Content-based analysis and indexing for speech, sound & multimedia

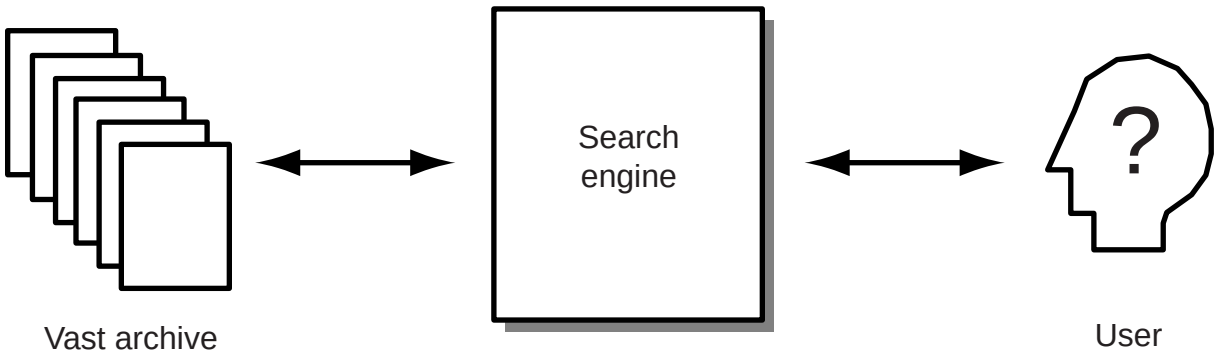
Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1. About content-based indexing
2. Related work
3. An overview of the project
4. Some specific pieces
5. Future plans

1 About content-based indexing

- **Problem: Automating search in large archives**



- **“Information retrieval” (IR)**
- **E.g.:**
  - searching the web
  - searching broadcast archives
  - automatic monitoring...

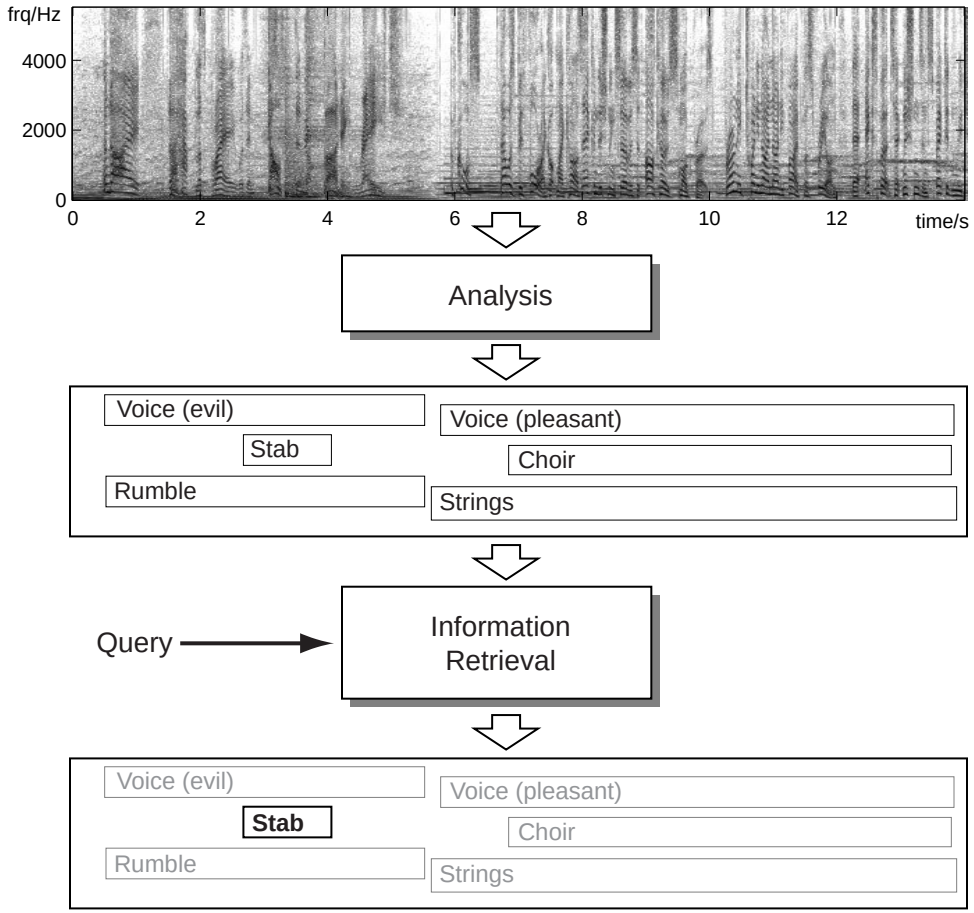
Varieties of Information Retrieval (IR)

