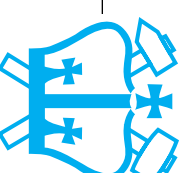


Lecture 12: Multimedia Indexing

- 1 Spoken document retrieval
- 2 Audio databases
- 3 Open issues

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

Columbia University Dept. of Electrical Engineering
Spring 2007



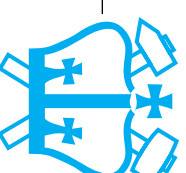
1 Spoken Document Retrieval (SDR)

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

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

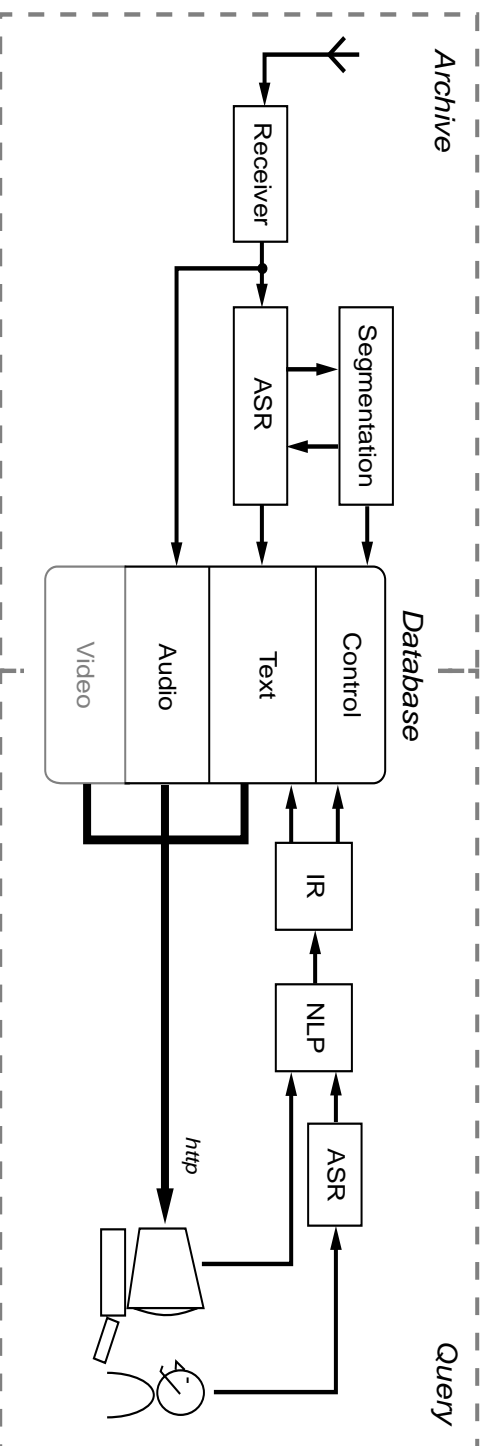
F4: AT THIS STAGE OF THE **ACCOUNTING FOR SEVENTY SCOTCH ONE** LEADER DANIEL ORTEGA IS IN SECOND PLACE THERE WERE TWENTY THREE PRESIDENTIAL CANDIDATES **OF THE ELECTION**

