Audio Signal Recognition for Speech, Music, and Environmental Sounds

1. **Pattern Recognition for Sounds**
2. **Speech Recognition**
3. **Other Audio Applications**
4. **Observations and Conclusions**

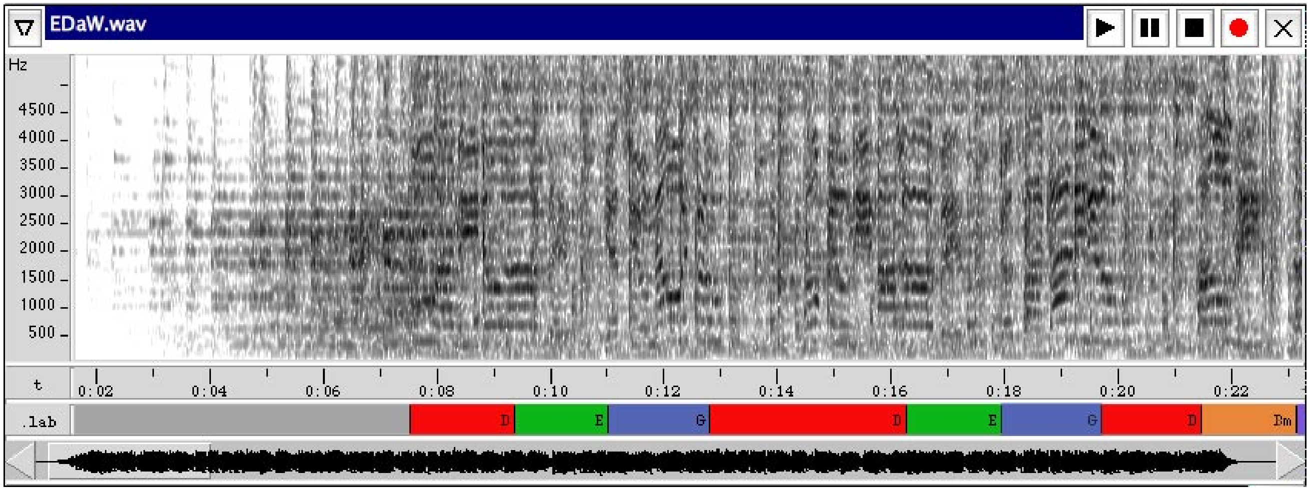
Dan Ellis <dpwe@ee.columbia.edu>

Laboratory for Recognition and Organization of Speech and Audio
(Lab**ROSA**)
Columbia University, New York
http://labrosa.ee.columbia.edu/

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

1 Pattern Recognition for Sounds

- **Pattern recognition is abstraction**

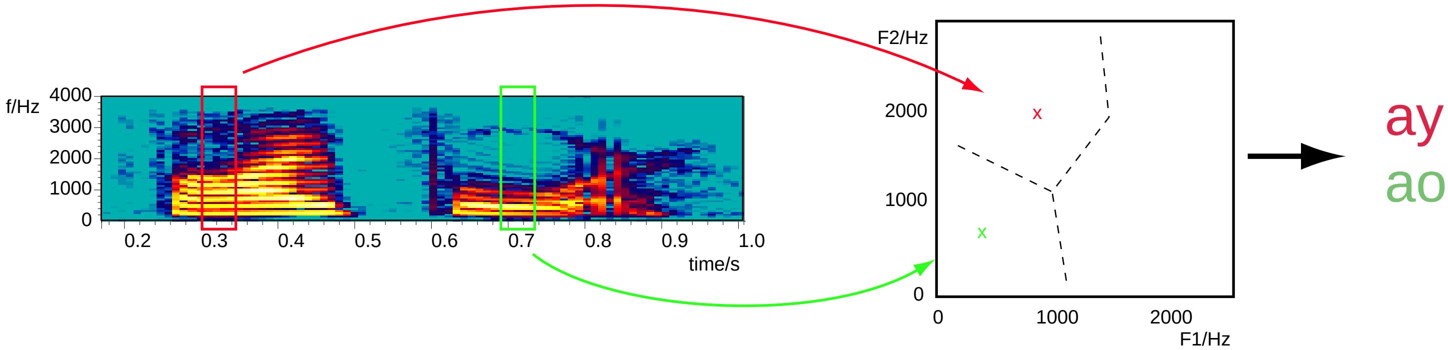


  - continuous signal → discrete labels
  - an essential part of understanding?
    "information extraction"

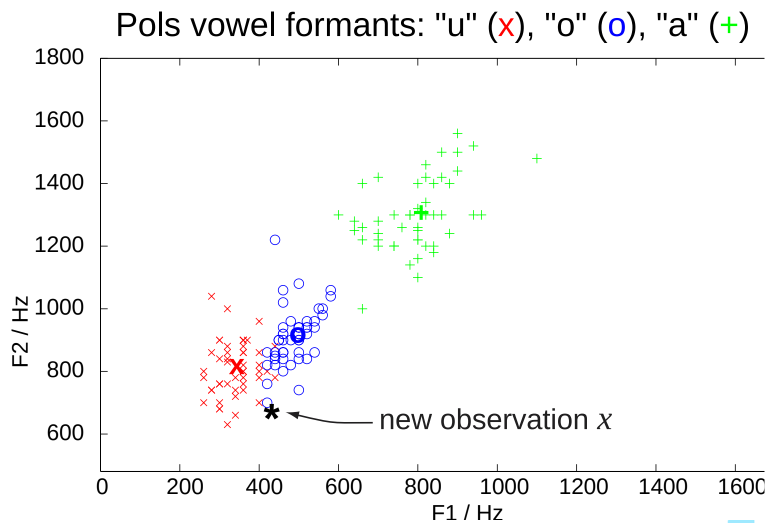
- **Sound is a challenging domain**
  - sounds can be highly variable
  - human listeners are extremely adept

