# VOICE/AUDIO INFORMATION RETRIEVAL: MINIMIZING THE NEED FOR HUMAN EARS

*Mark Clements*[1] *and Marsal Gavaldà*[2]

[1]Georgia Institute of Technology, Center for Signal and Image Processing, and [2]Nexidia, Inc.

Atlanta, Georgia, USA

clements@ece.gatech.edu, mgavalda@nexidia.com

# Perspective of Presentation

- Target: Information retrieval from speech and audio for the marketplace.
- Primary users: not engineers or speech scientists.
- Focus: one particular set of tools we have developed through academic/commercial partnership.
- Research is often request driven (what people want)
- 4 short demos

# Information retrieval from video and audio. Why is it important?

- There is rapid proliferation of unstructured media content (cheap storage, acquisition and, transmission).
- Rapid human scanning is extremely difficult for such material.
- Management of this data is spiraling out of control.
- Text management is much more mature.

# Text-based information retrieval

- Base-level paradigm:
  - Search for key words, phrases, with Boolean and proximity constraints.
  - Scan visually for results, eliminating false hits manually.
  - Refine search interactively, choosing new search terms and constraints.
  - Concentrate on first page or two of “hits.”
  - Quality of the experience:
    - Did I find anything that is useful?
    - Did I find everything that is useful?

# Key Measures (to a naïve user)

- Recall: if one particular item I wanted to find was there, what is the likelihood I found it? (This is often unknown to the user and a low value may not negatively impact the experience as long as there are many correct responses.)
- Precision: What is the likelihood a returned item was useful? (very obvious when in error).
- Search speed: very important to the user.
- Ordering by relevance: first few hits dominate the experience.
- Training of user: experience can influence what they look for if the process is interactive (do not look for the letter "e").

# More Advanced Searches

- Categorization of records: e.g., language, subject matter, urgency, spam.
- Similarity measures: LSA, clustering.
- Cross-language
- Concept spotting
- Integration of meta-data

Key feature of all of these: reduce the use of the eyes and cognitive processing as much as possible.

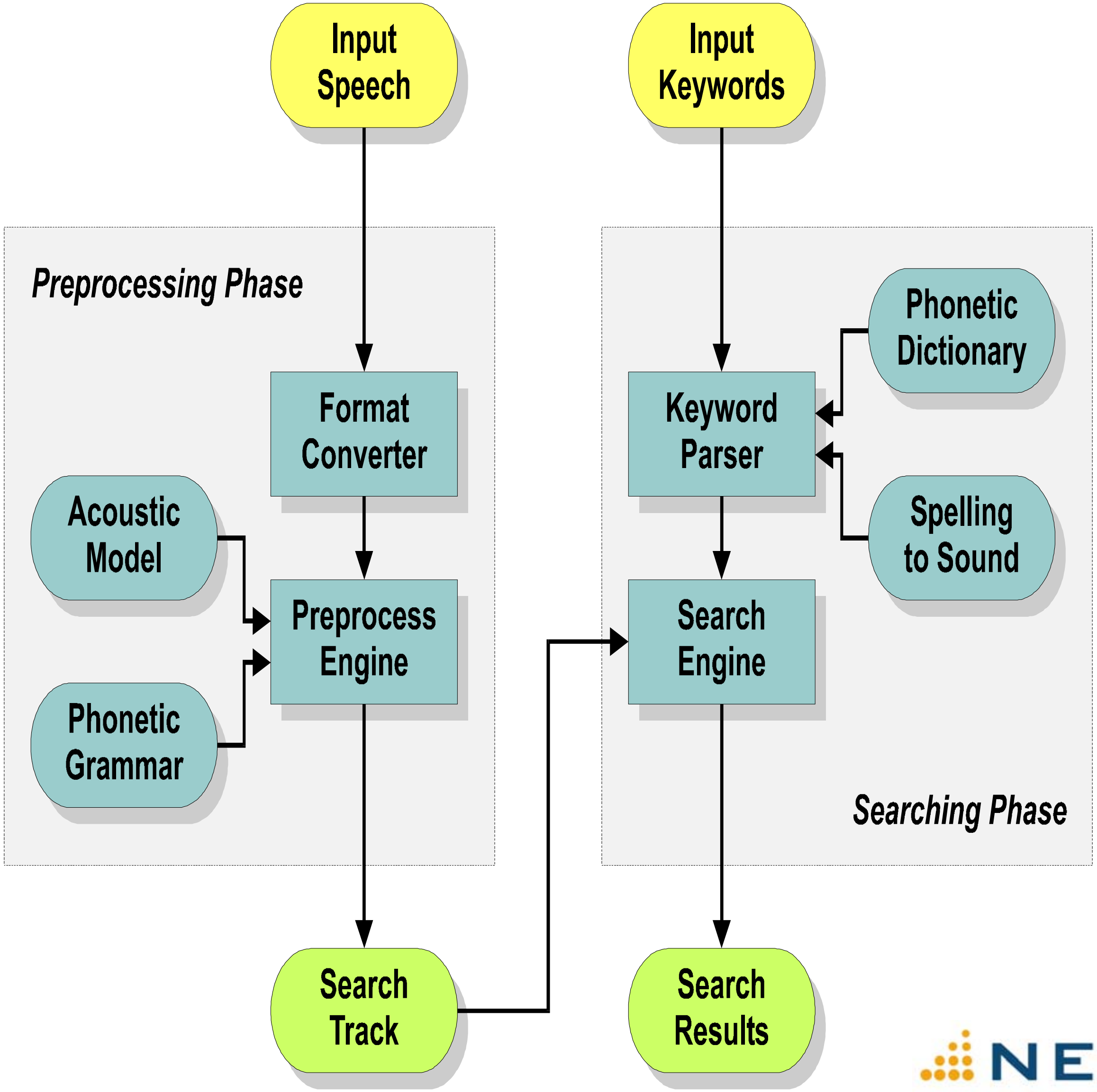
# Correlates to Speech/Audio

- Definition: information retrieval from speech/audio is anything that extracts information from a speech/audio record, thereby reducing the amount of listening that is required.
- Simple approaches:
  - Convert the speech/audio to text and utilize the tools of text-based retrieval.
  - Create a rich transcription of the data with annotations and expand the search criteria.
  - Perform word-spotting for base-level analysis, plus auxiliary analysis
- Issues:
  - Base-level analysis
  - Intelligence
  - Display methods: no matter how much processing is performed, the human observer is the ultimate consumer. The presentation is as important as the content.

