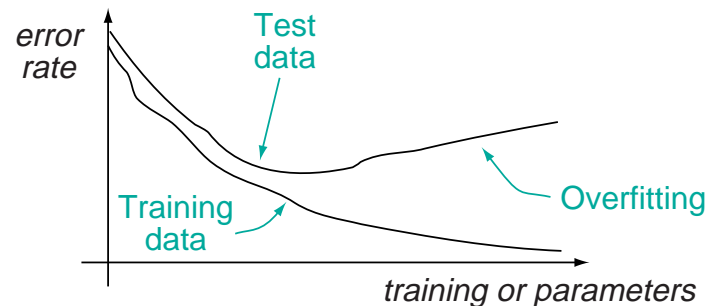

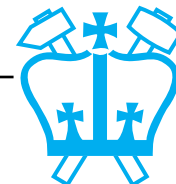## Contours @ 100 iterations

example...

Aside: Training and test data

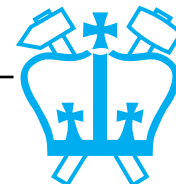- **A rich model can learn *every* training example (*overtraining*)**

- **But, goal is to classify new, unseen data i.e. *generalization***
  - sometimes use 'cross validation' set to decide when to stop training

- **For evaluation results to be meaningful:**
  - don't test with training data!
  - don't *train* on *test* data (even indirectly...)

Outline

1 **Music Content Analysis**

2 **Classification and Features**

3 **Statistical Pattern Recognition**

4 **Gaussian Mixtures and Neural Nets**

5 **Singing Detection**

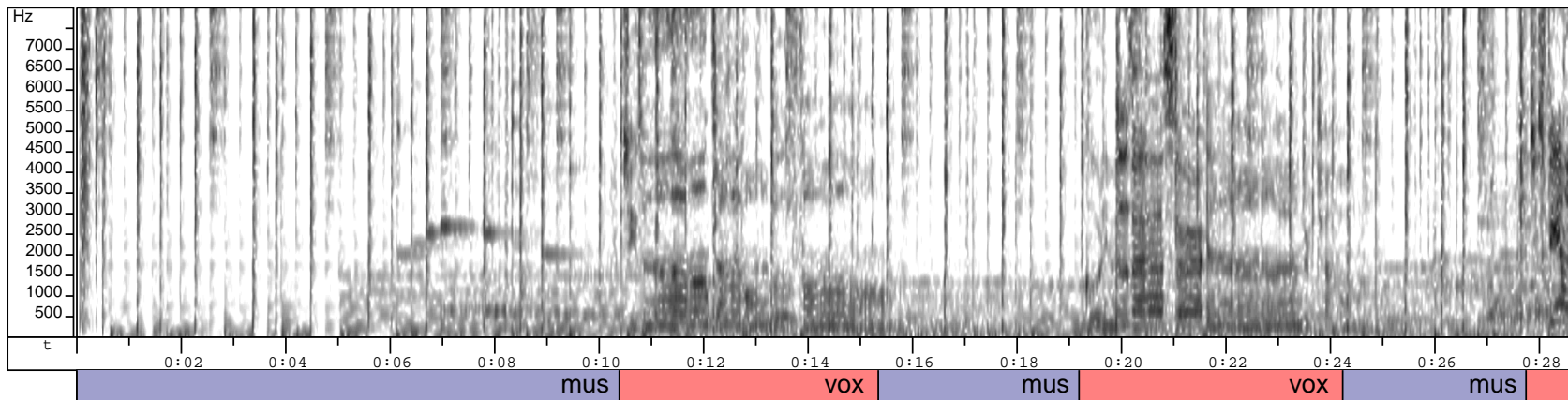- Motivation
- Features
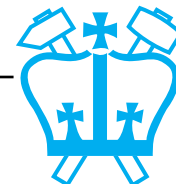- Classifiers

Singing Detection

(Berenzweig et al. '01)

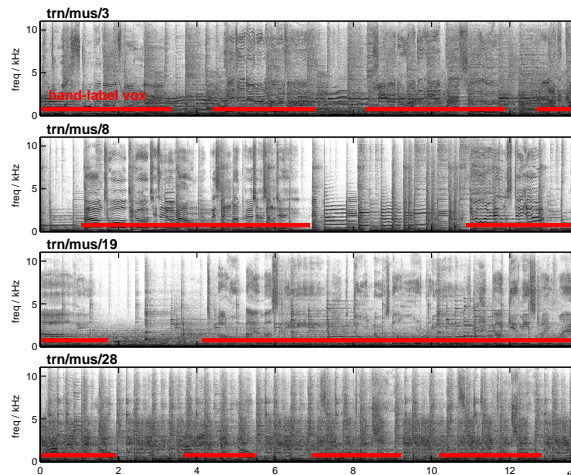- **Can we automatically detect when singing is present?**

File: /Users/dpwe/projects/aclass/aimee.wav

- for further processing (lyrics recognition?)
- as a song signature?
- as a basis for classification?