Pattern classification

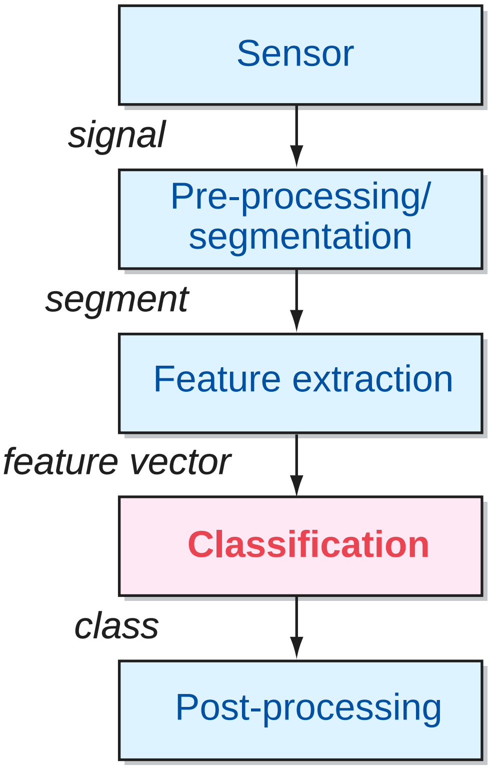
- **Classes are defined as distinct region in some feature space**
  - e.g. formant frequencies to define vowels



- **Issues**
  - finding segments to classify
  - transforming to an appropriate feature space
  - defining the class boundaries



Classification system parts



- STFT
- Locate vowels

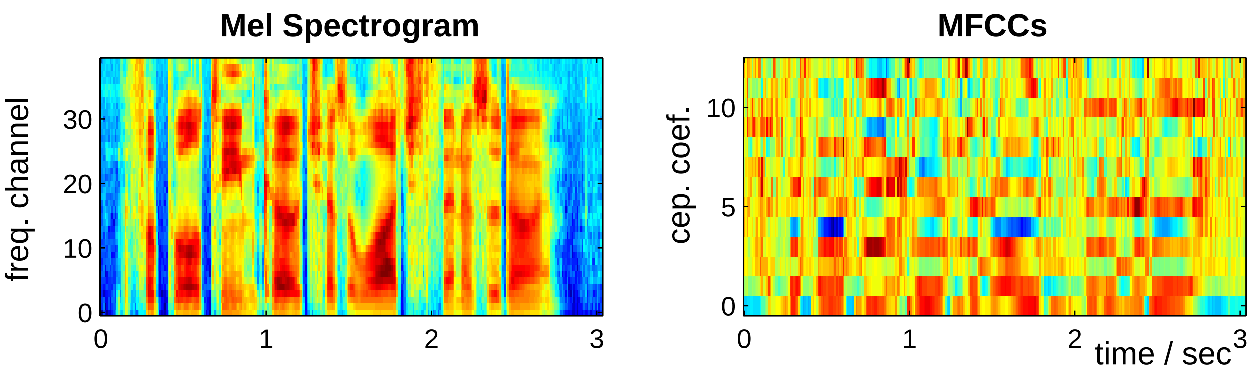
- Formant extraction

- Context constraints
- Costs/risk

Feature extraction

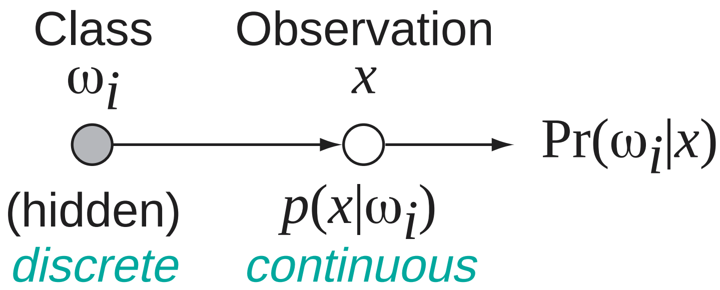
- **Feature choice is critical to performance**
  - make important aspects explicit, remove irrelevant details
  - 'equivalent' representations can perform very differently in practice
  - major opening for domain knowledge ("cleverness")

- **Mel-Frequency Cepstral Coefficients (MFCCs): Ubiquitous speech features**
  - DCT of log spectrum on 'auditory' scale
  - approximately decorrelated ...

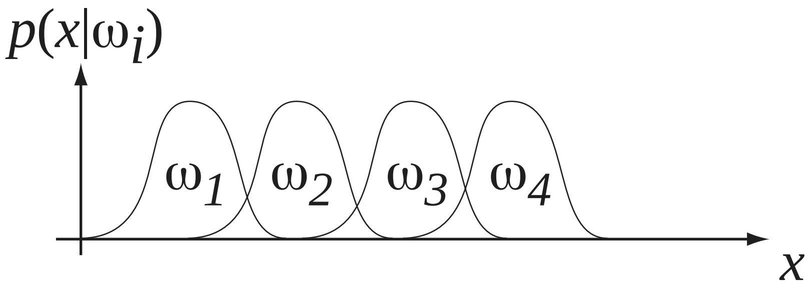


Statistical Interpretation

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



- **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)



Priors and posteriors

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

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)**

- **Minimize the probability of error by choosing *maximum a posteriori* (MAP) class:**

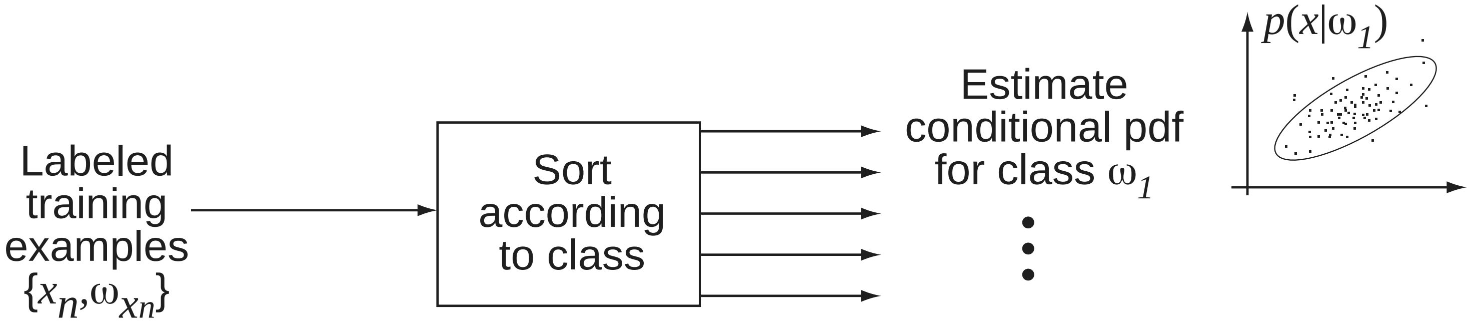
$$\hat{\omega} = \underset{\omega_i}{\operatorname{argmax}}\ Pr(\omega_i|x)$$