# Base-level Processing: Phonetic Word-Spotting

- Not the only approach, but for many scenarios, it is the most viable solution.
- Distinct from LVCSR-based word-spotting.
- Also distinct from classical word-model-based keyword spotting.
- Key descriptors:
  - Accuracy
  - Speed (preprocess and search)
  - Index size
  - Complexity
  - Training requirements
  - Latency
  - Flexibility (depth of search, vocabulary size, maintainability)

# Basic Flow Diagram



CSIP

NEXIDIA™

# Phonetic Word-Spotting

- No hard decisions are made; no pruning.
- Two partitions: ingest and search.
- Input can be  words in a lexicon, a spelling-to-sound rendering, a phonetic string, or an acoustic source. Lexicon supports multiple pronunciations.
- Most of the computation is ingest.
- Processing speeds: (Nov '07)
  - Ingest: 160 times faster than real time (single core); 1100 times (dual 2.66 GHz quad-core)
  - Search (3 styles)
    - 150,000 times real time on single core;  1,000,000 on 8-core
    - 500,000 times real time with different operating point (single core)
    - 5,300,000 times real time with another operating point (single core).

# Base-Level Accuracy

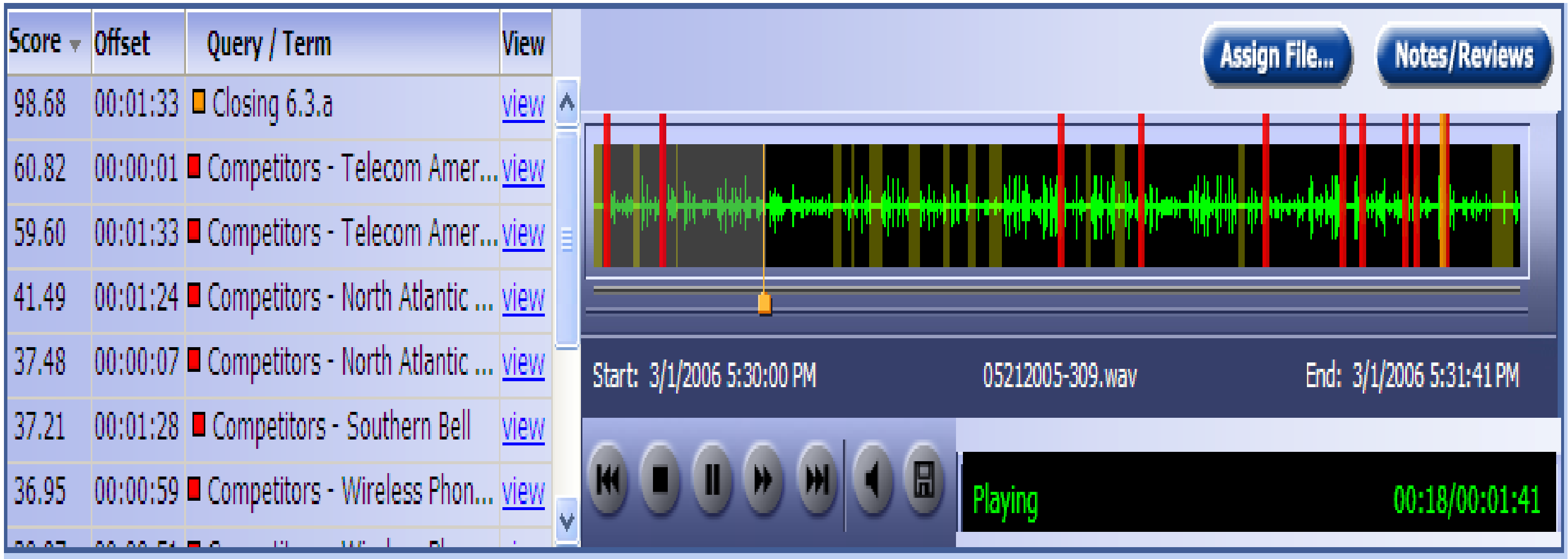
- **Arbitrary standard used**: Figure of merit (FOM): detection probability averaged over 0 to 10 false alarms per hour and averaged over 4-20 phoneme strings.
  - Broadcast news: 87% (language independent)
  - Switchboard: 77%
  - 2-bit 6KHz ADPCM: 69%
  - Production call-recorder: 56%
    - Difficult for professional transcriptionist
    - LVCSR WERs ~100%.
    - Still useful information.
  - 20 phoneme queries (multiple words, “social security reform”): >95% at one FA/hour on switchboard.

# Auxiliary Labeling

- Common front-end can be used for parallel detections:
  - Voice activity
  - Silence
  - Music
  - Speaker turns (esp., segmenting talkers in two-way voice signals)
  - Language and/or language family
  - Accent or dialect
  - Gender
  - DTMF and decoding
  - Number strings and decoding
  - Fax/modem signals
  - Other captured meta-data
  - Other requests
- All these “filters” add 10% to the preprocessing load.

# Display of Results

- In any system, the ultimate consumer of the information will be the user. Putting into a form he or she can manipulate for further exploration is important.



| Score | Offset | Query / Term | View |
|---|---|---|---|
| 98.68 | 00:01:33 | Closing 6.3.a | view |
| 60.82 | 00:00:01 | Competitors - Telecom Amer... | view |
| 59.60 | 00:01:33 | Competitors - Telecom Amer... | view |
| 41.49 | 00:01:24 | Competitors - North Atlantic ... | view |
| 37.48 | 00:00:07 | Competitors - North Atlantic ... | view |
| 37.21 | 00:01:28 | Competitors - Southern Bell | view |
| 36.95 | 00:00:59 | Competitors - Wireless Phon... | view |

