# Spoken Content Retrieval

## Beyond Cascading Speech Recognition and Text Retrieval

Lin-shan Lee and Hung-yi Lee

National Taiwan University

# Focus of this Tutorial

- New frontiers and directions towards the future of speech technologies
- Not skills and experiences in optimizing performance in evaluation programs

# Text Content Retrieval

# Spoken Content Retrieval

# Spoken Content Retrieval

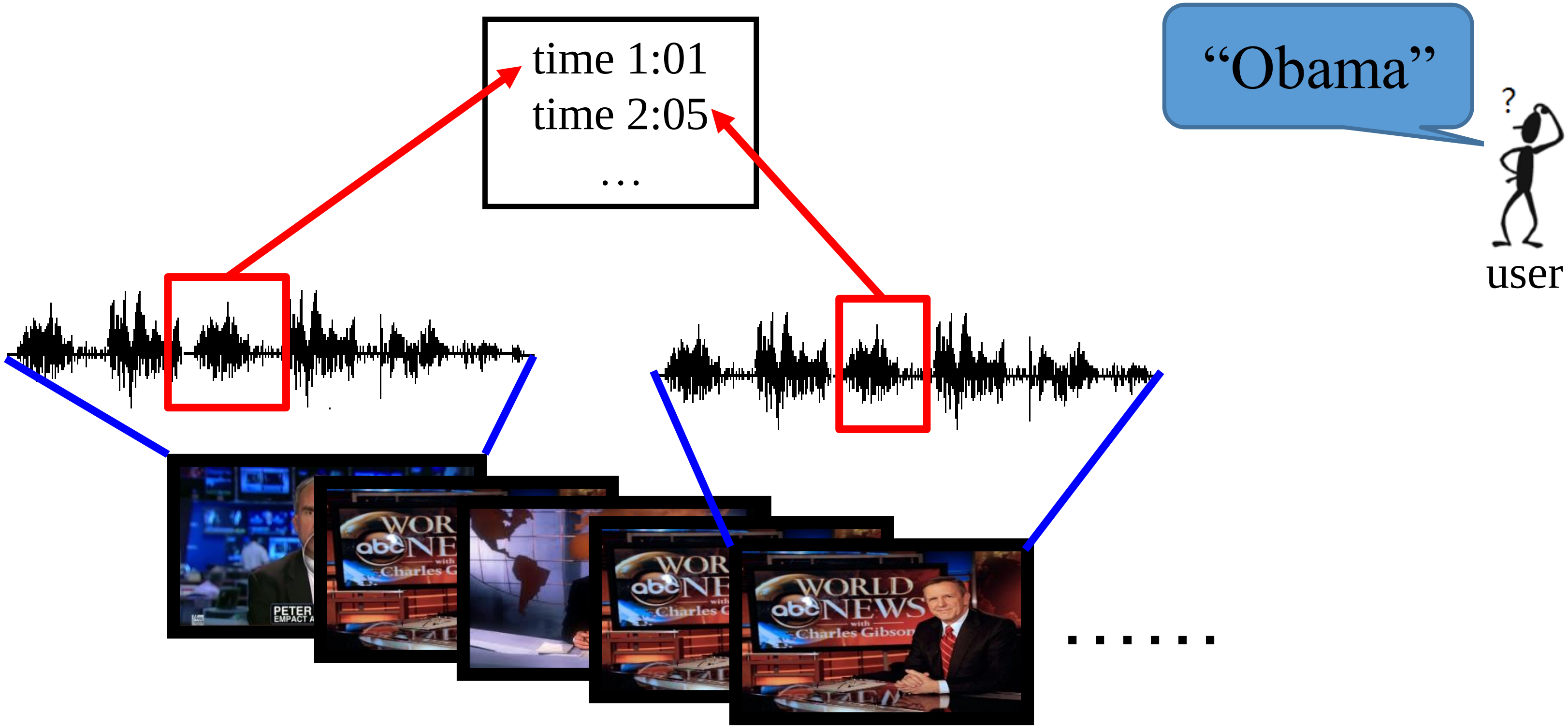
300 hrs multimedia is uploaded per minute. (2015.01)

1874 courses on coursera (2016.04)

- Nobody is able to go through the data.
- In these multimedia, the spoken part carries very important information about the content
- Spoken content retrieval: Machine listens to the data, and extract the desired information for each individual user.
  - Just as Google does on text data

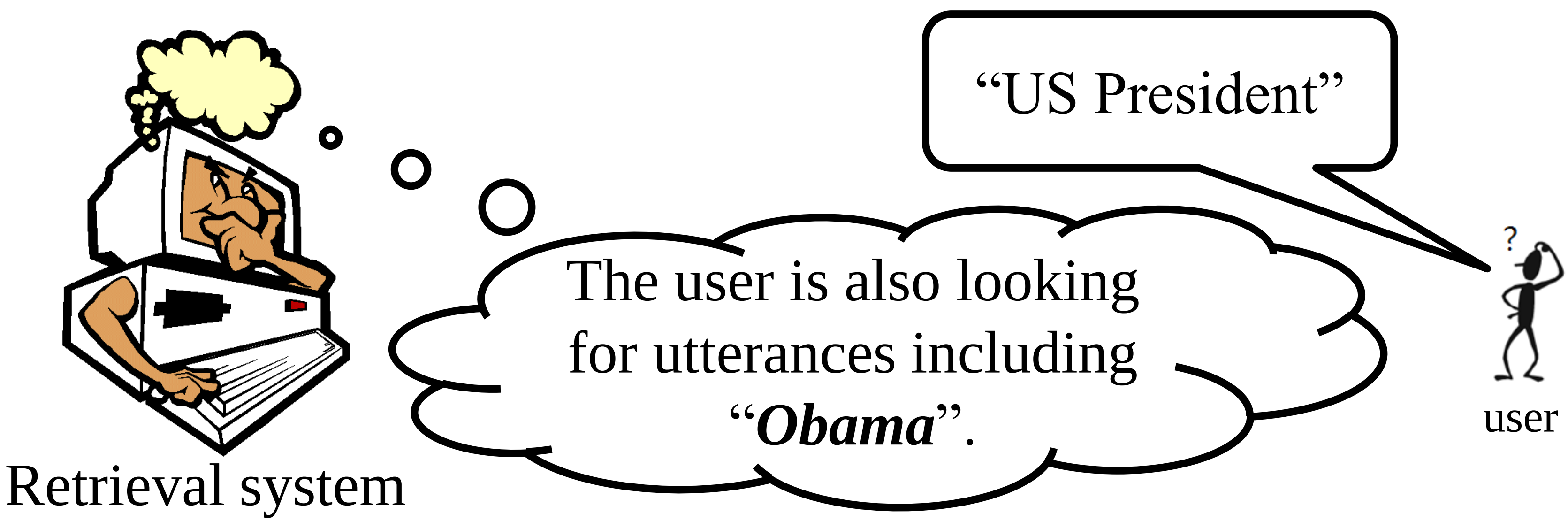
# Spoken Content Retrieval – Goal

- Basic goal: Identify the time spans that the query occurs in an audio database
  - This is called "***Spoken Term Detection***"



# Spoken Content Retrieval – Goal

- Basic goal: Identify the time spans that the query occurs in an audio database
  - This is called "***Spoken Term Detection***"
- Advanced goal: Semantic retrieval of spoken content



# It is natural to think ……

Spoken Content Retrieval

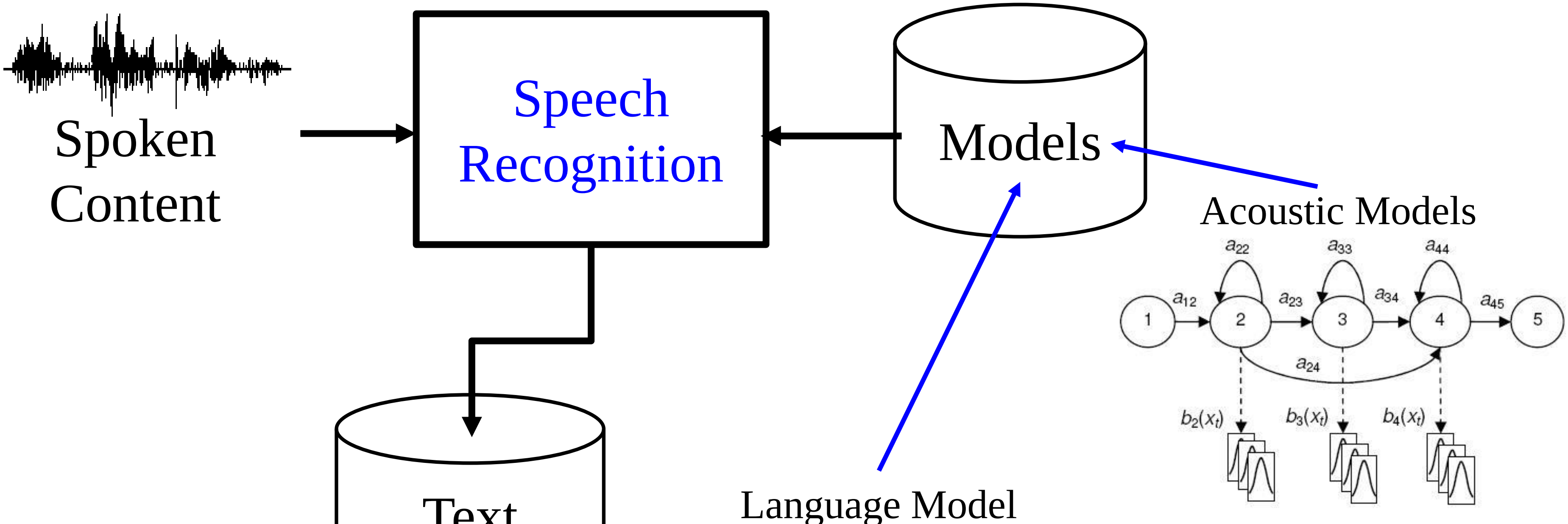
=

Speech Recognition

+

Text Retrieval

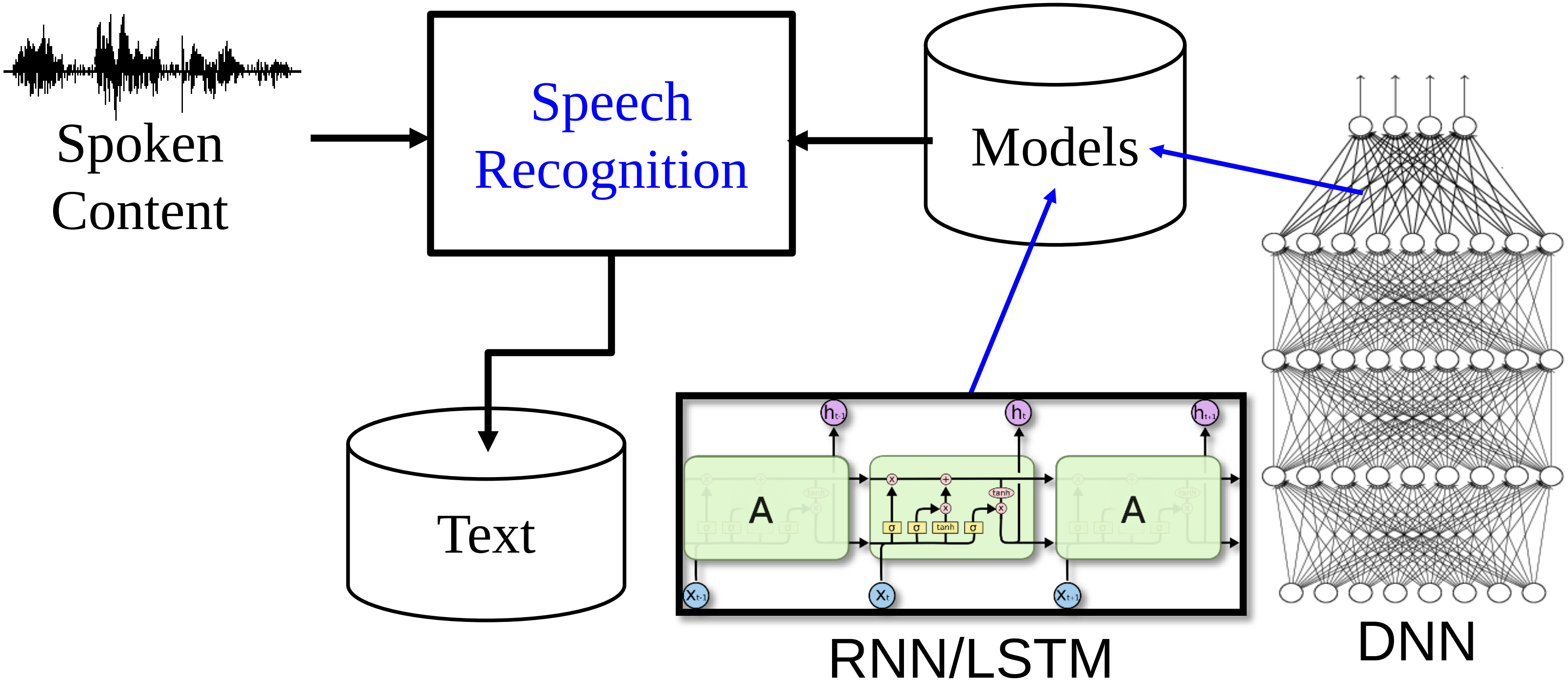
# It is natural to think ……




$$P(w_i|w_{i-(n-1)},\dots,w_{i-1}) = \frac{count(w_{i-(n-1)},\dots,w_{i-1},w_i)}{count(w_{i-(n-1)},\dots,w_{i-1})}$$

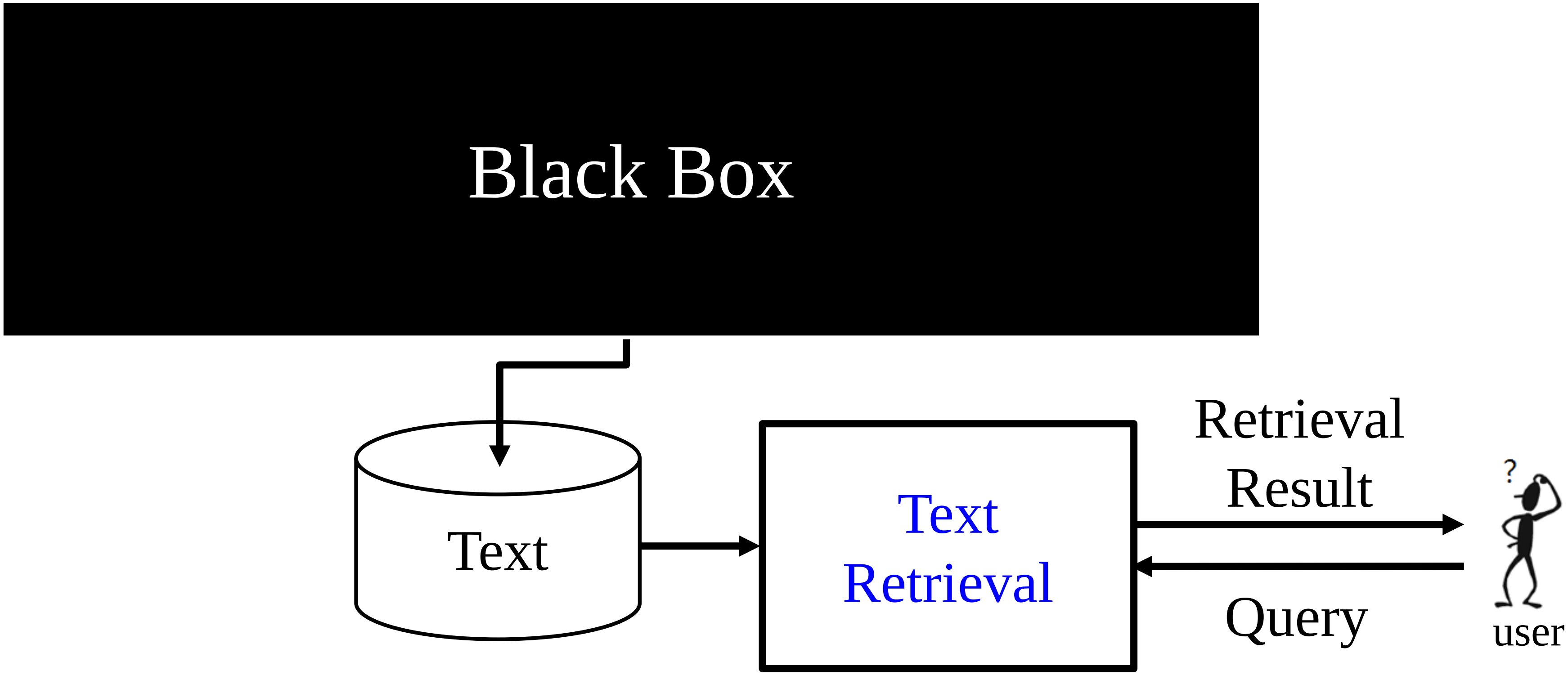
- Transcribe spoken content into text by speech recognition