Practical implementation

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

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

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



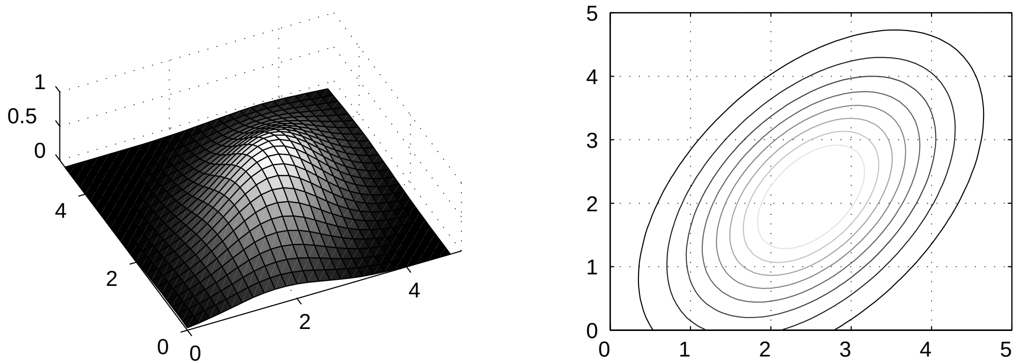
Gaussian models

- **Model data distributions via parametric model**
  - assume known form, estimate a few parameters
- **E.g. Gaussian in 1 dimension:**

$$p(x|\omega_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \cdot \exp\left[-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right]$$

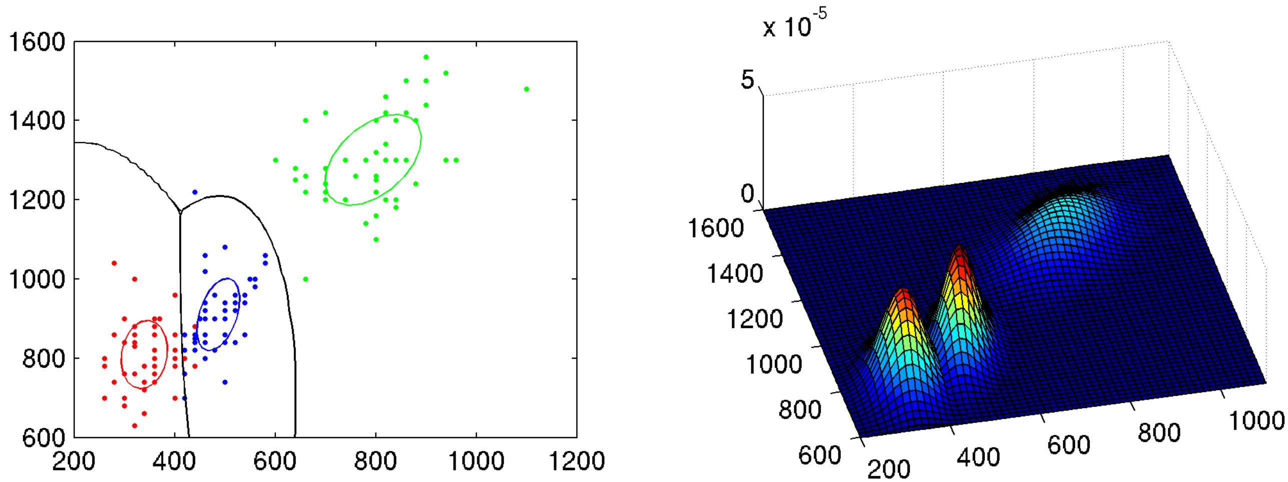
normalization

- **For higher dimensions, need mean vector μ_i and d x d covariance matrix Σ_i**



- **Fit more complex distributions with mixtures...**

Gaussian models for formant data



- **Single Gaussians a reasonable fit for this data**
- **Extrapolation of decision boundaries can be surprising**

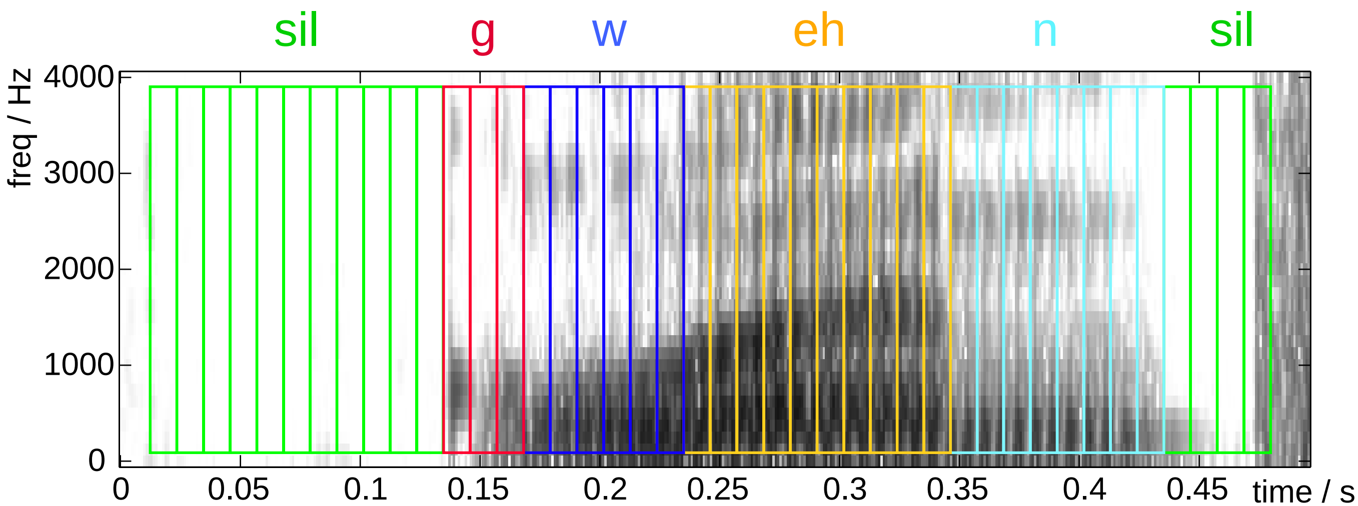
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Outline