Start: 3/1/2006 5:30:00 PM

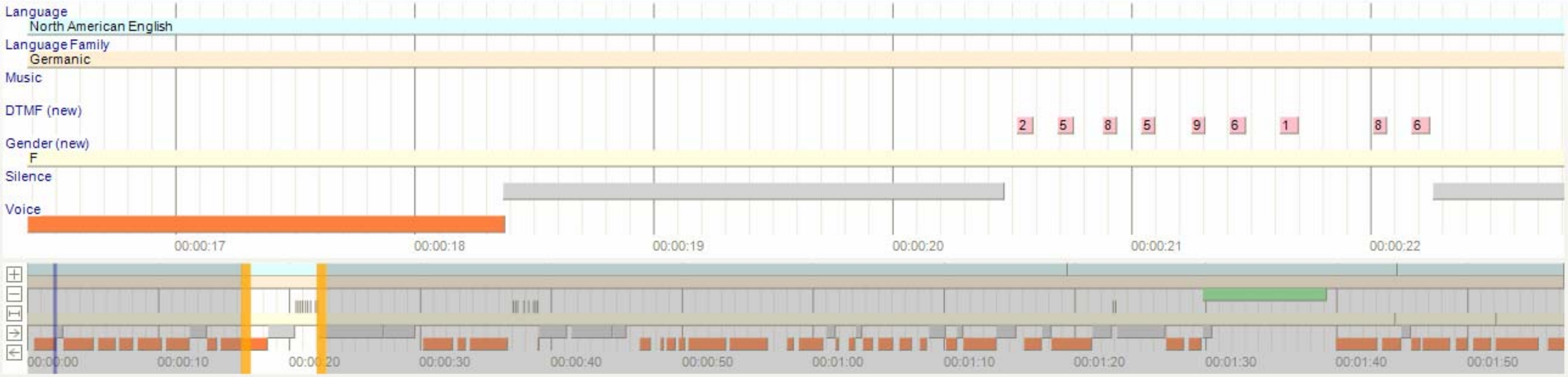
05212005-309.wav

End: 3/1/2006 5:31:41 PM

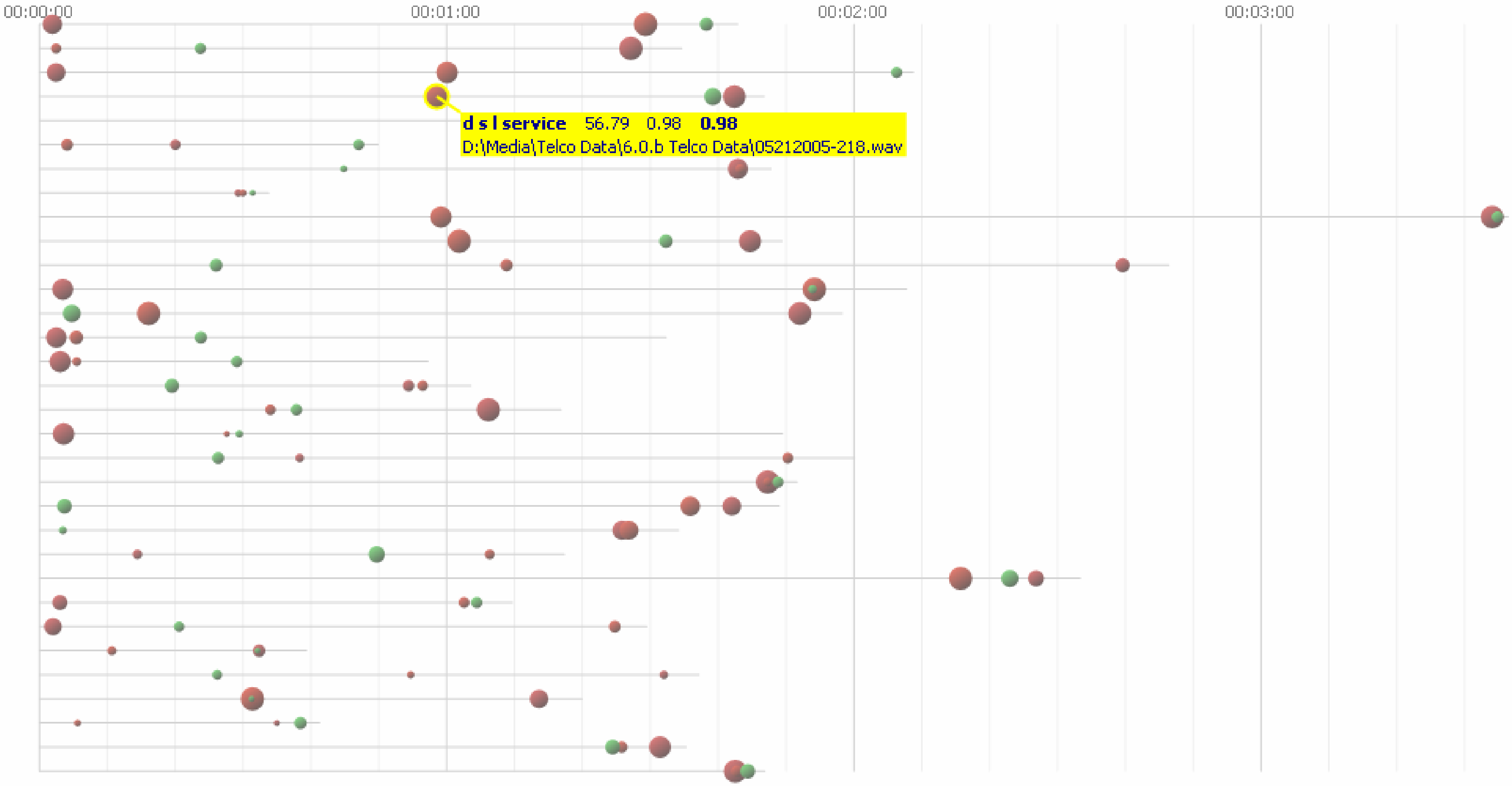
Playing

00:18/00:01:41

# Auxiliary Information Display (demo)



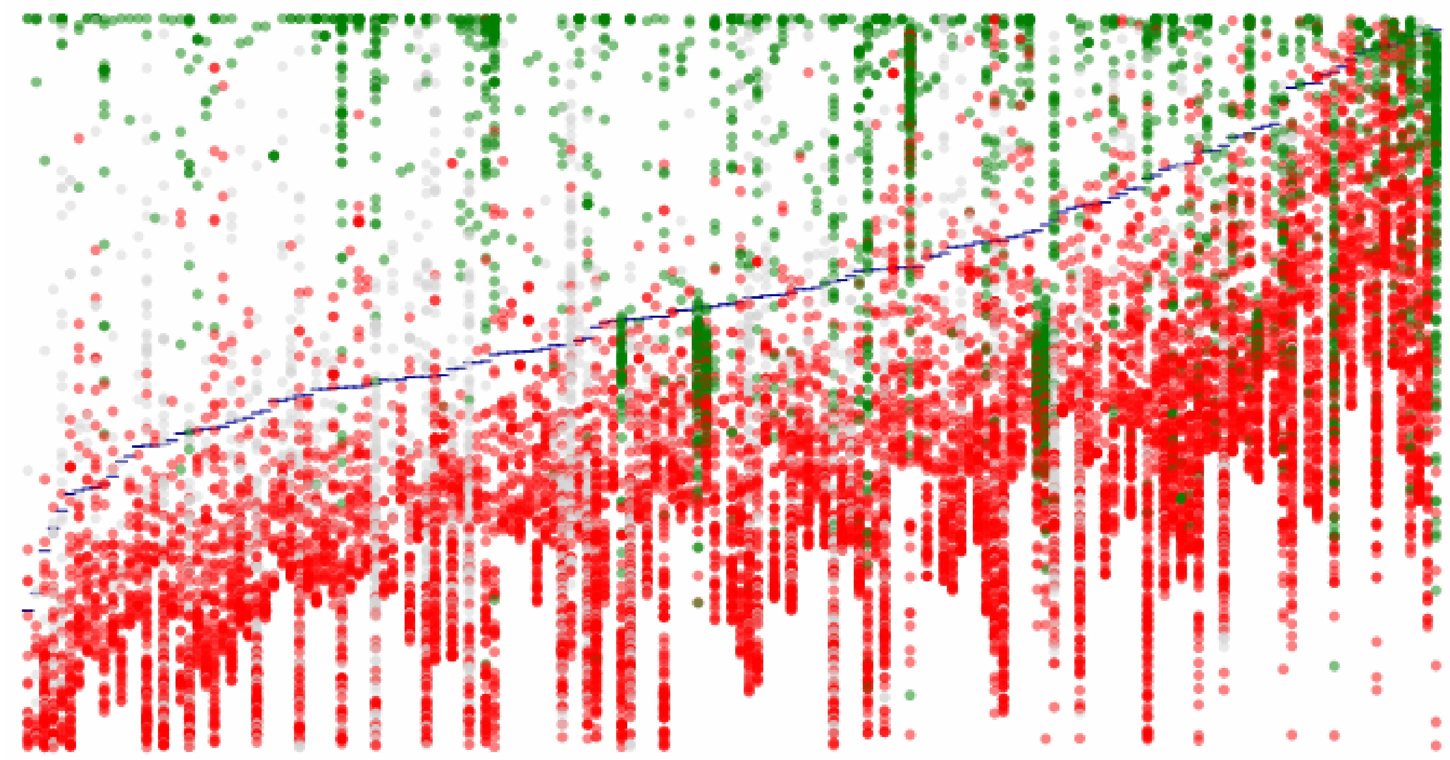
Example of a media file richly annotated by detector and classifier tracks such as language, language family, music, DTMF, gender, silence, and voice activity. The interactive control consists of two panels: the lower one offers a "bird's eye view" of the entire media file (about 2 minutes long in the example), whereas the top panel shows only the zoomed-in portion, corresponding to the segment delimited by the vertical orange bars in the lower panel (about 7 seconds long in the example).