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



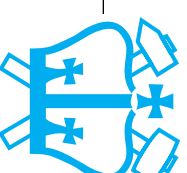
The THISL SDR system

- Original task: BBC newsroom support



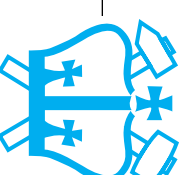
- **How to build the database:**

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



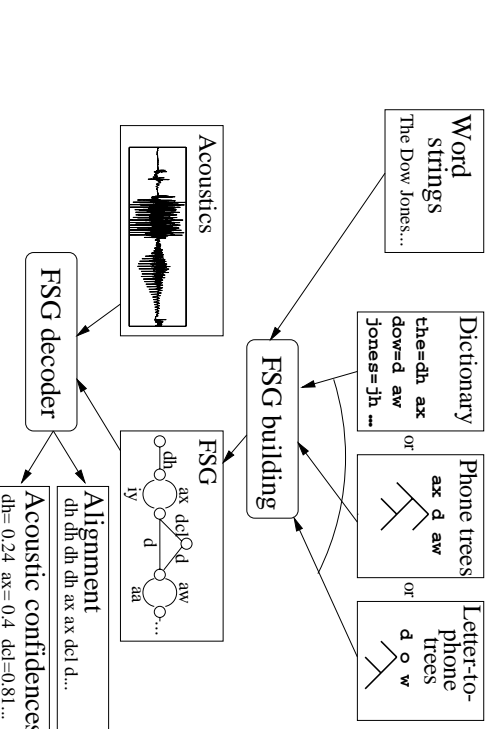
Building a new recognizer

- **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

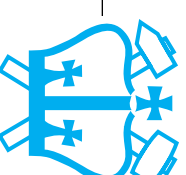


Vocabulary extension

- **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**

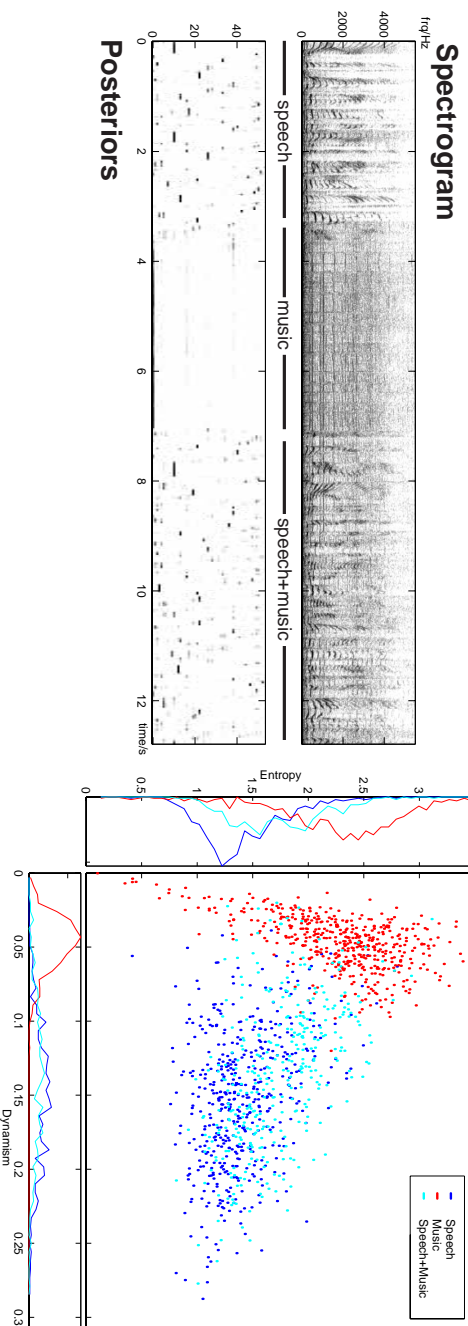


- grab from TV subtitles?

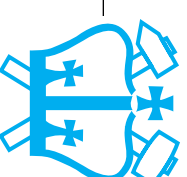


Audio segmentation

- 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



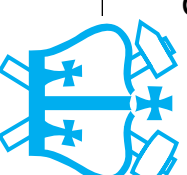
- e.g. dynamism $\frac{1}{N} \sum_n \sum_i [p(q_n^i) - p(q_{n-1}^i)]^2$



Information retrieval:

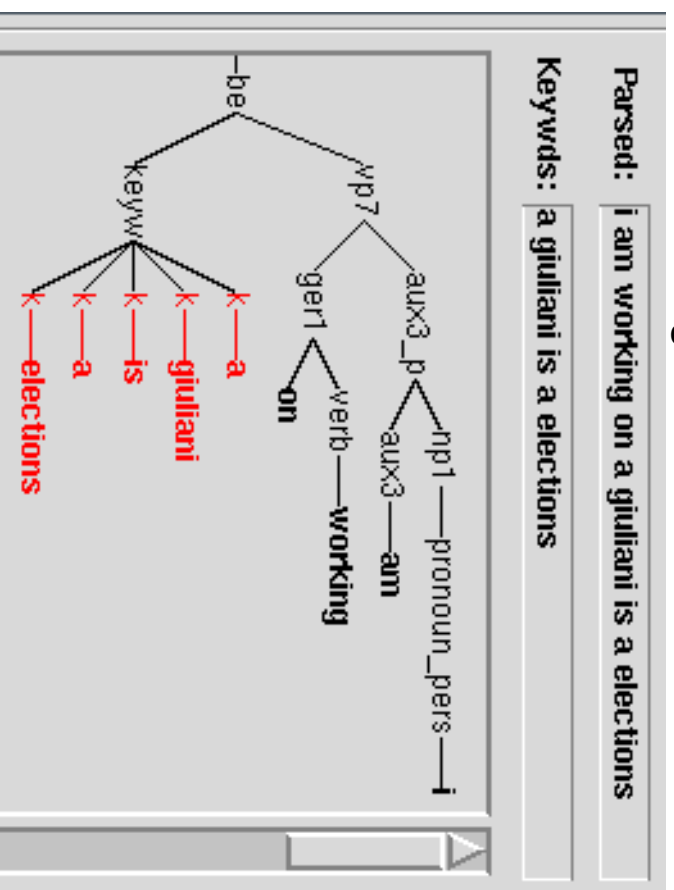
Text document IR

- 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

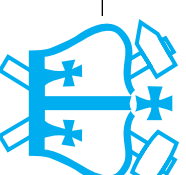


Queries in This!

- 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?