- **Many different search situations:**

Archive	*Queries*	*Technology*
Text	terms	Text IR (tf • idf, “term space”)
Speech	terms	ASR + Text IR
Multimedia	terms	Text IR on annotations
Images, video	examples/ sketches	Global image similarity metrics
Sound	examples/ categories	Global sound similarity metrics
Sound mixtures	**examples**	**object-based similarity**
	terms	**term-to-feature mapping**

- plus combinations (e.g. sound mixtures + video)

Content analysis of sound mixtures



- **Use local features to find individual objects**
- **Objects must mirror subjective experience**

Outline

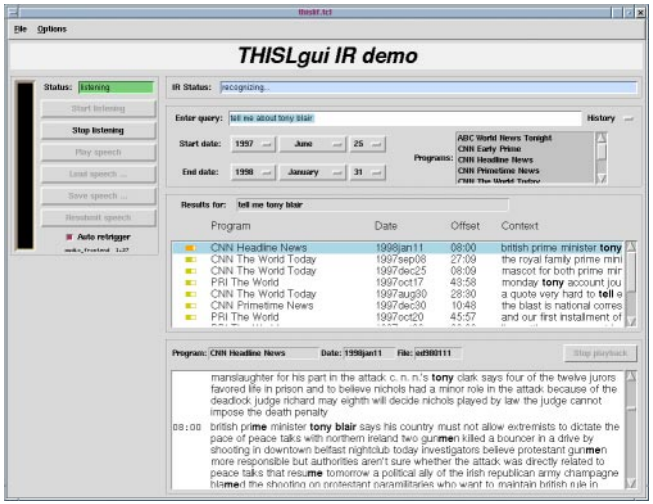
1. About content-based retrieval
2. **Related work**
   - Text-based IR
   - Spoken document retrieval
   - Image and video retrieval
   - Multimedia systems
   - Sound effects indexing
   - MPEG7 ‘metadata’
3. An overview of the project
4. Some specific pieces
5. Future plans

2 Text-based IR

- **e.g. Web search engines**
- **Metric:**
  term frequency • inverse document frequency
  - emphasizes unusual words
  - distances in Euclidean ‘term space’
- **Decomposition of documents into searchable atoms is (almost) trivial**
  - words are easily isolated, close to ideal terms
  - some problems, hence stemming

Spoken document retrieval

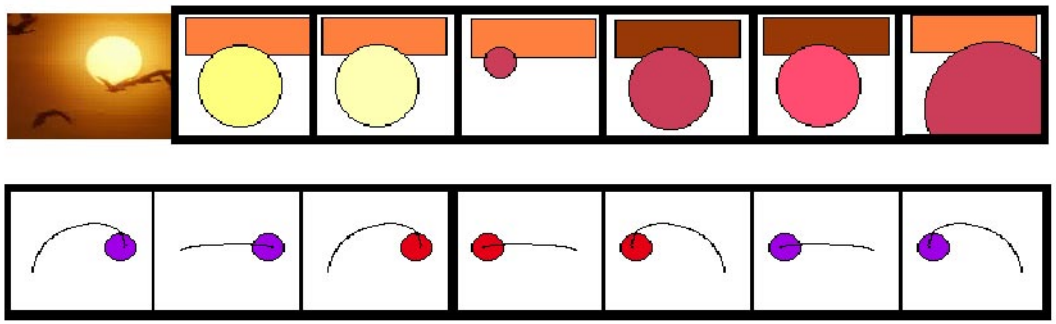
- **Information retrieval for speech recordings: Convert to text with speech recognition**
  - e.g. Thisl project (news broadcasts)



