EE E6820: Speech & Audio Processing & Recognition

Lecture 12: Audio Databases

1. **ASR wrap-up**
2. **Spoken document retrieval**
3. **General audio databases**
4. **Open issues**

Dan Ellis <dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/e6820/

Columbia University Dept. of Electrical Engineering
Spring 2003

Outline

1. **ASR wrap-up**
   - Discriminant modeling
   - Adaptation
   - Confidence measures
2. **Spoken document retrieval**
3. **General audio databases**
4. **Open issues**

1 ASR wrap-up: Discriminant models

- **EM training of HMMs is maximum likelihood**
  - i.e. set Θ for local $\max p(X_{trn} \mid \Theta)$
  - converges to Bayes optimal $\Theta \mid X_{trn}$ in the limit
- **Decision rule is** $\max p(X \mid M) \cdot p(M)$
  - training will increase $p(X \mid M_{correct})$
  - may also increase $p(X \mid M_{wrong})$ ...more?
- **Discriminant training tries directly to increase discrimination between right & wrong models**
  - e.g. Maximum Mutual Information (MMI)

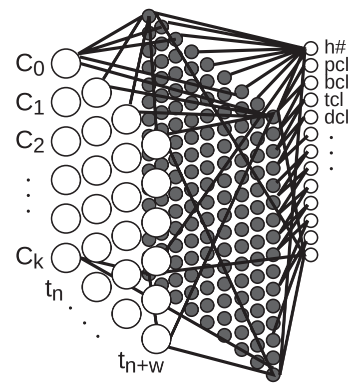
$$I(M_j ; X \mid \Theta) = E\left[\log \frac{p(M_j, X \mid \Theta)}{p(M_j \mid \Theta)\, p(X \mid \Theta)}\right]$$

$$= E\left[\log \frac{p(X \mid M_j, \Theta)}{\sum_k p(X \mid M_k, \Theta)\, p(M_k \mid \Theta)}\right]$$

Neural Network Acoustic Models

- **Single model generates posteriors directly for all classes at once = frame-discriminant**
- **Use regular HMM decoder for recognition**
  - set $b_i(x_n) = p(x_n|q^i) \propto p(q^i|x_n) / p(q^i)$
- **Nets are less sensitive to input representation**
  - skewed feature distributions
  - correlated features
- **Temporal context window allows net to 'see' feature dynamics:**

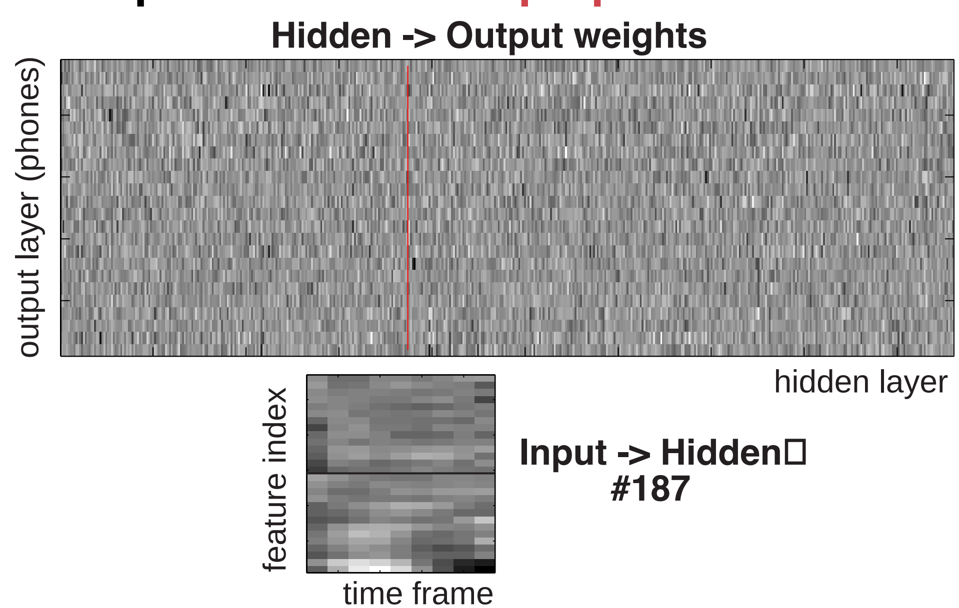
Feature calculation →



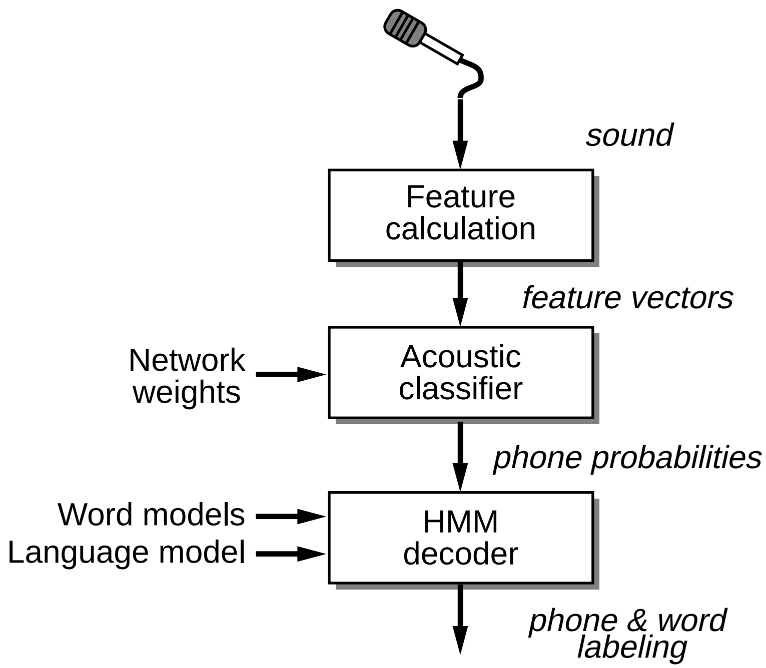
posteriors
$p(q^i \mid X)$

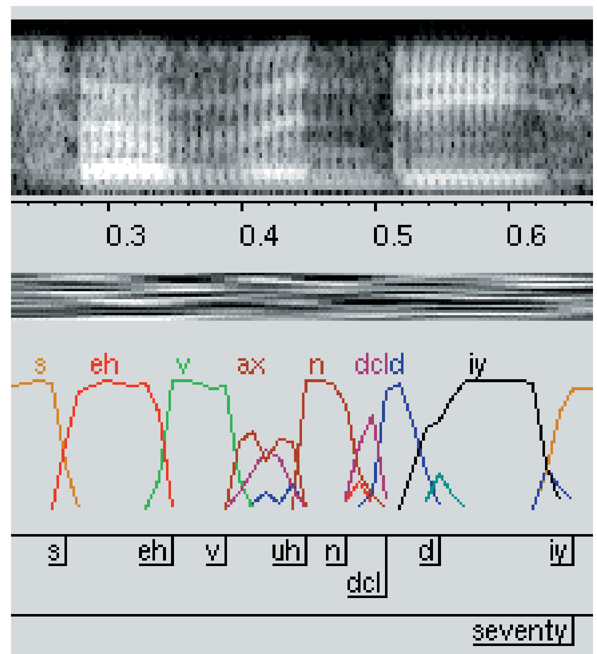
Neural nets: Practicalities

- **Typical net sizes:**
  - input layer: 9 frames x 9-40 features ~ 300 units
  - hidden layer: 100-8000 units, dep. train set size
  - output layer: 30-60 context-independent phones
- **Hard to make context dependent**
  - problems training many classes that are similar?
- **Representation is opaque:**