This! User Interface

File Options

thsi!.tcl

Date filters

Enter query: at one stories and bangladesh

Start date: 15 January 01

End date: 2004 December 31

Programs:
BBC1: Six O'Clock News
Radio 4: Midnight News
Radio 4: Six O'Clock News
Radio 4: The Today Programme

Speech input

Record speech
Stop recording
Play speech
Load speech ...
Save speech ...
Resubmit speech
Status: idle

Program

Results for: stories bangladesh

Program	Date	Offset	Context
Radio 4: Midnight News	1998nov21	11:35	the homes people in southern bangladesh
Radio 4: Midnight News	1999mar8	24:35	the british high commission in bangladesh
Radio 4: Six O'Clock News	1998may27	23:35	dhaka / many arrive in bangladesh with th
Radio 4: Six O'Clock News	1998sep17	14:47	. now's the floods in bangladesh start to re
BBC1: Six O'Clock News	1998sep17	16:23	is facing the people in bangladesh / decal
BBC1: Nine O'Clock News	1998sep03	05:59	the atlantic / and in bangladesh / tens of t
BBC1: Nine O'Clock News	1998sep17	01:11	/ arnie and a mere bangladesh goes down
BBC1: One O'Clock News	1997feb27	00:11	charlies among the poor of bangladesh / th
BBC1: Nine O'Clock News	1998sep11	21:11	it's now clear that much bangladesh that i

Program: Radio 4: Midnight News

Date: 1998nov21

File: r081121

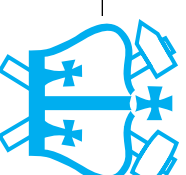
Stop playback

click-to-play

11: 33 and released and lunch and
11: 35 the leader the separatist group the p. k. k. . turkey has accused him of terrorist offences
11: 40 this city says it will not allow extradition two country which has the death penalty
11: 46 the homes people in southern bangladesh being evacuated is another cycle of moves towards the coast
11: 52 win speeds about seventy five miles now being reported . forecasters have a cyclone still gaining in
11: 58 and could have a coast to bangladesh or indeed any time in the next six hours
12: 03 and back up divisions and reports
12: 05 in the next few hours . with a cyclone braving all willie brown also in berthel to see . but barry would
12: 15 they still struggling to recover from its worst light on a recall would
12: 16 be to gain preparing a worse
12: 19 and available to people being evacuated from exposed area to

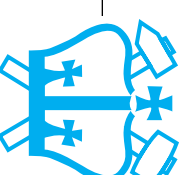
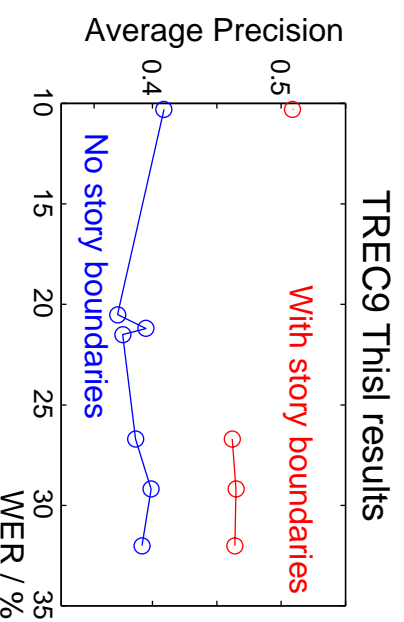
Program filter

Pauses & sentence breaks shown



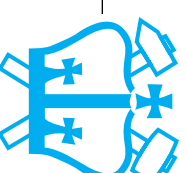
This! SDR performance

- **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...)



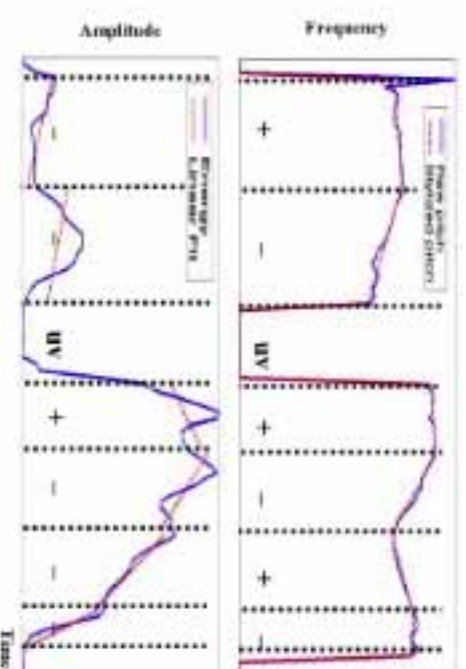
Speaker Identification

- Complement to speech recognition:
Identify the **speaker**, regardless of the **words**
- Different forms of the problem:
 - speaker **segmentation**
 - speaker **identification**
 - speaker **verification**
- **Factors:**
 - amount of **training** data (10 s .. 20 min)
 - amount of **test** data (3 s .. 5 min)
 - number of **competitors** (10 .. 500)
 - false accept vs. false reject
- **Standard baseline**
 - large “universal background model” (**UBM**)
(e.g. 2000 mixture GMM on MFCCs)
 - likelihood ratio to speaker-specific model