# It is natural to think ……



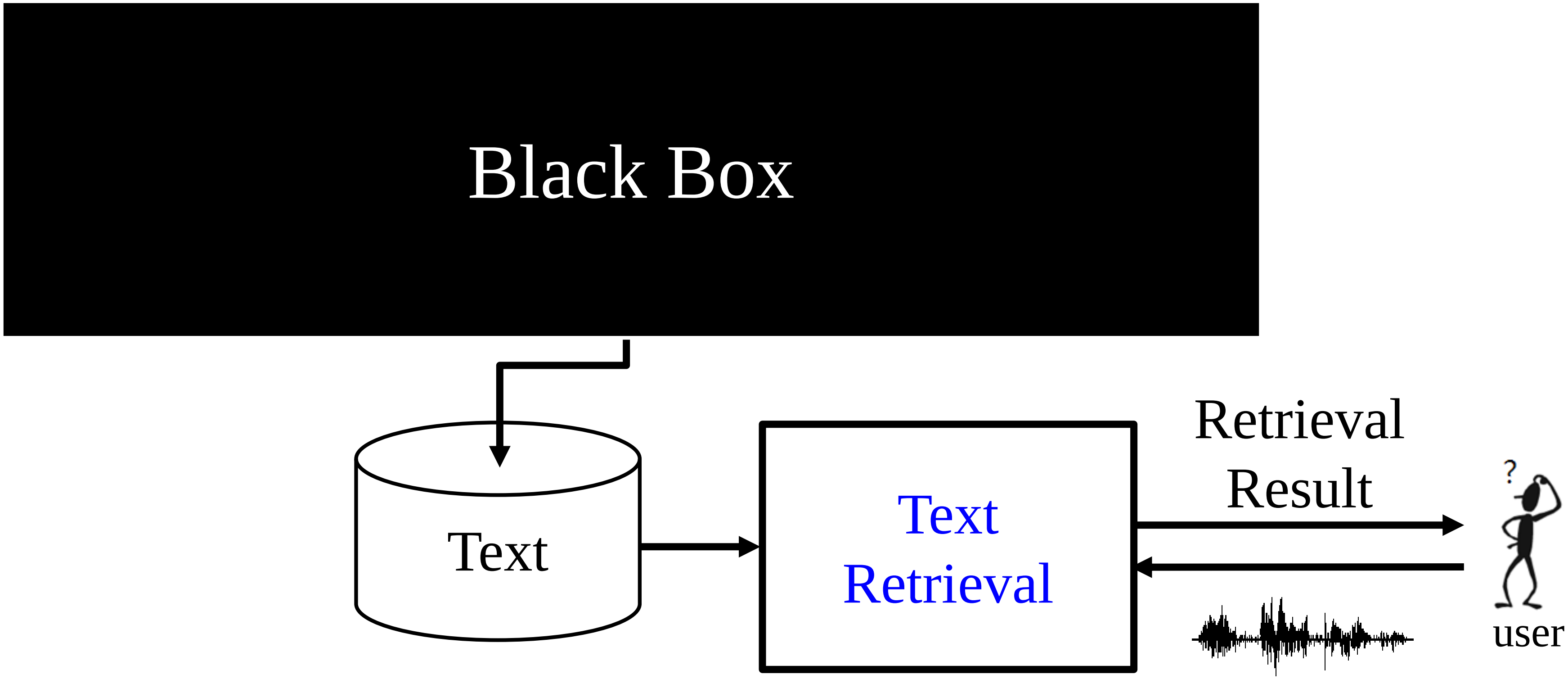

- Transcribe spoken content into text by speech recognition

# It is natural to think ……




- Transcribe spoken content into text by speech recognition
- Use text retrieval approaches to search over the transcriptions

# It is natural to think ……




- For spoken queries, transcribe them into text by speech recognition.

# Our point in this tutorial

Spoken Content Retrieval

$\cancel{=}$

Speech Recognition

+

Text Retrieval

# Outline

- Introduction: Conventional Approach:

  *Spoken Content Retrieval =*

  *Speech Recognition + Text Retrieval*

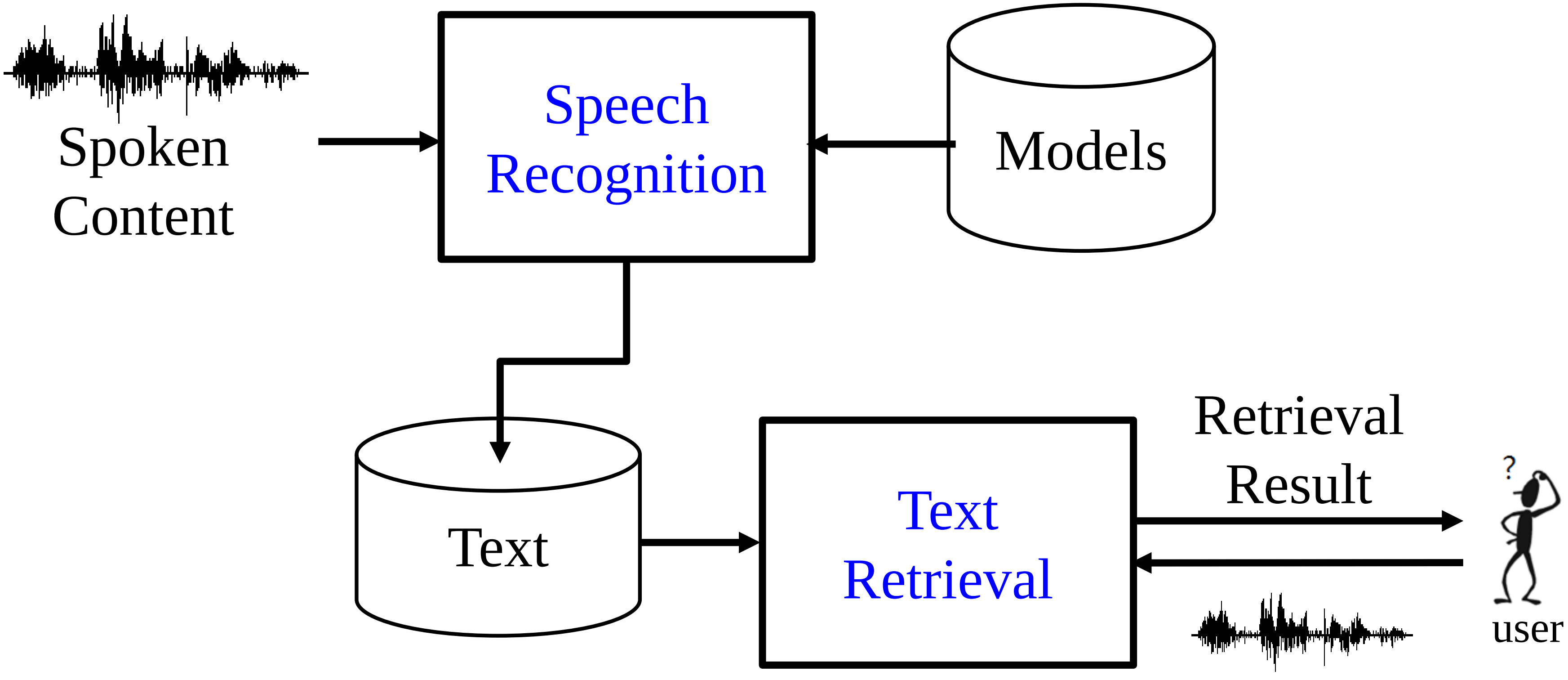
- Core: Beyond Cascading Speech Recognition and Text Retrieval
  - Five new directions

*Introduction:*

*Spoken Content Retrieval =*

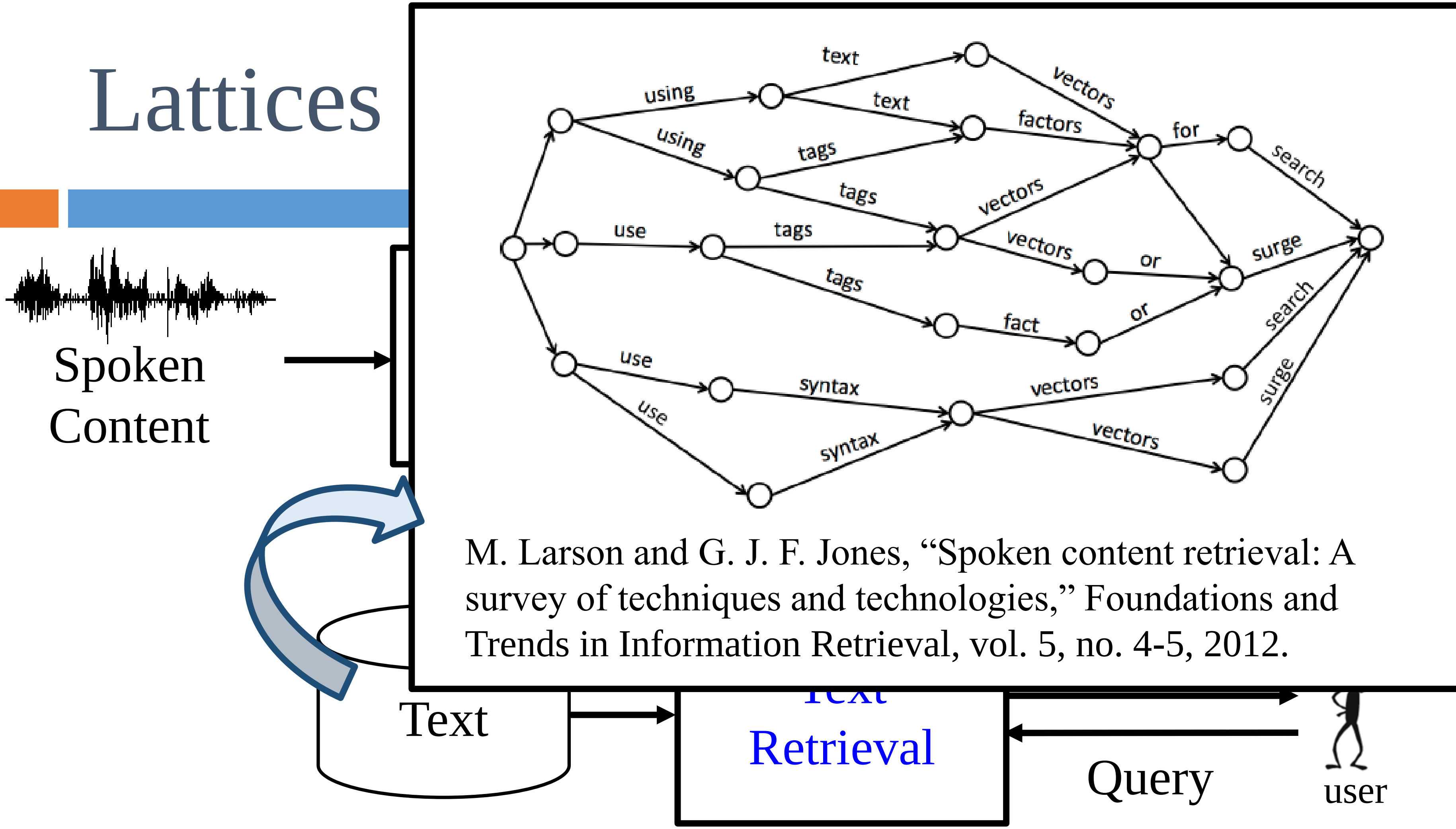
*Speech Recognition + Text Retrieval*

# It is natural to think ……




Speech Recognition always produces errors.

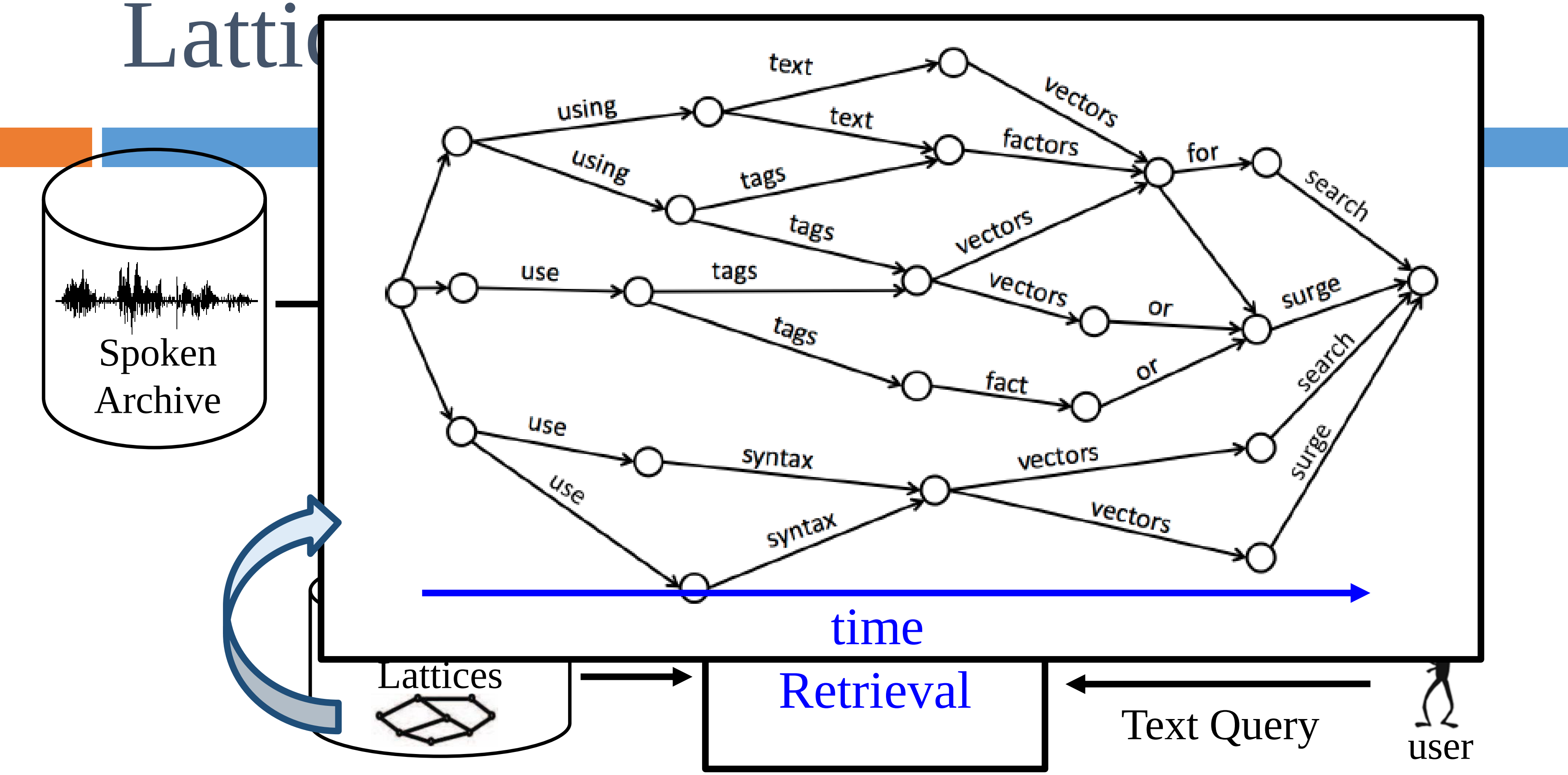
# Lattices




**Lattices**

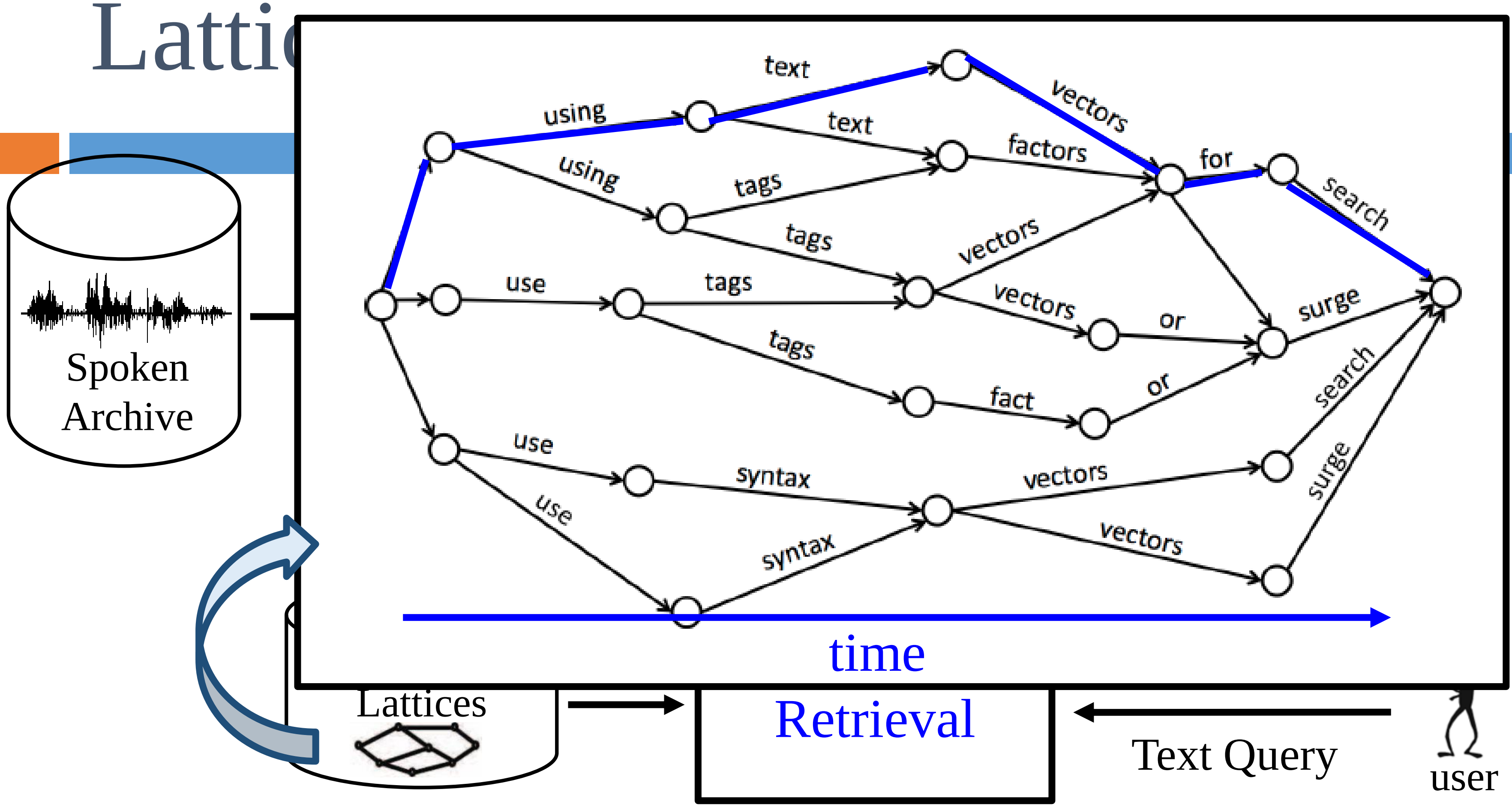
- Keep most possible recognition output
- Each path has a weight (confidence to be correct)