Recap: Recognizer Structure





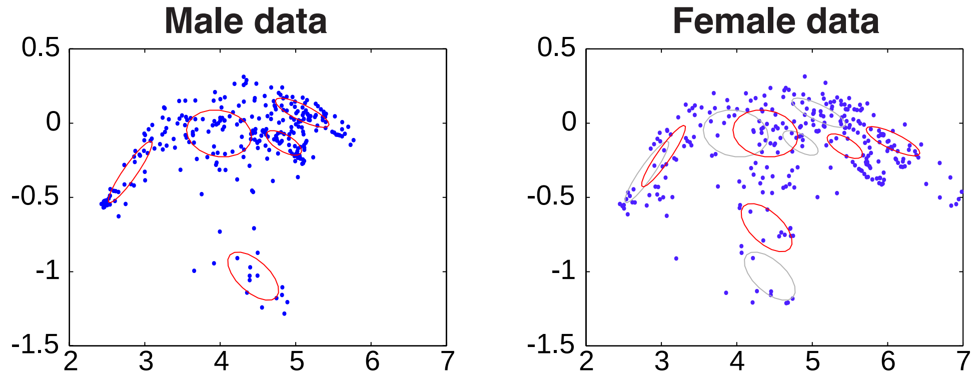
- **Now we have it all!**

Model adaptation

- **Practical systems often suffer from mismatch**
  - test conditions are not like training data: accent, microphone, background noise ...
- **Desirable to continue tuning during recognition = adaptation**
  - but: no 'ground truth' labels or transcription
- **Assume that recognizer output is correct; Estimate parameters from those labels**
  - like Viterbi training
  - can iterate until convergence
- **Normally have little adaptation data**
  - want to adapt using data from one speaker/ condition only (clustering)
  - hence, can estimate only a few parameters (not update whole acoustic model)

Maximum-Likelihood Linear Regression

- **Model mismatch as an affine transform:**



- **Estimate matrix to transform means of GMM components**

$$\hat{\mu}_j = A \cdot \mu_j + b$$

  - maximum-likelihood solution via EM over adaptation data
  - ML only works for transforming model, not data

- **Typically 10-20% relative WER reduction**
  - given enough data from one condition

Recognizer outputs

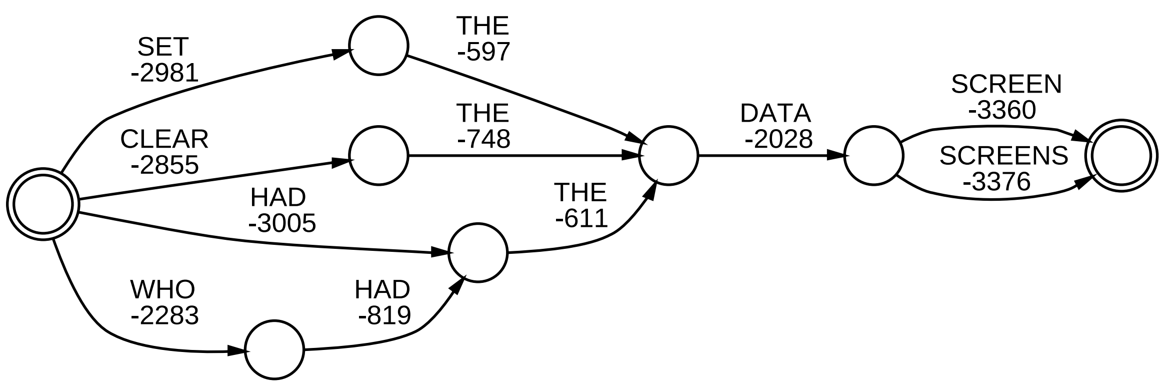
- **Simple recognizer output:**
  most probable word sequence (1-best)

  ```
  SET THE DATA SCREEN
  ```

- **Other possible outputs:**
  - *N*-best word sequences (with likelihoods)

    ```
    -9076  SET THE DATA SCREEN
    -9092  SET THE DATA SCREENS
    -9158  CLEAR THE DATA SCREEN
    ```

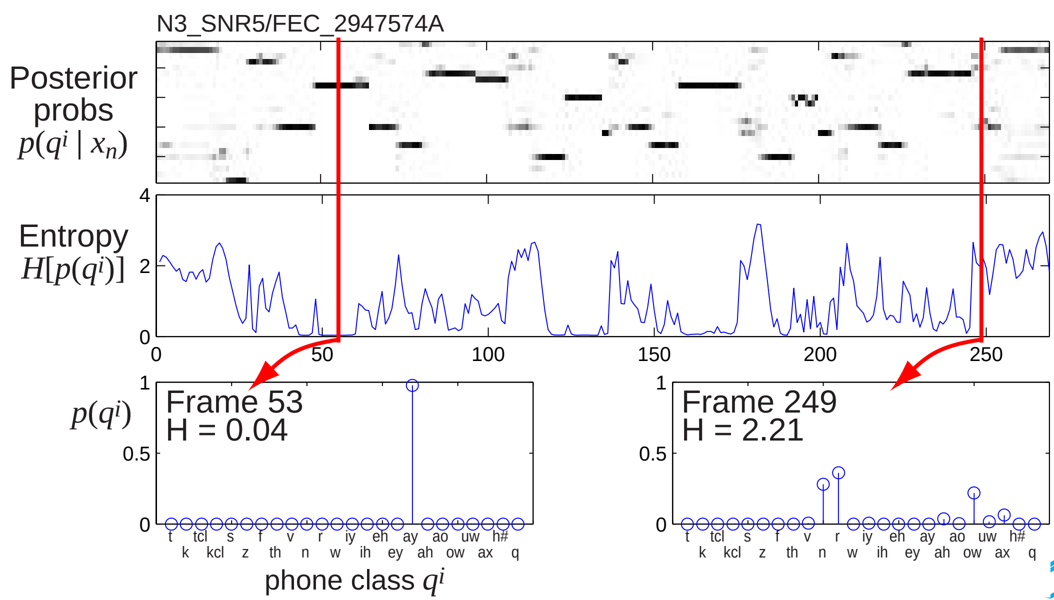
  - word graph/lattice (but: LM)



post process with different constraints

Confidence measures

- **Can we tell which words might be wrong?**
- **Use** $p(X, Q|M_j) = \prod_n p(x_n|q_n)p(q_n|q_{n-1})$
  - good for comparing M_js, not observations X
- **Discriminative measures are better**
  - e.g. entropy $H[p(q^i)] = -\sum_i p(q^i)\log p(q^i)$