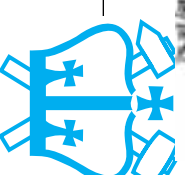
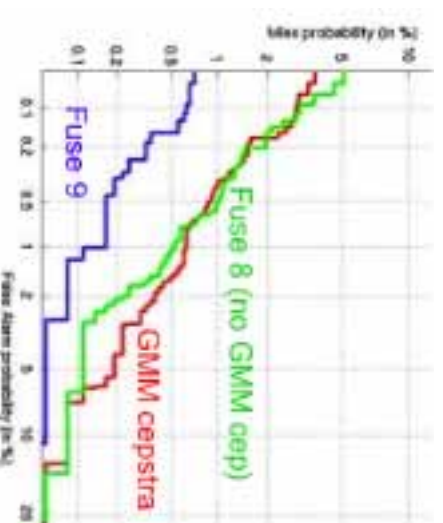


“Super Speaker ID”

- MFCC features don't capture ‘high level’ info
- 2002 JHU project to investigate new features
 - e.g. combined pitch/energy contour sequences:
 - also phone ftrs...
- Favorable **fusion** with standard baseline



<http://www.cisp.jhu.edu/ws02/groups/supersid/>



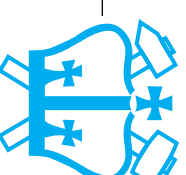
Outline

1 Spoken Document Retrieval

2 Audio databases

- Nonspeech audio retrieval
- Personal audio archives

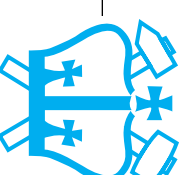
3 Open issues



2

Real-world audio

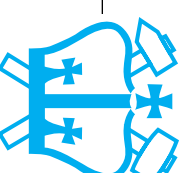
- **Speech is only part of the audio world**
 - word transcripts are not the whole story
- **Large audio datasets**
 - movie & TV soundtracks
 - events such as sports, news 'actualities'
 - situation-based audio 'awareness'
 - personal audio recording
- **Information from sound**
 - speaker identity,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



Multimedia Description: MPEG-7

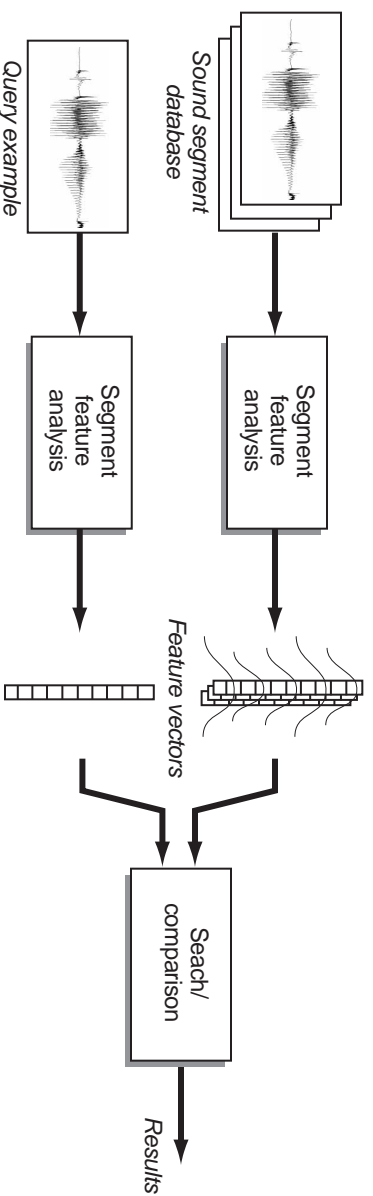
- MPEG has produced standards for audio / video **data compression** (MPEG-1/2/4)
- MPEG-7 is a standard for **metadata**: describing multimedia content
 - because search and retrieval are so important
- Defines descriptions of time-specific tags, ways to define categories, specific category instances
- + **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>

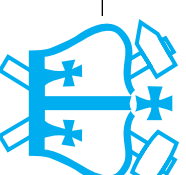


Muscle Fish “SoundFisher”

- 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”

SoundFisher by Muscle Fish – demoSounds

Database Navigate Fields Preferences Sound Help

Query

keyword	contains	animal
OR	contains	barn
sampleRate	equals	44100 Hz
creation date	is after	01/01/1996 m/d/y

More fields

Find sounds similar to

filename	keyword	comment	duration	pi
goose1.au	animal,goose		1.24	343
goose2.au	animal,goose	typical	2.57	275
goose3.au	animal,goose		1.12	278

More weights

Similarity weighted by

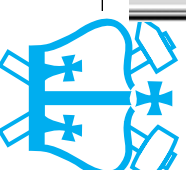
duration	0.0
loudness	0.0
<i>More weights</i>	