# Latti




- ➢Horizontal scale is the time
- ➢Each path is a possible recognition result

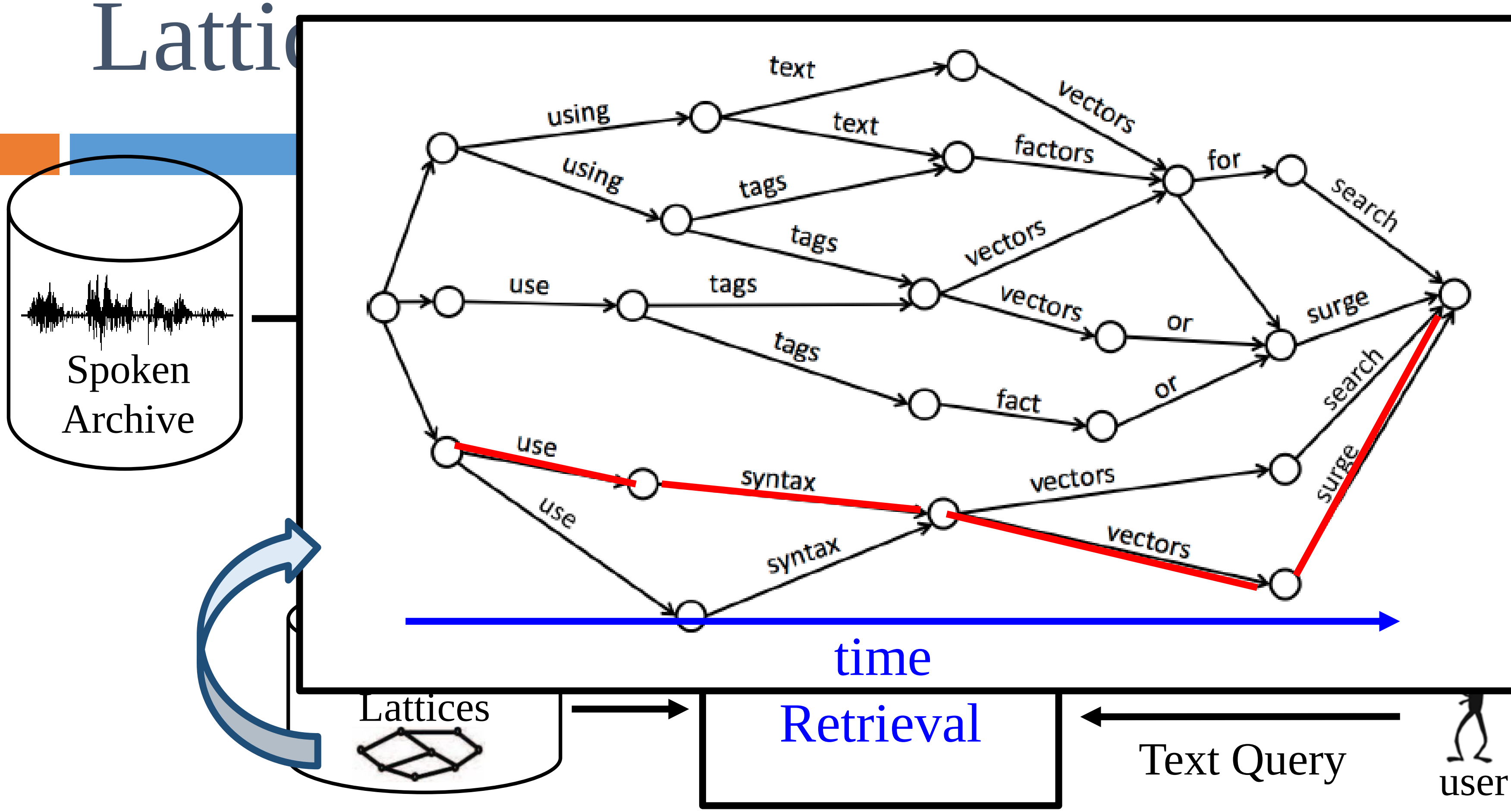
# Latti




➢Horizontal scale is the time

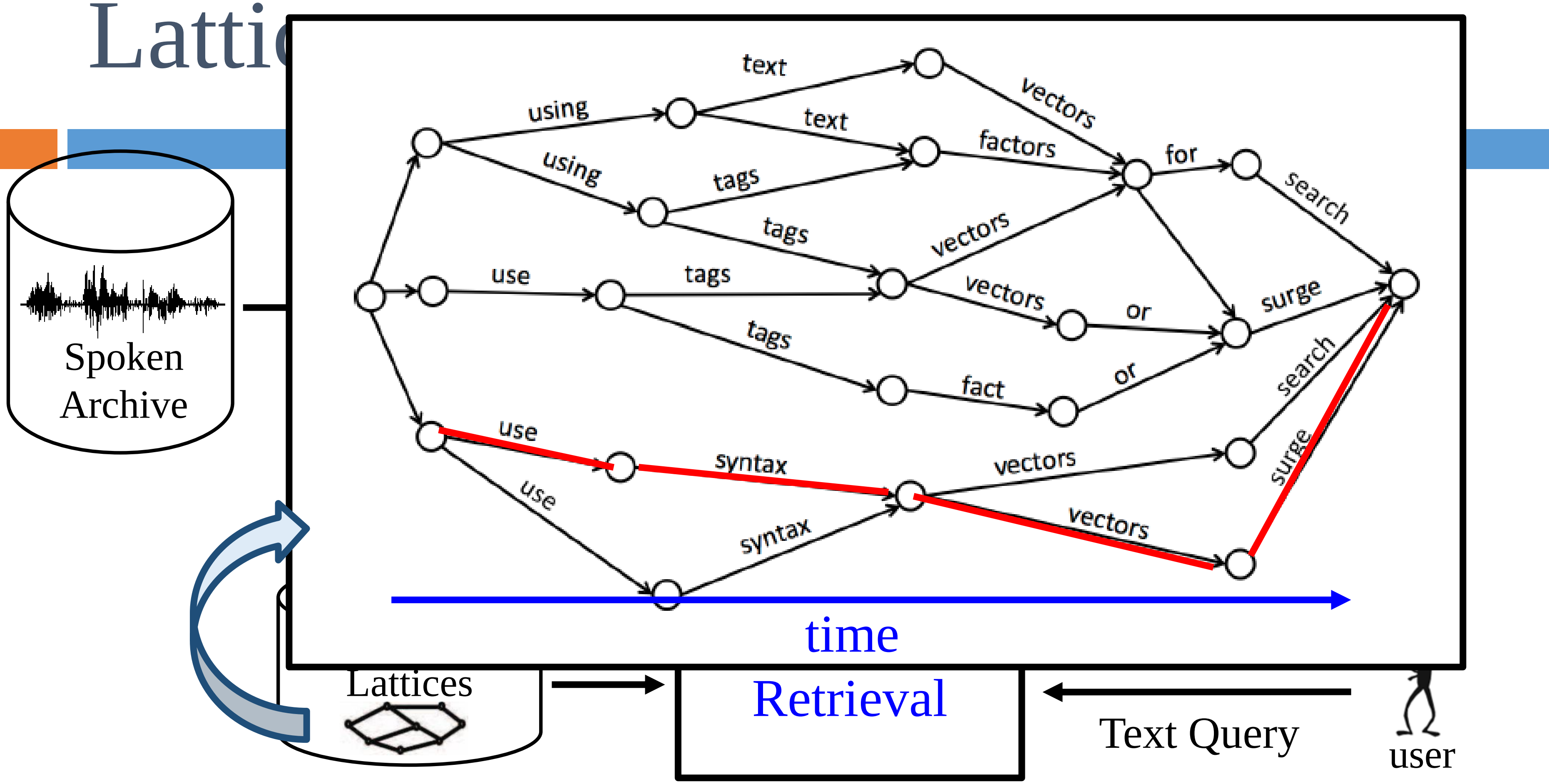
➢Each path is a possible recognition result

# Latti




- Horizontal scale is the time
- Each path is a possible recognition result

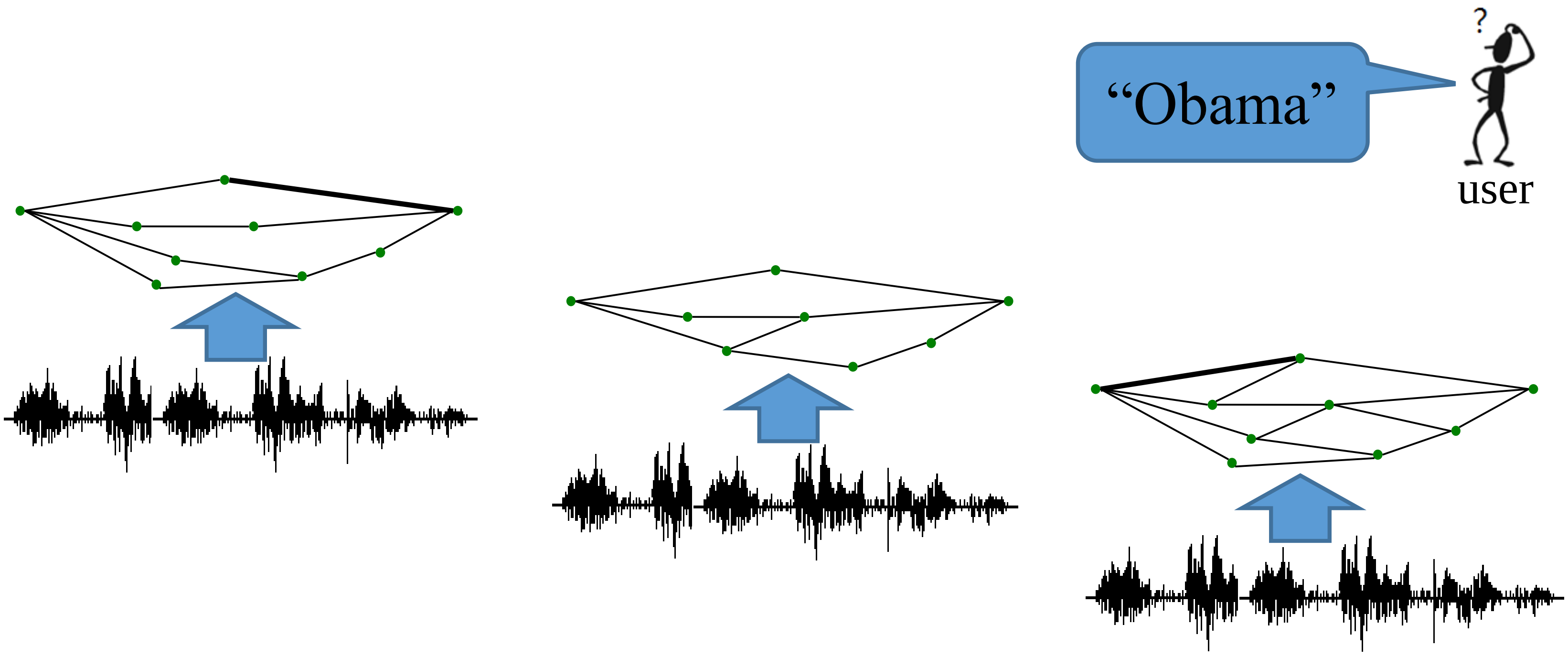
# Latti




- ➢Higher probability to include the correct words
- ➢More noisy words included inevitably
- ➢Higher memory/computation requirements

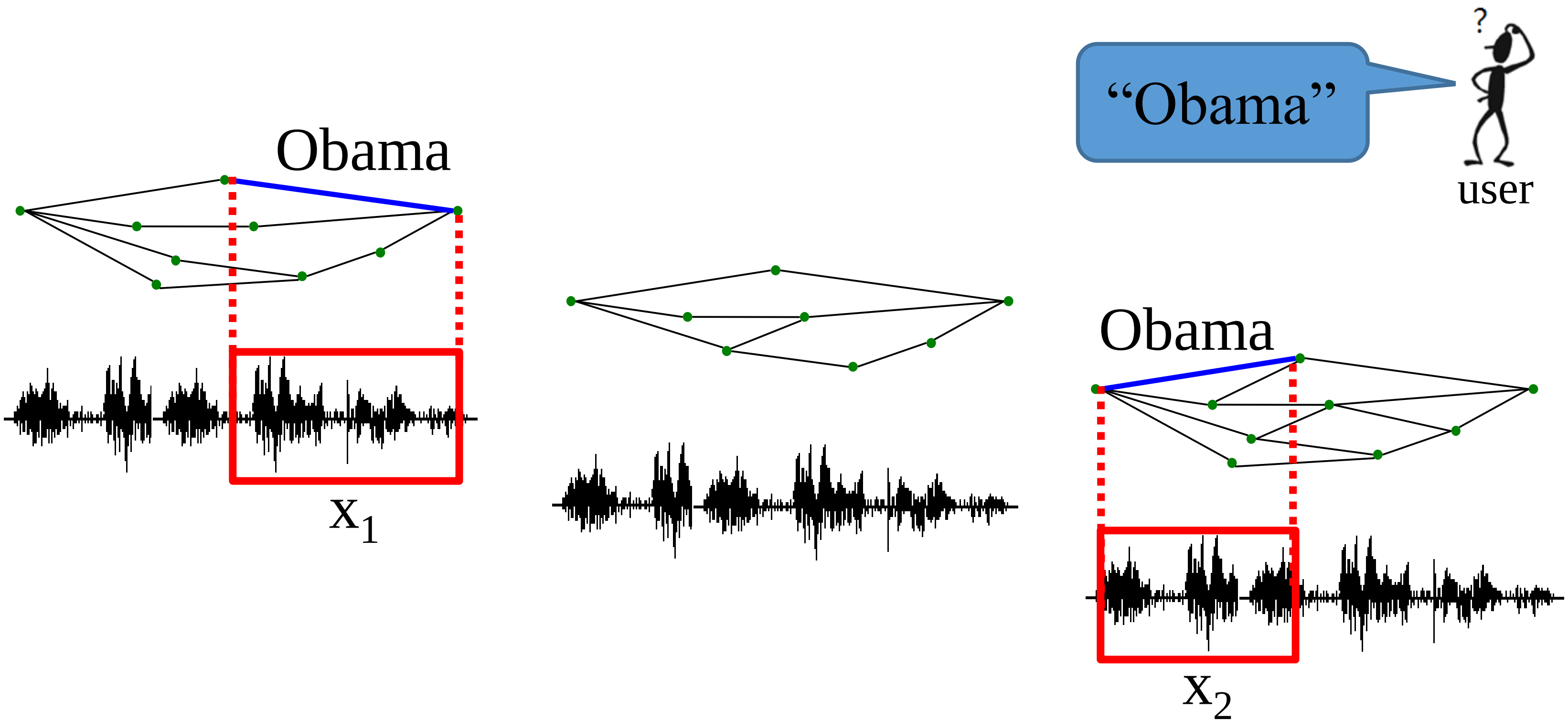
# Searching over Lattices

- Consider the basic goal: Spoken Term Detection



# Searching over Lattices

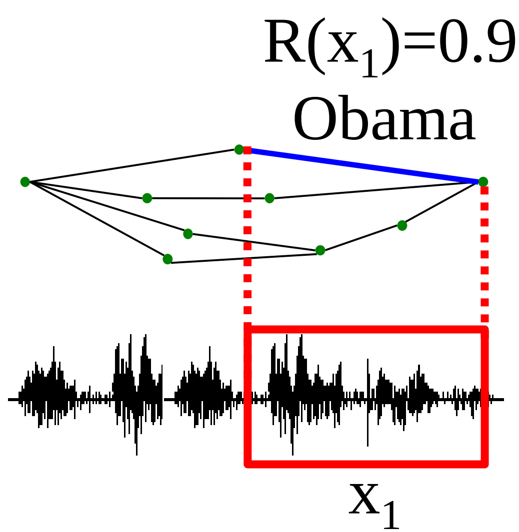
- Consider the basic goal: Spoken Term Detection
  - Find the arcs hypothesized to be the query term

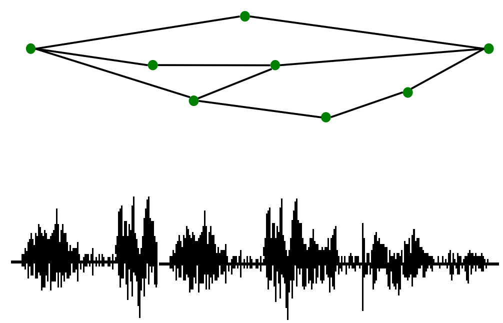


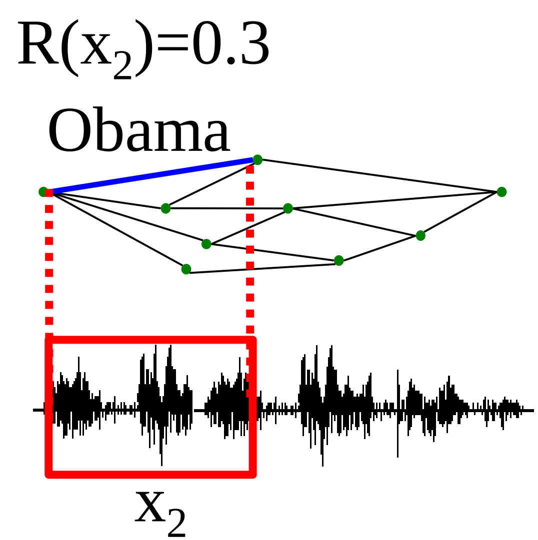
# Searching over Lattices

- Consider the basic goal: Spoken Term Detection
  - ***Posterior probabilities*** computed from lattices used as confidence scores

Two ways to display the results: unranked and ranked.