State of the art recognition

- **e.g. 1999 NIST 'Broadcast News' recognition**
  - real TV and radio news broadcasts
  - 100,000 word vocabulary
  - 150 hours of transcribed training data
- **Features of best systems (LIMSI, Cambridge)**
  - context-dependent GMM acoustic models: 3,500 states; 300,000 Gaussians
  - speaker adaptation in training and test
  - segmentation and clustering...
  - discriminative training?
- **Performance:**
  - overall: WER ~14% (300x real time)
  - studio speech: 8%
  - 'Fast' system ~17% (10x real time)

2 Spoken Document Retrieval (SDR)

- **20% WER is horrible for transcription**
  - is it good for anything else?
- **Information Retrieval (IR)**
  - TREC/MUC ‘spoken documents’
  - tolerant of word error rate, e.g.:

F0: THE VERY EARLY RETURNS OF THE NICARAGUAN PRESIDENTIAL ELECTION SEEMED TO FADE BEFORE THE LOCAL MAYOR ON A LOT OF LAW

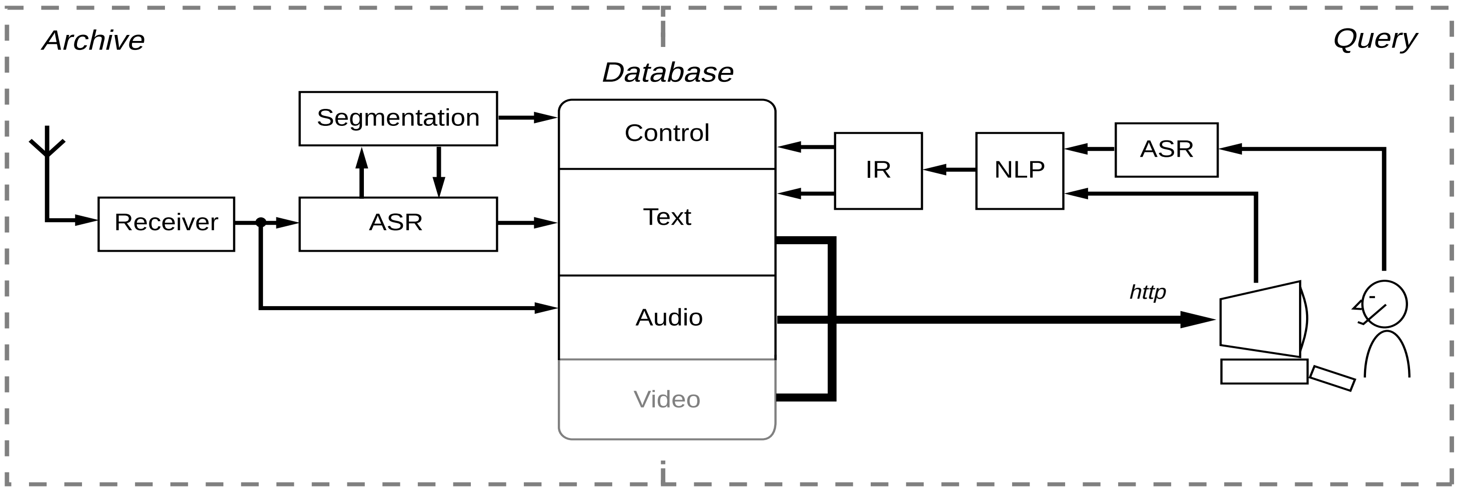
F4: AT THIS STAGE OF THE ACCOUNTING FOR SEVENTY SCOTCH ONE LEADER DANIEL ORTEGA IS IN SECOND PLACE THERE WERE TWENTY THREE PRESIDENTIAL CANDIDATES OF THE ELECTION

F5: THE LABOR MIGHT DO WELL TO REMEMBER THE LOST A MAJOR EPISODE OF TRANSATLANTIC CONNECT TO A CORPORATION IN BOTH CONSERVATIVE PARTY OFFICIALS FROM BRITAIN GOING TO WASHINGTON THEY WENT TO WOOD BUYS GEORGE BUSH ON HOW TO WIN A SECOND TO NONE IN LONDON THIS IS STEPHEN BEARD FOR MARKETPLACE

- **Promising application area**
  - document retrieval already hit-and-miss
  - plenty of untranscribed material

The THISL SDR system

- **Original task: BBC newsroom support**



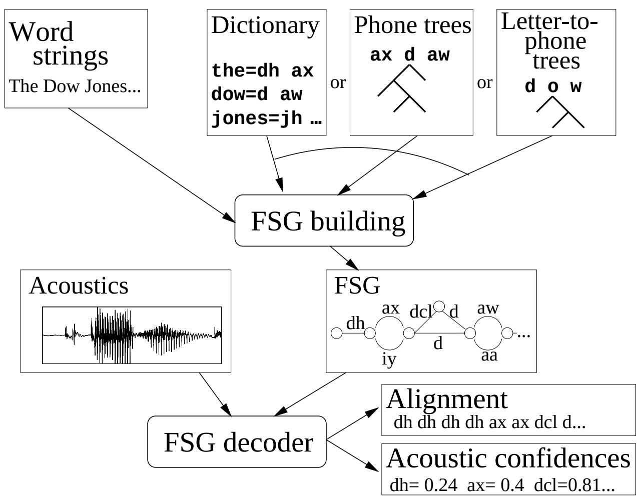
- **How to build the database:**
  - automatically record news programs 'off air'
  - several hours per day → > 3,000 hrs
  - run recognition the whole time
  - problems storing audio!

Building a new recognizer

- **No models available for BBC English**
  - need to develop a new recognizer based on US English Broadcast News, read British English...
- **Training set:**
  Manual transcription of 40 hours of news
  - word-level transcription takes > 10x real-time
  - Viterbi training, starting from read speech model
- **Language model:**
  200M words of US & UK newspaper archives
- **Dictionary:**
  Standard UK-English + extensions
  - many novel & foreign words

Vocabulary extension

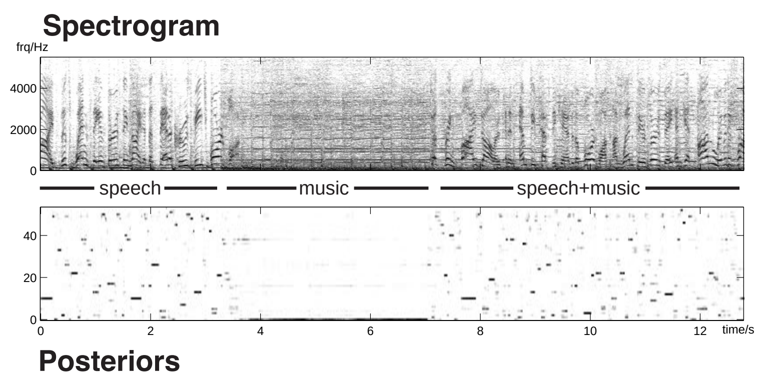
- News always has novel words
- **Starting point: Text-to-speech rules**
  - speech synthesizers' rules for unknown words
  - but novel words are often foreign names
- **Sources to identify new words**
  - BBC 'house style' information
- **Choose model by single acoustic example**

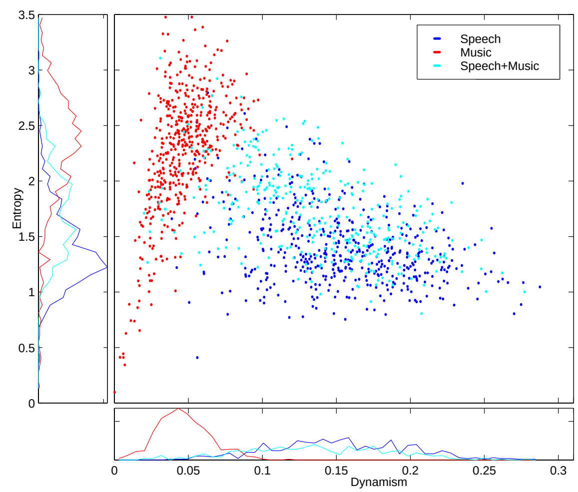


  - grab from TV subtitles?