Add selected current sounds Remove

Search Find top 20 items

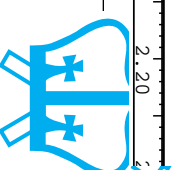
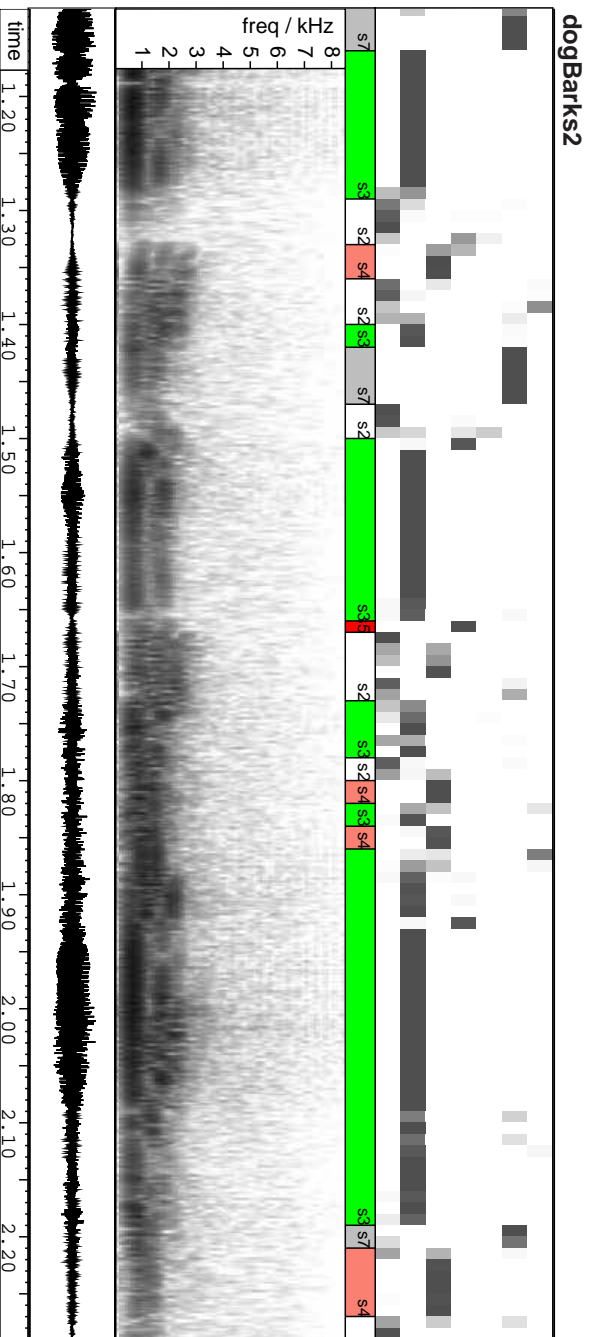
Current sounds

likelihood	filename	keyword	comment	duration	pitch	loudness	brightness
1.00	goose1.au	animal,goose		1.24	343.23	-22.45	120
1.00	goose2.au	animal,goose	typical	2.57	275.34	-17.65	186
1.00	goose3.au	animal,goose		1.12	278.01	-32.90	233
1.00	horse1.au	animal,horse		3.42	223.86	-11.47	43
1.00	horse2.au	animal,horse		5.01	287.61	-12.32	42
1.00	horse3.au	animal,horse		1.03	156.48	-45.50	32
1.00	horse4.au	animal,horse		1.69	190.48	-16.60	64



