Combining Low Resource and High Resource Acoustic Modeling in Spoken Term Detection

Richard Rose, Atta Norouzian, and Sina Hamidi

July 17, 2012

McGill University Dept. of ECE

Outline

- Motivation

- Introduction: ASR Based Spoken Utterance Retrieval (SUR) and Spoken Term Detection (STD)
  - Reminder: Important issues in search
  - Another Reminder: How lattice based STD systems work

- Combining High Resource and Low Resource Models – for Low Resource Verification of Spoken Terms
  - Graph Based Re-ranking of Retrieved Intervals [Chen et al, 2011] [Hsu et al, 2007] …

  … using measures of interval self-similarity [Jansen et al, 2010] [Parks and Glass, 2005]

Motivation

- ASR based ***high resource*** spoken term detection …
  - Accurate, fast, and robust within a given domain

    … but needs a range of resources for system configuration.

    … and not easy to adapt to new domains
- Alternative ***low resource*** approaches …
  - Can provide orthogonal knowledge sources

    … features, models, learning formalisms, …
  - Can benefit from high resource ASR

    … “richer” audio segments, pseudo labeling, …
- Costs and benefits associated with ***levels of supervision***:
  - **Language Expert** – Lexicons and sub-word inventories
  - **Domain Expert** – Associate Acoustics with actions / concepts
  - **Corpus Expert -** Parallel Speech and Text

Reminder: Important Issues for Search

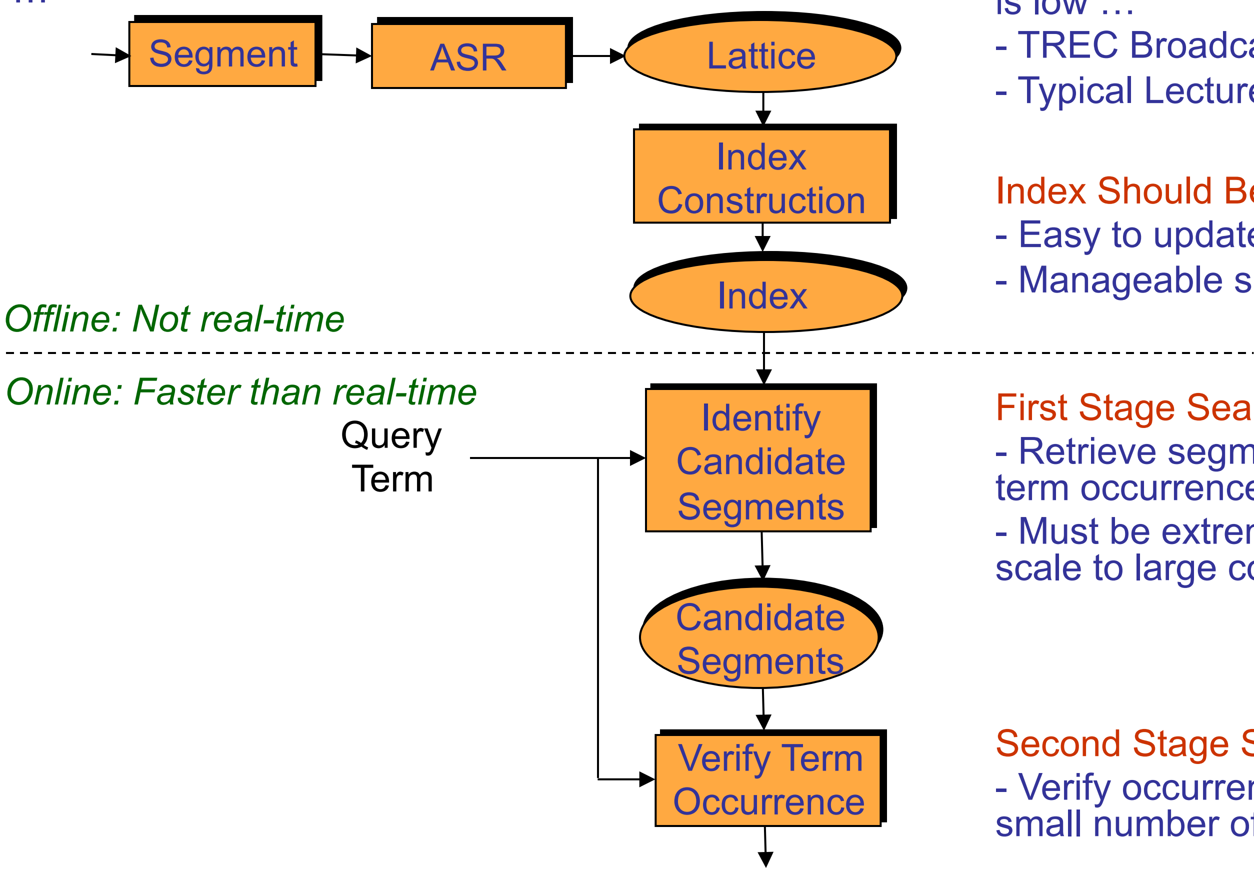
- Large Collections
  - Search performed at ***thousands of times faster than real time***
- Open Vocabulary Queries
  - Most informative search terms are often unseen
- Accuracy
  - Performance requirements can be difficult to define
- Difficult Task Domains
  - Informal ***spontaneous speech*** from ***specialized topics***
- Lack of Linguistic Resources
  - Limited or no available ***domain specific*** text or speech data