- **Speech recognition errors not the limiting factor**
  - TREC-98 results: average precision 0.5→0.4
- **Output should be original audio**
  - best not to show the recognizer output!

Image and video retrieval

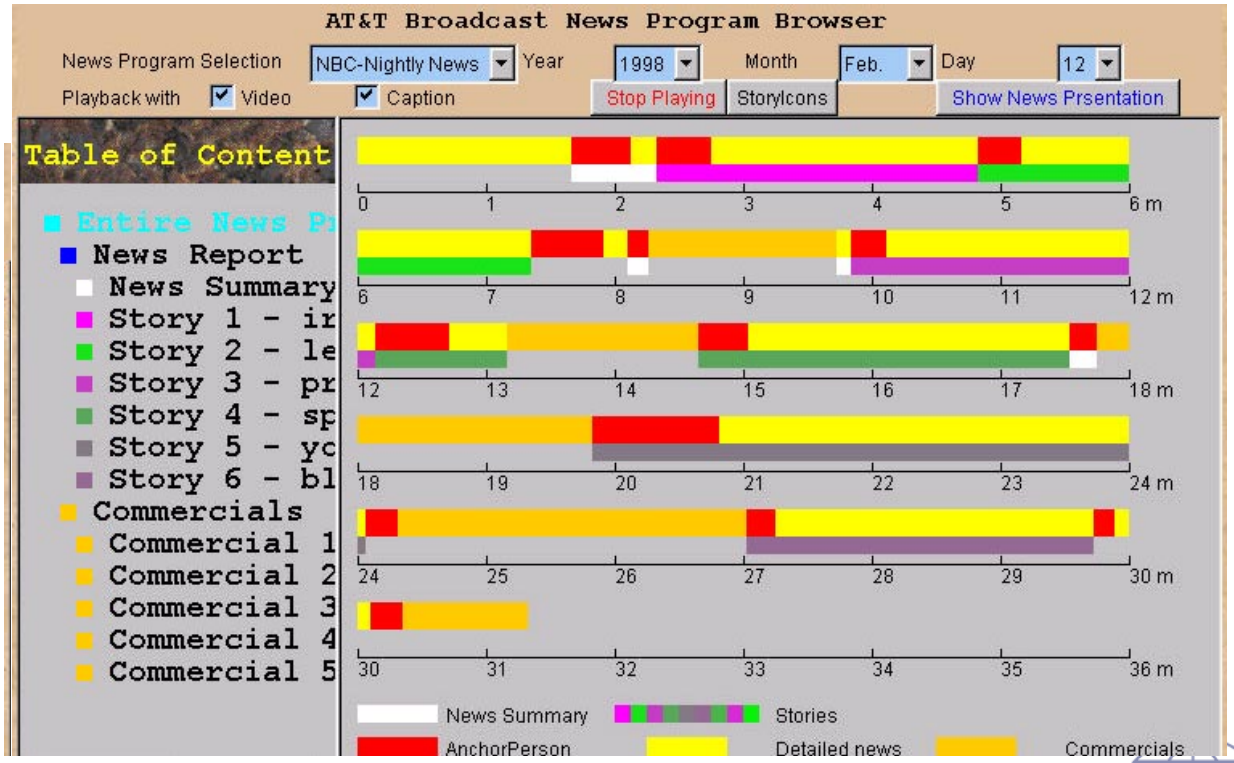
- **e.g. Query By Image Content (QBIC) (IBM 1995)**
  - templates, color, texture
- **VideoQ (Columbia 1999)**
  - sketching for images and video
  - color, shape, size, position, motion



- **Image 'objects'?**
  - analog of terms in text
  - acquired by unsupervised clustering
  - object frequency • inverse image frequency?