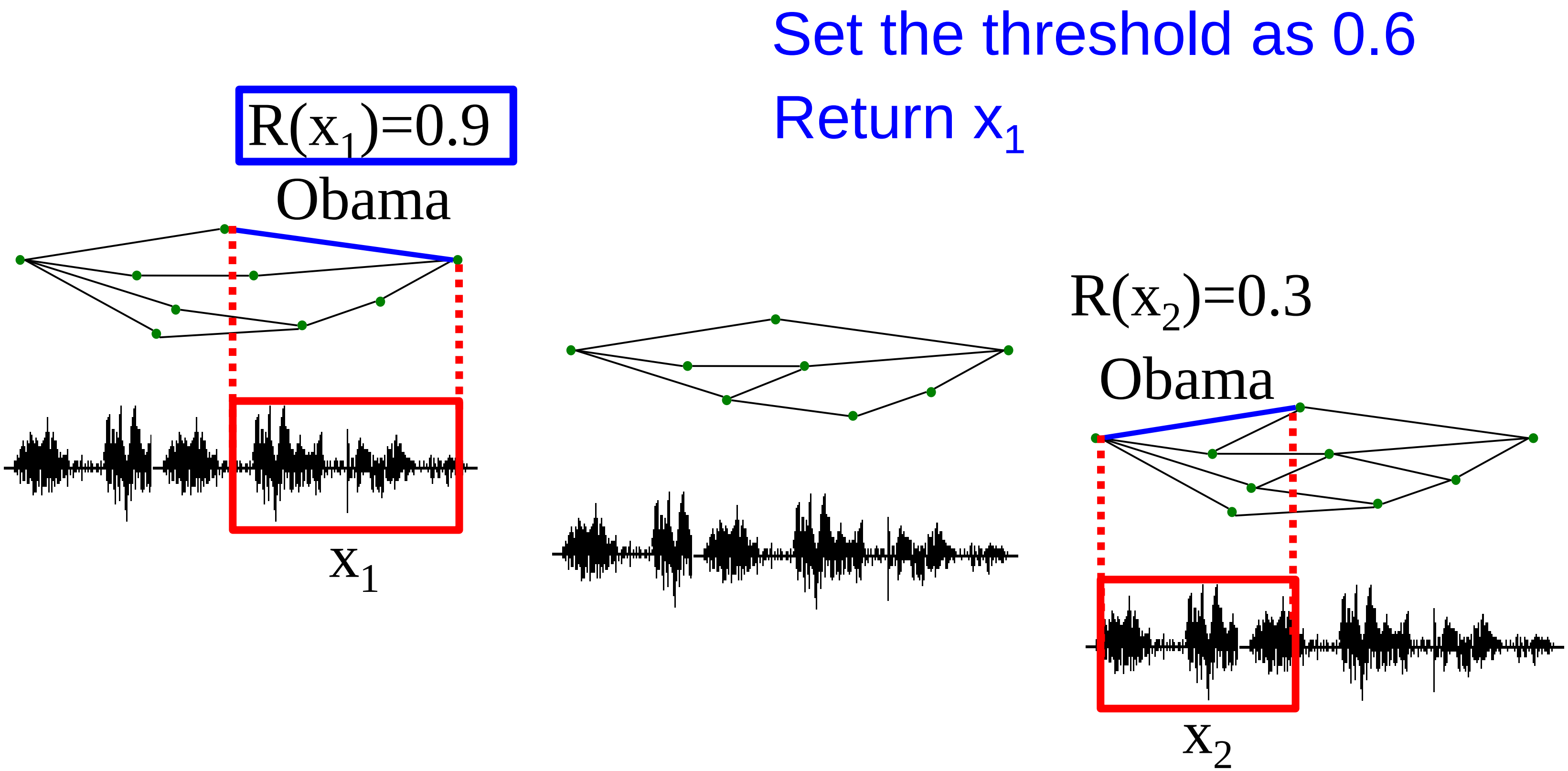




# Searching over Lattices

- Consider the basic goal: Spoken Term Detection
  - Unranked: Return the results with the scores higher than a threshold

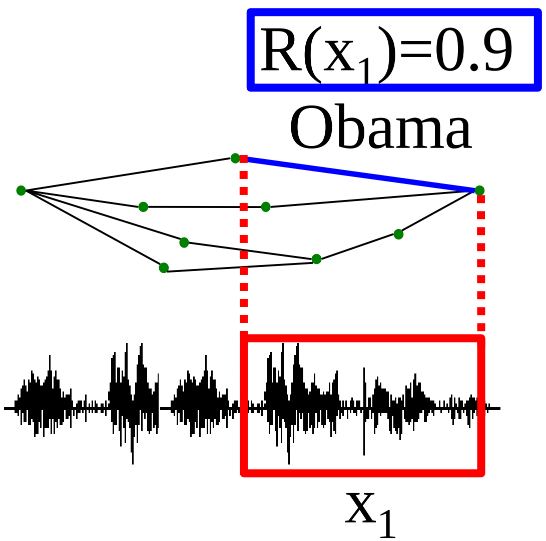


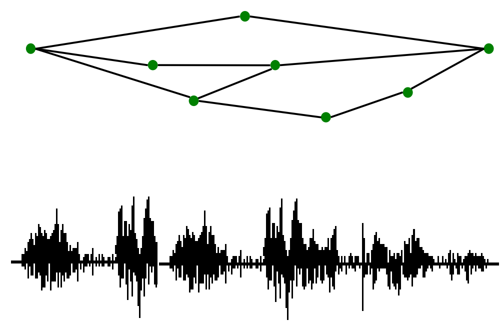
# Searching over Lattices

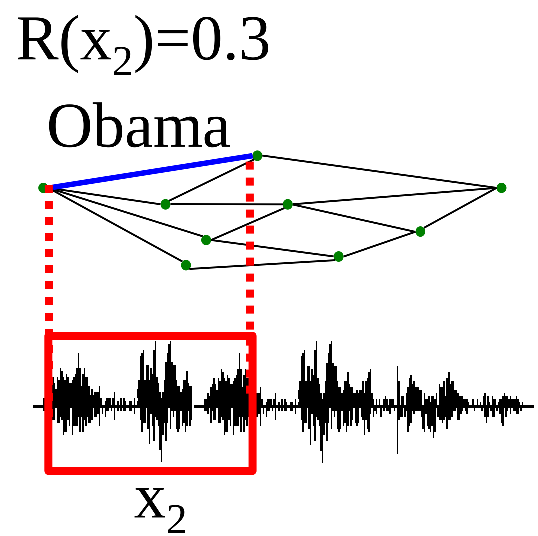
- Consider the basic goal: Spoken Term Detection
  - Unranked: Return the results with the scores higher than a threshold

The threshold can be determined automatically and query specific.

[Miller, Interspeech 07][Can, HLT 09][Mamou, ICASSP 13][Karakos, ASRU 13][Zhang, Interspeech 12][Pham, ICASSP 14]

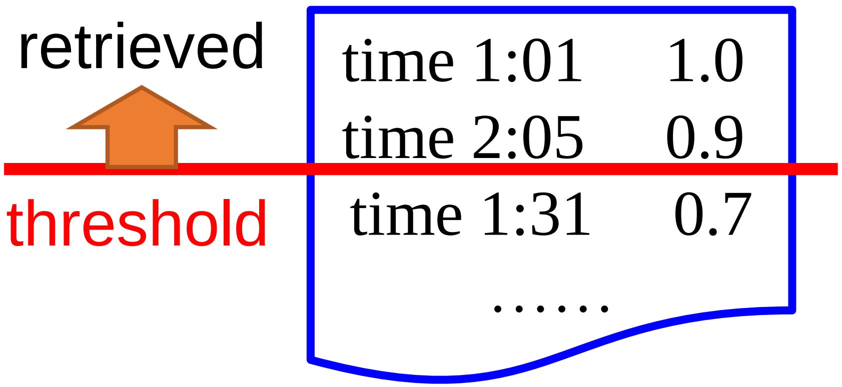




# Actual Term Weighted Value (ATWV)

- Evaluating unranked result




$$ATWV = 1 - P_{miss} - \beta P_{FA}$$

$$P_{miss} = 1 - N_{correct} / N_{ref}$$

$$P_{FA} = N_{spurious} / N_{NT}$$

$N_{ref}$: number of times the query term appears in audio database

$N_{correct}$: the number of retrieved objects that are actually correct

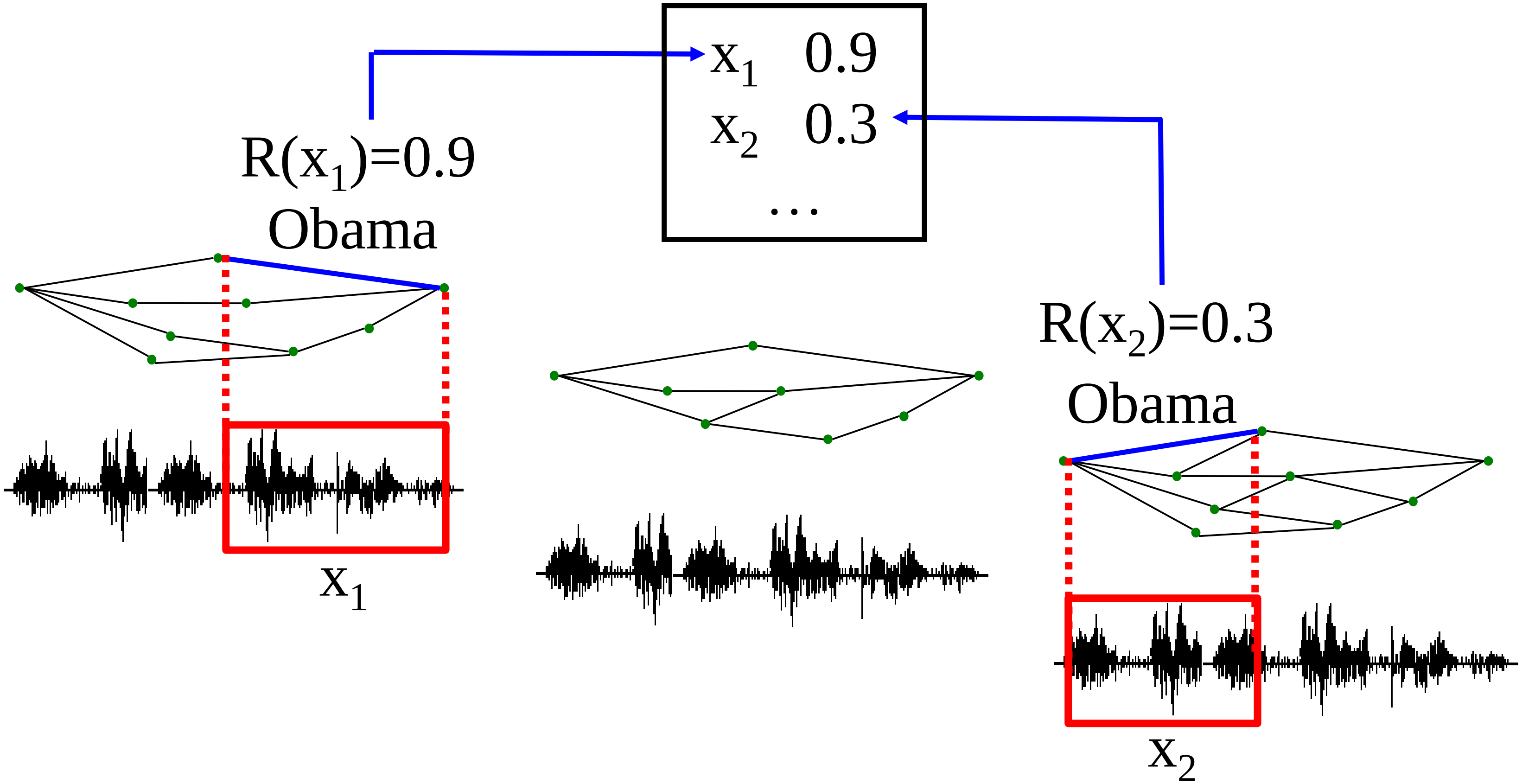
$N_{spurious}$: the number of retrieved objects that are incorrect

$N_{NT}$: audio duration (in seconds) – $N_{ref}$

Maximum Term Weighted Value (MTWV): tune the threshold to obtain the best ATWV

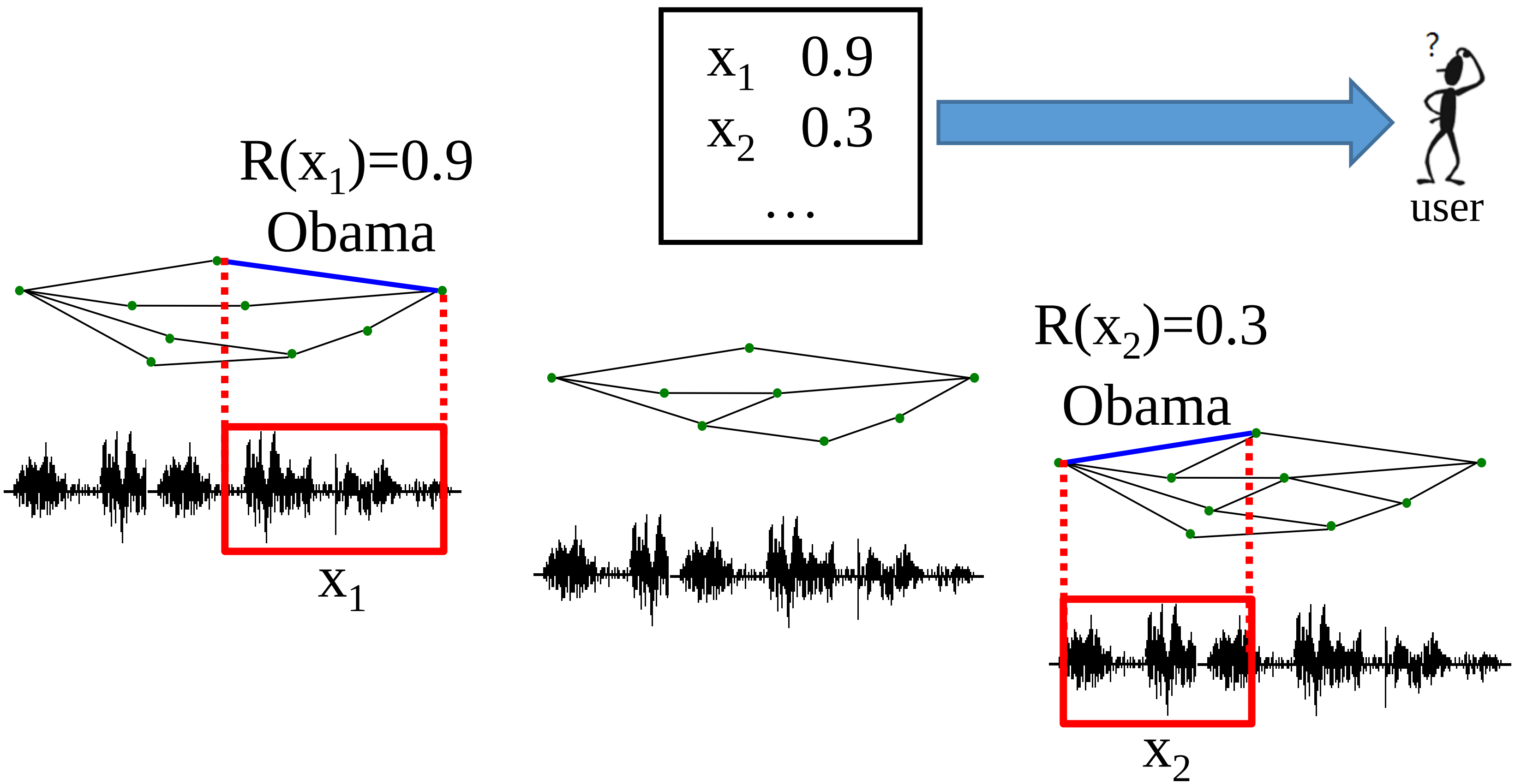
# Searching over Lattices

- Consider the basic goal: Spoken Term Detection
  - Ranked: results ranked according to the scores



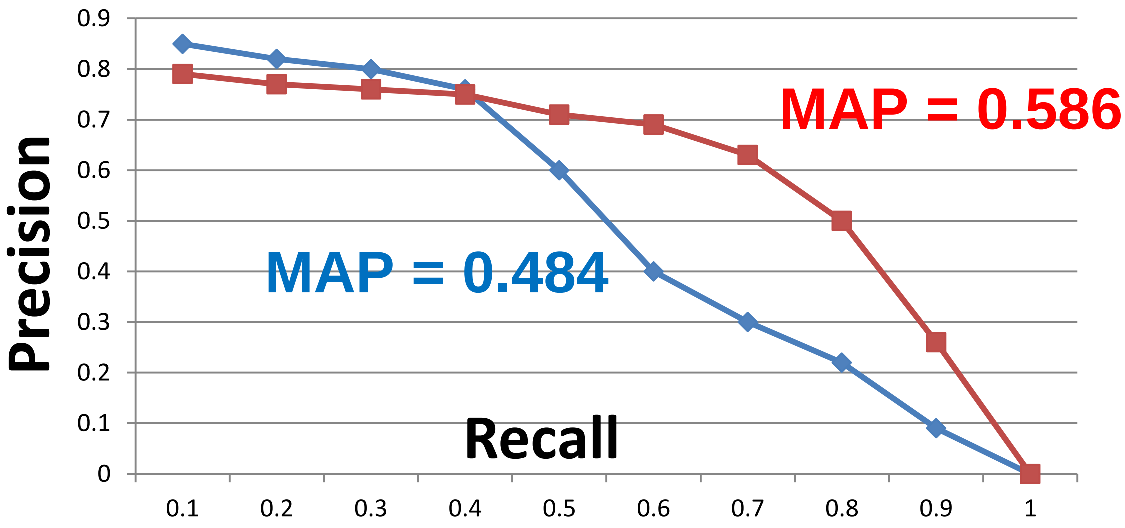
# Searching on Lattices

- Consider the basic goal: Spoken Term Detection
  - Ranked: The results are ranked according to the scores



# Mean Average Precision (MAP)

- Evaluating ranked list
- area under recall-precision curve
  - Recall: percentage of ground truth results retrieved
  - Precision: percentage of retrieved results being correct
  - Higher threshold gives higher precision but lower recall, etc.