Another Reminder: Lattice-Based STD

Audio: lecture, phone call, meeting, news broadcast, …

Audio Segments: <1 minute

Lattice produced for each audio segment



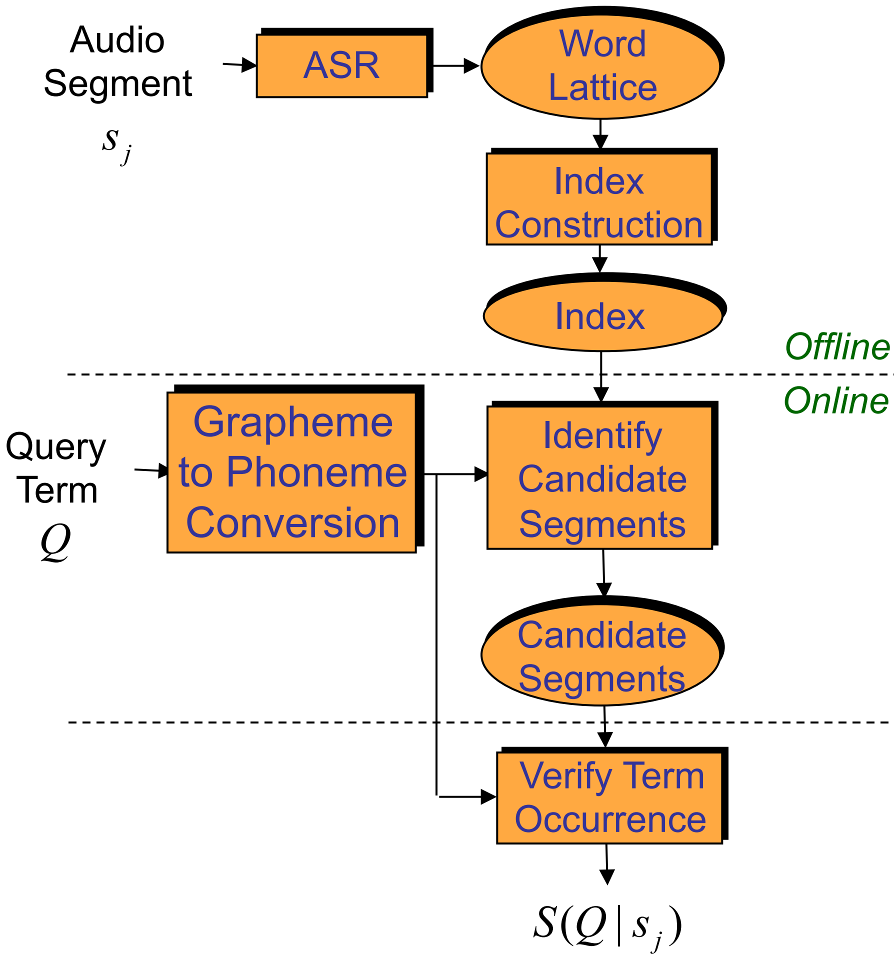
Lattice vs. Single Best ASR String:
Important when word accuracy (WAC) is low …
- TREC Broadcast News WAC = 90%
- Typical Lecture Speech WAC = 58%