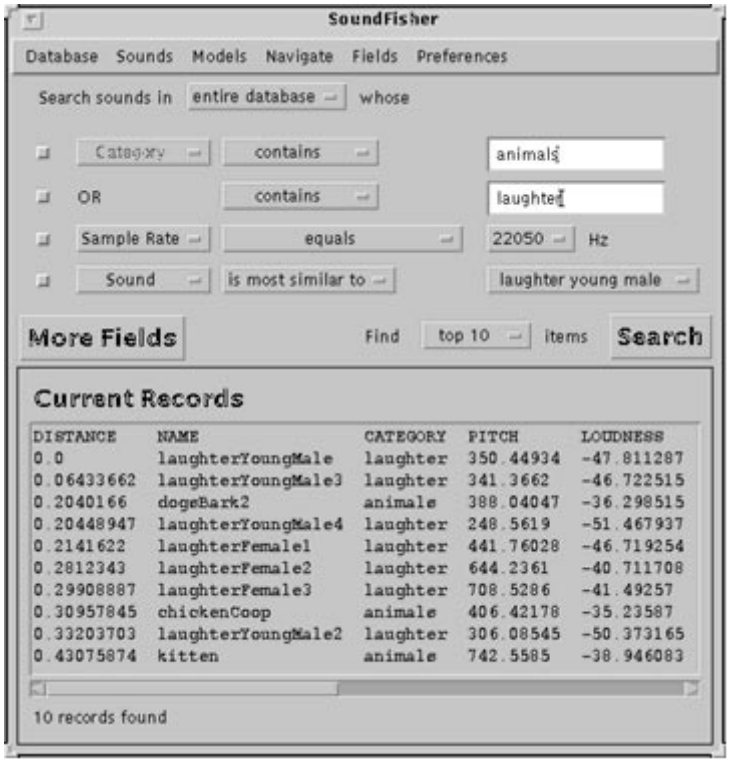
Multimedia systems

- **Informedia (CMU, 1996-)**
  - ASR + video cuts + OCR of screen + IR
- **AT&T multilevel structuring**
  - exploit knowledge of genre (TV news shows)
  - multiple special-purpose information sources



Sound effects indexing

- **Muscle Fish “SoundFisher”**
  - browser for sound-effects archives
  - define multiple ‘perceptual’ feature dimensions
  - no attempt to separate objects in mixtures