1. **Pattern Recognition for Sounds**
2. **Speech Recognition**
    - How it’s done
    - What works, and what doesn’t
3. **Other Audio Applications**
4. **Observations and Conclusions**

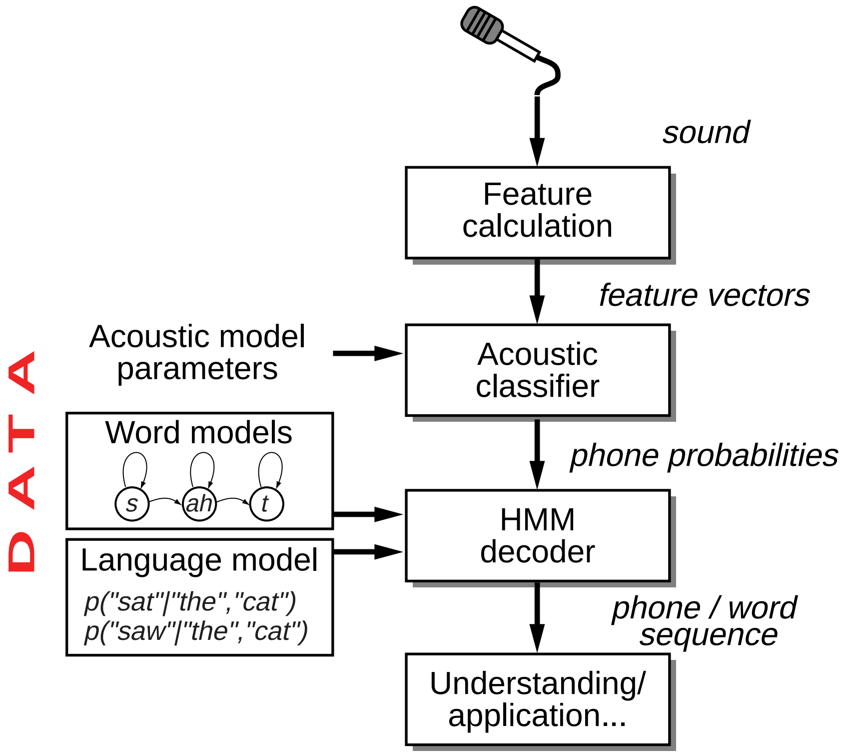
2 How to recognize speech?

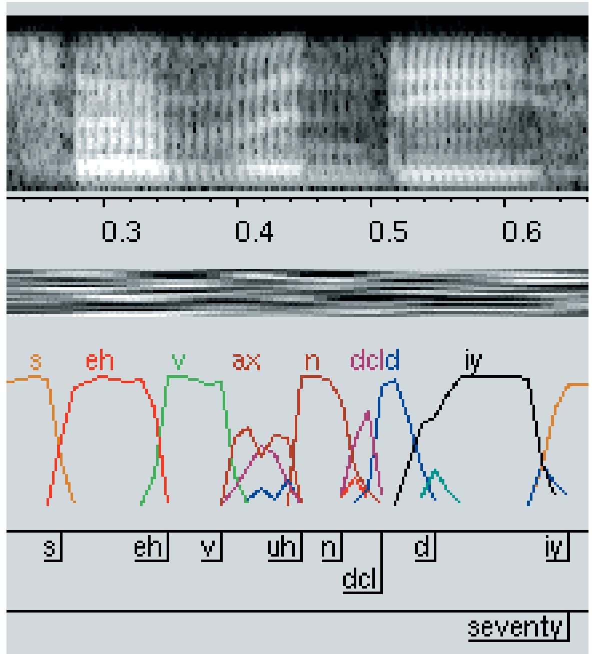
- **Cross correlate templates?**
  - waveform?
  - spectrogram?
  - time-warp problems
- **Classify short segments as phones (or ...), handle time-warp later**
  - model with slices of ~ 10 ms
  - pseudo-piecewise-stationary model of words:



Speech Recognizer Architecture

- **Almost all current systems are the same:**

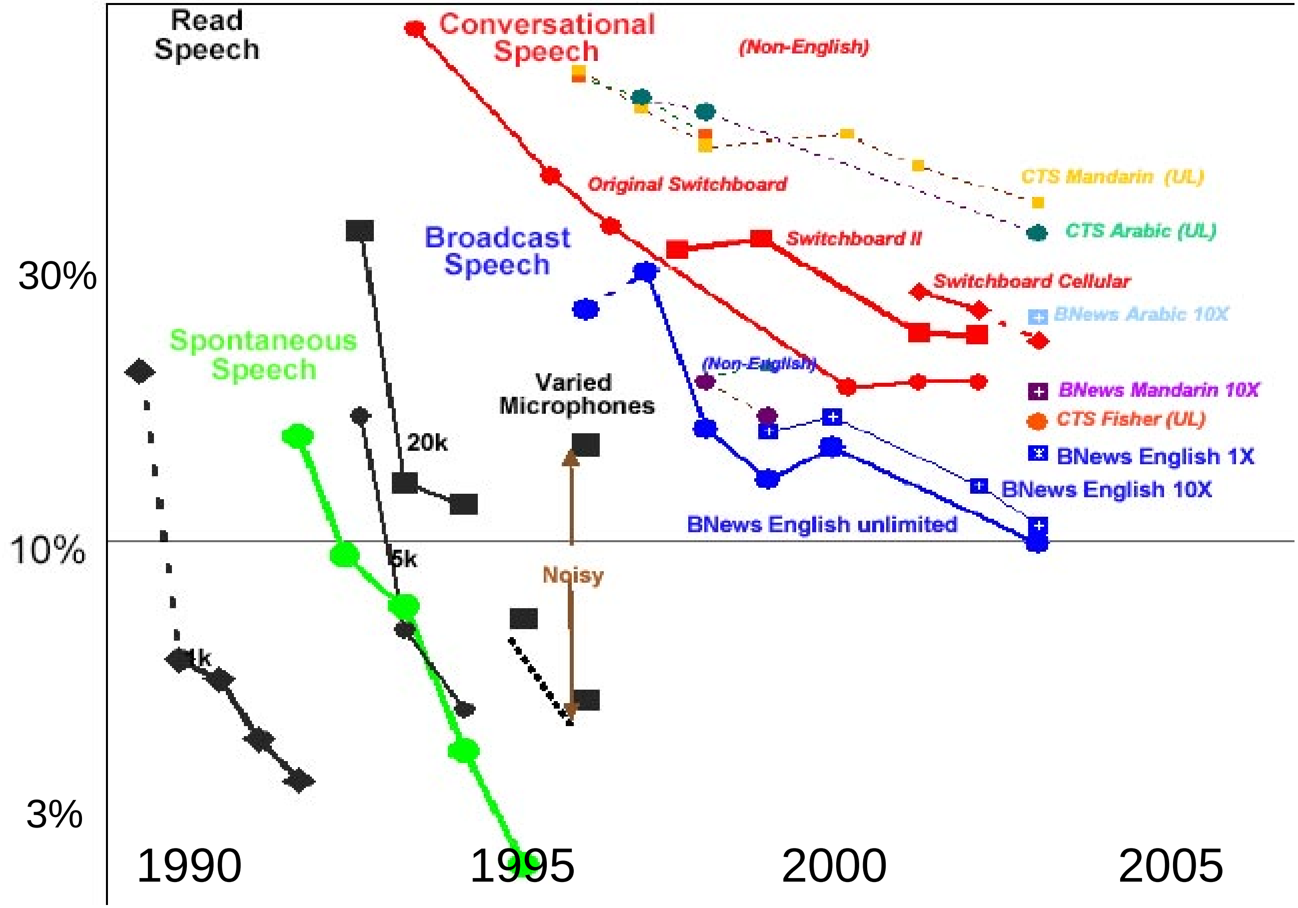




- **Biggest source of improvement is increase in training data**
  - .. along with algorithms to take advantage

Speech: Progress

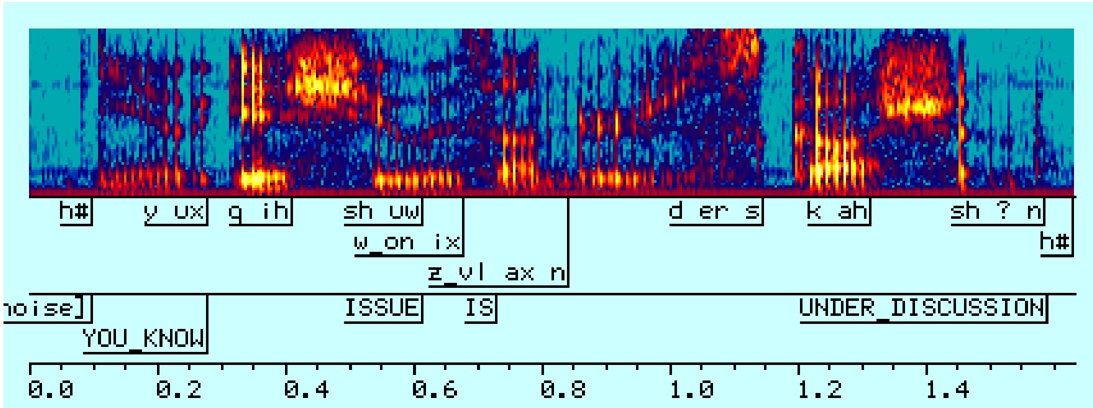
- **Annual NIST evaluations**



  - steady progress (?), but still order(s) of magnitude worse than human listeners

Speech: Problems

- **Natural, spontaneous speech is weird!**



  - coarticulation
  - deletions
  - disfluencies

  → is word transcription even a sensible approach?

- **Other major problems**
  - speaking style, rate, accent
  - environment / background...

Speech: What works, what doesn't

- **What works: Techniques:**
  - MFCC features + GMM/HMM systems trained with Baum-Welch (EM)
  - Using lots of training data

  Domains:
  - Controlled, low noise environments
  - Constrained, predictable contexts
  - Motivated, co-operative users
- **What doesn't work: Techniques:**
  - rules based on 'insight'
  - perceptual representations (except when they do...)

  Domains:
  - spontaneous, informal speech
  - unusual accents, voice quality, speaking style
  - variable, high-noise background / environment

Outline

1. Pattern Recognition for Sounds
2. Speech Recognition
3. **Other Audio Applications**
   - Meeting recordings
   - Alarm sounds
   - Music signal processing
4. Observations and Conclusions

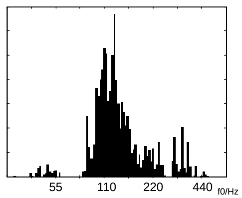
Other Audio Applications: ICSI Meeting Recordings corpus

3

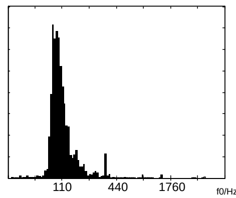
- **Real meetings, 16 channel recordings, 80 hrs**

  - released through NIST/LDC

- **Classification e.g.: Detecting emphasized utterances based on f_0 contour** (Kennedy & Ellis '03)
  - per-speaker normalized f0 as unidimensional feature → simple threshold classification



Speaker 1

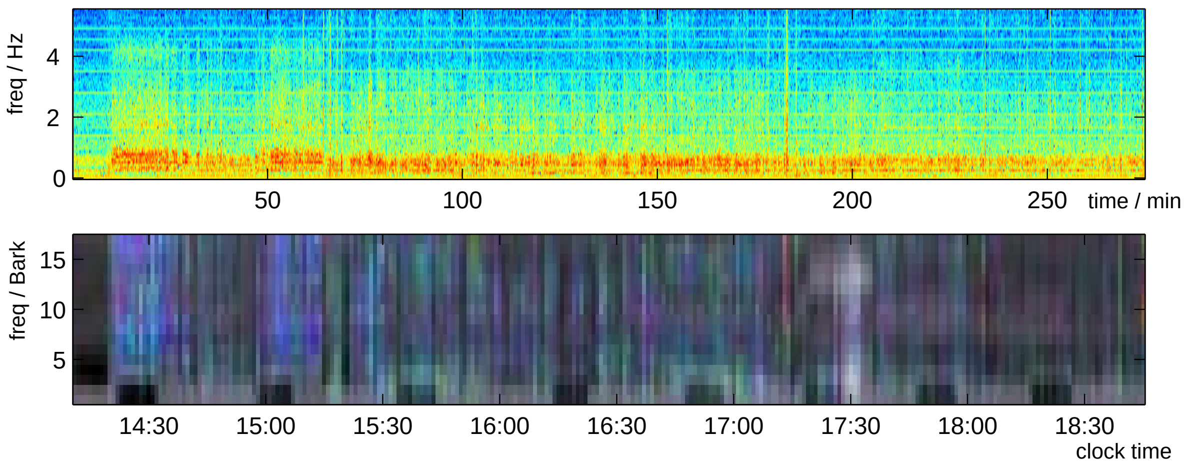


Speaker 2

Personal Audio

- **LifeLog / MyLifeBits / Remembrance Agent:**
  - easy to record everything you hear