Another view into a search result set. In this case, each horizontal line corresponds to a media file (thus the different lengths) and the hits are depicted by a sphere whose radius is proportional to the confidence score and whose color is determined by the query (only two in this case).

# Confidence Scoring

- Setting a confidence score with any detection strategy is often difficult.
- Score normalization and consistent thresholding requires additional training step training.
- Scoring/thresholding based on query length, content, and quality of signal.
- Demo shows some results.

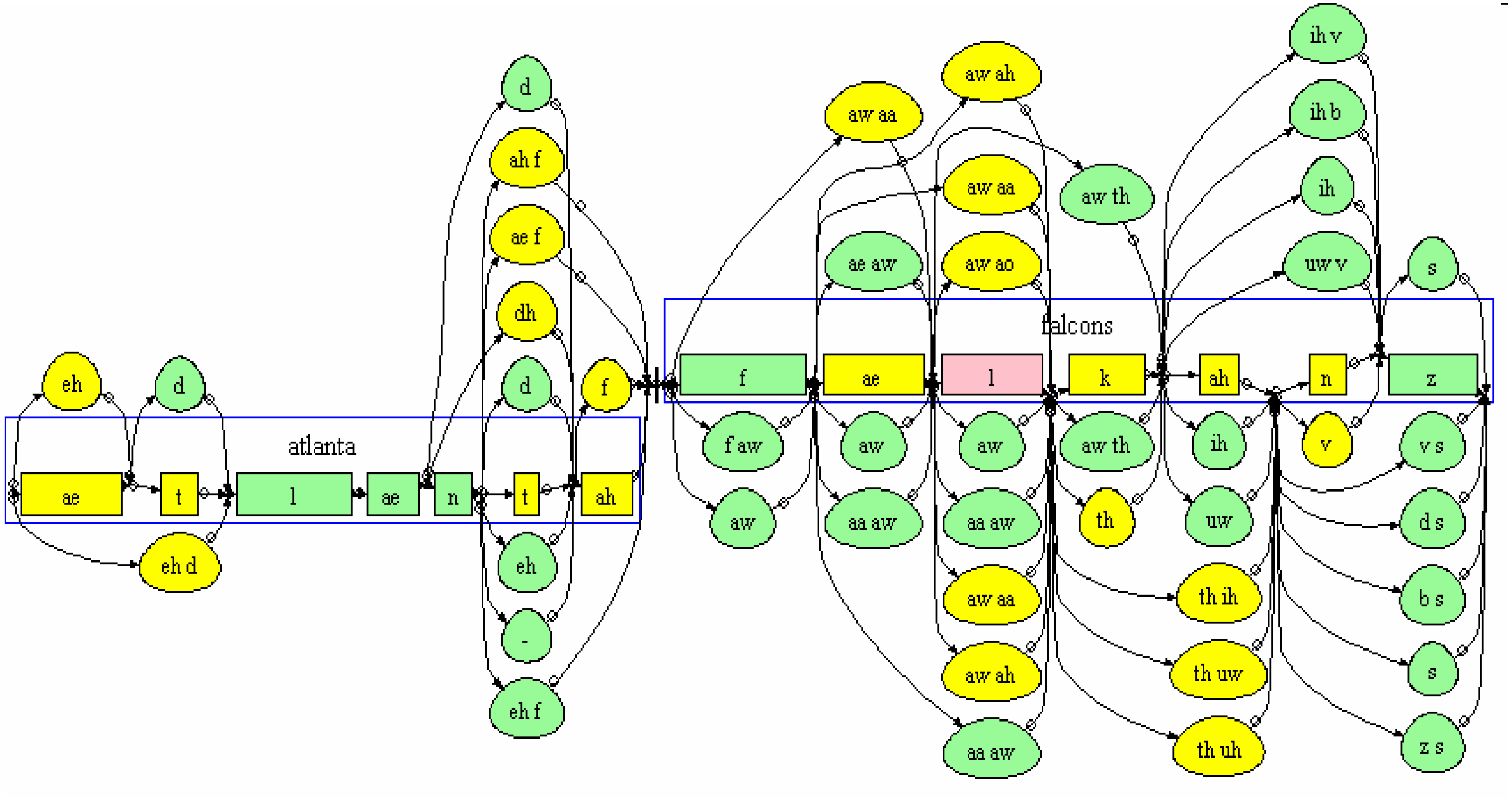


A more advanced visualization of search result sets which includes "dispositioning" information, i.e., truth marks. In this case, each dot corresponds to a hit, color-coded according to its dispositoned value: red for a false alarm, green for a true positive, and gray for a phonetic partial match (as in matching "sixty" when searching for "sixteen"). Each column (X-axis) corresponds to a search term, and the Y-axis corresponds to the confidence score. In this view, the search terms are sorted by their automatic threshold (horizontal blue line), an attempt to automatically separate true hits from false alarms.