# Examples of Lattice Indexing Approaches

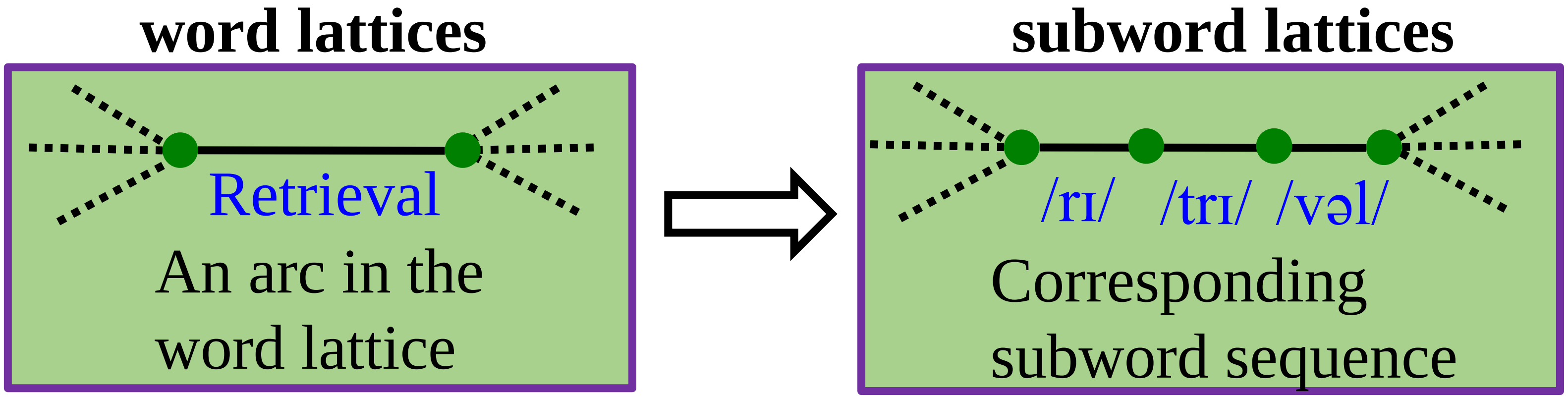
- Position Specific Posterior Lattices (PSPL)[Chelba, ACL 05][Chelba, Computer Speech and Language 07]
- Confusion Networks (CN)[Mamou, SIGIR 06][Hori, ICASSP 07][Mamou, SIGIR 07]
- Time-based Merging for Indexing (TMI)[Zhou, HLT 06][Seide, ASRU 07]
- Time-anchored Lattice Expansion (TALE)[Seide, ASRU 07][Seide, ICASSP 08]
- WFST: directly compile the lattice into a weighted finite state transducer [Allauzen, HLT 04][Parlak, ICASSP 08][Can, ICASSP 09][Parada, ASRU 09]

# Out-of-Vocabulary (OOV) Problem

- Speech recognition is based on a lexicon
- Words not in the lexicon can never be transcribed
- Many informative words are out-of-vocabulary (OOV)
  - Many query terms are new or special words or named entities

# Subword-based Retrieval

- All OOV words composed of subword units
  - Generate subword lattices
    - Transform word lattices into subword lattices




    - Can also be directly generated by speech recognition using subword-based lexicon and language model

# Subword-based Retrieval

- Subword-based retrieval
  - Generate subword lattices
  - Transform user query into subword sequence
    - Obama → /au/ /ba/ /mə/
  - Text retrieval techniques equally useful except based on subword lattices and subword query
    - Replace words by subword units
  - OOV words can be retrieved by matching over the subword units without being recognized

# Subword-based Retrieval - Frequently Used Subword Units

- Linguistically motivated units
  - phonemes, syllables/characters, morphemes, etc.

  [Ng, MIT 00][Wallace, Interspeech 07][Chen & Lee, IEEE T. SAP 02]

  [Pan & Lee, ASRU 07][Meng, ASRU 07][Meng, Interspeech 08]

  [Mertens, ICASSP 09][Itoh, Interspeech 07][Itoh, Interspeech 11]

  [Pan & Lee, IEEE T. ASL 10]

- Data-driven units
  - particles, word fragments, phone multigrams, morphs, etc.

  [Turunen, SIGIR 07] [Turunen, Interspeech 08]

  [Parlak, ICASSP 08][Logan, IEEE T. Multimedia 05]

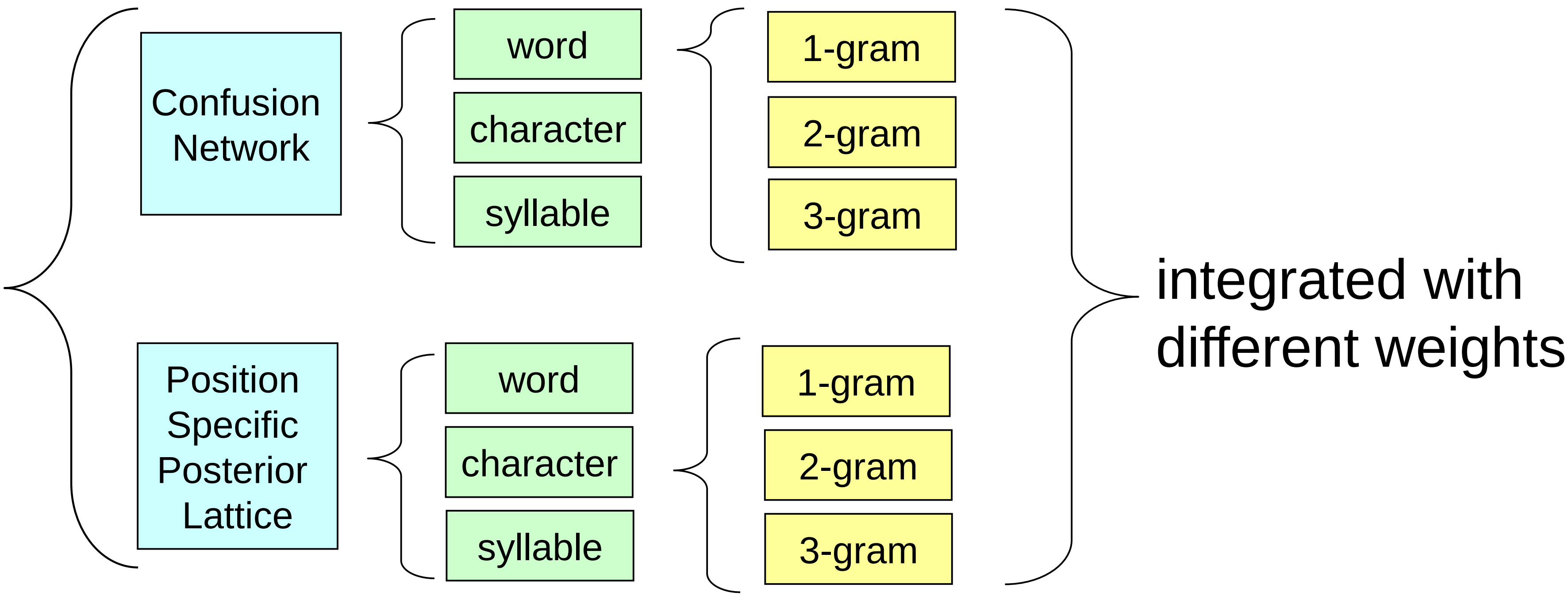
  [Gouvea, Interspeech 10][Gouvea, Interspeech 11][Lee & Lee, ASRU 09]

# Integrating Different Clues from Recognition

- Similar to system combination in ASR
- Consistency very often implies accuracy
  - Integrating the outputs from different recognition systems [Natori, Interspeech 10]
  - Integrating results based on different subword units [S.-w. Lee, ICASSP 05][Pan & Lee, Interspeech 07][Meng, Interspeech 10][Itoh, Interspeech 11]
  - Weights of different clues estimated by optimizing some retrieval related criteria [Meng & Lee, ICASSP 09][Chen & Lee, ICASSP 10][Meng, Interspeech 10][Wollmer, ICASSP 09]

# Integrating Different Clues from Recognition

- Weights for Integrating 1,2,3-grams for different word/subword units and different indices

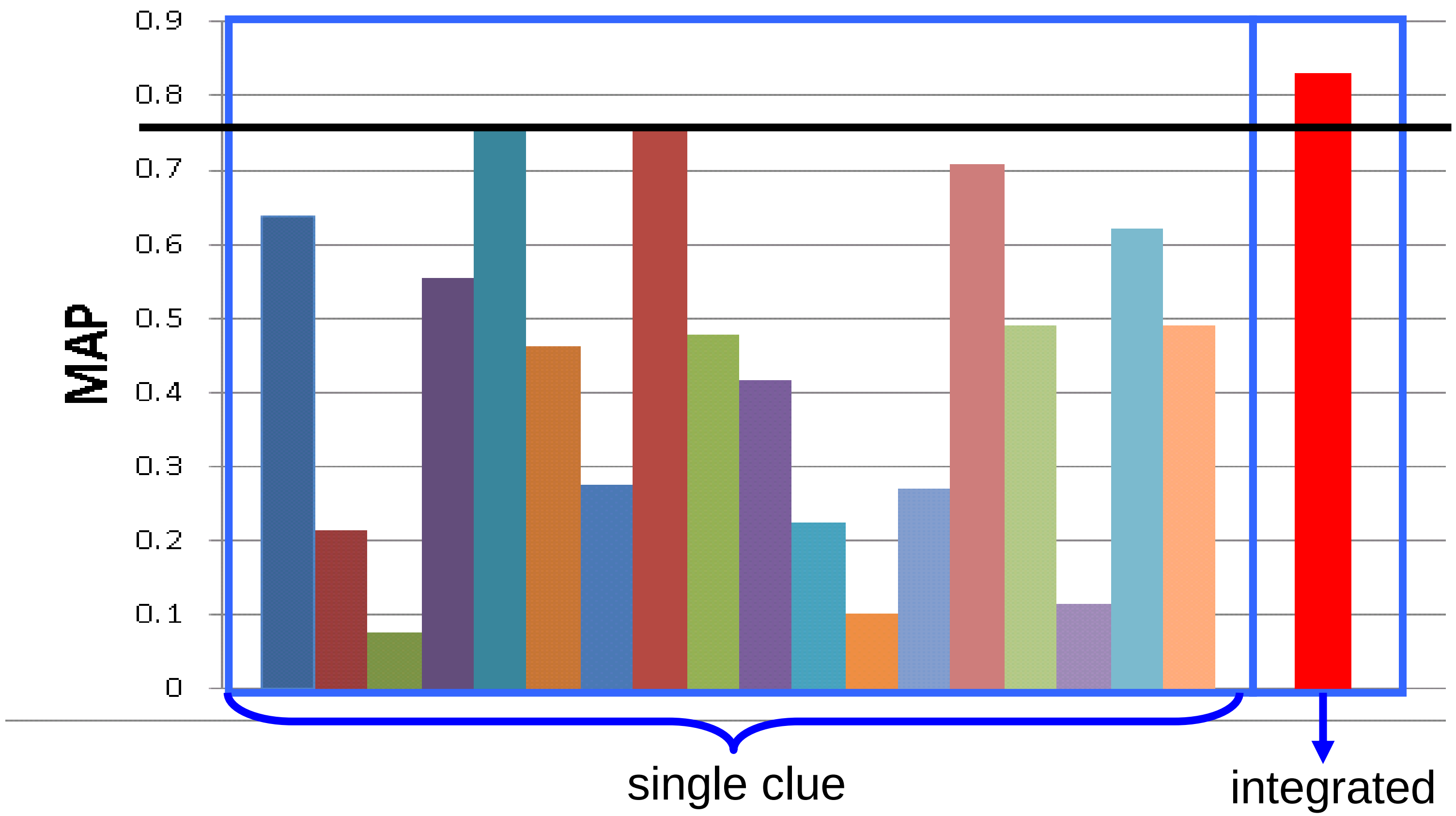



maximizing the lower bound of MAP by SVM-MAP

[Meng & Lee, ICASSP 09]

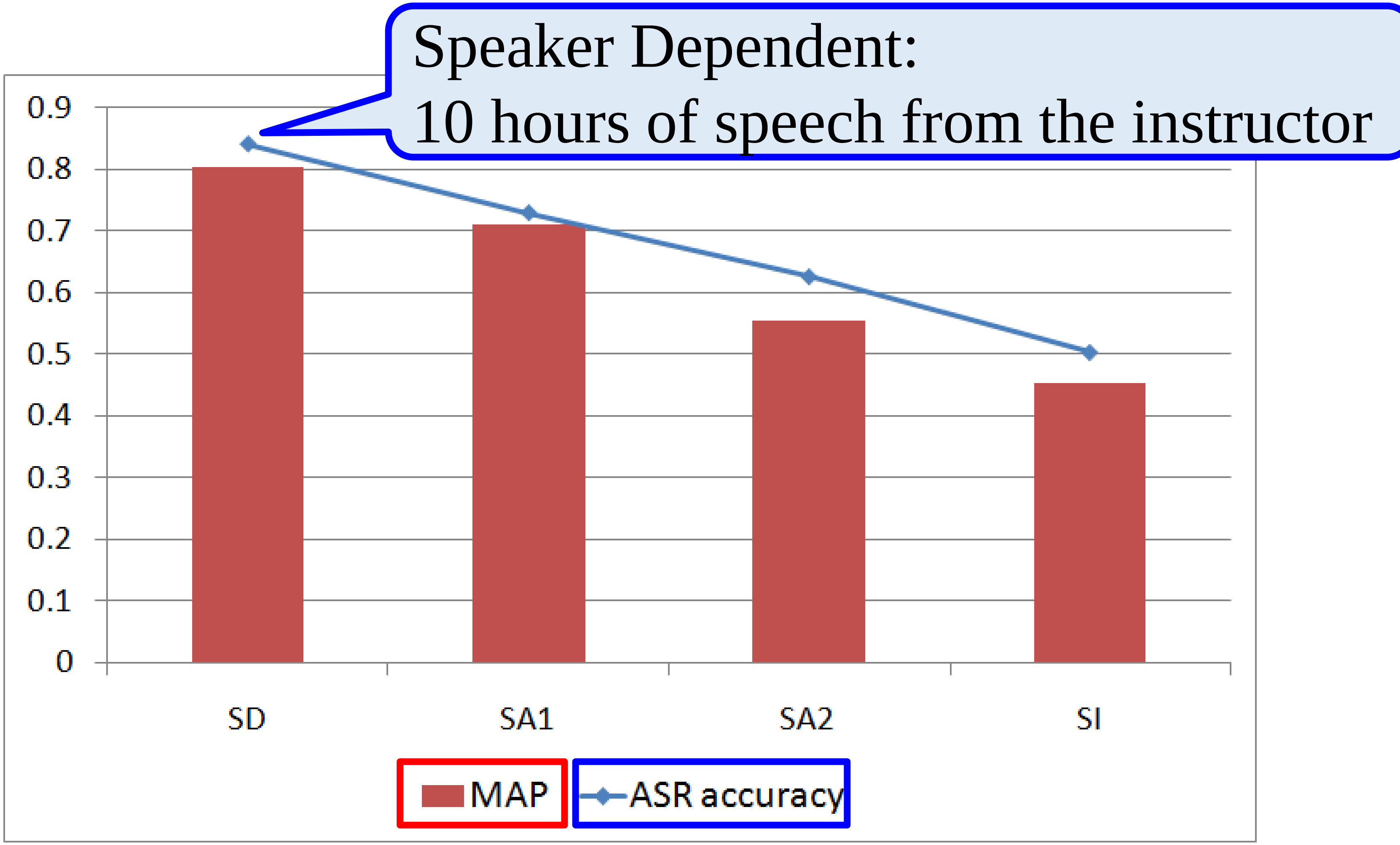
# Training Retrieval Model Parameters

- Integrating different n-grams, word/subword units and indices [Meng & Lee, ICASSP 09] [Chen & Lee, ICASSP 10]