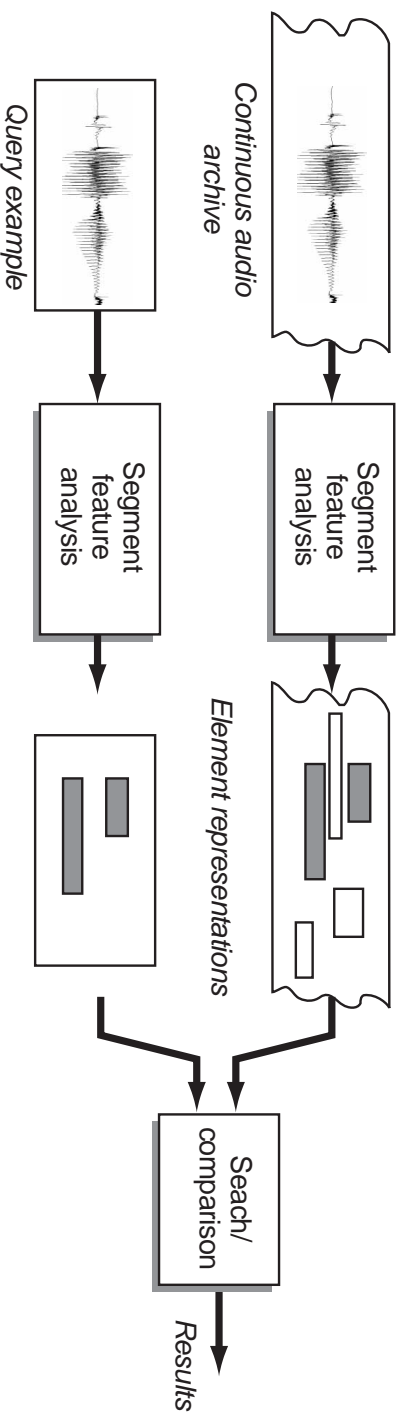
HMM modeling of nonspeech

- 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

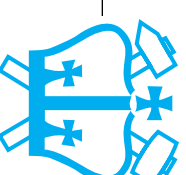


Indexing for soundtracks

- 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:**

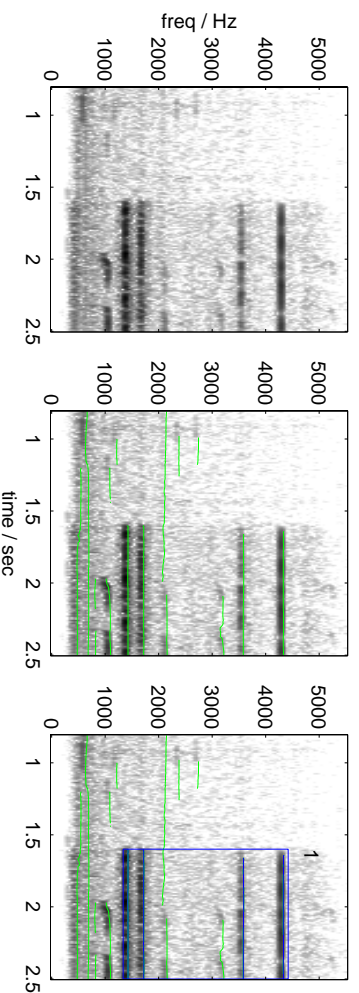


- analyze sound into source-related elements
- perform search & match in that domain

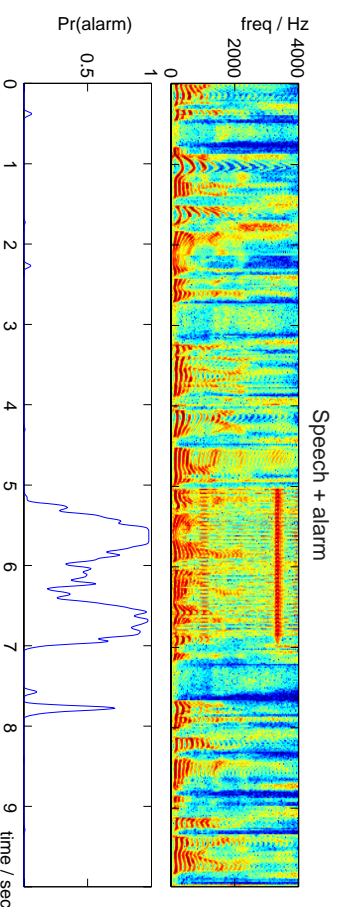


Alarm sound detection

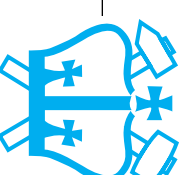
- **Alarm** sounds have particular **structure**
 - people 'know' them when they hear them'
- Isolate alarms in sound **mixtures**



- sinusoid peaks have invariant properties

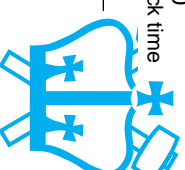
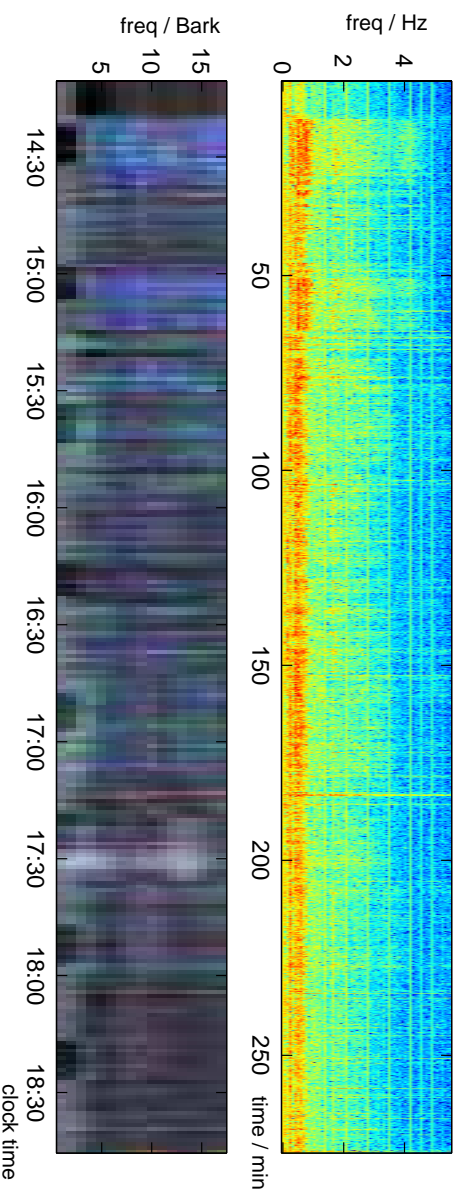