Index Should Be:
- Easy to update for new segments
- Manageable size

First Stage Search:
- Retrieve segments that are “rich” in term occurrences
- Must be extremely fast if we are to scale to large collections

Second Stage Search:
- Verify occurrence of query term for small number of candidate segments

Example of a Lattice-Based STD



ASR

- WAC = 56.5%, OOV Rate=11.2%

Index Size

- Subword based index: 42,000 index entries per hour of audio

Retrieval of Segments for I.V. Terms

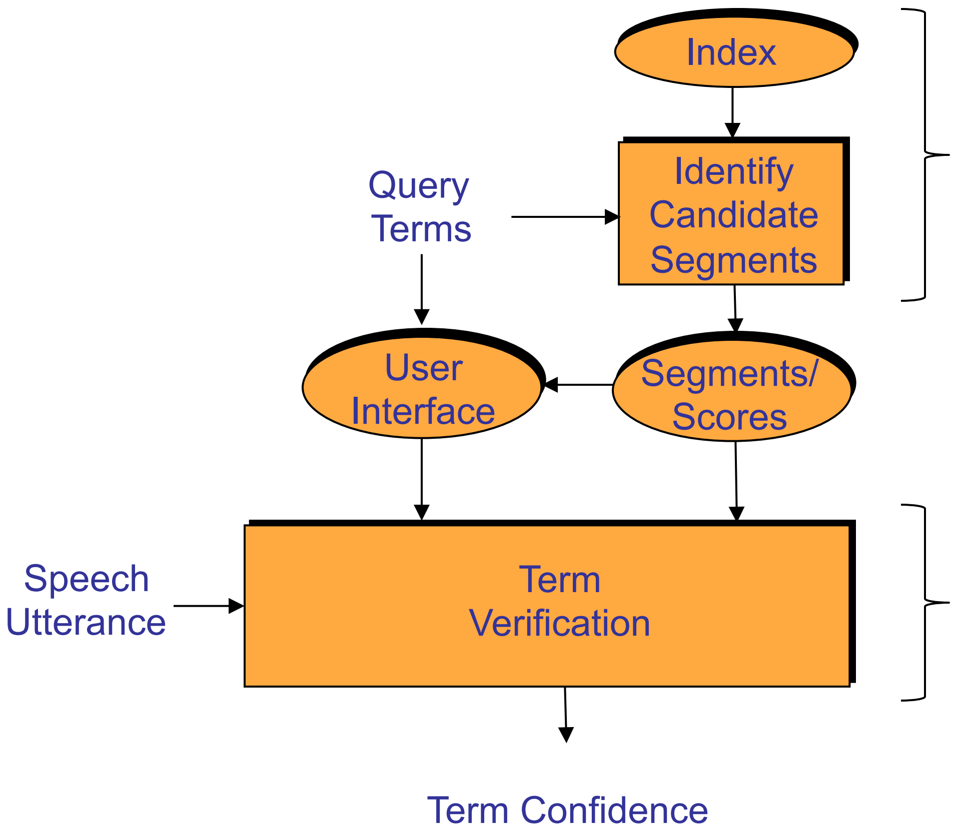
- 88% recall of segments containing V_i
- 65% precision for retrieved segments

In-Vocabulary Query Term Detection

- 81% prob. of detection at 10 false alarms generated per hour

Combing High Resource and Low Resource Models

- Low resource detection of candidates produced by high resource STD system:



High Resource: Lattice Based STD
Fast Search
Increased Segment "Richness"
"Tokenization" of Term Occurences
Labels for "Pseudo-Supervision"

Low Resource: Term Detection
Candidate Re-Ranking
Term Discovery
Out-of-Vocabulary Detection
Domain Adaptation

Low Resource Models – Benefits of Incorporating Complementary Knowledge Sources

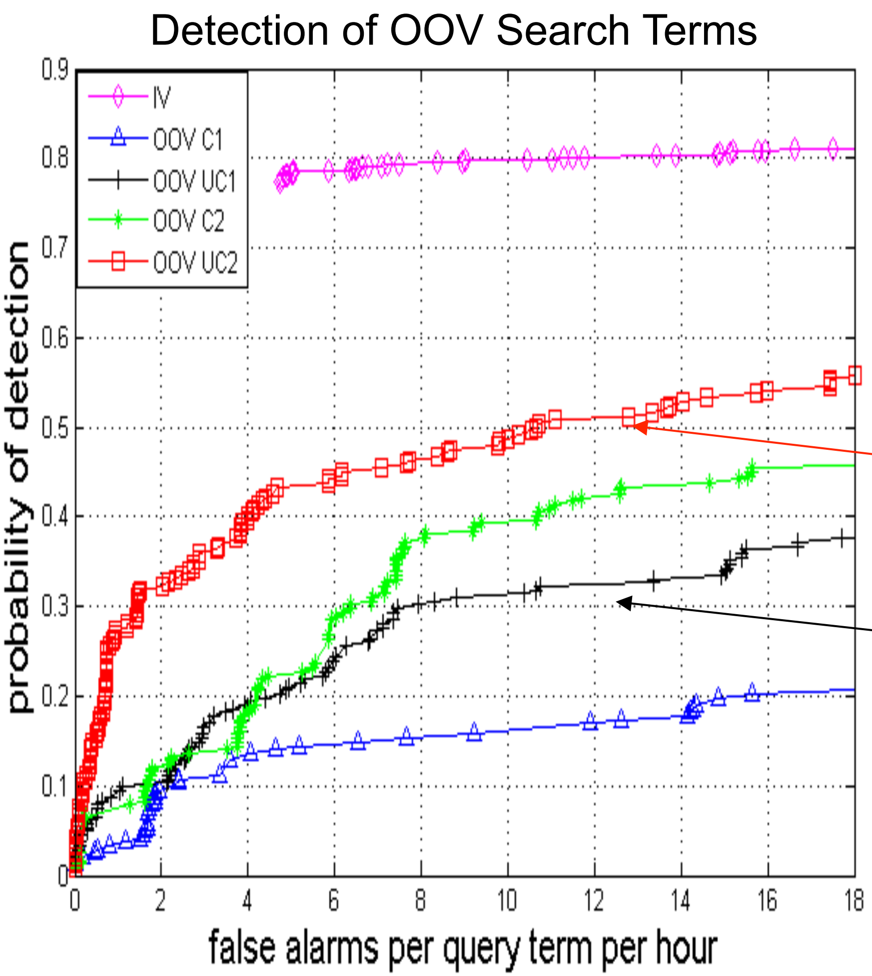
- High Resource: Lattice based STD
  - Speed – Search with lattice based indexing can be ***thousands of times faster than real time***
  - Accuracy – ***Increases richness*** of audio segments for search
  - Context – Incorporates acoustic and language context
  - Pseudo-Labels – STD can provide labels for unsupervised training: ***pseudo relevance feedback***
- Low Resource:
  - Diverse Acoustic ***Feature Representations***
  - Diverse ***Modeling Formalisms***
  - Specialized Domains – OOV search terms
  - Learn from behavior of high resource system

Low Resource Knowledge

- Alternate Features:
  - Phone posteriograms
  - Point process patterns
- Alternate Modeling Formalisms
  - Kernel based classifiers [Jansen et al, 2011], SVM, MLP, …
- Unsupervised / Semi-supervised Learning:
  - Pseudo-relevance feedback [Shen et al, 2005][Lee et al, 2010]
- Context Based Constraints
  - Measures of acoustic self-similarity - acoustic dotplots
  - Random walk based graph re-ranking