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

Singing Detection: Requirements

- **Labeled training examples**
  - 60 x 15 sec. radio excerpts
  - hand-mark sung phrases

- **Labeled test data**
  - several complete tracks from CDs, hand-labelled

- **Feature choice**
  - Mel-frequency Cepstral Coefficients (MFCCs) popular for speech; maybe sung voice too?
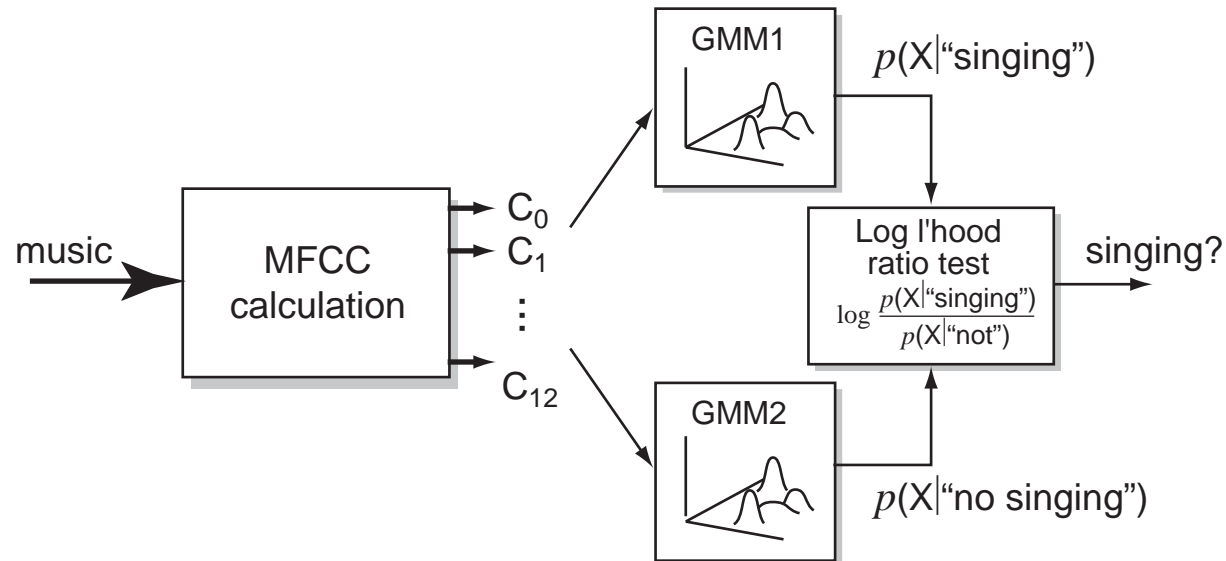 - separation of voices? temporal dimension?

- **Classifier choice**
  - MLP Neural Net
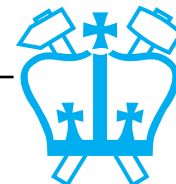 - GMMs for singing / music
  - SVM?
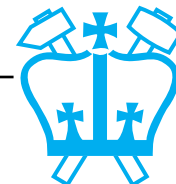
GMM System