Audio segmentation

- **Broadcast audio includes music, noise etc.**
- **Segmentation is important for recognition**
  - speaker identity tagging, model adaptation
  - excluding nonspeech segments
- **Can use generic models of similarity/difference**
- **Look at statistics of speech model output**





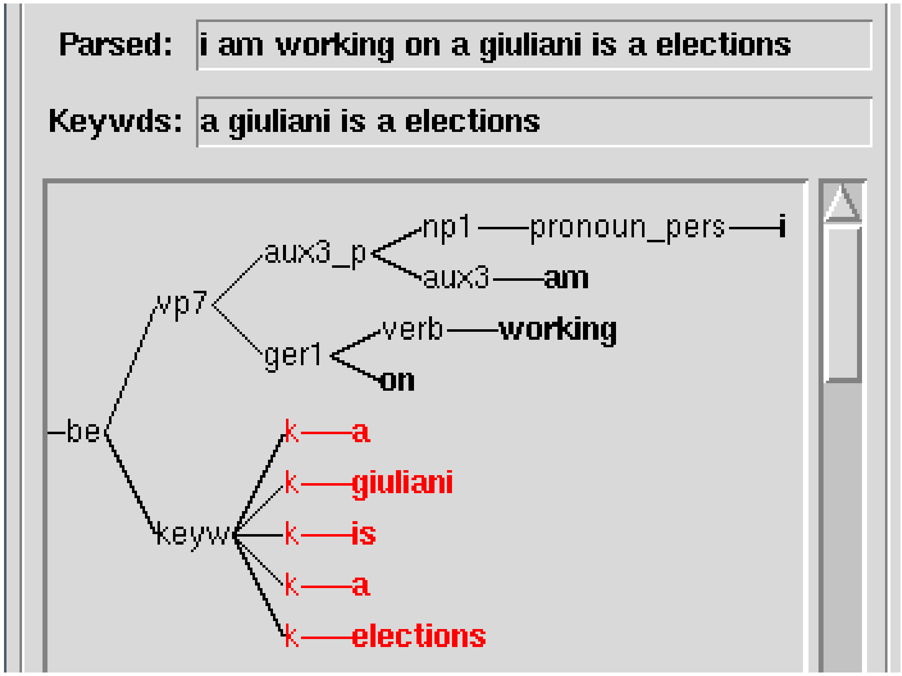
- e.g. dynamism $\frac{1}{N}\sum_n \sum_i [p(q_n^i) - p(q_{n-1}^i)]^2$

Information retrieval: Text document IR

- **Given query terms T_q, document terms $T_{D(i)}$ how to find and rank documents?**
- **Standard IR uses 'inverted index' to find:**
  - one entry per term T_D, listing all documents $D(i)$ containing that term
- **Documents are ranked using "tf • idf"**
  - **tf** (term frequency) = how often term is in doc
  - **idf** (inverse document frequency) = how many (how few) docs contain term
- **Performance measures**
  - precision: (correct found)/(all found)
  - recall: (correct found)/(all correct)
  - mean reciprocal rank - for specific targets

Queries in Thisl

- **Original idea: speech in, speech out**
- **Try to 'understand' queries**
    - hand-built grammar:



    - .. but keywords better

- **Phonetic matching with speech input**
    - search 'phone lattice' recognizer output?