Low Resource Features: Phone Posteriorgrams

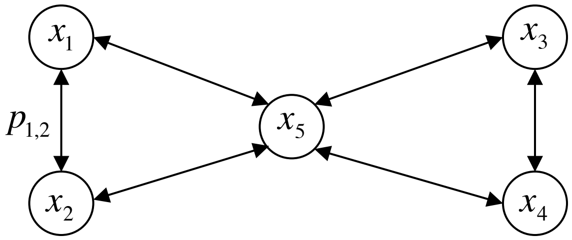
- Unconstrained phone posterior-gram decoder used to increase richness of candidates produced by high resource STD system



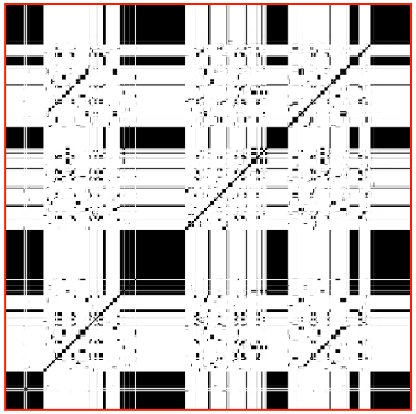
- Unconstrained phone recognition:
  - Hybrid HMM/NN decoder
  - Phone posterior features
  - Find query term phone expansion in decoded phone sequence
- Improves "richness" of segments produced by STD system for OOV words
- Is a relatively "weak" knowledge source when scoring "un-filtered" segments

Low Resource Context Constraints: Graph Based Re-Ranking

- High Resource: Treat scores from spoken term detection as node potentials [Chen et al, 2011]