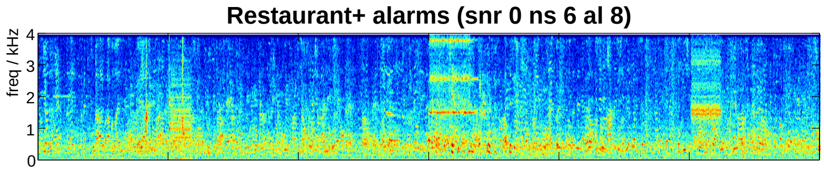
- **Then what?**
  - prohibitive to review
  - applications if access easier?
- **Automatic content analysis / indexing...**



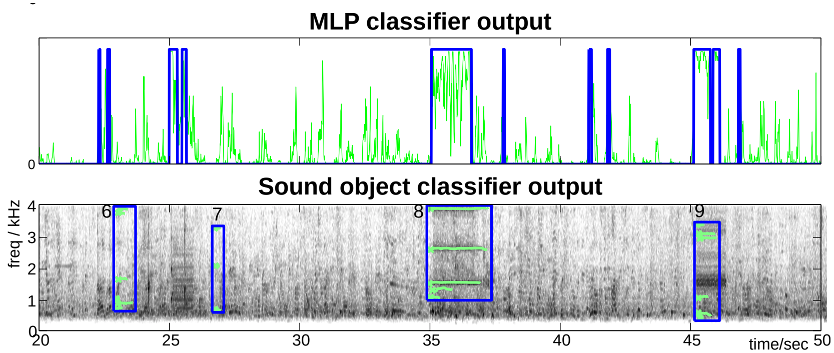
  - find features to classify into e.g. locations

Alarm sound detection

- **Alarm sounds have particular structure**
  - clear even at low SNRs
  - potential applications...



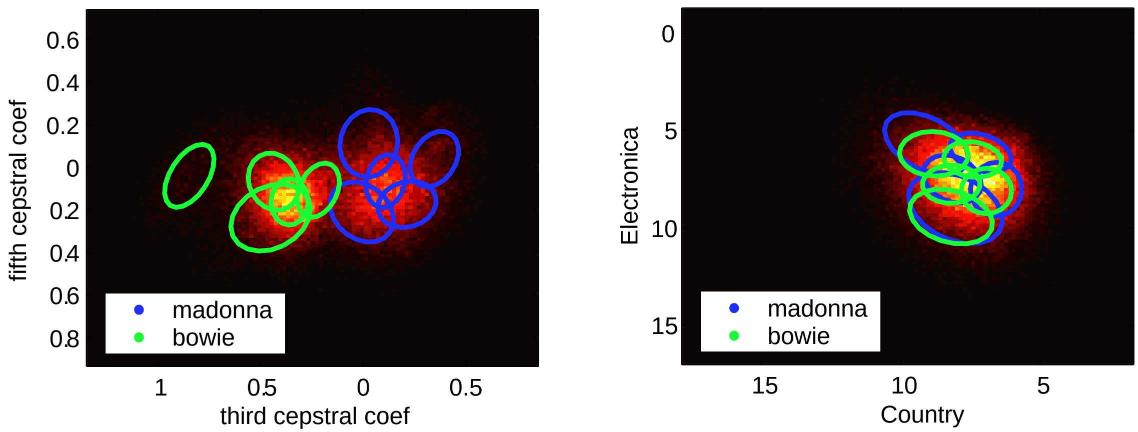
- **Contrast two systems:** (Ellis '01)
  - standard, global features, $P(X|M)$
  - sinusoidal model, fragments, $P(M,S|Y)$



  - error rates high, but interesting comparisons...

Music signal modeling

- **Use “machine listener” to navigate large music collections**
    - e.g. unsigned bands on MP3.com
- **Classification to label:**
    - notes, chords, singing, instruments
    - .. information to help cluster music
- **“Artist models” based on feature distributions**



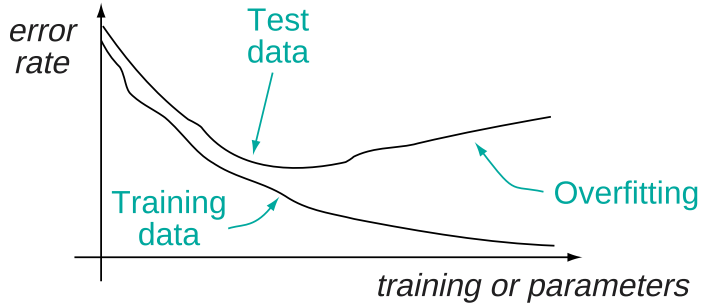
- measure similarity between users’ collections and new music? (Berenzweig & Ellis ’03)

Outline

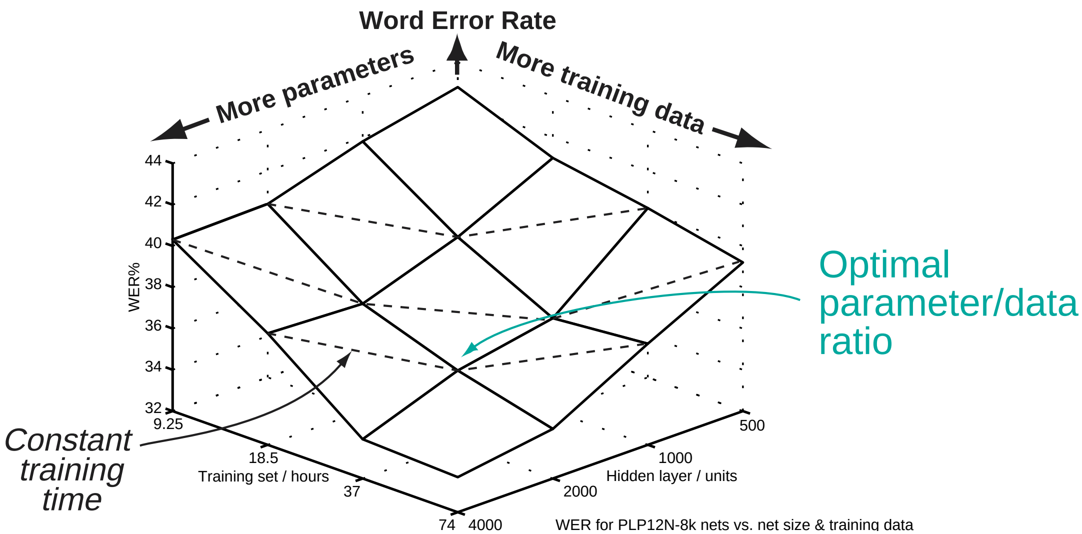
1. Pattern Recognition for Sounds
2. Speech Recognition
3. Other Audio Applications
4. **Observations and Conclusions**
   - Model complexity
   - Sound mixtures