Thisl User Interface

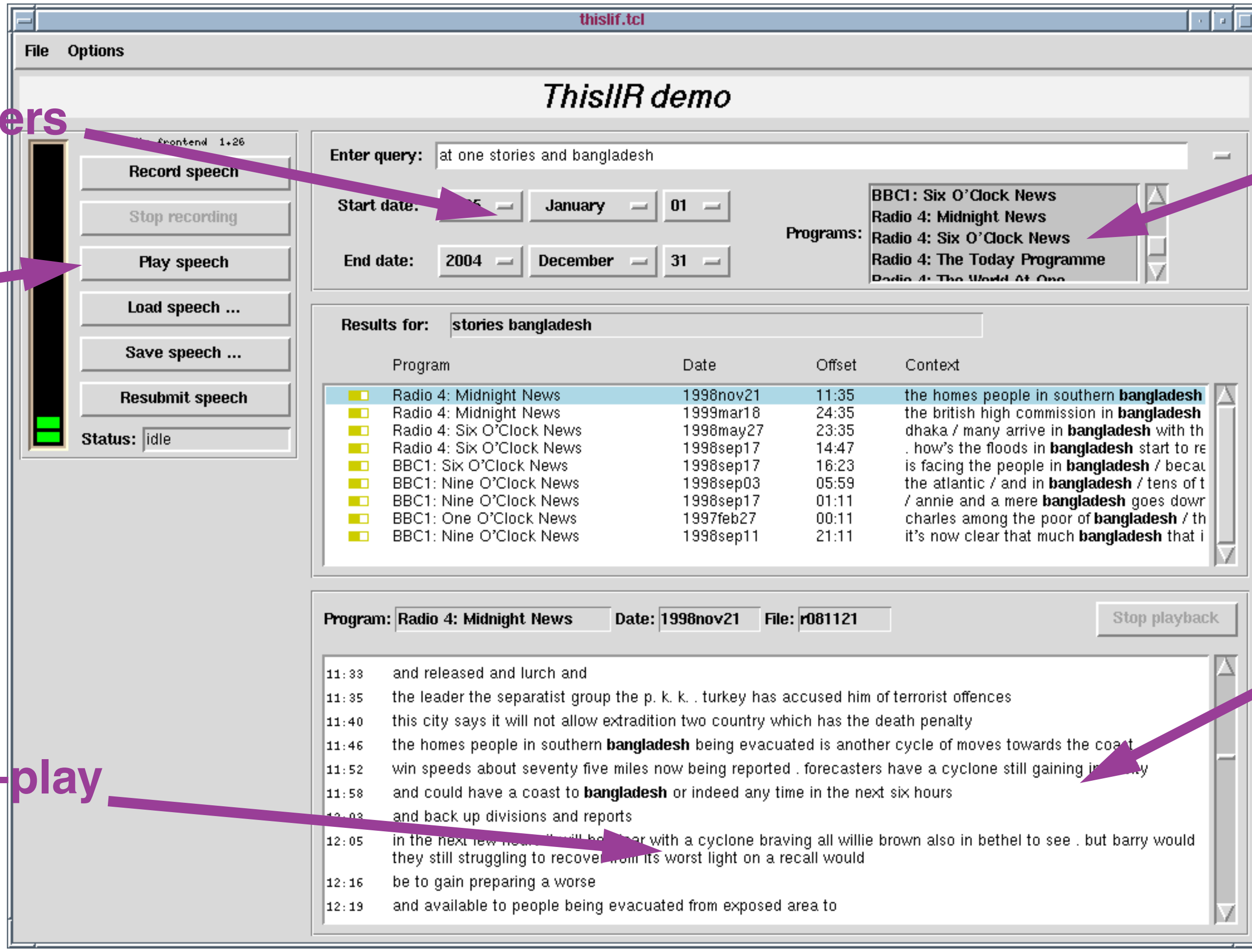
Date filters

Program filter

Speech input

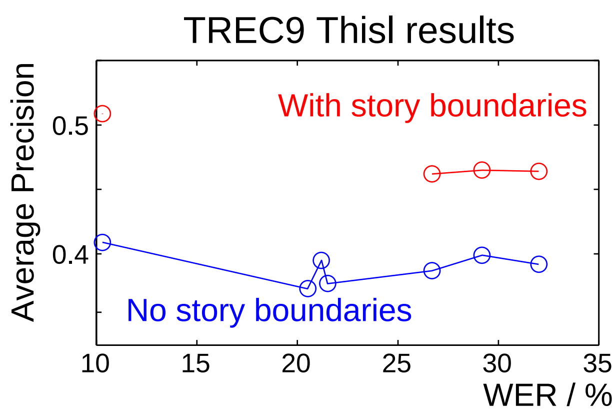
Pauses & sentence breaks shown

click-to-play



Thisl SDR performance

- **NIST Text Retrieval Conference (TREC), Spoken Documents track**
  - 500 hours of data → need fast recognition
  - set of ‘evaluation queries’ + relevance judgments
- **Components tried in different combinations**
  - different speech transcripts (subtitles, ASR)
  - different IR engines & query processing
- **Performance of systems**
  - ASR less important than IR (query expansion...)



Outline

1. ASR wrap-up

2. Spoken Document Retrieval

3. **General audio databases**
   - Nonspeech audio retrieval
   - Audio mixtures & CASA

4. Open issues

3 Real-world audio

- **Speech is only part of the audio world**
  - word transcripts are not the whole story
- **Large audio datasets**
  - movie & TV soundtracks
  - events such as sports, news ‘actualities’
  - situation-based audio ‘awareness’
  - personal audio recording
- **Information from sound**
  - speaker identity, mood, interactions
  - ‘events’: explosions, car tires, bounces...
  - ambience: party, subway, woods
- **Applications**
  - indexing, retrieval
  - description/summarization
  - intelligent reaction

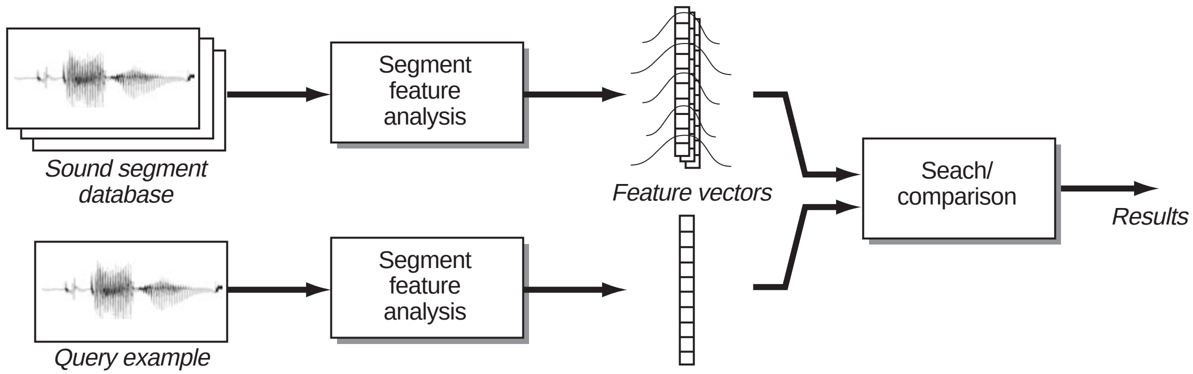
Multimedia Description: MPEG-7

- **MPEG has produced standards for audio / video data compression (MPEG-1/2/4)**
- **MPEG-7 is a standard for metadata: describing multimedia content**
  - because search and retrieval are so important
- **Defines descriptions of time-specific tags, ways to define categories, specific category instances**
- **+ Preliminary feature definitions e.g. for audio:**
  - spectrum: centroid, spread, flatness
  - harmonicity: degree, stability
  - pitch, attack time, melody structure ...

http://www.darmstadt.gmd.de/mobile/MPEG7/Documents.html

Muscle Fish "SoundFisher"