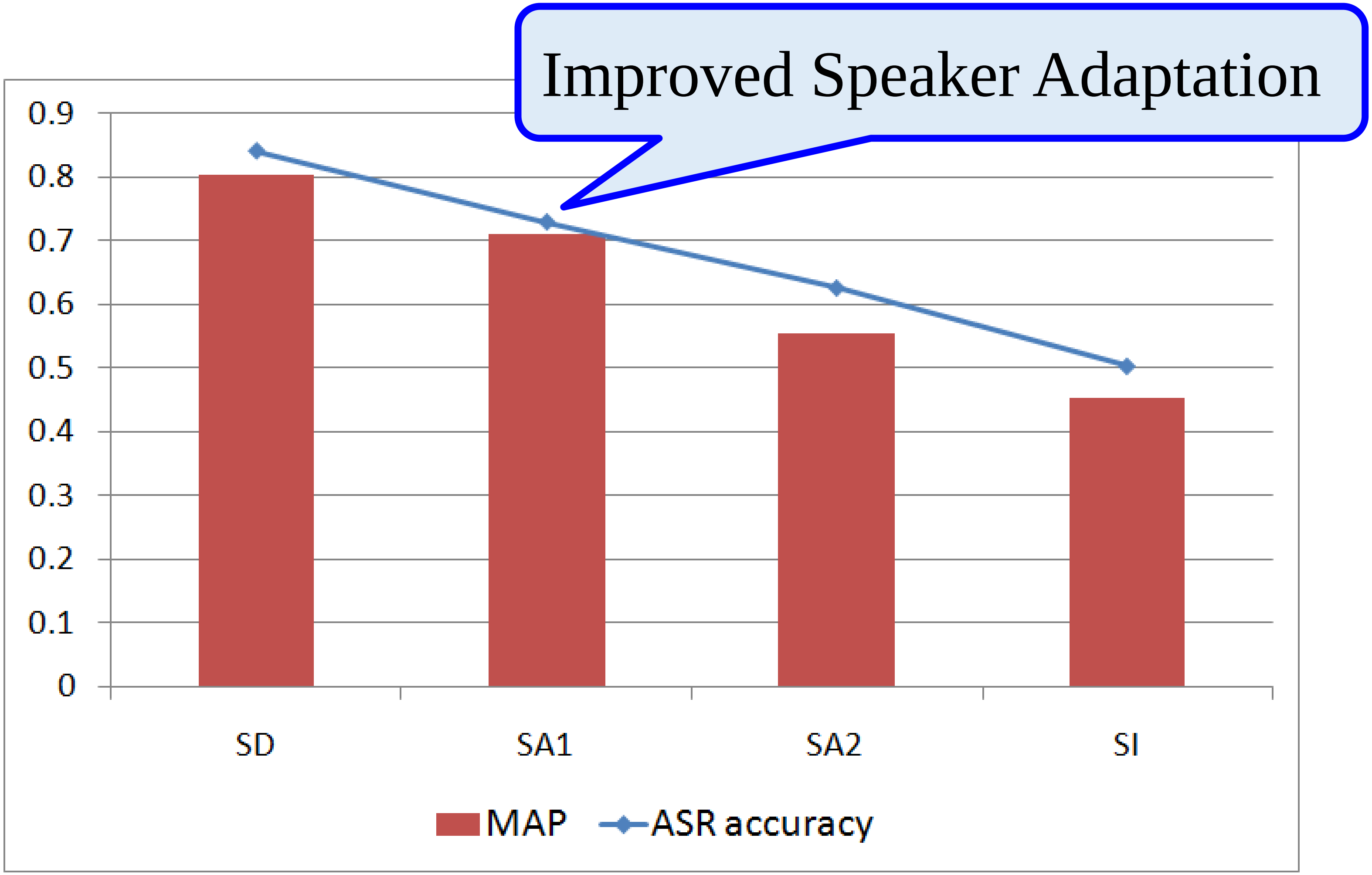
# ASR Accuracy v.s. Retrieval Performance

- Spoken Term Detection, Lectures



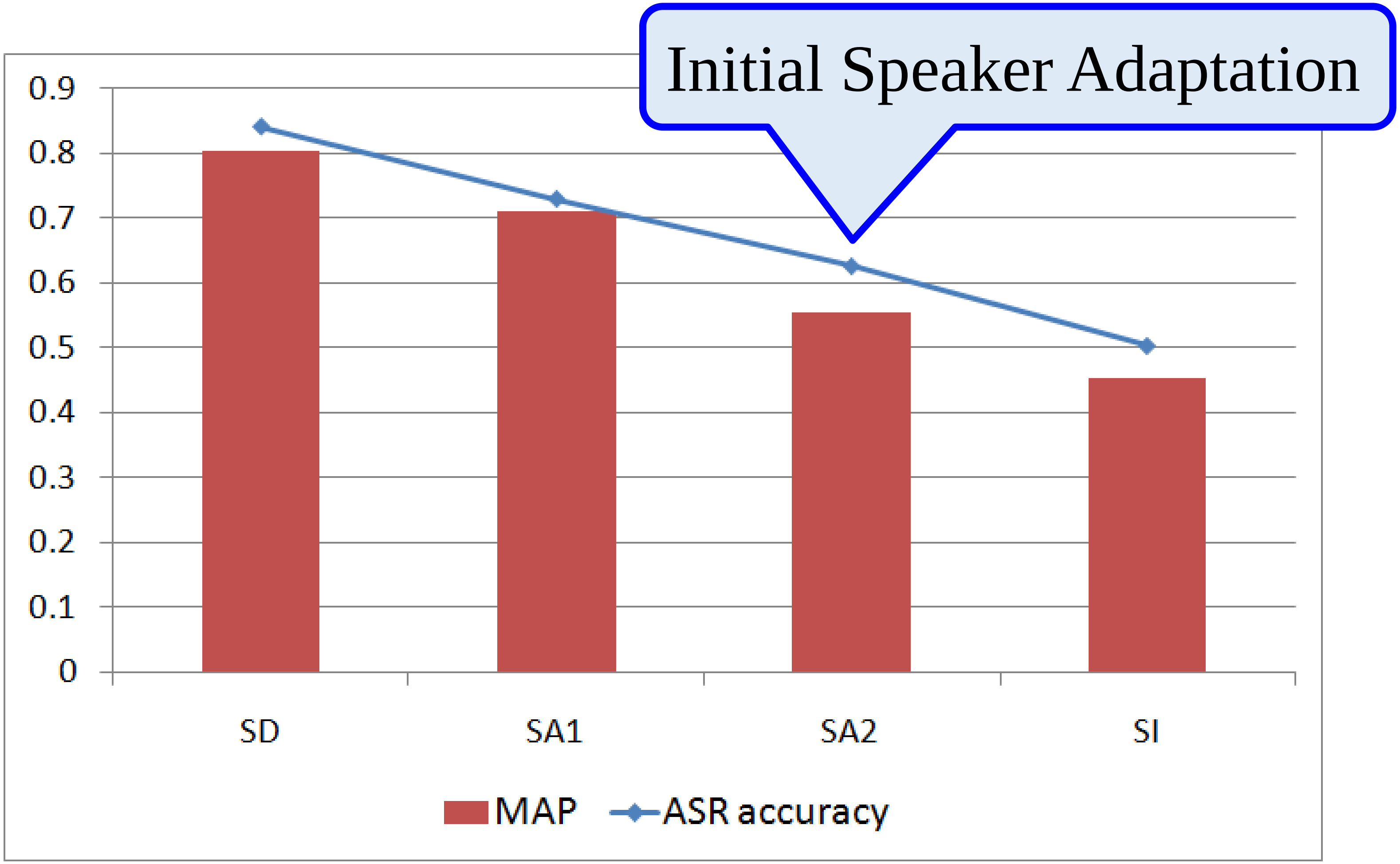
# ASR Accuracy v.s. Retrieval Performance

- Spoken Term Detection, Lectures



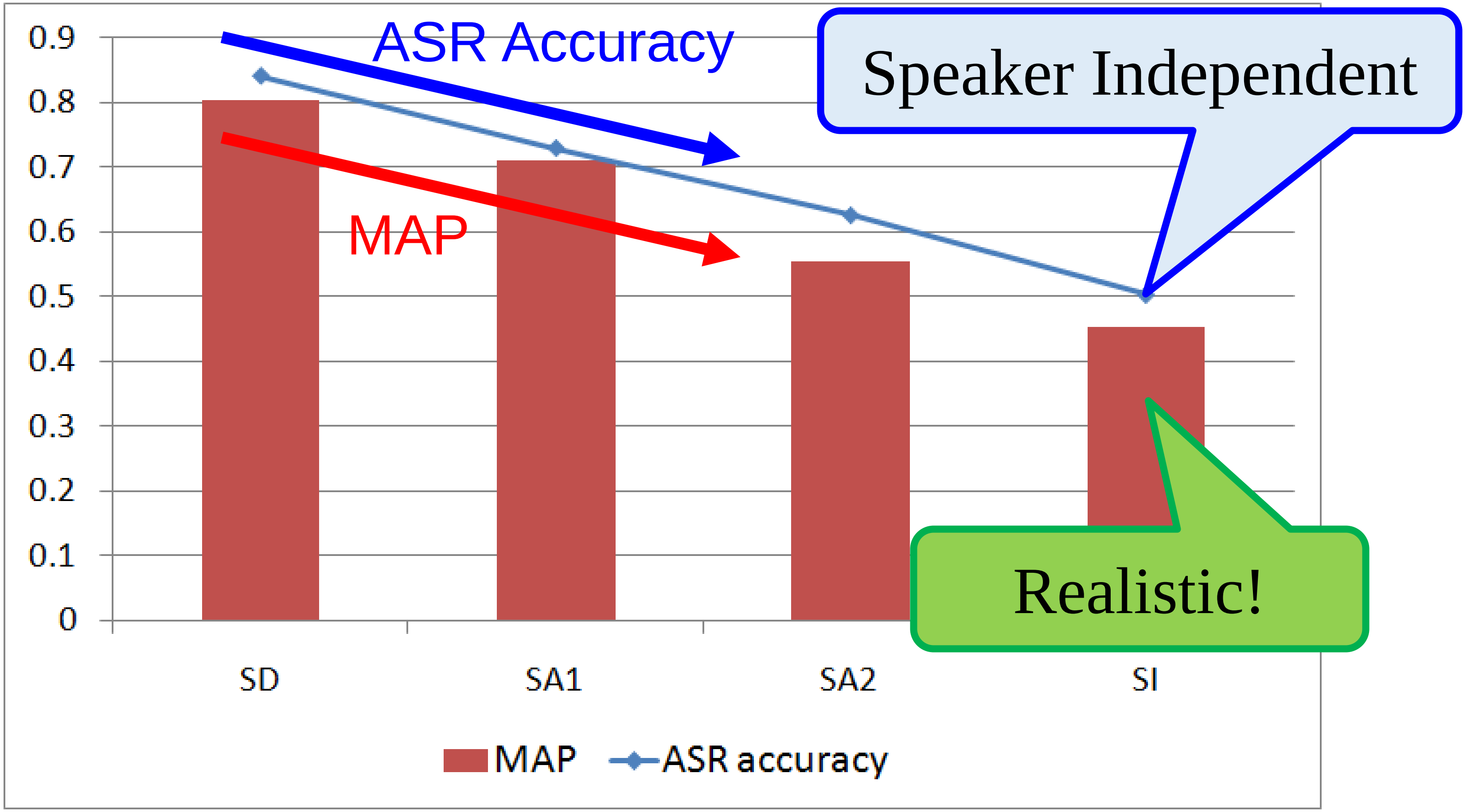
# ASR Accuracy v.s. Retrieval Performance

- Spoken Term Detection, Lectures



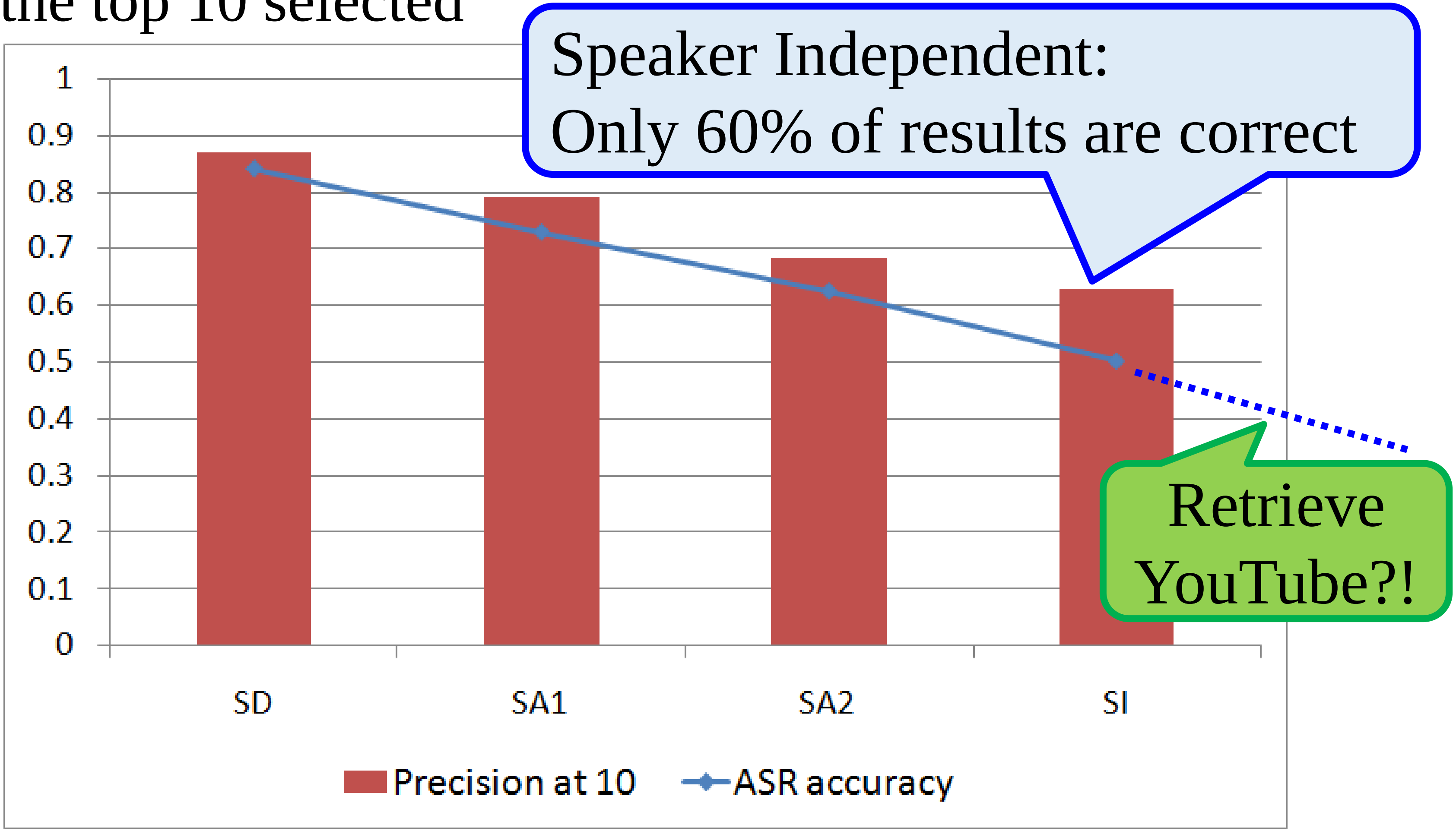
# ASR Accuracy v.s. Retrieval Performance

- Spoken Term Detection, Lectures



# ASR Accuracy v.s. Retrieval Performance

- Precision at 10: Percentage of the correct items among the top 10 selected



# Is the problem solved?

- Did lattices solve the problem?
  - Need high quality recognition models to produce better lattices and accurately estimate the confidence scores
  - Spoken content over the Internet is produced in different languages on different domains in different parts of the world under varying acoustic conditions
  - High quality recognition models for such content doesn't exist yet
- Retrieval performance limited by ASR accuracy

# Is the problem solved?

- Desired spoken content retrieval
  - Less constrained by ASR accuracy
  - Existing approaches limited by ASR accuracy because of the cascading of speech recognition and text retrieval
- Go beyond the cascading concept

# Our point in this tutorial

Spoken Content Retrieval

$\cancel{=}$

Speech Recognition

+

Text Retrieval

# Core:

# Beyond Cascading Speech Recognition and Text Retrieval

# New Directions

1. Modified ASR for Retrieval Purposes
2. Incorporating Those Information Lost in ASR
3. No Speech Recognition!
4. Special Semantic Retrieval Techniques for Spoken Content
5. Spoken Content is Difficult to Browse!