4 Observations and Conclusions: Training and test data

- **Balance model/data size to avoid overfitting:**



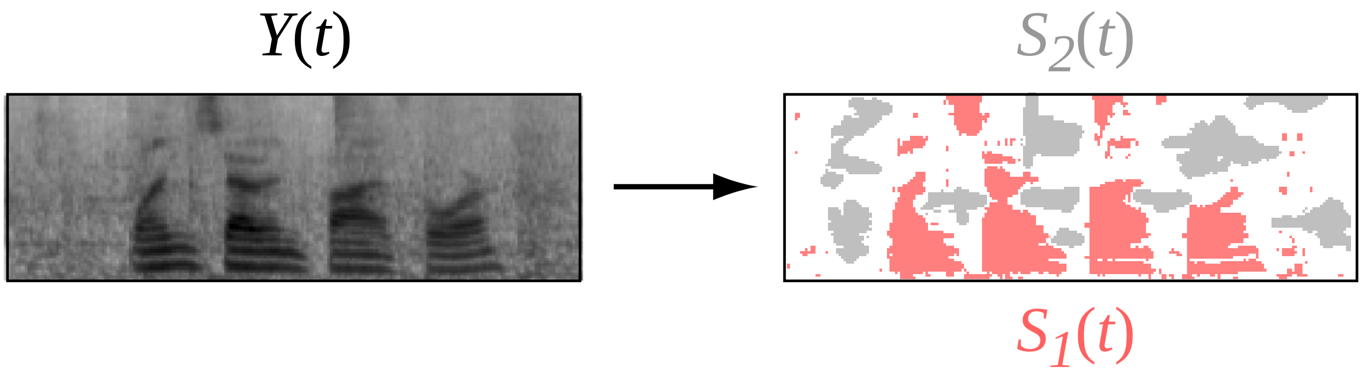
- **Diminishing returns from more data:**



Beyond classification

- "**No free lunch**":
  Classifier can only do so much
  - always need to consider other parts of system
- **Features**
  - impose ceiling on system performance
  - improved features allow simpler classifiers
- **Segmentation / mixtures**
  - e.g. speech-in-noise:
    only subset of feature dimensions available

  → missing-data approaches...



Summary

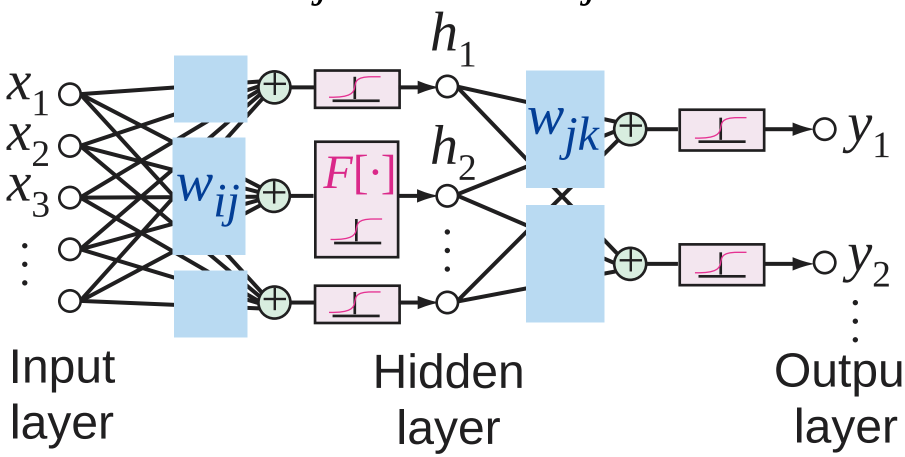
- **Statistical Pattern Recognition**
  - exploit training data for probabilistically-correct classifications
- **Speech recognition**
  - successful application of statistical PR
  - .. but many remaining frontiers
- **Other audio applications**
  - meetings, alarms, music
  - classification is information extraction
- **Current challenges**
  - variability in speech
  - acoustic mixtures

Extra slides

Neural network classifiers

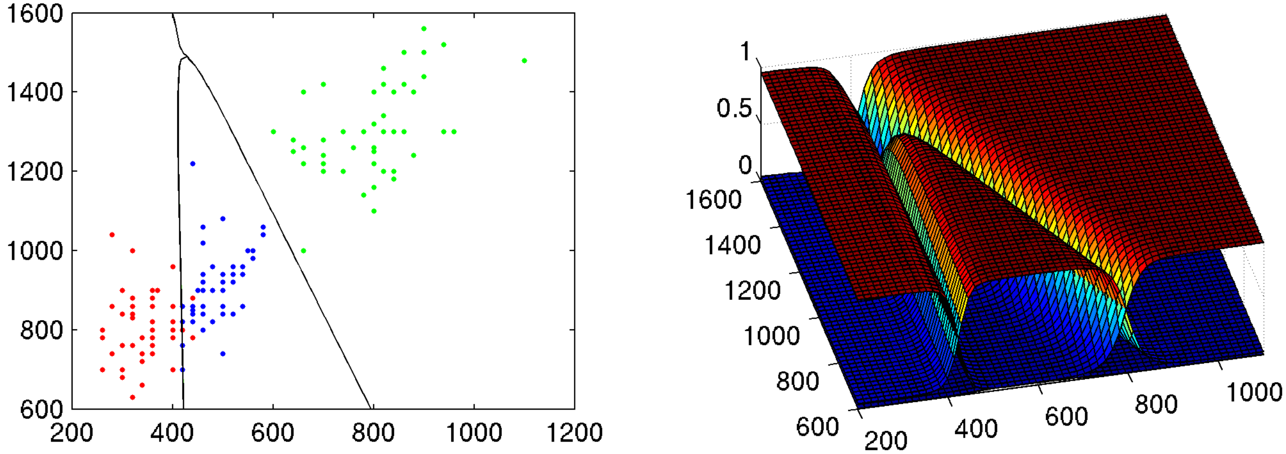
- **Instead of estimating $p(x|\omega_i)$ and using Bayes, can also try to estimate posteriors $Pr(\omega_i|x)$ directly (the decision boundaries)**
- **Sums over nonlinear functions of sums give a large range of decision surfaces...**
- **e.g. Multi-layer perceptron (MLP):**