# Pronunciation Optimization

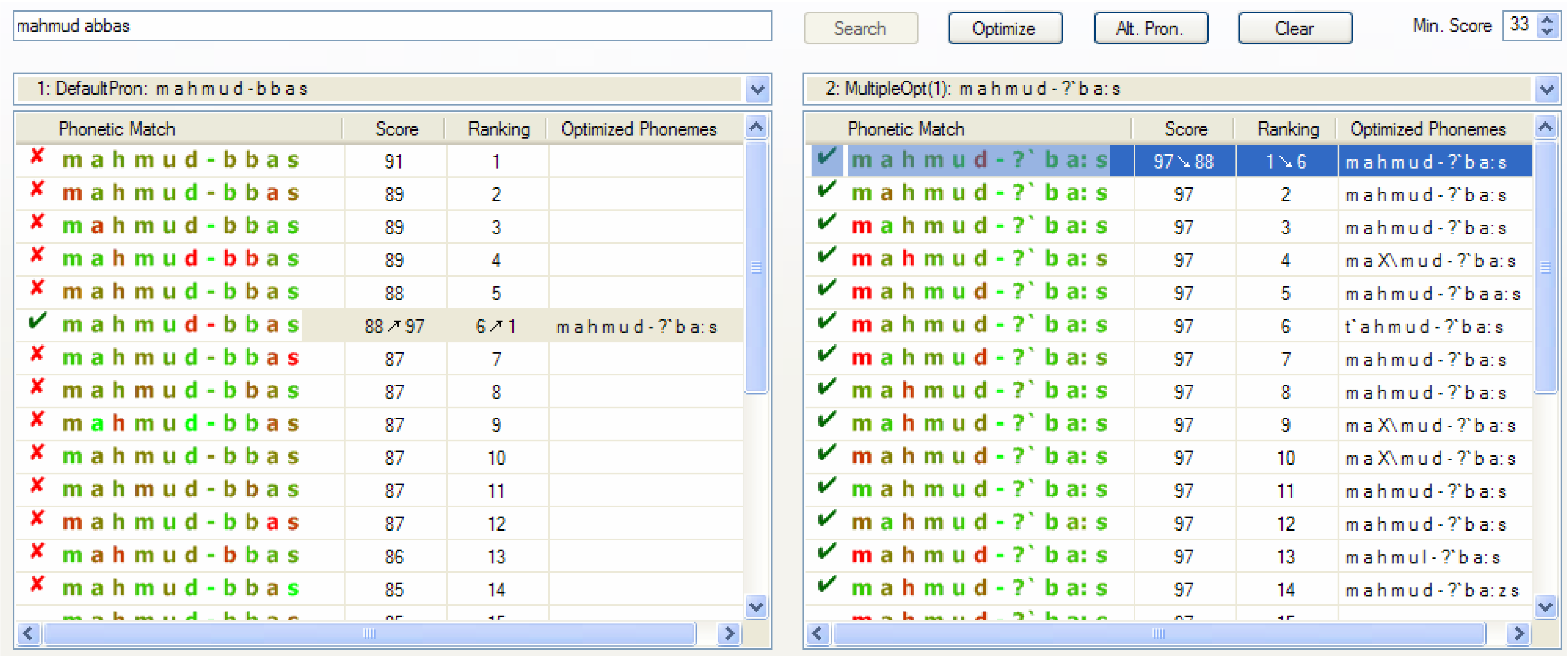
- Phonetic analysis/search requires accurate renderings of the queries.
- Often users have trouble knowing what they really want, especially if it is not one's native language.
- Finding exemplars of what you want helps.
- Networks can then be constructed for finding better matches.
- Interactive demo: uses good and bad exemplars for discriminative training.

# Example for “Atlanta Falcons”



Example of a phoneme lattice such as the ones that underlie the computation of alternative sequence of phonemes used for pronunciation optimization and repeated search.

# Pronunciation Optimization



1: DefaultPron: mahmud-bbas

| | Phonetic Match | Score | Ranking | Optimized Phonemes |
|---|---|---|---|---|
| ✗ | mahmud-bbas | 91 | 1 | |
| ✗ | mahmud-bbas | 89 | 2 | |
| ✗ | mahmud-bbas | 89 | 3 | |
| ✗ | mahmud-bbas | 89 | 4 | |
| ✗ | mahmud-bbas | 88 | 5 | |
| ✓ | mahmud-bbas | 88 ↗ 97 | 6 ↗ 1 | mahmud-?`ba:s |
| ✗ | mahmud-bbas | 87 | 7 | |
| ✗ | mahmud-bbas | 87 | 8 | |
| ✗ | mahmud-bbas | 87 | 9 | |
| ✗ | mahmud-bbas | 87 | 10 | |
| ✗ | mahmud-bbas | 87 | 11 | |
| ✗ | mahmud-bbas | 87 | 12 | |
| ✗ | mahmud-bbas | 86 | 13 | |
| ✗ | mahmud-bbas | 85 | 14 | |

2: MultipleOpt(1): mahmud-?`ba:s

| | Phonetic Match | Score | Ranking | Optimized Phonemes |
|---|---|---|---|---|
| ✓ | mahmud-?`ba:s | 97 ↘ 88 | 1 ↘ 6 | mahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 2 | mahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 3 | mahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 4 | maX\mud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 5 | mahmud-?`baa:s |
| ✓ | mahmud-?`ba:s | 97 | 6 | t`ahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 7 | mahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 8 | mahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 9 | maX\mud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 10 | maX\mud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 11 | mahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 12 | mahmud-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 13 | mahmul-?`ba:s |
| ✓ | mahmud-?`ba:s | 97 | 14 | mahmud-?`ba:zs |

Example of interactive dispositioning that leads to a more accurate sequence of phonemes to represents the query at hand. In this case, after marking the 6th hit as correct, the engine proposes a variation in the phoneme sequence that allows the user to find many more occurrences of the target search term.

# “Find it Again”

- Similar to pronunciation optimization based one a single utterance.
- The user finds a phrase of interest and wishes to find it again.
- An approximate phonetic rendering must have been performed to constrain the problem. Otherwise it is just an audio query (later).
- Tends to find the same or similar speakers of the phrase.

# Clip Spotting

- Monitoring of archives (or real-time) for instances of acoustic events.
- Examples: music clips, sound bites, sound effects, automated messages.
- Applications: royalties, compliance, plagiarism, filtering, etc.
- PAT files do not specifically encode such events, but the rendering is sufficiently consistent as to allow rapid retrieval of events if more than 2 seconds is clipped.
- Search speed is the same as for speech, and no additional preprocessing needs to be done.

# Voice User Interface

- Simple approach: phonetically decode input query.
  - 1-best string
  - N-best
  - Consensus networks
- Perform word recognition on input query (not used due to usual problems with OOV inputs)
- Performance tested by clipping segments from switchboard and performing same FOM tests as before → 40% reduction in FOM on all three methods.
- Single speaker dependent data, same speaker queries, live feedback → less than 10% FOM reduction.

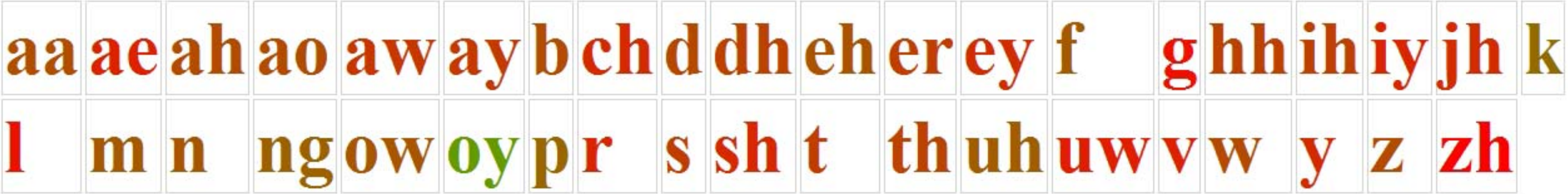
# Query Building

- *Query languages,* designed for information retrieval from databases, are widely available.
- Many flavors exist based on the contents of unstructured data.
- Word-spotting adds an additional level of noise, given imperfect decisions. Confidence levels must be combined with Boolean logic, equivalence classes, and proximities.
  - Interactive tools have been developed that evaluate the effectiveness of a query based on word content and reliability of the query terms (e.g., two phoneme terms are not good). Non speech events can be folded in (e.g., DTMF).
  - LSI with constraints and confidence scoring can also be used for classification of records for later retrieval. Also use one voice document to retrieve another.