- **Separate models for** $p(x|sing)$, $p(x|no\ sing)$
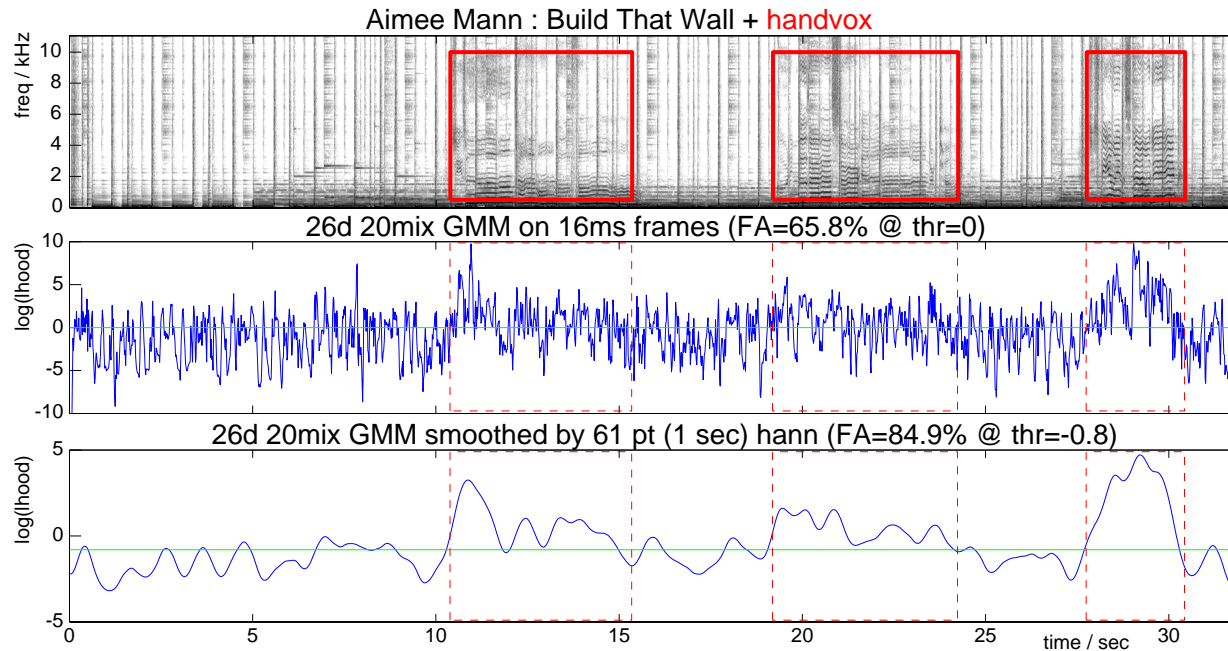 - combined via likelihood ratio test

- **How many Gaussians for each?**
  - say 20; depends on data & complexity

- **What kind of covariance?**
  - diagonal (spherical?)

Lab
ROSA
Laboratory for the Recognition and
Organization of Speech and Audio

GMM Results

- **Raw and smoothed results (Best FA=84.9%):**

Aimee Mann : Build That Wall + handvox

26d 20mix GMM on 16ms frames (FA=65.8% @ thr=0)
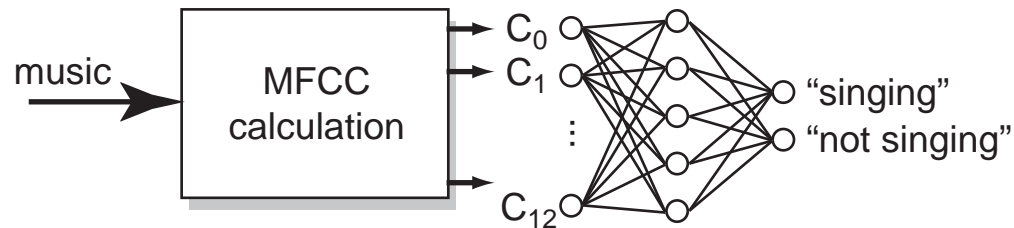
26d 20mix GMM smoothed by 61 pt (1 sec) hann (FA=84.9% @ thr=-0.8)

- **MLP has advantage of discriminant training**

- **Each GMM trains only on data subset
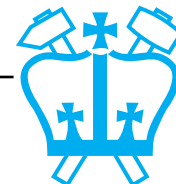 → faster to train? (2 x 10 min vs. 20 min)**

Lab ROSA
Laboratory for the Recognition and
Organization of Speech and Audio
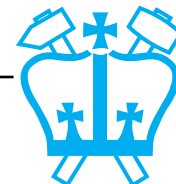
MLP Neural Net