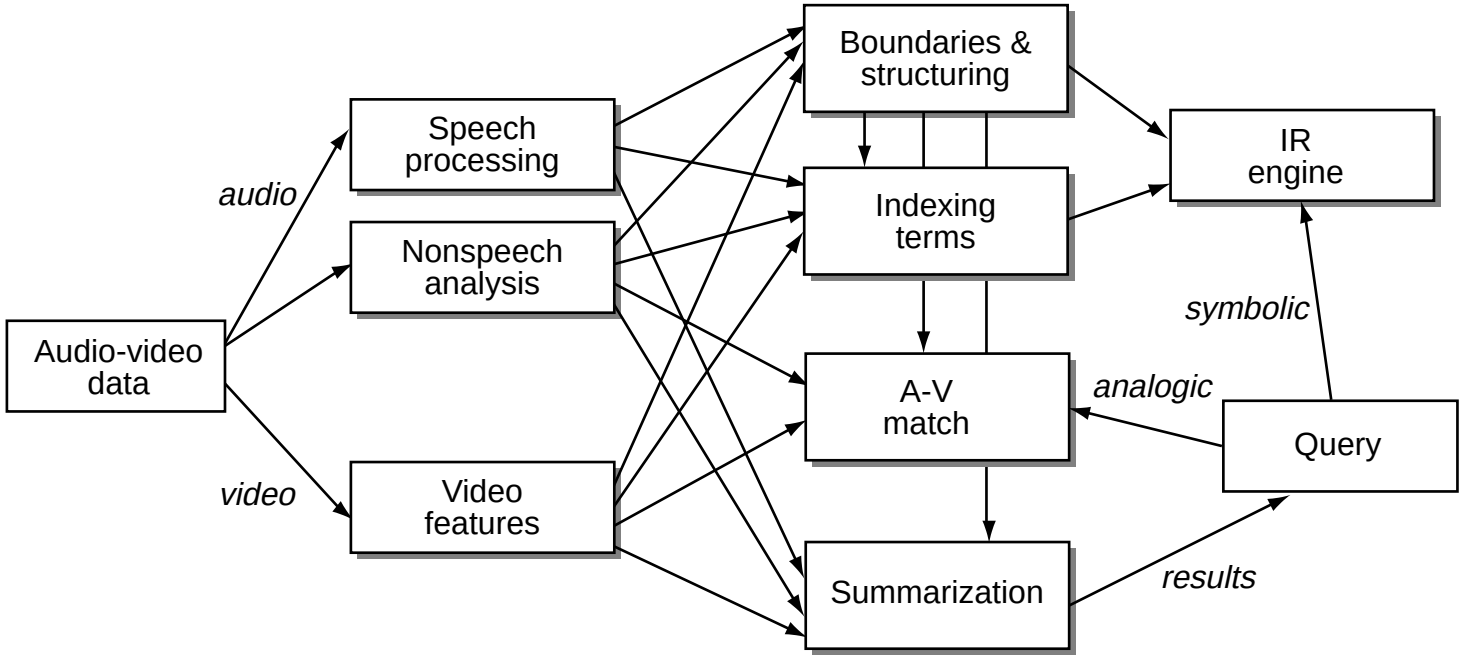
MPEG-7 'Metadata'

- **MPEG is known for audio/video *compression* standards;**
  also developing standards for use in *search and indexing*
- **MPEG-7 will be a standard format for *metadata*:**
  Well-defined categories for content description
- **Mostly just framework, some actual categories**
- **How to *derive* descriptors from content is not specified**

Outline

1. About content-based retrieval
2. Related work
3. **An overview of the project**
   - Boundaries & structuring
   - Query forms
   - Summarization
   - Evaluation
4. Some specific pieces
5. Future plans

3 Audio-video content-based retrieval: System overview



- **Fusion of audio + video (+...?) information**
- **Different query forms**

Boundaries and structuring

- **Multimedia documents lack structure**
- **Changes relatively easy to *detect***
  - if we don't have to *characterize* the change
- **Audio and video are complementary**
- **Boundaries define structure e.g. stories**
- **May be able to identify genre based on structure pattern (TV, news, interviews, sports)**
  - notice *repetition* of particular segments (title sequences, commercials etc.)

Forms of query

- **Traditional term-based**
  - mapping of terms to audio/video features?
  - ... plus all the usual lexical ambiguities
  - literal vs. thematic terms
- **Similarity e.g. by example**
  - easy once you have initial hits/documents
  - but: which aspects of the example?
- **User-provided example e.g. a ‘sketch’**
  - better idea of which parts of a sketch are salient and which to ignore
  - audio sketches?
  - spoken words?

Summarization of results

- **Multimedia 'hits' are hard to present**
  - multi-media → many aspects
  - some are intrinsically temporal
- **Video presentation**
  - salient stills/story board
  - sped-up video
- **Spoken content**
  - textual summarization based on salience & recognizer confidence
  - audio selection based on prosodic cues
- **Audio content**
  - choosing 'distinctive' events
  - visual representation?
  - timescale modification?