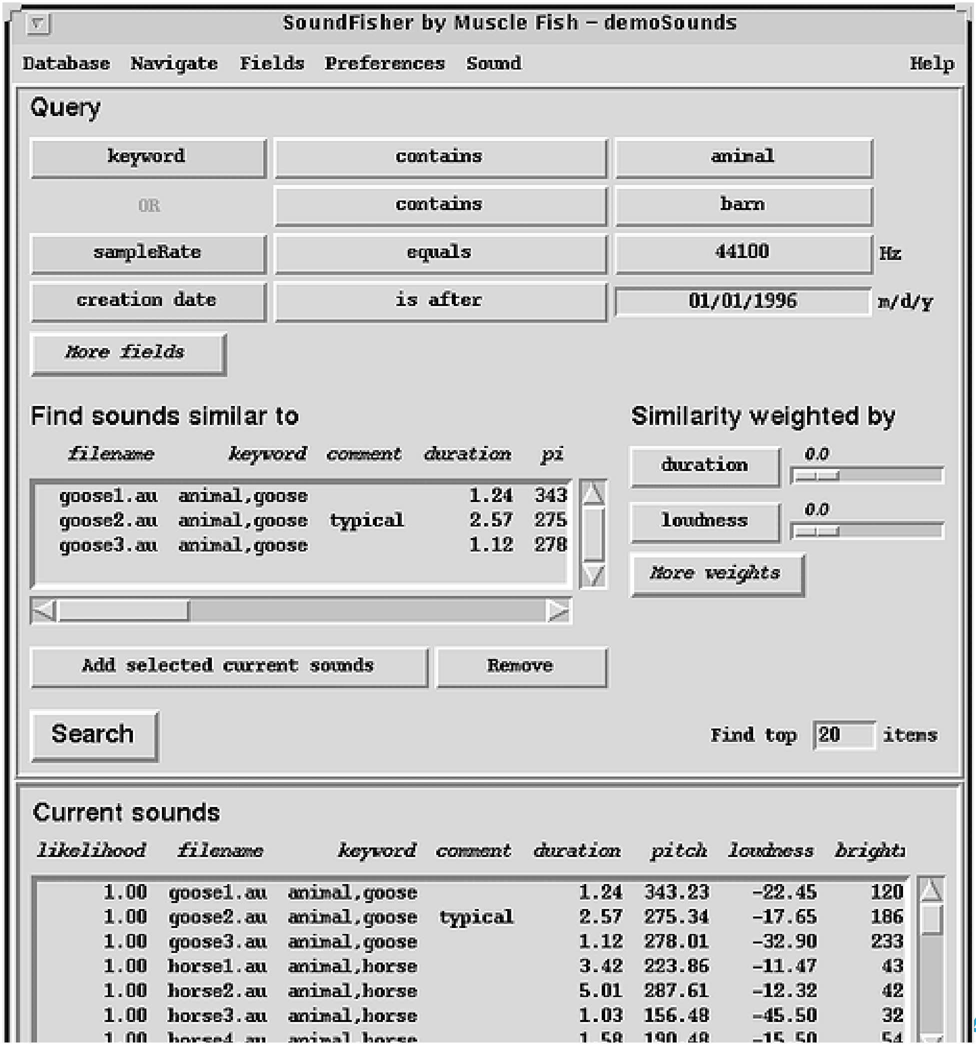
- **Access to sound effects databases**



- **Features (time series contours):**
  - loudness, brightness, pitch, cepstra
- **Query-by-example**
  - direct correlation of contours (normalized/not)
  - comparison of value histograms (time-collapsed)
- **Always global features**
  - a mixture of two sounds looks like neither

SoundFisher user interface

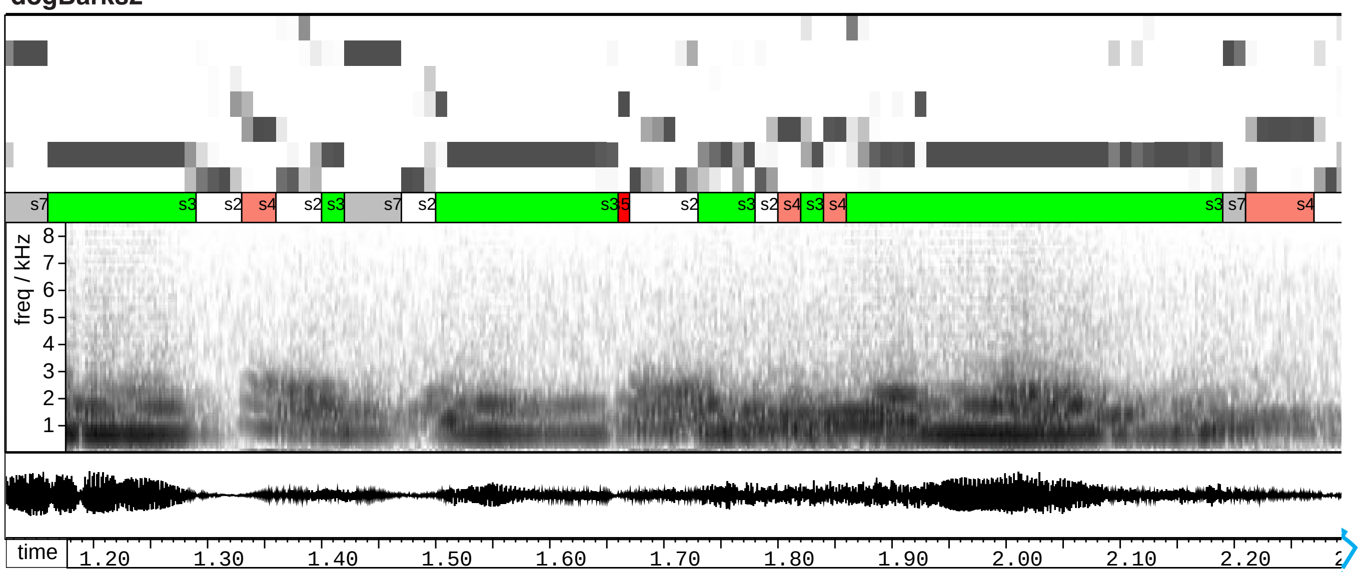
- **Principle query mechanism is "sounds like"**



HMM modeling of nonspeech

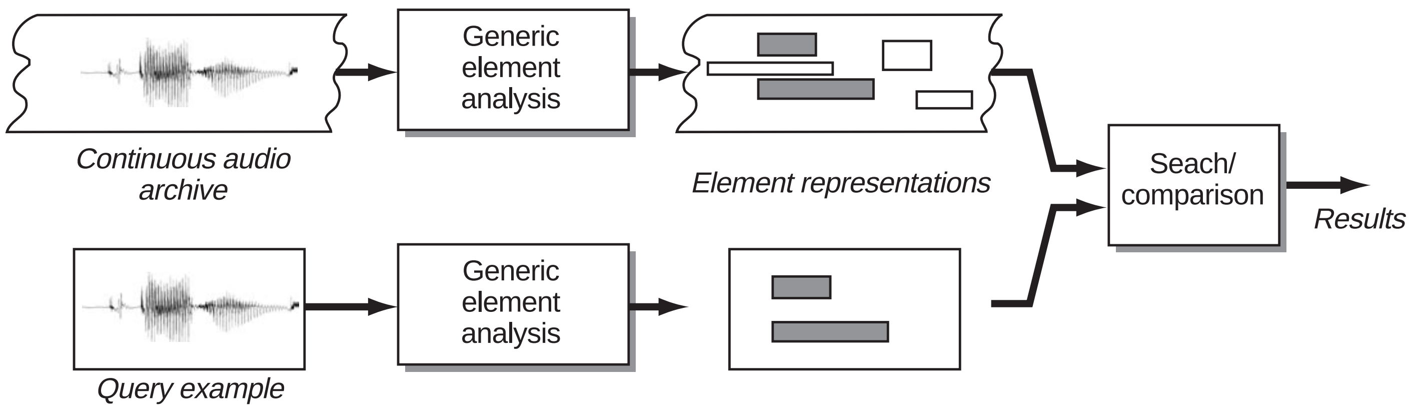
- **No sub-units defined for nonspeech sounds**
  - but can still train HMMs with EM
- **Final states depend on EM initialization**
  - labels / clusters
  - transition matrix
- **Have ideas of what we'd like to get**
  - investigate features/initialization to get there

dogBarks2



Indexing for soundtracks

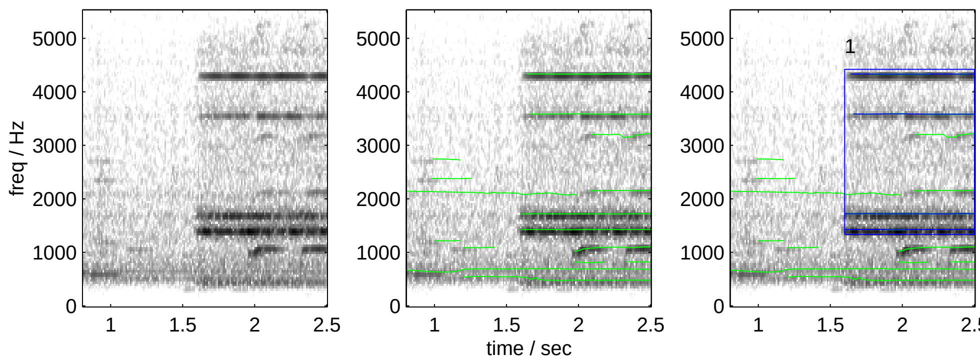
- **Any real-world audio will have multiple simultaneous sound sources**
- **Queries typically relate to one source only**
  - not a source in a particular context
- **Need to index accordingly:**