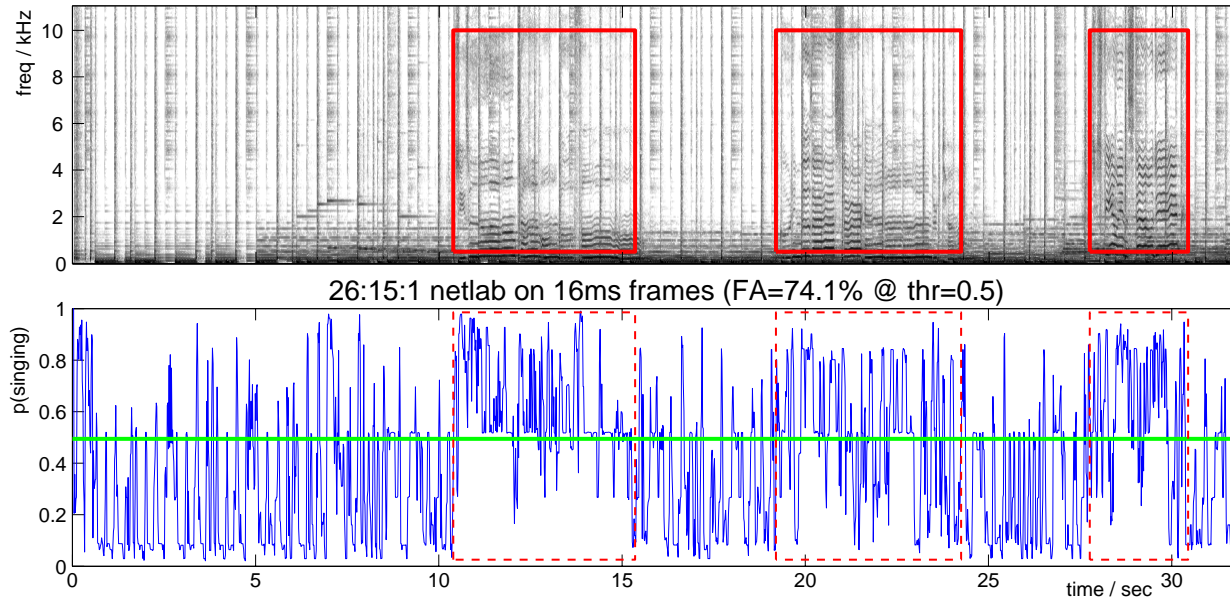- **Directly estimate** $p(\text{singing} \mid x)$

  - net has 26 inputs (+Δ), 15 HUs, 2 o/ps (26:15:2)

- **How many hidden units?**
  - depends on data amount, boundary complexity

- **Feature context window?**
  - useful in speech

- **Delta features?**
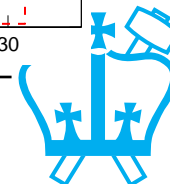 - useful in speech
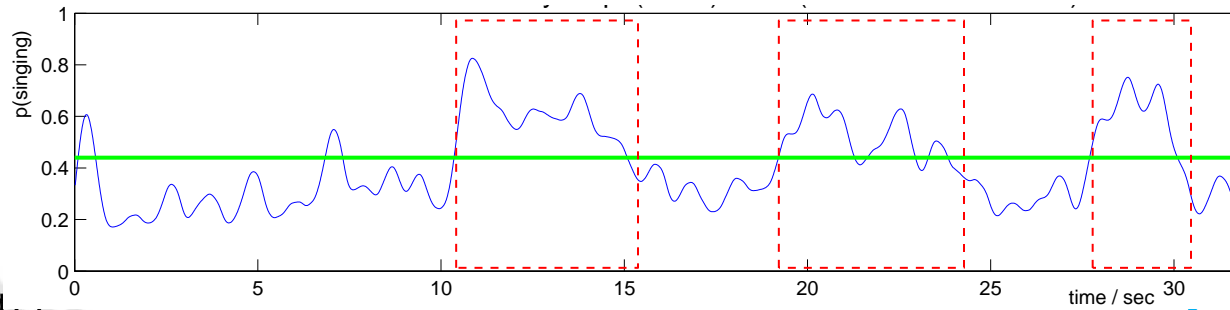
- **Training parameters...**