Evaluation

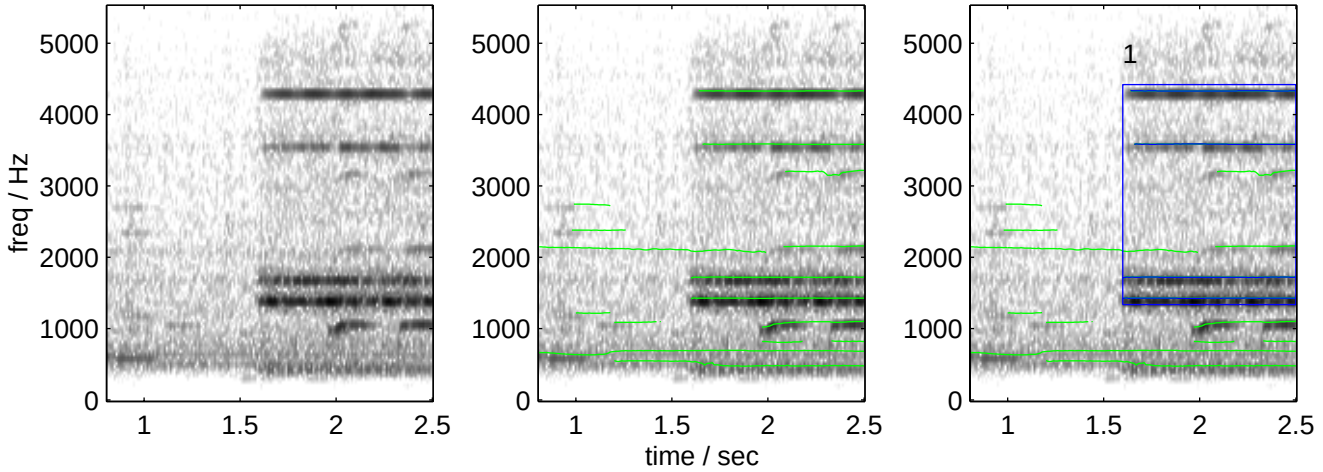
- **Multimedia IR is an emerging field**
  - no consensus on what the task really is
  - no common evaluation metrics
- **Evaluation is critical**
  - sanity check on progress
  - affords ‘fundability’
- **How to do it?**
  - quantitative tests e.g. datasets and queries
  - qualitative user evaluation
- **Prototype demos**
  - de rigueur...
  - also provide input to design:
    what kind of queries will people really ask?

Outline

1. About content-based retrieval
2. Related work
3. An overview of the project
4. **Some specific pieces**
   - Object-based audio analysis
   - Speech recognition for retrieval
   - Music processing
   - Machine learning of terms
5. Future plans

4 Object-based audio analysis: Computational Auditory Scene Analysis

- **Deconstructing sound mixtures**

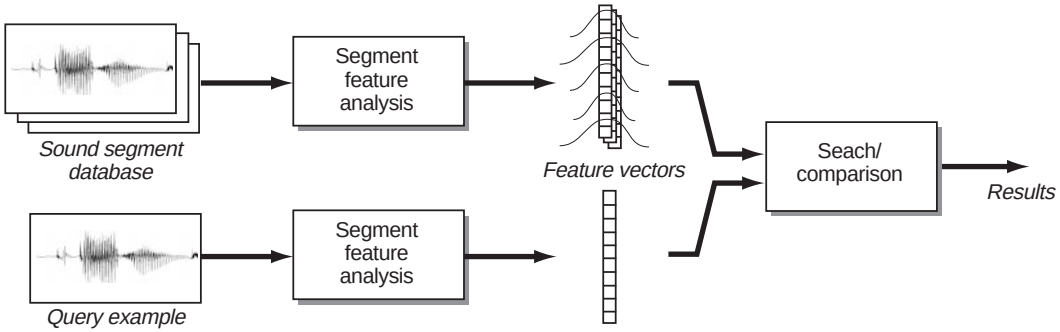


  - representation of energy in time-frequency
  - formation of atomic elements
  - grouping by common properties (onset &c.)