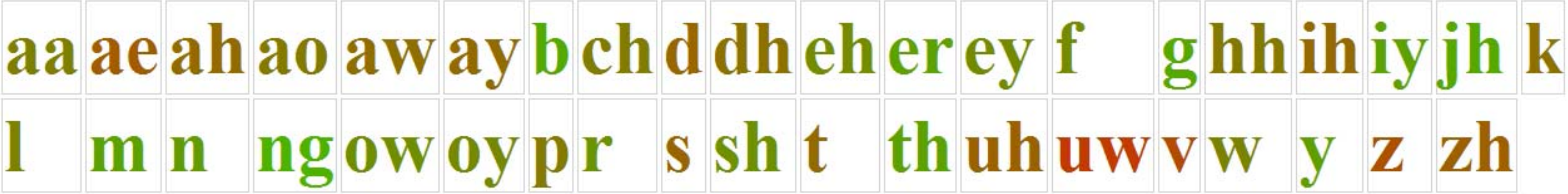
# Language Proficiency Assessment

- It is *information retrieval* in that it is taking human listeners out of the loop to find proficient talkers of a language.
- Candidates call into an automated, interactive client, and generate speech from a script.
- Speech is synchronized and assessed for disfluencies and pronunciation.
- Rank orderings form this procedure and human listening assessments are very similar.

# Demonstration of Assessment Tool



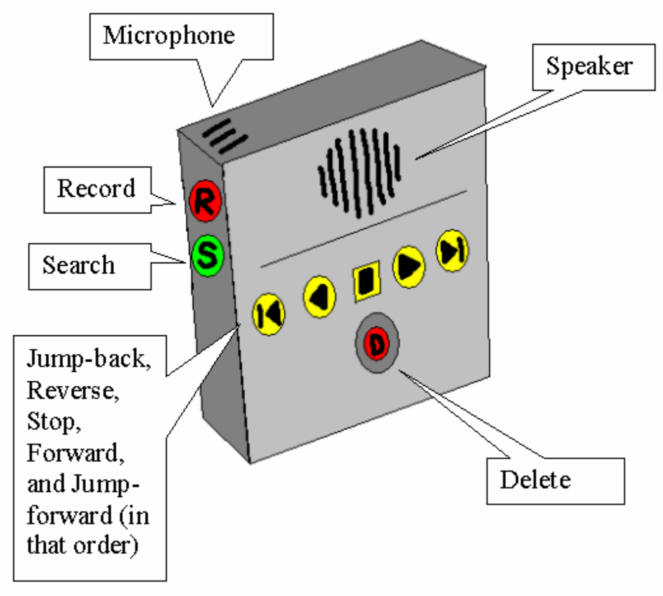
| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aa | ae | ah | ao | aw | ay | b | ch | d | dh | eh | er | ey | f | g | hh | ih | iy | jh | k |
| l | m | n | ng | ow | oy | p | r | s | sh | t | th | uh | uw | v | w | y | z | zh | |

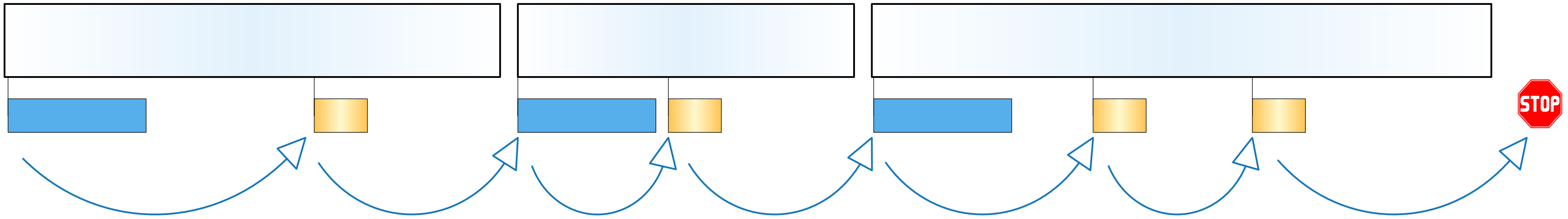


| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aa | ae | ah | ao | aw | ay | b | ch | d | dh | eh | er | ey | f | g | hh | ih | iy | jh | k |
| l | m | n | ng | ow | oy | p | r | s | sh | t | th | uh | uw | v | w | y | z | zh | |

Sample output from the phonetic scoring portion of a language assessment application. The tables show the average score for each phoneme across all the words in the script, color-coded in a continuous gradient from pure red to pure green. The top table corresponds to a low-scoring speaker, the bottom table to a high-scoring speaker.

# Personal Information Retrieval


Microphone
Speaker
Record
R
Search
S
Jump-back, Reverse, Stop, Forward, and Jump-forward (in that order)
D
Delete



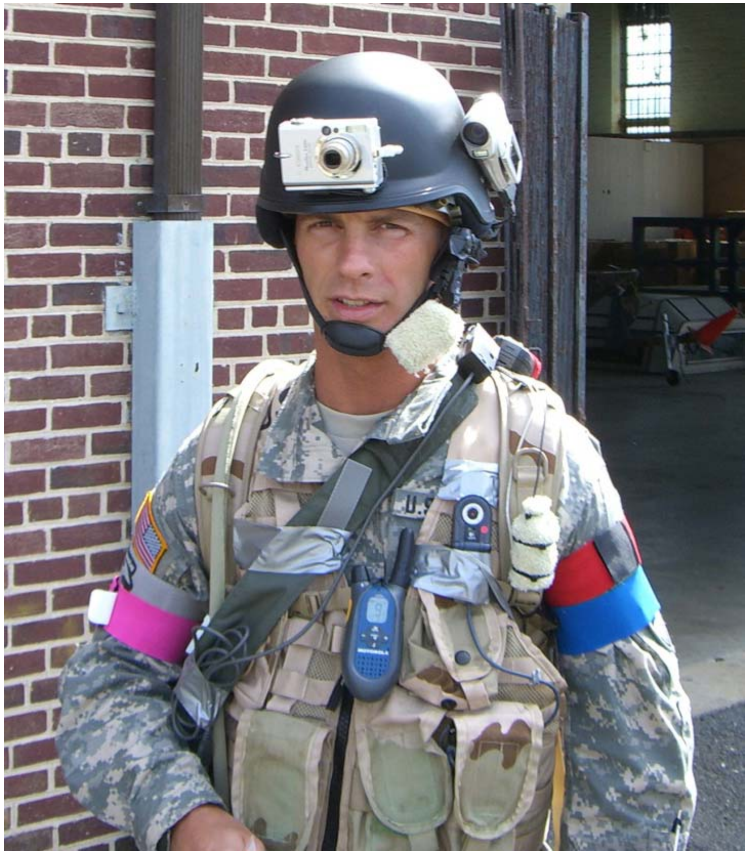
Sketch of the device for note retrieval for the blind (top), and example of the navigational steps to find the occurrences of “phone” (bottom).