# Overview Paper

- Lin-shan Lee, James Glass, Hung-yi Lee, Chun-an Chan, "Spoken Content Retrieval —Beyond Cascading Speech Recognition with Text Retrieval," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.9, pp.1389-1420, Sept. 2015
- http://speech.ee.ntu.edu.tw/~tlkagk/paper/Overview.pdf
- This tutorial includes updated information after this paper is published.

# New Direction 1:

## Modified ASR
## for Retrieval Purposes

# Retrieval Performance v.s. Recognition Accuracy

- Intuition: Higher recognition accuracy, better retrieval performance
  - Not always true!

In Taiwan, the need of …

Recognition System A

Recognition System B

In Taiwan, a need of …

In Thailand, the need of …

Same recognition accuracy

# Retrieval Performance v.s. Recognition Accuracy

- Intuition: Higher recognition accuracy, better retrieval performance
  - Not always true!

In Taiwan, the need of …

Recognition System A

Recognition System B

In Taiwan, a need of …

Not important for retrieval

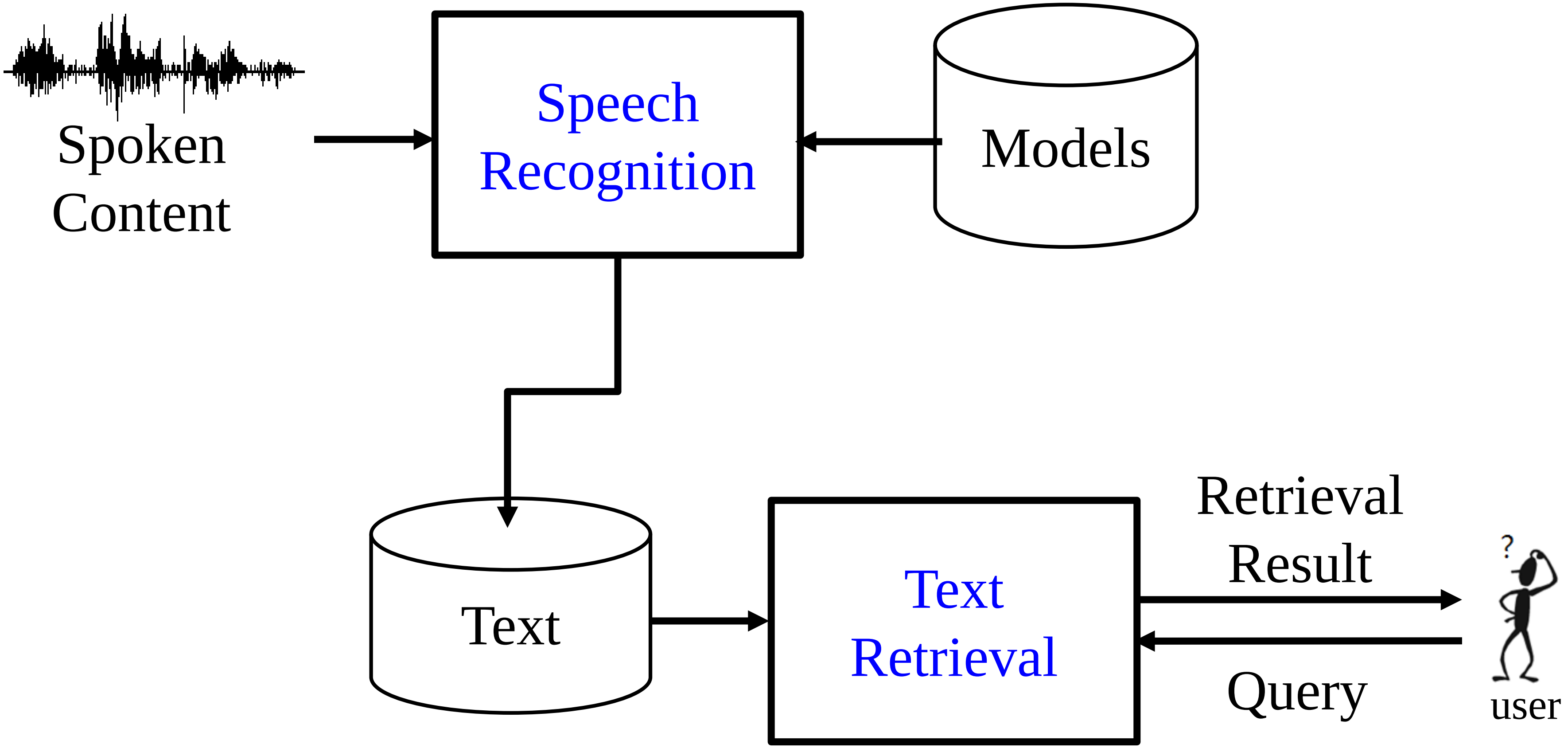
In Thailand, the need of …

Serious problem for retrieval

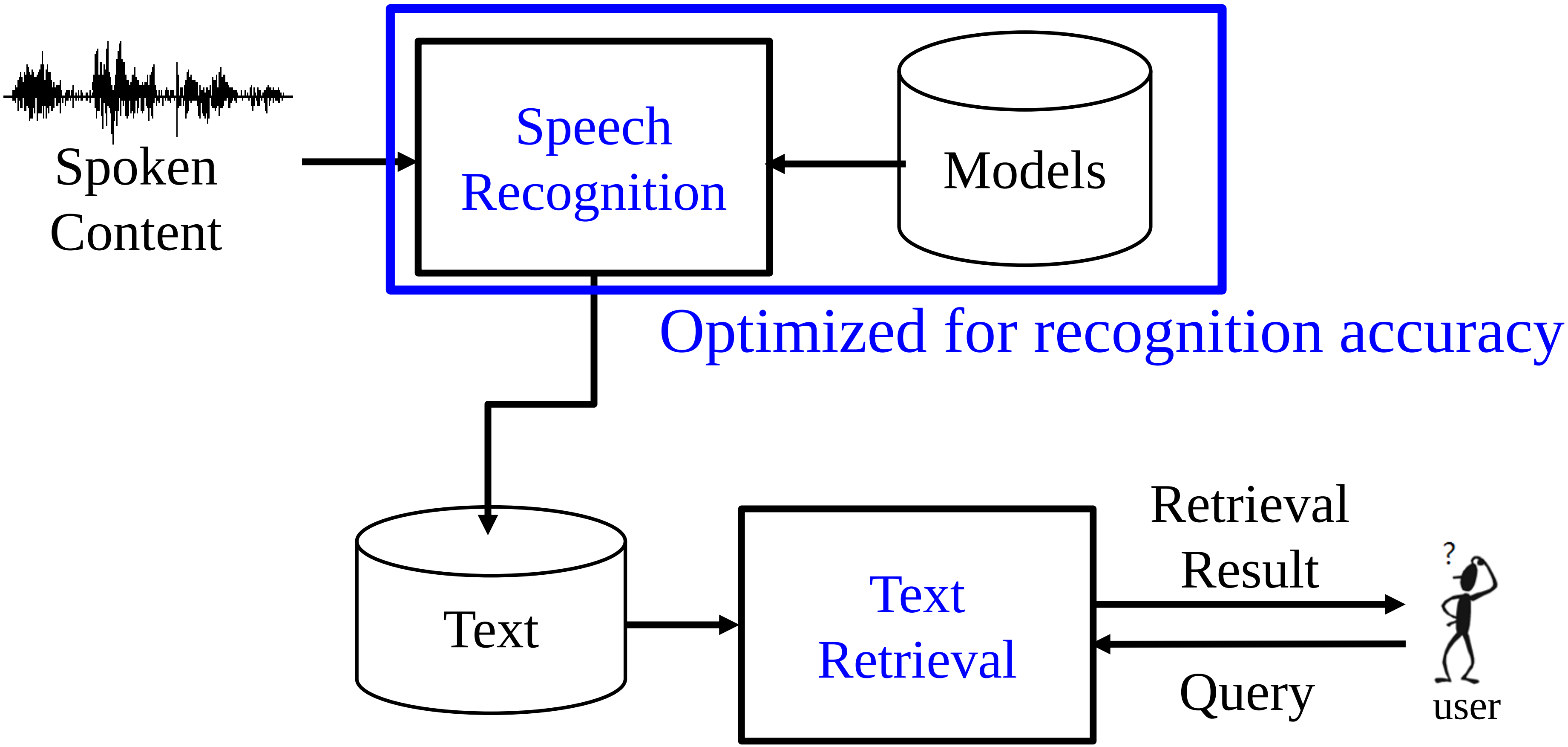
# Retrieval Performance v.s. Recognition Accuracy

- Retrieval performance is more correlated to the ASR errors of name entities than normal terms [Garofolo, TREC-7 99][L. van der Werff, SSCS 07]
- Expected error rate defined on lattices is a better predictor of retrieval performance than one-best transcriptions [Olsson, SSCS 07]
  - lattices used in retrieval
- For retrieval, substitution errors have more influence than insertions and deletions [Johnson, ICASSP 99]
- The language models reducing ASR errors do not always yield better retrieval performance [Cui, ICASSP, 13][Shao, Interspeech, 08][ Wallace, SSCS 09]
  - Query terms usually topic-specific with lower n-gram probabilities

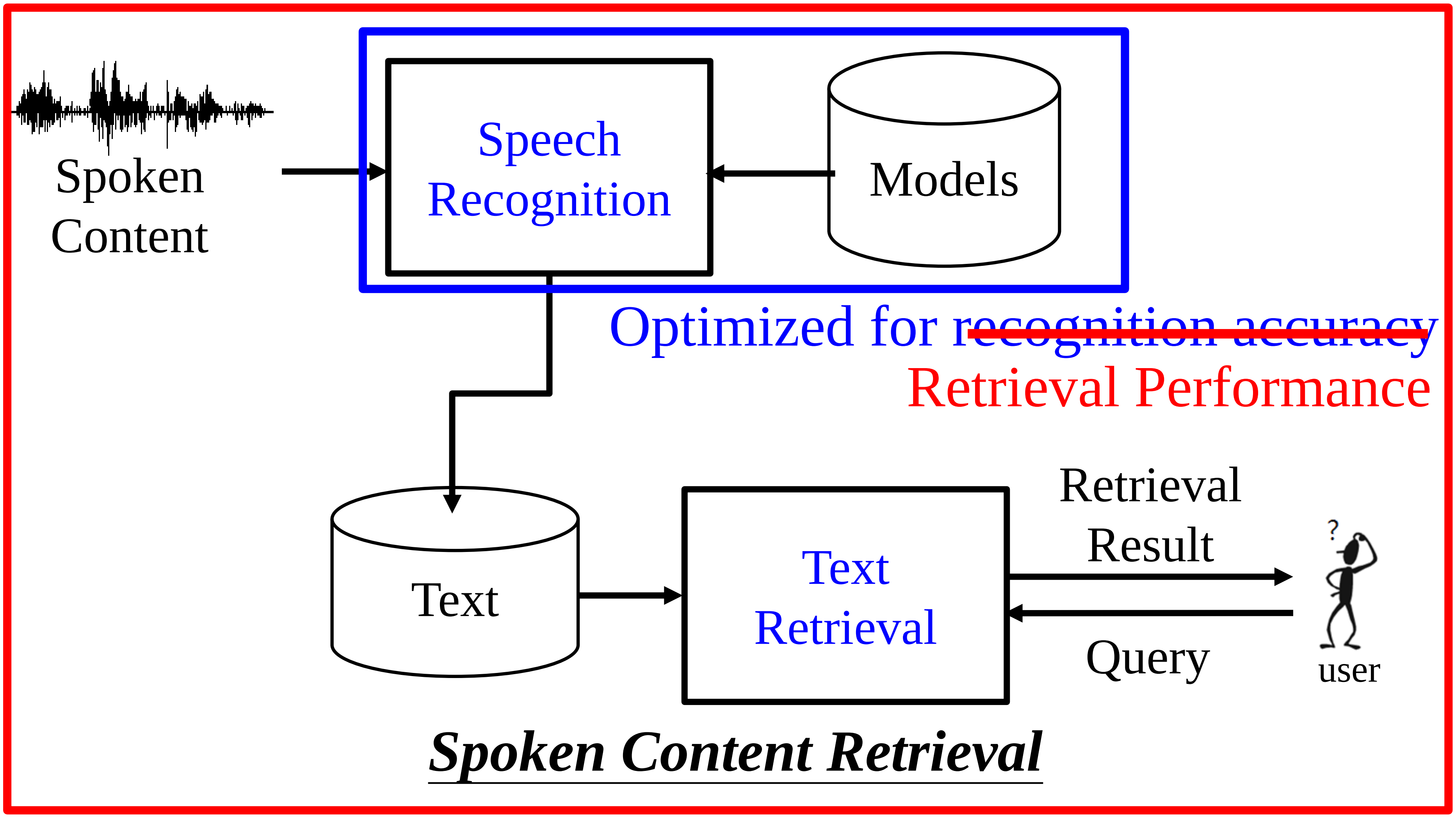
# ASR models learned by Optimizing Retrieval Performance



# ASR models learned by Optimizing Retrieval Performance



# ASR models learned by Optimizing Retrieval Performance



# New Direction 1-1:

## Modified ASR
## for Retrieval Purposes

Acoustic Modeling

# Acoustic Modeling

- Acoustic Model Training

$$\hat{\theta} = arg \max_{\theta} F(\theta)$$

$\theta$: acoustic model parameters

$F(\theta)$: objective function

The objective function $F(\theta)$ usually defined to optimize ASR accuracy

Design a new objective function for optimizing retrieval performance.

# Acoustic Modeling

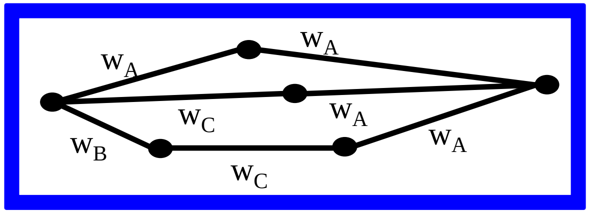
$$\hat{\theta} = arg \max_{\theta} F(\theta)$$

- Objective Function for optimizing ASR performance

$$F(\theta) = \boxed{\sum_{u}} \sum_{s_u \in \boxed{L(u)}} A(r_u, s_u) P_\theta(s_u|u)$$

Summation over all the utterances u in the training data

➢L(u): all the word sequence in the lattice of x




lattice of utterance u

# Acoustic Modeling

$$\hat{\theta} = arg \max_{\theta} F(\theta)$$

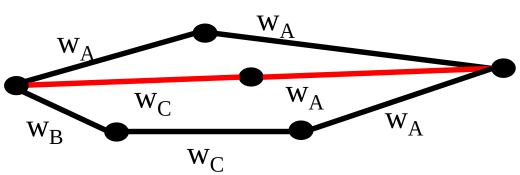
- Objective Function for optimizing ASR performance

$$F(\theta) = \sum_{u} \sum_{\boxed{s_u} \in L(u)} \boxed{A(r_u, s_u)} \boxed{P_\theta(s_u|u)}$$

- $s_u$: a word sequence in the lattice of x
- $P_\theta(s_u|u)$: posterior probability of word sequence $s_u$ given acoustic model $\theta$
- $A(r_u, s_u)$: the accuracy of word or phoneme sequence $s_u$ comparing with reference $r_u$

$\theta$ can be HMM or DNN

MCE, MPE, sMBR




lattice of utterance u

# Acoustic Modeling

$$\hat{\theta} = arg\max_{\theta} F(\theta)$$

- Objective Function for optimizing ~~ASR~~ retrieval performance

$$F(\theta) = \sum_{u} \sum_{s_u \in L(u)} \boxed{A(r_u, s_u)} P_\theta(s_u|u)$$

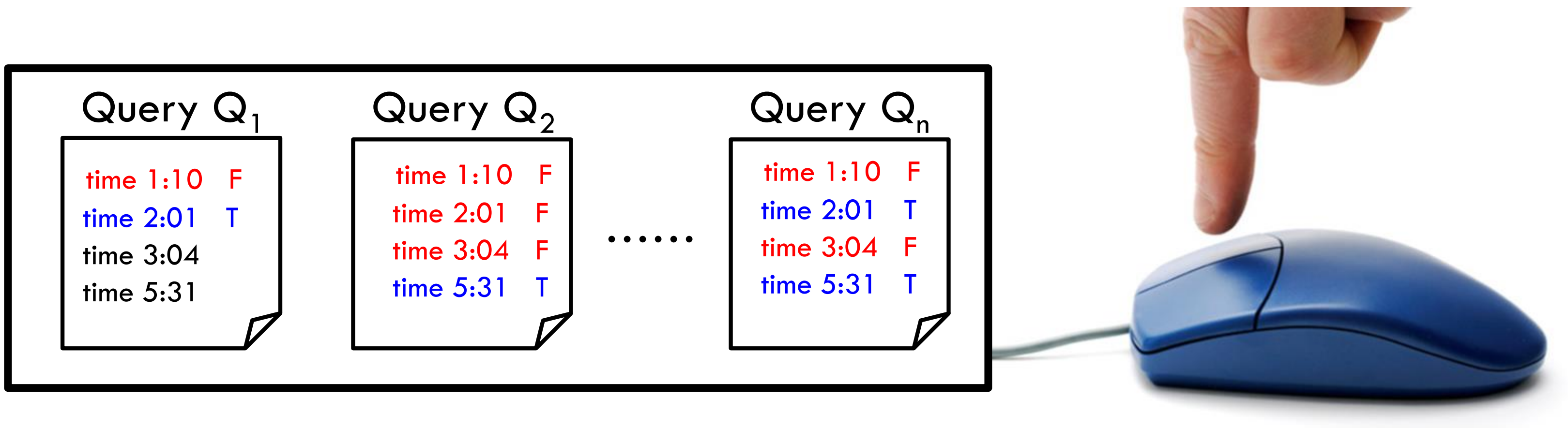
If the possible query terms are known in advance, they can be weighted higher in $A(r_u, s_u)$