- cepstral coefficients are easy to model



Personal Audio

- LifeLog / MyLifeBits / Remembrance Agent:
Easy to record **everything** you hear
- Then what?
 - prohibitively time consuming to search
 - but .. applications if access easier
- Automatic content analysis / indexing...



Segmenting Personal Audio

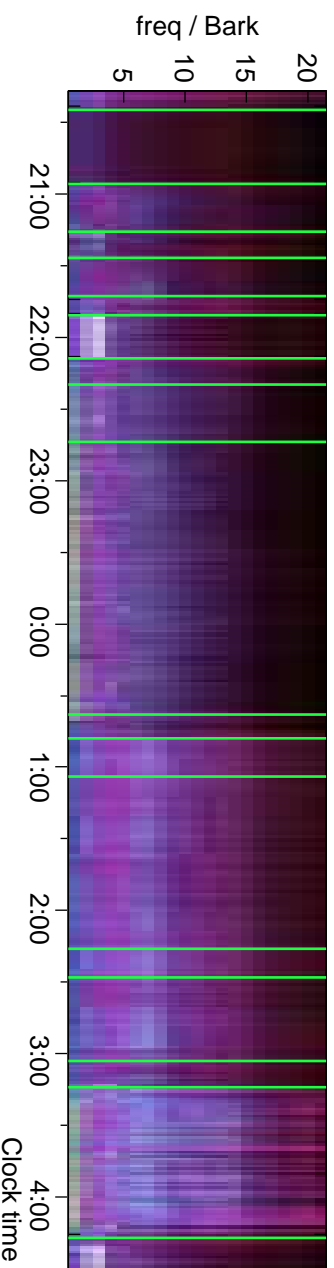
- First step: **segment** into consistent 'episodes'

- Variety of **features**:

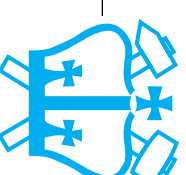
- regular spectrum
 - auditory spectrum
 - MFCCs
 - subband entropy
- $$A[n, j] = \sum_{k=0}^{N_{FT}/2+1} w_{jk} X[n, k]$$
- $$H[n, j] = - \sum_{k=0}^{N_{FT}/2+1} \frac{w_{jk} X[n, k]}{A[n, j]} \cdot \log \left(\frac{w_{jk} X[n, k]}{A[n, j]} \right)$$

- Mean/variance over 1 min segments + BIC

Hand-marked
boundaries

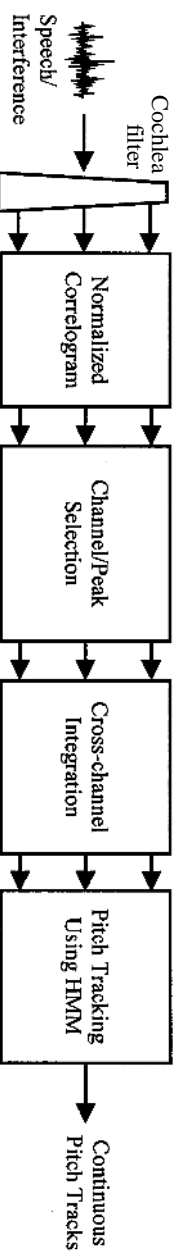


- **Best: 84% correct detect @ 2% false alarm**
 - mean audspec energy + entropy

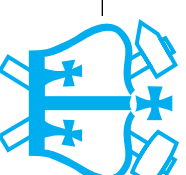
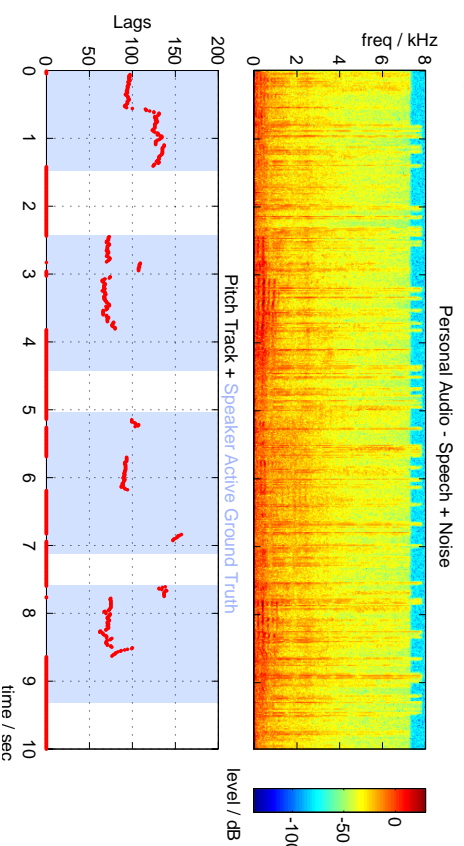