MLP Results

- **Raw net outputs on a CD track (FA 74.1%):**
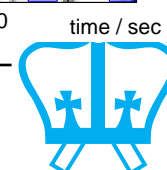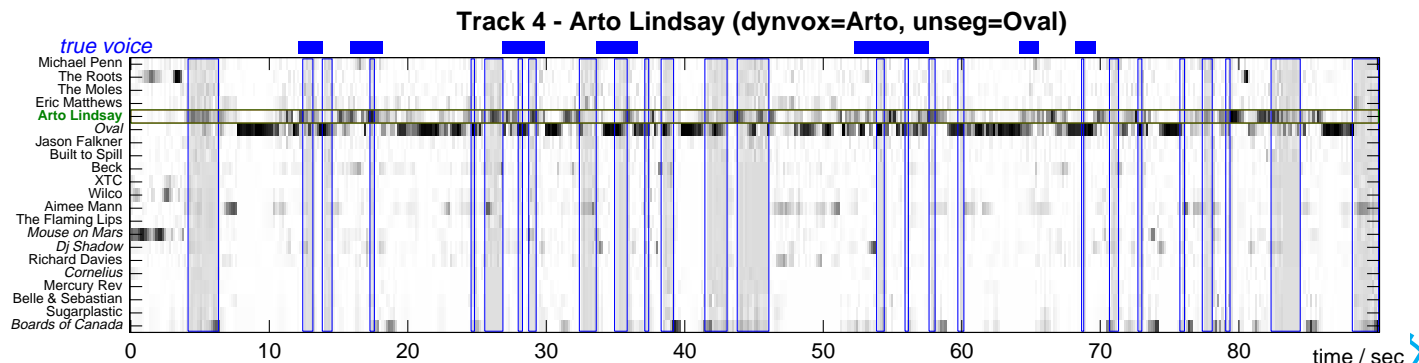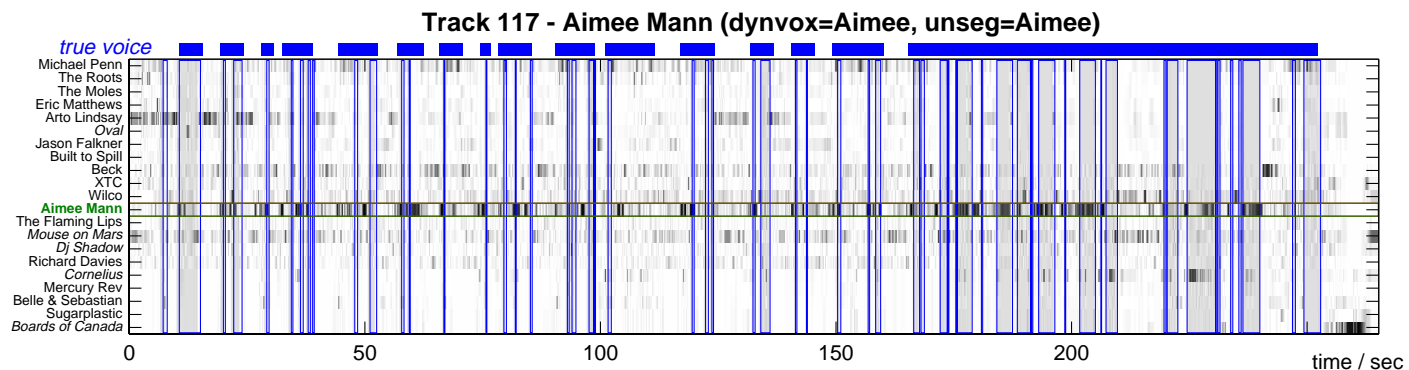

Aimee Mann : Build That Wall + handvox

- **Smoothed for continuity: best FA = 90.5%**

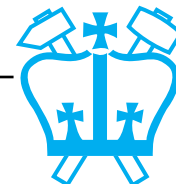# Artist Classification
(Berenzweig et al. 2002)

- **Artist label as available stand-in for genre**

- **Train MLP to classify frames among 21 artists**

- **Using only "voice" segments:**
  Song-level accuracy improves 56.7% → 64.9%

Track 117 - Aimee Mann (dynvox=Aimee, unseg=Aimee)

Track 4 - Arto Lindsay (dynvox=Arto, unseg=Oval)

Summary

- **Music content analysis:
  Pattern classification**

- **Basic machine learning methods:
  Neural Nets, GMMs**

- **Singing detection: classic application**

but... the time dimension?

References

A.L. Berenzweig and D.P.W. Ellis (2001)
> "Locating Singing Voice Segments within Music Signals",
> Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio, Mohonk NY, October 2001.
> http://www.ee.columbia.edu/~dpwe/pubs/waspaa01-singing.pdf

R.O. Duda, P. Hart, R. Stork (2001)
> *Pattern Classification, 2nd Ed.*
> Wiley, 2001.

E. Scheirer and M. Slaney (1997)
> "Construction and evaluation of a robust multifeature speech/music discriminator",
> Proc. IEEE ICASSP, Munich, April 1997.
> http://www.ee.columbia.edu/~dpwe/e6820/papers/ScheiS97-mussp.pdf