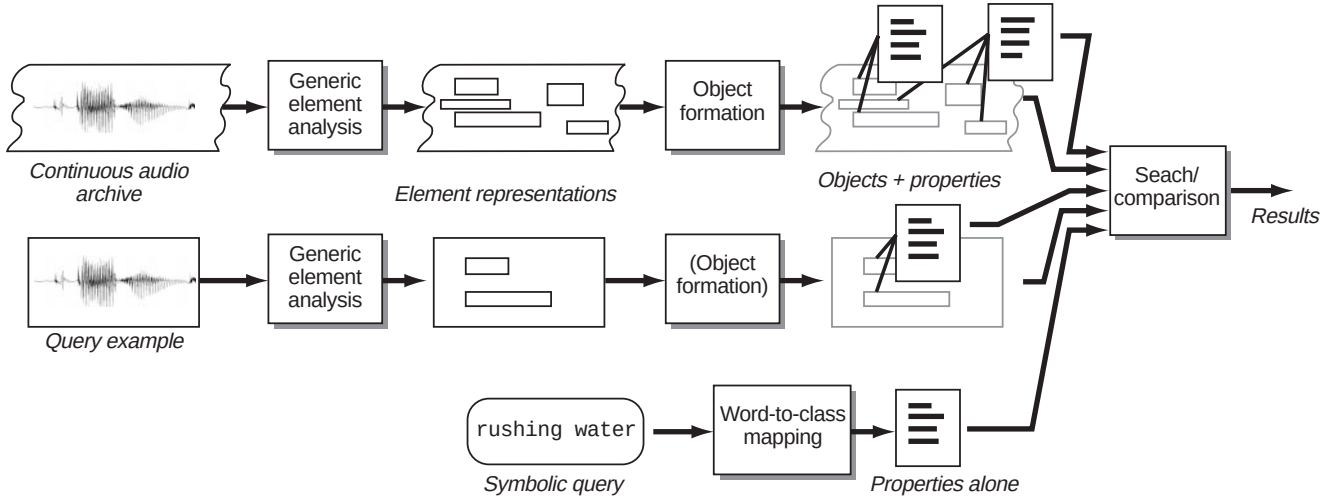
- **Ambiguous/noisy sounds need more...**
  - top-down constraints
  - multiple alternative hypotheses

Retrieval of sound objects

- **Muscle Fish system uses global features:**



- **Mixtures → need elements & objects:**



- features are calculated over grouped subsets

Speech recognition for retrieval

- **Words are not enough;**
  Confidence-tagged alternate word hypotheses
- **Other useful information:**
  - speaker change detection
  - speaker characterization
  - phrasing & timing
  - prosodic features
- **Integration with other analyses**
  - segmentation for adaptation
  - nonspeech events to ignore
  - video-derived information?

Music processing

- **Music is a highly-structured special case**
- **Need to detect it at the least**
- **Algorithms to extract special information**
  - melody, harmony, rhythm
  - instrument identification
  - genre classification
- **Body of existing research...**

Machine learning of patterns & terms

- **What can you do with a large unlabeled training set (e.g. multimedia clips from the web)?**
  - bootstrap learning: look for common patterns
  - have to learn generalizations in parallel: e.g. self-organizing maps, EM HMMs
  - post-filtering by humans may find 'meaning' in clusters
- **Associated text annotations provide a very small amount of labeling**
  - .. but for a very large number of examples – sufficient to obtain purchase?
  - maximize label utility through NLP-type operations (expansion, disambiguation etc.)
  - goal is automatic term-to-feature mapping for term-based content queries

Outline

1. About content-based retrieval
2. Background technologies
3. An overview of the project
4. Some specific pieces
5. **Future plans**

5 Future plans

- **Obtain funding:**
  - Thisl follow-on with the EU?
  - NSF: sound IR, also audio-video (with Zakhor)
  - other sources?
- **Choose a task and an archive**
  - multimedia clips on the web
  - existing archives e.g. taped UCB lectures
  - speech/broadcast archives
  - meeting recorder
- **Begin developing features**
  - computational auditory scene analysis
  - .. need to apply to large corpora
- **Online demo ASAP?**
  - to help clarify the problem