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



- **Problem is finding the weights w_{ij} ... (*training*)**

Neural net classifier



- **Models** boundaries**, not density** $p(x|\omega_i)$
- Discriminant **training**
  - concentrate on boundary regions
  - needs to see all classes at once

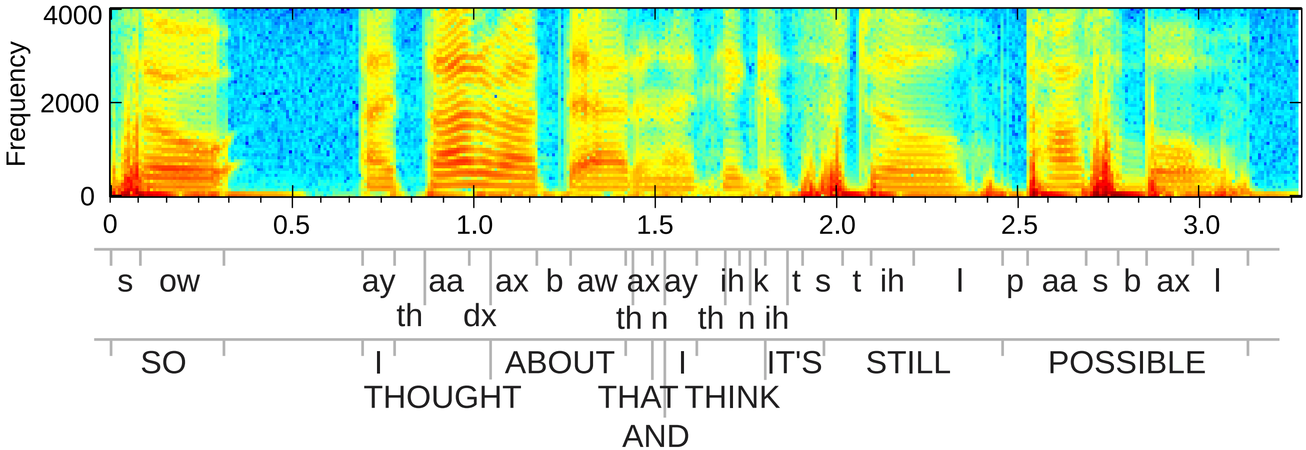
Why is Speech Recognition hard?

- **Why not match against a set of waveforms?**
  - waveforms are never (nearly!) the same twice
  - speakers minimize information/effort in speech
- **Speech variability comes from many sources:**
  - speaker-dependent (SD) recognizers must handle within-speaker variability
  - speaker-independent (SI) recognizers must also deal with variation between speakers
  - all recognizers are afflicted by background noise, variable channels

→ **Need recognition models that:**
  - generalize i.e. accept variations in a range, and
  - adapt i.e. 'tune in' to a particular variant

Within-speaker variability

- **Timing variation:**
  - word duration varies enormously



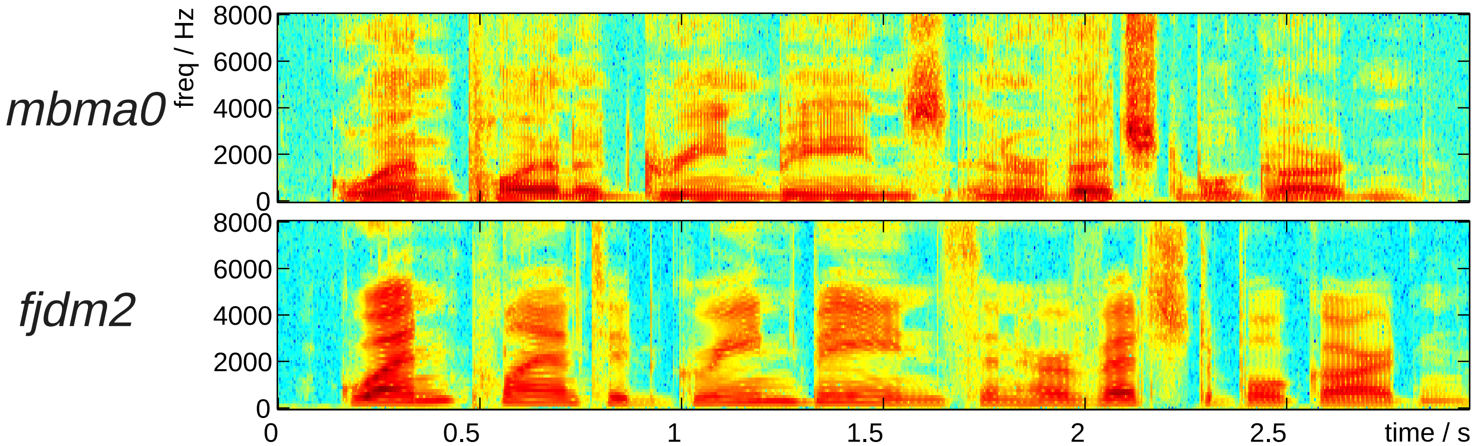
  - fast speech ‘reduces’ vowels

- **Speaking style variation:**
  - careful/casual articulation
  - soft/loud speech

- **Contextual effects:**
  - speech sounds vary with context, role:
    “How **do** you **do**?”

Between-speaker variability

- **Accent variation**
  - regional / mother tongue
- **Voice quality variation**
  - gender, age, huskiness, nasality
- **Individual characteristics**
  - mannerisms, speed, prosody



Environment variability

- **Background** noise
  - fans, cars, doors, papers
- **Reverberation**
  - 'boxiness' in recordings
- **Microphone channel**
  - huge effect on relative spectral gain