# Portable Device for Voice Notes

- Developed for the Visually Impaired (with cooperation of the US Veteran’s Administration).
- Current devices merely record messages, and the user must remember which one to retrieve.
- New device allows searching of contents for words and phrases, with queries input by voice.
- Archives are small (only a few hours, generally) so retrieval is very accurate.
- Speaker adaptive models can be applied to speaker dependent input.
- Case studies show significantly improved information retrieval times over existing tools (recipes, addresses, meetings, etc.)

# Access to Reports from Patrols

- Chest mic
- Close-talking mic
- High-resolution camera
- MPEG4 camcorder
- GPS/Compass/Altimeter
- 6 Bluetooth accelerometers



# Example of Multimedia Data Display

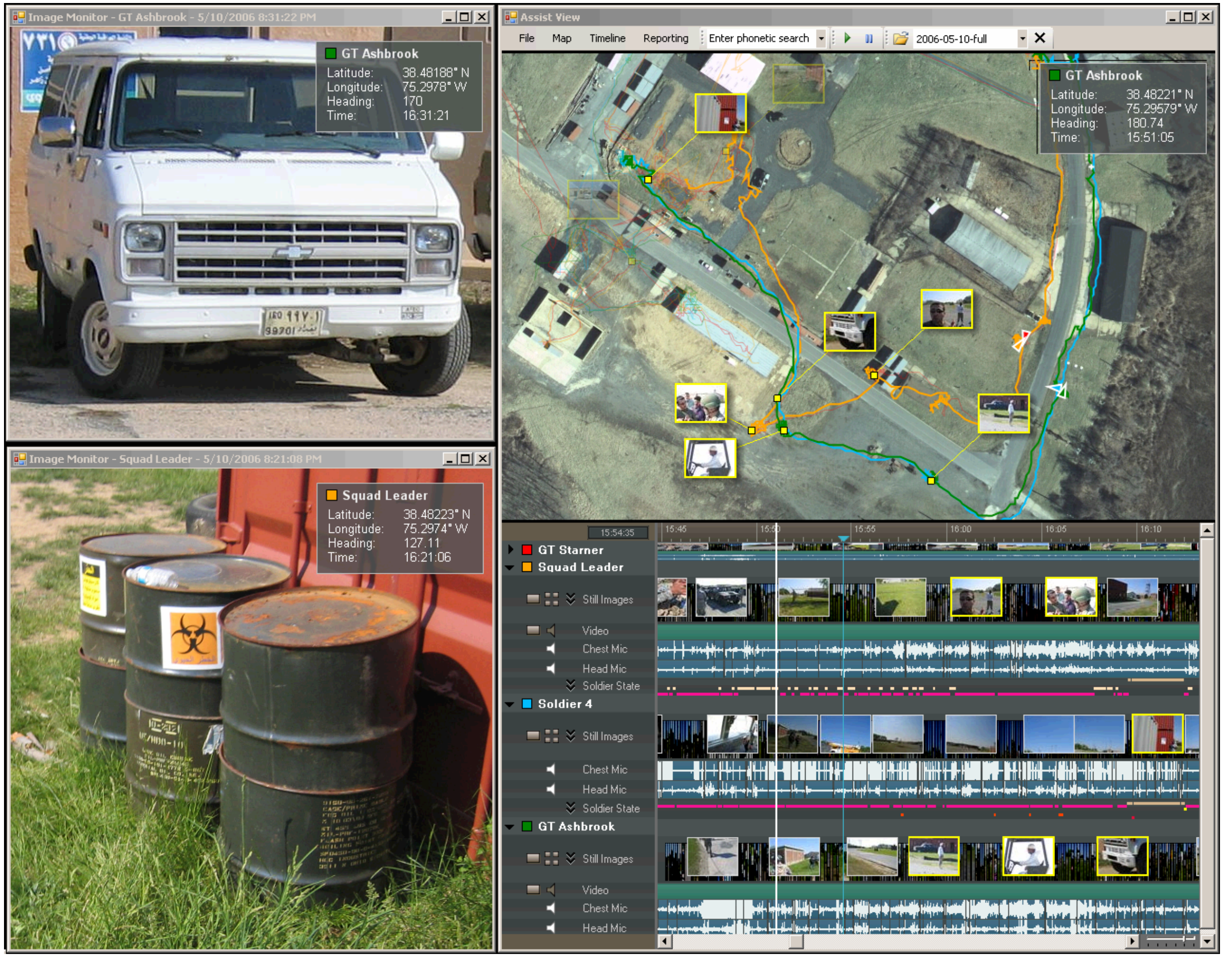


Image Monitor - GT Ashbrook - 5/10/2006 8:31:22 PM

GT Ashbrook
Latitude: 38.48188° N
Longitude: 75.2978° W
Heading: 170
Time: 16:31:21

Image Monitor - Squad Leader - 5/10/2006 8:21:08 PM

Squad Leader
Latitude: 38.48223° N
Longitude: 75.2974° W
Heading: 127.11
Time: 16:21:06

Assist View

GT Ashbrook
Latitude: 38.48221° N
Longitude: 75.29579° W
Heading: 180.74
Time: 15:51:05

# Accessing Patrol Data by Audio

- One access method is through the audio tracks.
- Patrol agent generates voice annotations describing key events, names, places, and even unfamiliar words.
- Retrieval based on phonetic queries found to be useful.

# Major Challenges

- Cross language search:
  - Many examples exist of mixed audio: multiple languages being spoken in a conversation; foreign terms inserted which use phones not in that language; corruption of a pronunciation by a non-native speaker.
  - Ideal solution: phonetically render all speech using expanded IPA-type phoneme set. Currently unsolved due to wide acoustic differences within IPA equivalence classes.