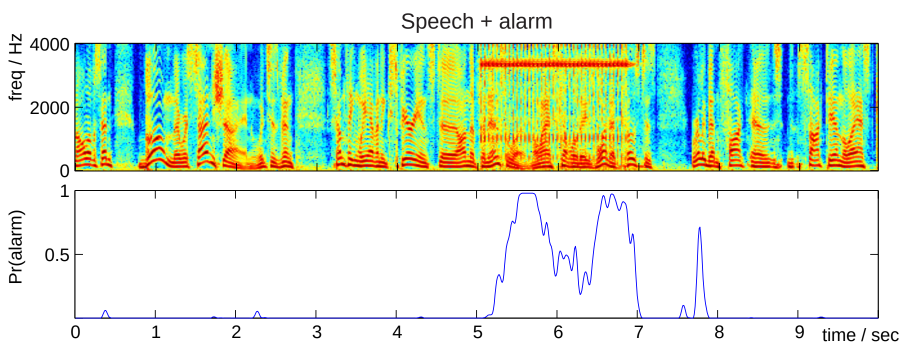
  - analyze sound into source-related elements
  - perform search & match in that domain

Alarm sound detection

- **Alarm sounds have particular structure**
  - people 'know them when they hear them'
- **Isolate alarms in sound mixtures**



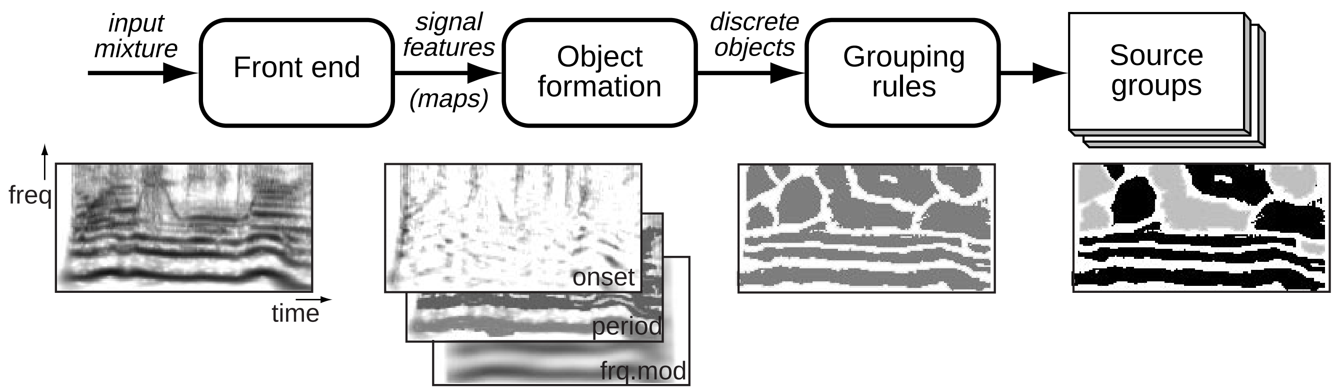
  - sinusoid peaks have invariant properties



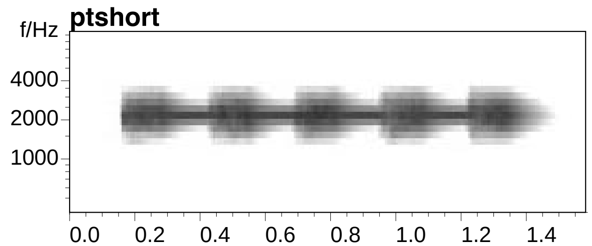
  - cepstral coefficients are easy to model

Computational Auditory Scene Analysis

- **Real-world sounds come from multiple sources**
- **Psychoacoustics gives 'scene analysis' cues:**
  - common onset
  - common harmonicity
- **Suitable for computational models:**