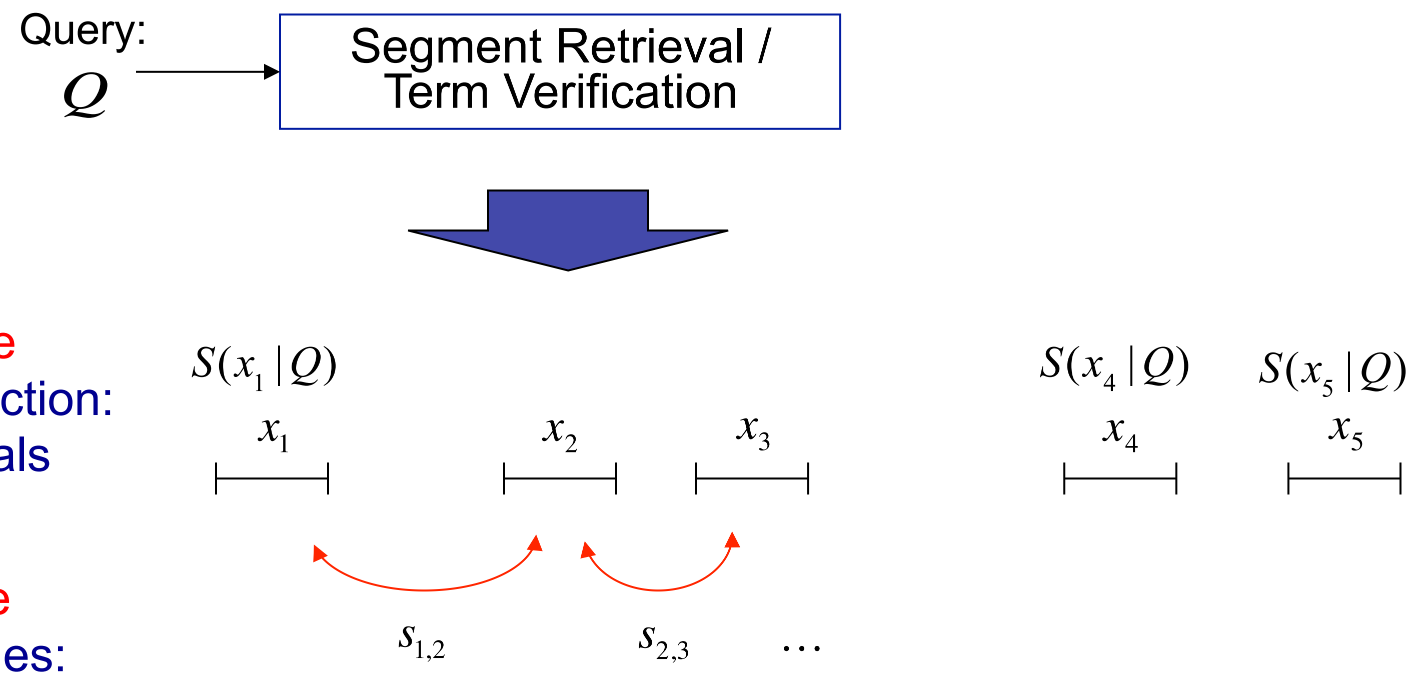


- Low Resource: Use posteriorgram based dot-plots to discover self-similar intervals [Jansen et al, 2010] [Parks and Glass, 2005]



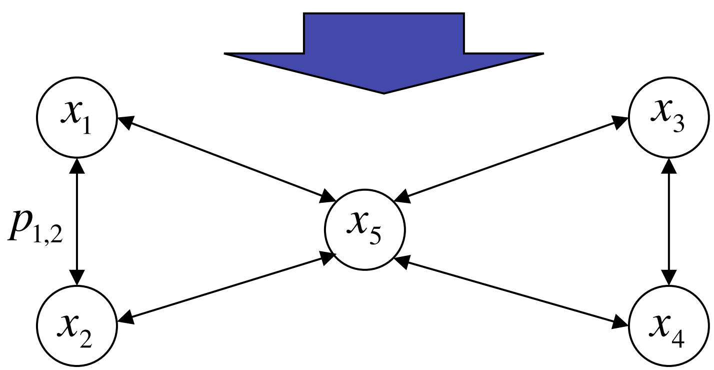
- Graph-Based Re-Ranking: Allows interval similarity to constrain relationship among scores [Hsu et al 2007]
  - Increase / decrease confidence of hypothesized terms based on similarity to other hypotheses
  - Discover new hypotheses?

Graph Based Re-Ranking



High Resource
Spoken Term Detection:
Detected Intervals

Low Resource
Interval Similarities:
Acoustic Dot Plots



Constraints
Graph Re-ranking

Arc weights:

$$p_{i,j} = \frac{s_{i,j}}{\sum_k s_{i,k}}$$

Normalized Scores:

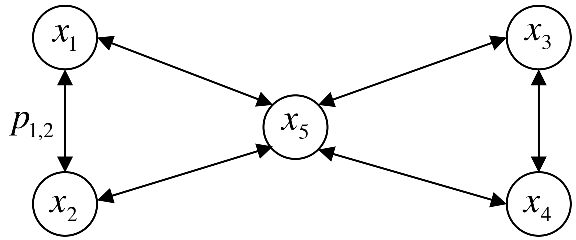
$$u_i = \frac{S(x_i \mid Q)}{\sum_{x_i} S(x_i \mid Q)}$$

Graph Based Re-Ranking for Spoken Term Detection

- Treat scores from spoken term detection as node potentials [Chen et al, 2011]

Arc weights:

$$p_{i,j} = \frac{s_{i,j}}{\sum_k s_{i,k}}$$



Normalized Scores:

$$u_i = \frac{S(x_i \mid Q)}{\sum_{x_i} S(x_i \mid Q)}$$

- Random Walk – State probabilities $\mathbf{v}_k = v_k(1), \ldots, v_k(N)$ at step k [Hsu et al, 2007]