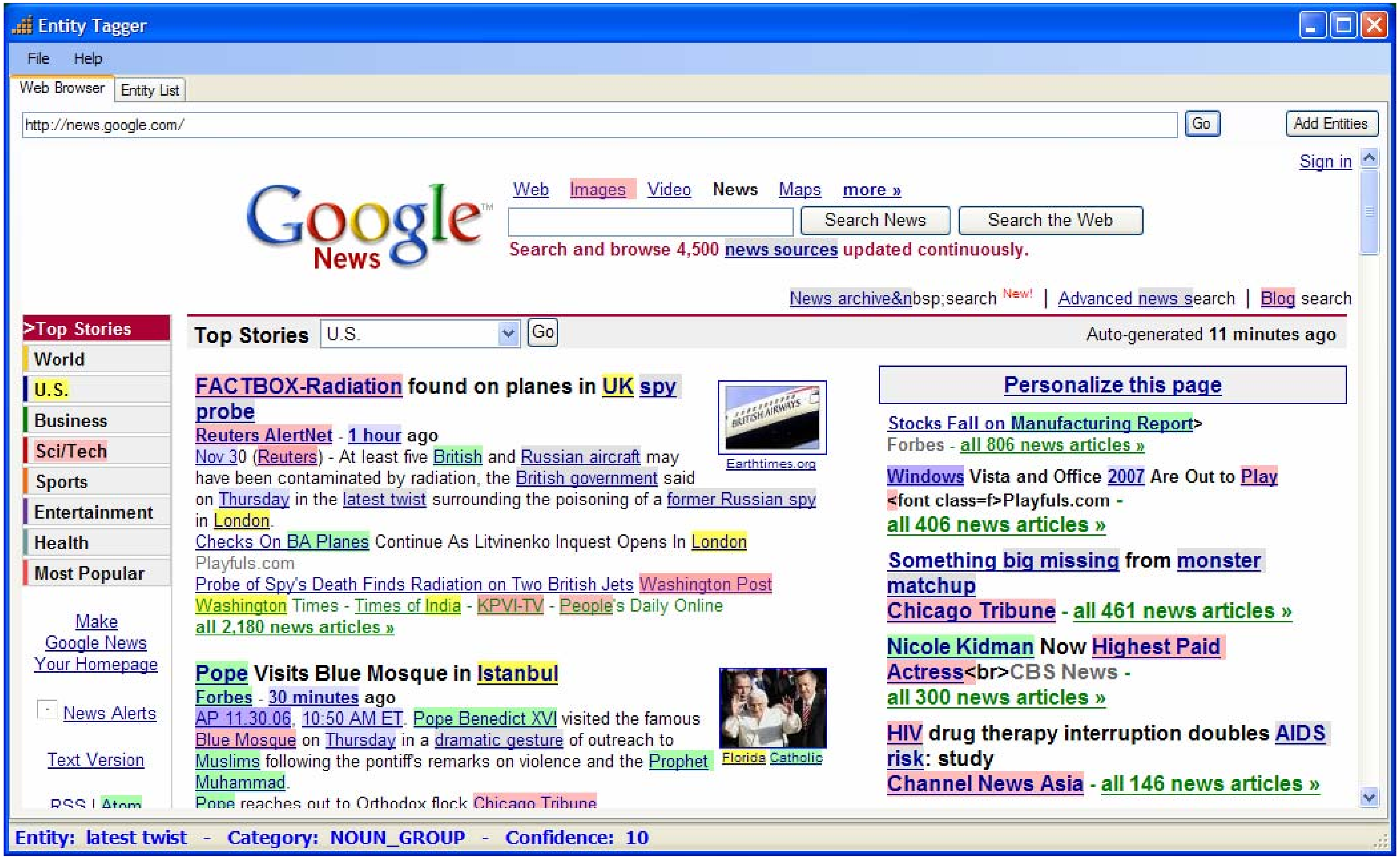
# Cross-Acoustical Search

- Often, audio data is mixed in acoustic conditions.
  - Conference calls with mix of landline, mobile, and speaker phones.
  - Broadcast news with studio anchors, field reporters, call-ins.
- Confidence measures are based on query content and length as well as acoustical conditions.
  - Suboptimal solution: detect condition changes and switch models.
  - More rigorous approach needed.

# Search Speed

- Brute force searching will never be fast enough for large data-sets (the same with text – e.g., “grep” the Library of Congress)
- Must use pre-searching, reverse indexing, caching, crawling for large search sets.
- Current implementations involve extracting pre-searched terms from earlier searches, associated text, standard sets, and new arrivals.

# Entity Tagging for Pre-searching News Stories



Search and browse 4,500 news sources updated continuously.

News archive search New! | Advanced news search | Blog search

>Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health
Most Popular

Make Google News Your Homepage

□ News Alerts

Text Version

RSS | Atom

**Top Stories** U.S. Go Auto-generated **11 minutes ago**

**FACTBOX-Radiation found on planes in UK spy probe**
Reuters AlertNet - **1 hour** ago
Nov 30 (Reuters) - At least five British and Russian aircraft may have been contaminated by radiation, the British government said on Thursday in the latest twist surrounding the poisoning of a former Russian spy in London.
Checks On BA Planes Continue As Litvinenko Inquest Opens In London
Playfuls.com
Probe of Spy's Death Finds Radiation on Two British Jets Washington Post
Washington Times - Times of India - KPVI-TV - People's Daily Online
all 2,180 news articles »

Earthtimes.org

**Pope Visits Blue Mosque in Istanbul**
Forbes - 30 minutes ago
AP 11.30.06, 10:50 AM ET. Pope Benedict XVI visited the famous Blue Mosque on Thursday in a dramatic gesture of outreach to Muslims following the pontiff's remarks on violence and the Prophet Muhammad.
Pope reaches out to Orthodox flock Chicago Tribune

Florida Catholic

**Personalize this page**

Stocks Fall on Manufacturing Report>
Forbes - all 806 news articles »

**Windows Vista and Office 2007 Are Out to Play**
<font class=f>Playfuls.com -
all 406 news articles »

**Something big missing from monster matchup**
Chicago Tribune - all 461 news articles »

**Nicole Kidman Now Highest Paid Actress**<br>CBS News -
all 300 news articles »

**HIV drug therapy interruption doubles AIDS risk: study**
Channel News Asia - all 146 news articles »

**Entity: latest twist - Category: NOUN_GROUP - Confidence: 10**

CSIP

NEXIDIA™

# Direct Search Strategies

- Use meta-data as much as possible (do not search for “conservation of momentum” in courses on poetry.)
- 150,000 times faster than real-time allows searching of only 200 hours with 5 second response time (1440 hours 8-core).
- One algorithm produces a speed increase to (single core) 500,000 x r.t. with 1-2% decrease in FOM. Still not suitable for thousands of hours.
- Higher speed, however, produces false hits very rapidly.

# Direct Search, Low-False-Alarm-Rate Operating Point

- To directly search large data-sets, FA rate must be low and speed must be high.
- New procedure operates at 5.3 M x real time (2.7 second response time for 4000 hours of data, 6 seconds for 8760 hours = 1 year)
- False alarm rates set from .001 to .01 per hour.
- Accuracy (Nov 07) at .01: 12 phonemes - 35%, 20 phonemes (“conservation of momentum”) – 85% on telephony.
- Question: how many hits does one need?

# Integration of Other Methods

- Key is to use the best of both; direct integration gets the worst – slow and with FA’s.
- Multistage search – refine putative hits with additional processing.
- Rescoring based on multiple estimates.

# Conclusions

- Though development is early compared to text-based retrieval, the marketplace is responding to what can be done.
- Different conditions (acoustics, language, constraints) dictate different solutions.
- Audio itself contains additional information that can also be used effectively.