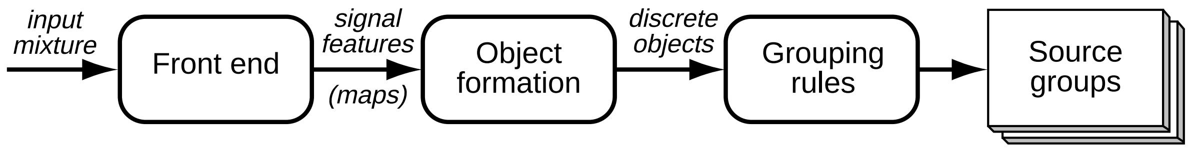


- **What about masking & 'illusions'?**

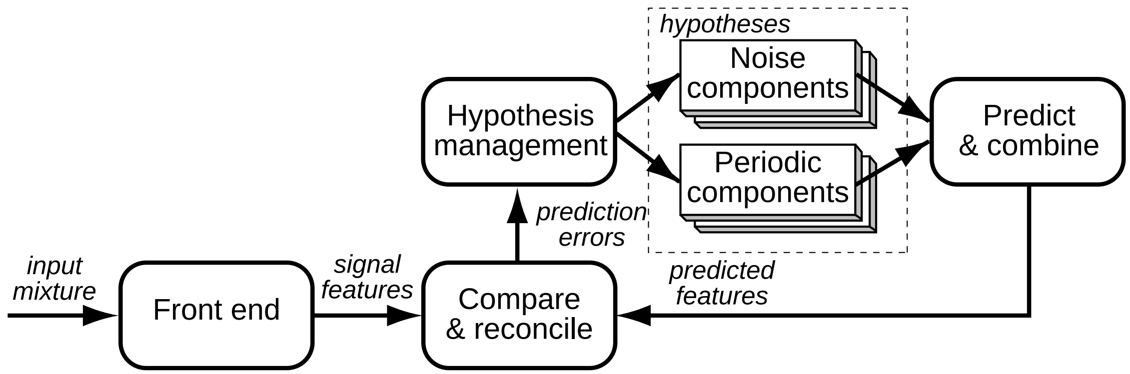


Top-down CASA

- **Data-driven (bottom-up) fails for noisy, ambiguous sounds (most mixtures!)**

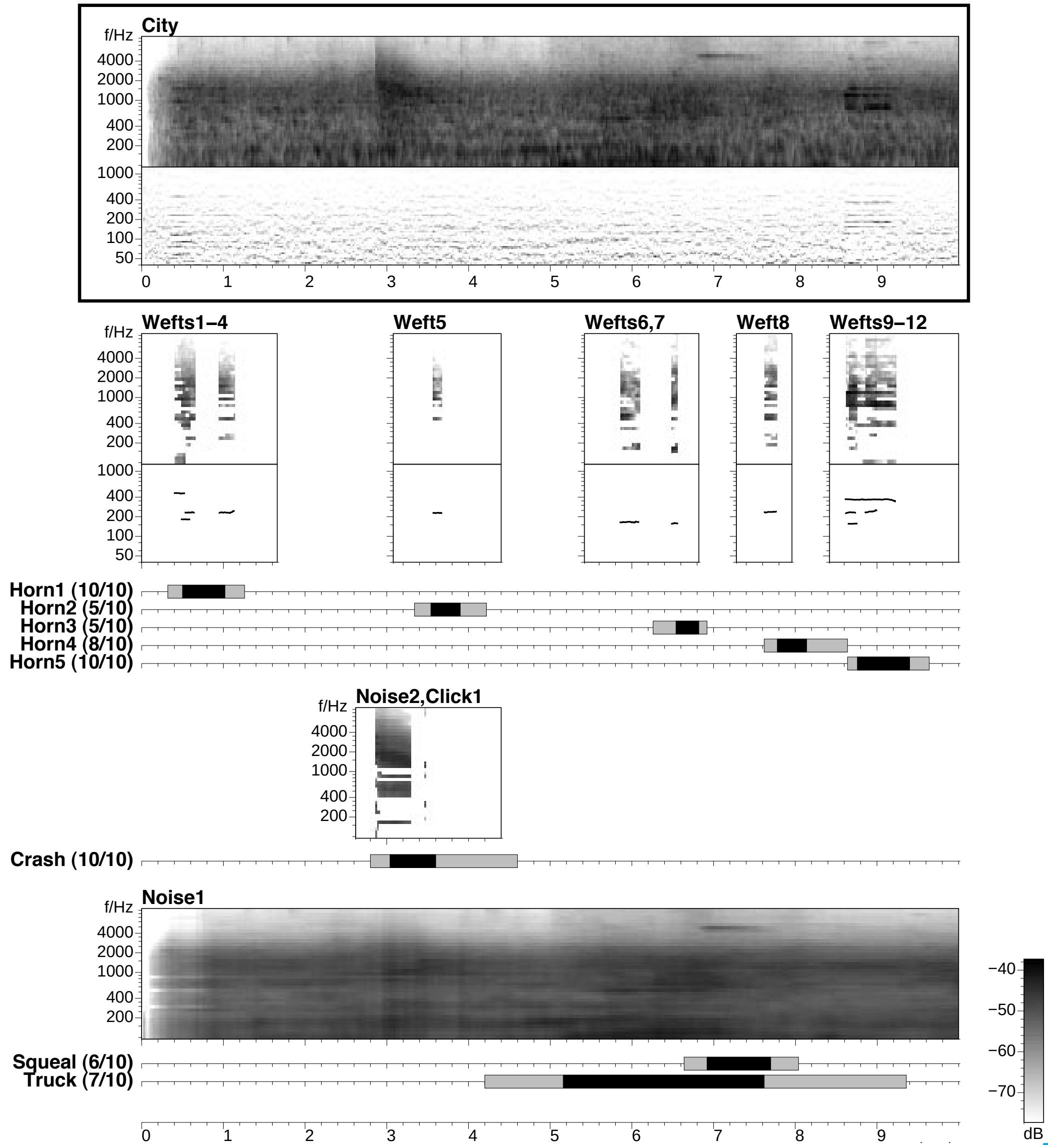


- **Need top-down constraints:**



  - vocabulary of generic elements to fit sound
    ... bottom of a hierarchy?
  - account for entire scene
  - driven by prediction failures
  - pursue alternative hypotheses

CASA example



Outline

1. **ASR wrap-up**
2. **Spoken document retrieval**
2. **General audio databases**
4. **Open issues**
   - Speech recognition
   - Sound source separation
   - Information extraction & visualization
   - Learning from audio

4 Open issues 1: Speech recognition

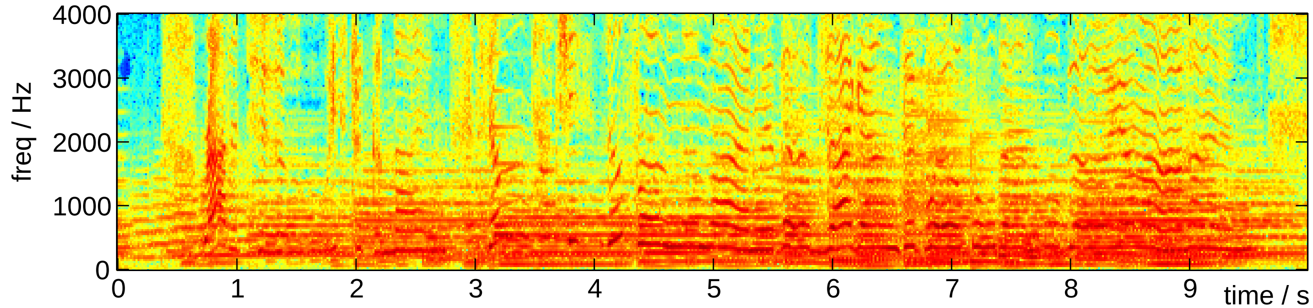
- **Speech recognition is good & improving**
  - but rate of improvement has slowed:
    BN WER: 1997=22% 1998=16% 1999=14%
  - limits to benefits of additional data?
- **Problem areas:**
  - noisy speech (meetings, cellphones)
  - informal speech (casual conversations)
  - speaker variations (style, accent)
- **Is the current approach correct?**
  - MFCC-GMM-HMM systems are optimized
  - new approaches can't compete
  - but: independence, classifier, HMMs...

Open issues 2: Sound mixtures

- **Real-world sound always consists of mixtures**
  - we experience it in terms of separate sources
  - 'intelligent' systems must do the same
- **How to separate sound sources?**
  - exact decomposition ('blind source separation')
  - extract cues
  - overlap, masking
    → top-down approaches, analysis-by-synthesis
- **How to represent & recognize sources?**
  - which features, attributes?
  - hierarchy of general-to-specific classes...

Open issues 3: Information & visualization

- **Spectrograms are OK for speech, often unsatisfactory for more complex sounds**



  - frequency axis, intensity axis – time axis?
  - separate spatial, pitch, source dimensions

- **Visualization may not be possible .. but helps us think about sound features**

- **Different representations for different aspects**
  - best for speech, music, environmental, ...

Open issues 4: Learning from audio

- **HMMs (EM, Baum-Welch etc.) have had a huge impact on speech, handwriting ...**
  - very good for optimizing models
  - little help for determining model structure
- **Applicable to other audio tasks?**
  - e.g. textures, ambience, vehicles, instruments
- **Problems:**
  - finding the right model structures
  - constraining what the models learn: initial clustering, target labelling
- **How to leverage large databases, bulk audio**
  - unsupervised acquisition of classes, features
  - the analog of infant development

Outline

1. ASR details
2. ASR in practice
2. Real-world audio
4. Open issues

Course retrospective

Fundamentals

L1: **DSP**

L2: **Acoustics**

L3: **Pattern recognition**

L4: **Auditory perception**

Audio processing

L5: **Signal models**

L6: **Music analysis/ synthesis**

L7: **Audio compression**

L8: **Spatial sound & rendering**

Speech recognition

L9: **Speech features**

L10: **Sequence recognition**

L11: **Recognizer training**

L12: **Systems & applications**

Summary

- **Large Vocabulary speech recognition**
  - dictation is a limited domain
  - noisy recognition useful for indexing
  - \+ speech understanding?

- **Recognizing nonspeech audio**
  - lots of other kinds of acoustic events
  - speech-style recognition can be applied

- **Open questions**
  - lots of things that we don't know