$$v_k(j) = \alpha \sum_{i \in B_j} v_{k-1}(i) p_{i,j} + (1-\alpha) u(j)$$

… Bias towards nodes (intervals) with high STD scores by adjusting α

Graph Based Re-Ranking for Spoken Term Detection

- Random Walk – State probabilities $\mathbf{v}_k = v_k(1), \ldots, v_k(N)$ at step k [Hsu et al, 2007]

$$v_k(j) = \alpha \sum_{i \in B_j} v_{k-1}(i) p_{i,j} + (1-\alpha) u(j)$$

- Random Walk Steady State:

$$v_\pi(j) = \alpha \sum_{i \in B_j} v_\pi(i) p_{i,j} + (1-\alpha) u(j)$$

- Steady State Solution for $\mathbf{v}_\pi$ - Largest eigenvector of $\mathbf{R}$:

$$\mathbf{v}_\pi^T = \mathbf{v}_\pi^T \mathbf{R}$$

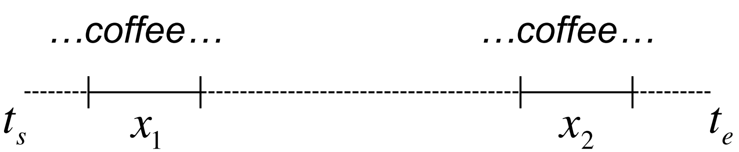
- These state probabilities are used as the "constrained" scores for spoken terms

Electrical & Computer Engineering

Interval Similarity: Acoustic Dot-Plots

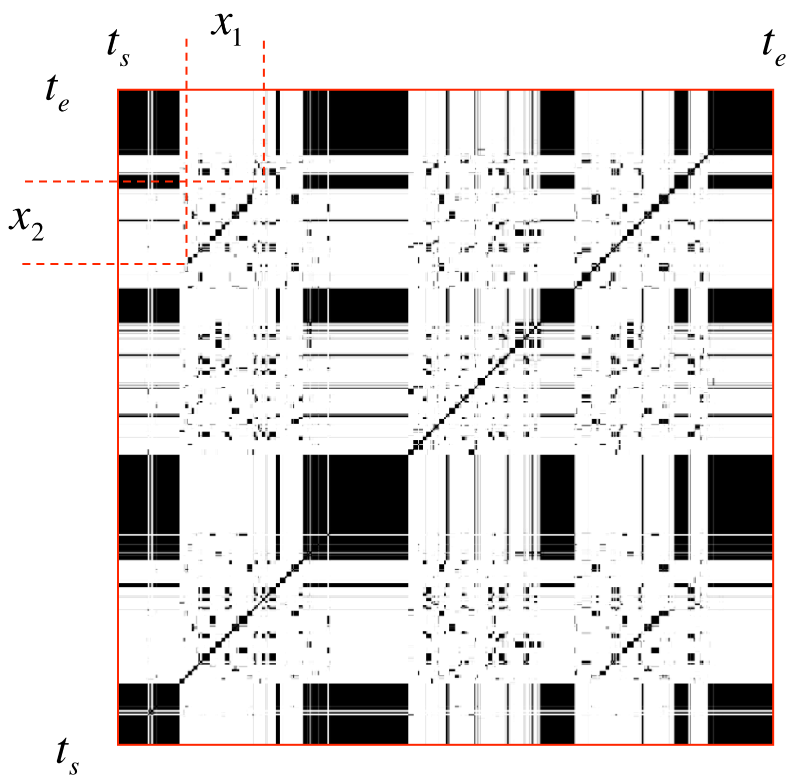
- Posterior-gram Dot-plot : [Jansen et al, 2010] [Parks and Glass, 2005]

- Locate "self-similar" regions:



... from thresholded posteriorgram images [Jansen et al, 2010]

- Retain interval "matches" whose distances exceed a threshold



Taken with no permission what so ever from [Jansen et al, 2010]

Electrical & Computer Engineering

Anecdotal Experiment for a Lecture Speech Task

- Given ASR based STD system and a lecture speech utterance
- Segment continuous utterance into ~2 – ~30 sec. segments
- Generate hypothesized occurrences (intervals and scores) for a single query term (*“penicillin”*) in all segments
- Obtain posterior-gram based dot-plots for each pair of acoustic segments and find self-similar intervals
- Perform graph re-ranking of hypothesized term occurrences