Detecting Speech Segments

- **Segments with speech are most interesting**
 - high noise defeats Voice Activity Detection
- **Voice Pitch as the strongest cue?**
 - periodicity + speech dynamics
 - need noise-robust pitch tracker



- **Improved detection in noise**



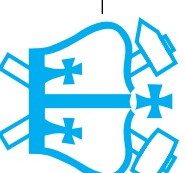
Outline

1 Spoken document retrieval

2 Audio databases

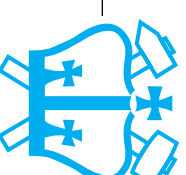
3 Open issues

- Speech recognition
- Sound source separation
- Information extraction & visualization
- Learning from audio



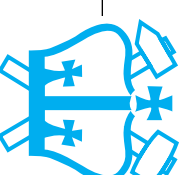
Open issues 1: Speech recognition

- **Speech recognition is good & improving**
 - but *reaching asymptote*: BN WER:
1997=22% 1999=14% 2004=9%
 - training data: 1999=150h 2004=3500h
- **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...



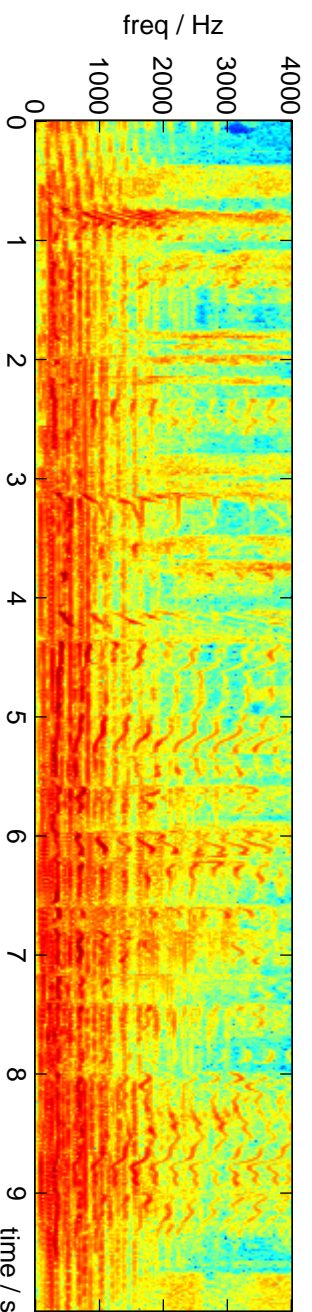
Open issues 2: Sound mixtures

- **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...

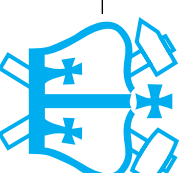


Open issues 3: Information & visualization

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

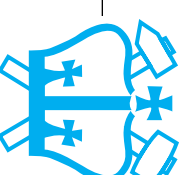


- frequency axis, intensity axis – time axis?
- separate spatial, pitch, source dimensions
- **Visualization** may not be possible
.. but helps us think about sound features
- **Different representations for different aspects**
 - best for speech, music, environmental, ...



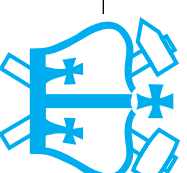
Open issues 4: Learning from audio

- **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 Spoken Document Retrieval
- 2 Audio Databases
- 3 Open issues



Course retrospective

Fundamentals

L1:	L2:	L3:	L4:
DSP	Acoustics	Pattern recognition	Auditory perception

Audio processing

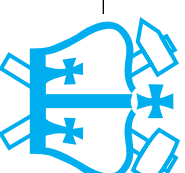
L5:	L6:
Signal models	Music analysis/synthesis

L7:	L8:
Audio compression	Spatial sound & rendering

Applications

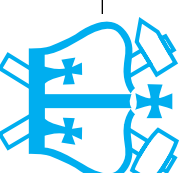
L9:	L10:
Speech recognition	Music retrieval

L11:	L12:
Signal separation	Multimedia indexing



Summary

- **Large Vocabulary speech recognition**
 - errors are OK for indexing
 - .. but still needs controlled audio quality
- **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



Final Presentations

- **Two Sessions**
Thursday April 26th, 10:00-12:30
Thursday May 3rd, 10:00-12:30
- **20 minute slots**
e.g. 15 minute talk, 5 min Qs / discussion
- **Background!**
Results!
Examples!
+ Discussion!
- **Special AV requirements?**
Let me know...