Single Term STD from Lecture Speech

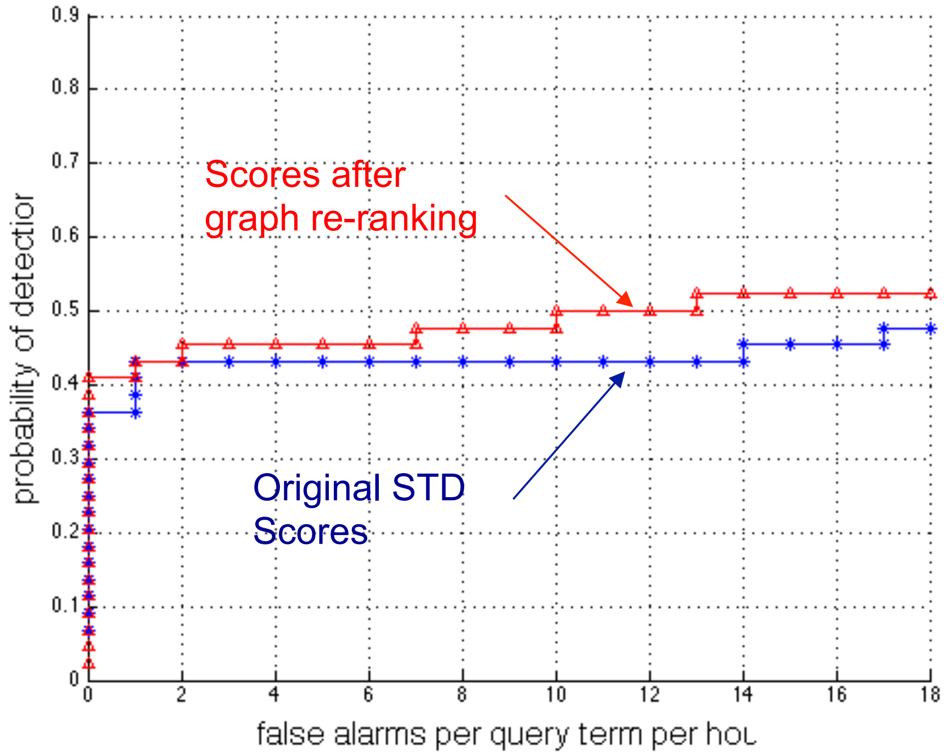
- A single ~1 hour chemistry lecture – 167 segments
- ASR: WAC – 58%, OOV Rate (type) – 11.2%
- Spoken term detection:
  - Pick a single ***in-vocabulary*** term: "*penicillin*"
    - 59 actual occurrences
  - Retrieved hypothesis:
    - 340 total hypothesized occurrences – 45 correct and 295 false alarms
- Plot Receiver operating curve (ROC) for STD scores before and after re-ranking

Graph Re-ranking for a Lecture Speech Task

- Compute posterior-grams
  - Obtained for 167 segments
- Compute Dot-plots for each segment pair
  - discovered N=5700 connected intervals (graph states)
- STD hypothesized query terms occurrences are "connected" by dot-plots:
  - 328 STD hypotheses (out of 340 total) overlapped by at least one dot plot detected interval
  - 44 out 45 correct STD hypotheses overlap dot-plot detected interval

Graph Re-ranking for a Lecture Speech Task

- Receiver operating curve (ROC) for STD scores before and after re-ranking:



- Shows a small but significant increase improvement in detection performance

Electrical & Computer Engineering

Summary

- Graph re-ranking is an interesting formalism for
  - constraining events produced by STD
  - Using low resource non-parametric measures of self similarity
- Further Work
  - Perform a complete experiment – Full inventory of search terms
  - Investigate empirical issues for random walk
  - Incorporate zero resource dot plots [Jansen et al, 2012]
  - Investigate other notions of self-similarity in graph re-ranking
  - Term Discovery?