W-MCE, [Fu, ASRU 07][Weng, Interspeech 12][Weng, ICASSP, 13]

keyword-boosted sMBR [Chen, Interspeech 14]

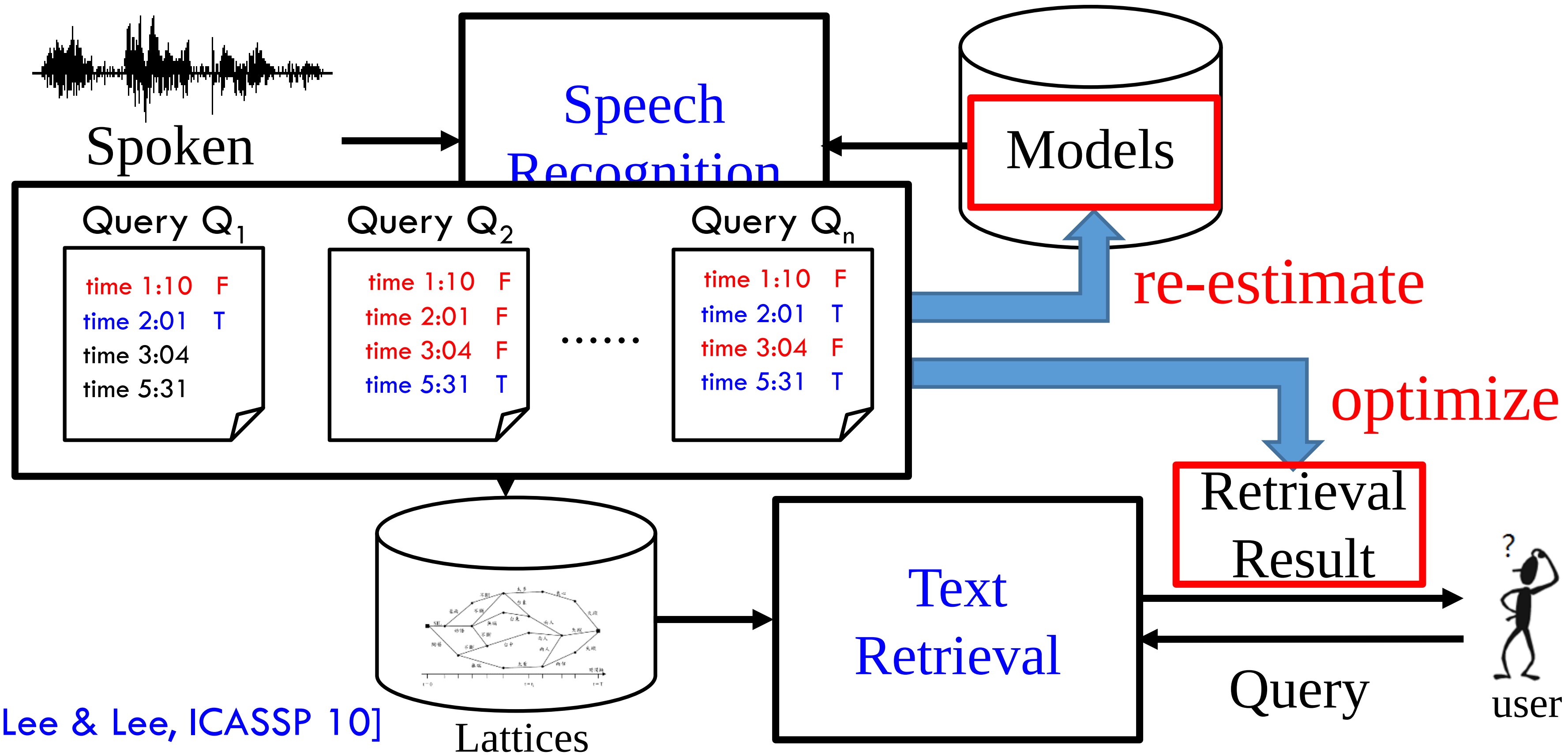
# Training Data collected from User

- In most cases, the query terms are not known in advance
- Collect feedback data on-line
  - Use the information to optimize search engines




  - Feedback can be ***implicit***

# ASR models learned by Optimizing Retrieval Performance




[Lee & Lee, ICASSP 10]
[Lee & Lee, Interspeech 10]
[Lee & Lee, SLT 10]
[Lee & Lee, IEEE T. ASL 12]

# Updated Retrieval Process

- Each retrieval result x has a confidence score R(x)
- R(x) depends on the recognition model θ
  - R(x) should be R(x;θ)

Re-estimate recognition model θ → Update the scores R(x; θ) → The retrieval results can be re-ranked.

Considering some retrieval criterion

# Basic Form

$$\hat{\theta} = \arg\max_{\theta} F(\theta)$$

- Basic Form:

$$F(\theta) = \sum_{\boxed{x_+}} \boxed{R(x_+;\theta)} - \sum_{\boxed{x_-}} \boxed{R(x_-;\theta)}$$

$x_+$ : a positive example

$x_-$ : a negative example

$R(x_+;\theta)$ : confidence score of the positive example

$R(x_-;\theta)$ : confidence score of the negative example

# Basic Form

$$\hat{\theta} = \arg\max_{\theta} F(\theta)$$

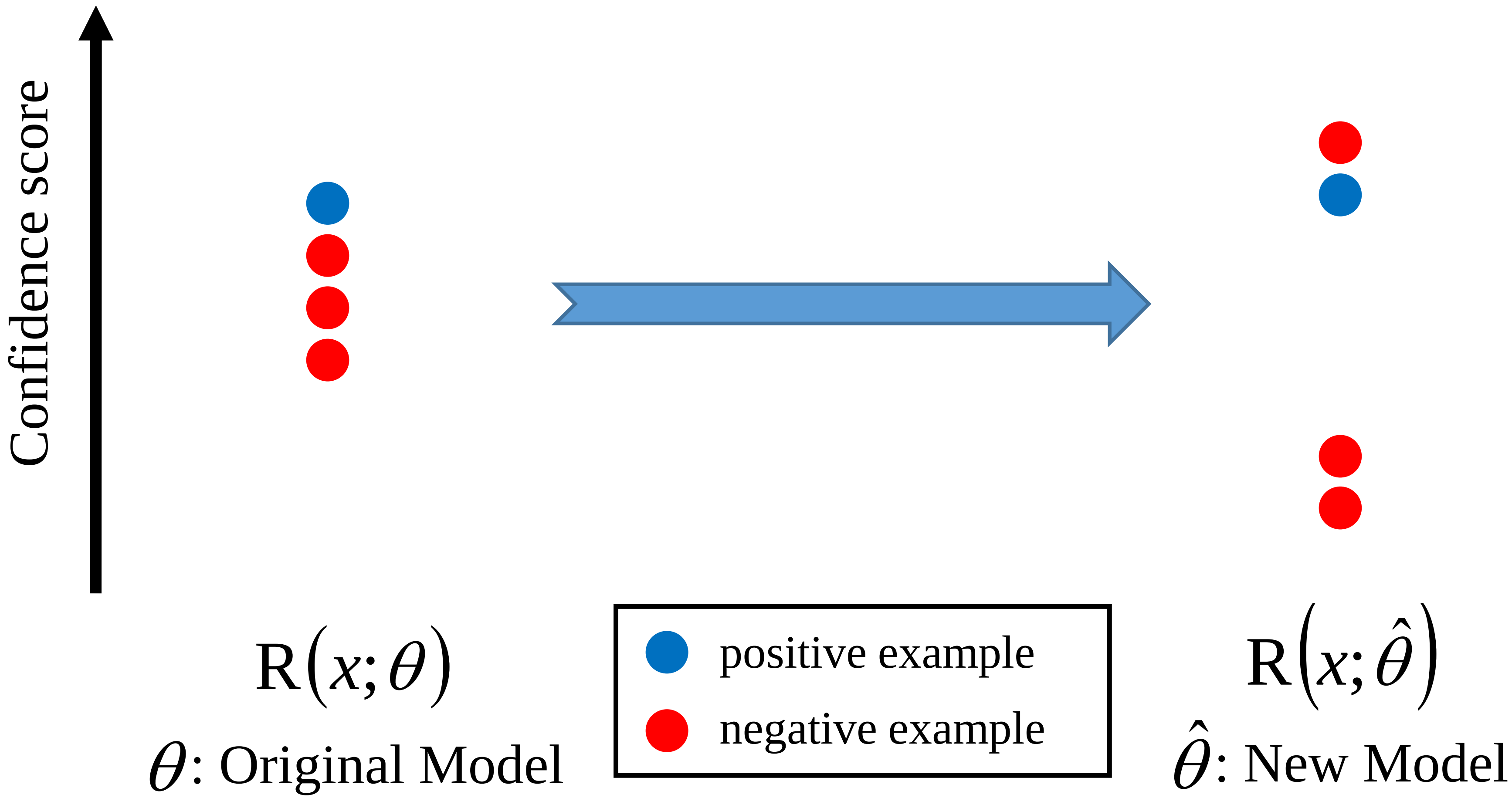
- Basic Form:

$$F(\theta) = \boxed{\sum_{x_+} R(x_+;\theta)} - \boxed{\sum_{x_-} R(x_-;\theta)}$$

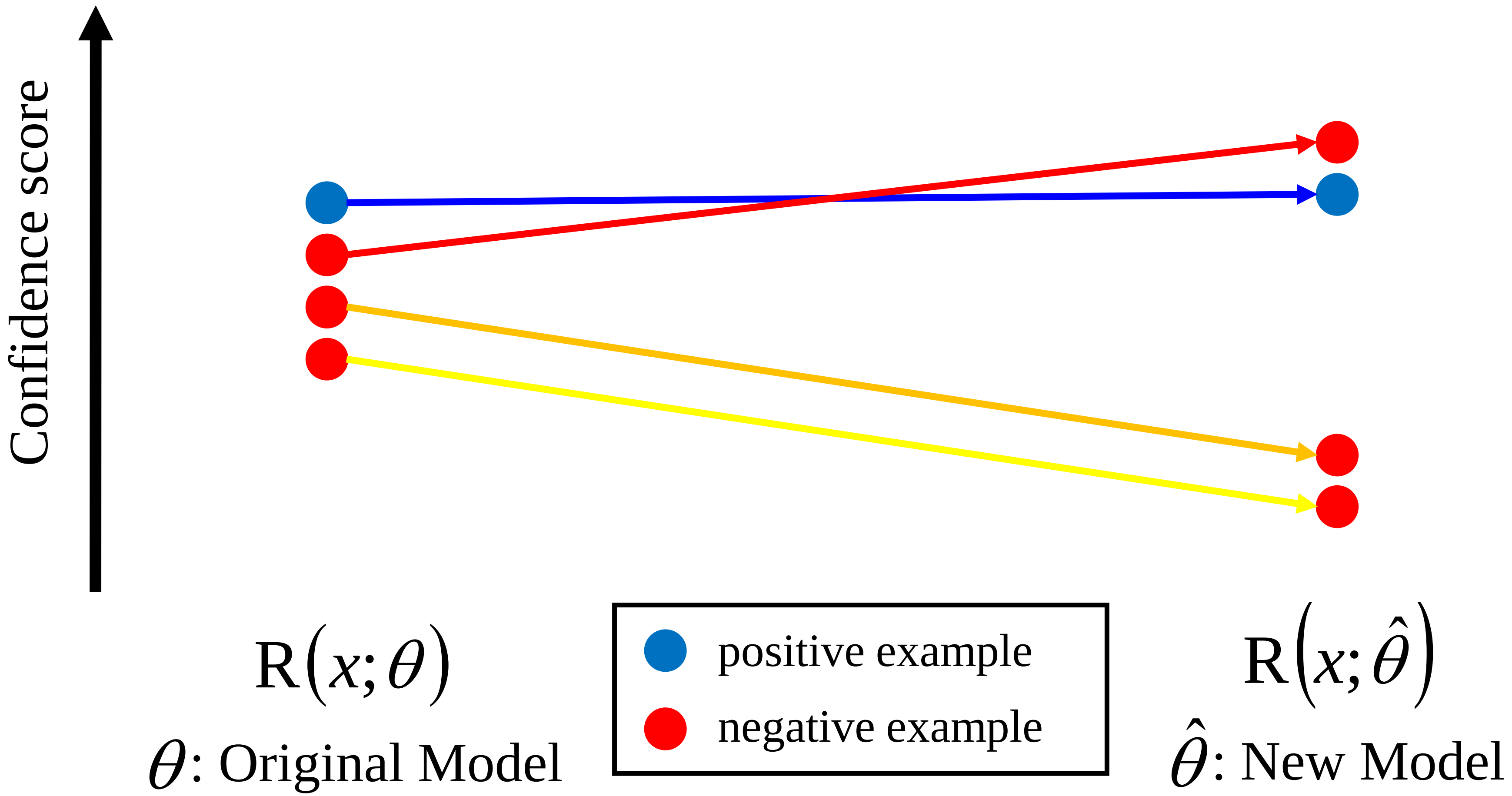
Increase the confidence scores of the positive examples

Decrease the confidence scores of the negative examples

# Consider Ranking

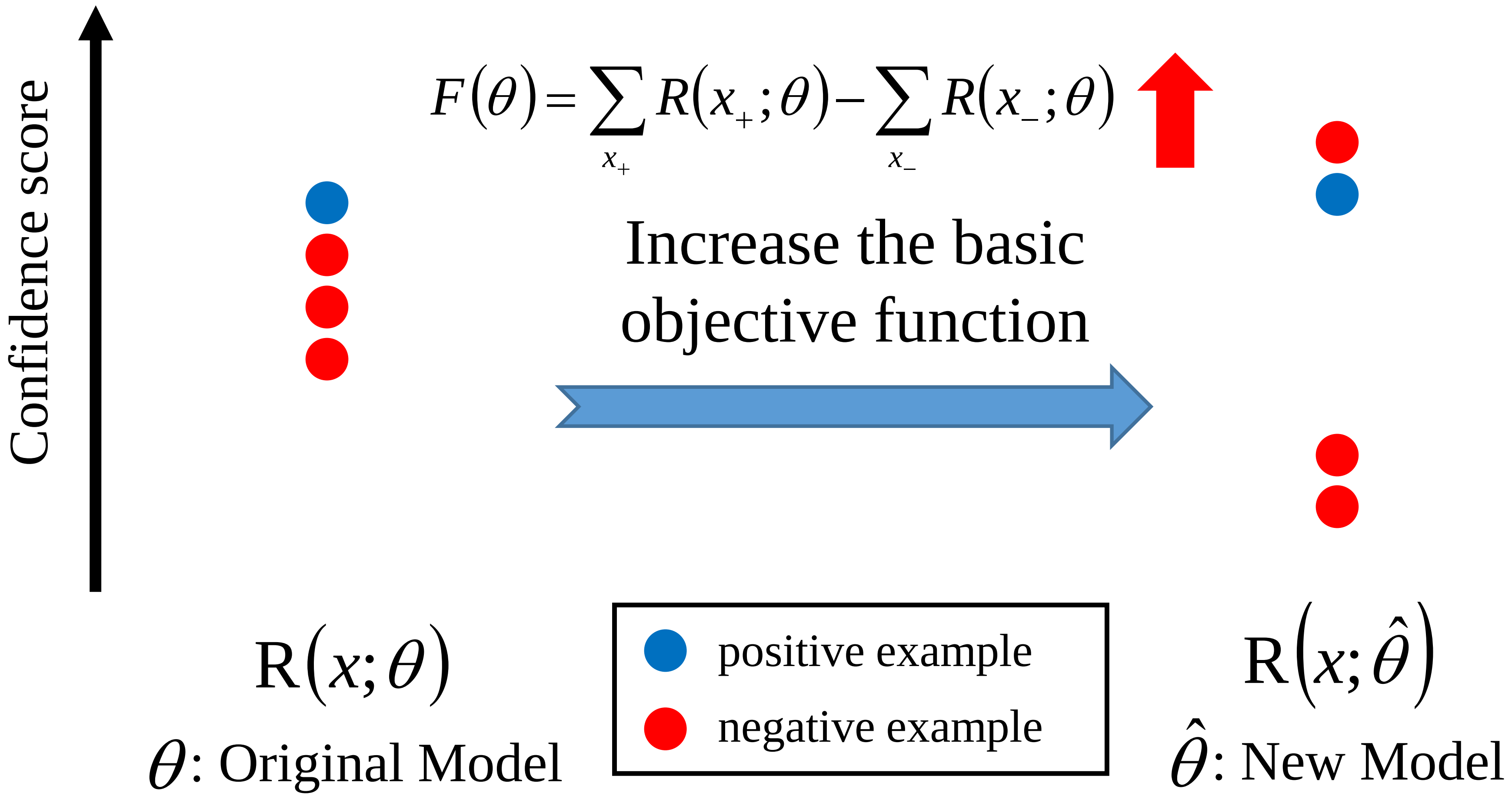


# Consider Ranking



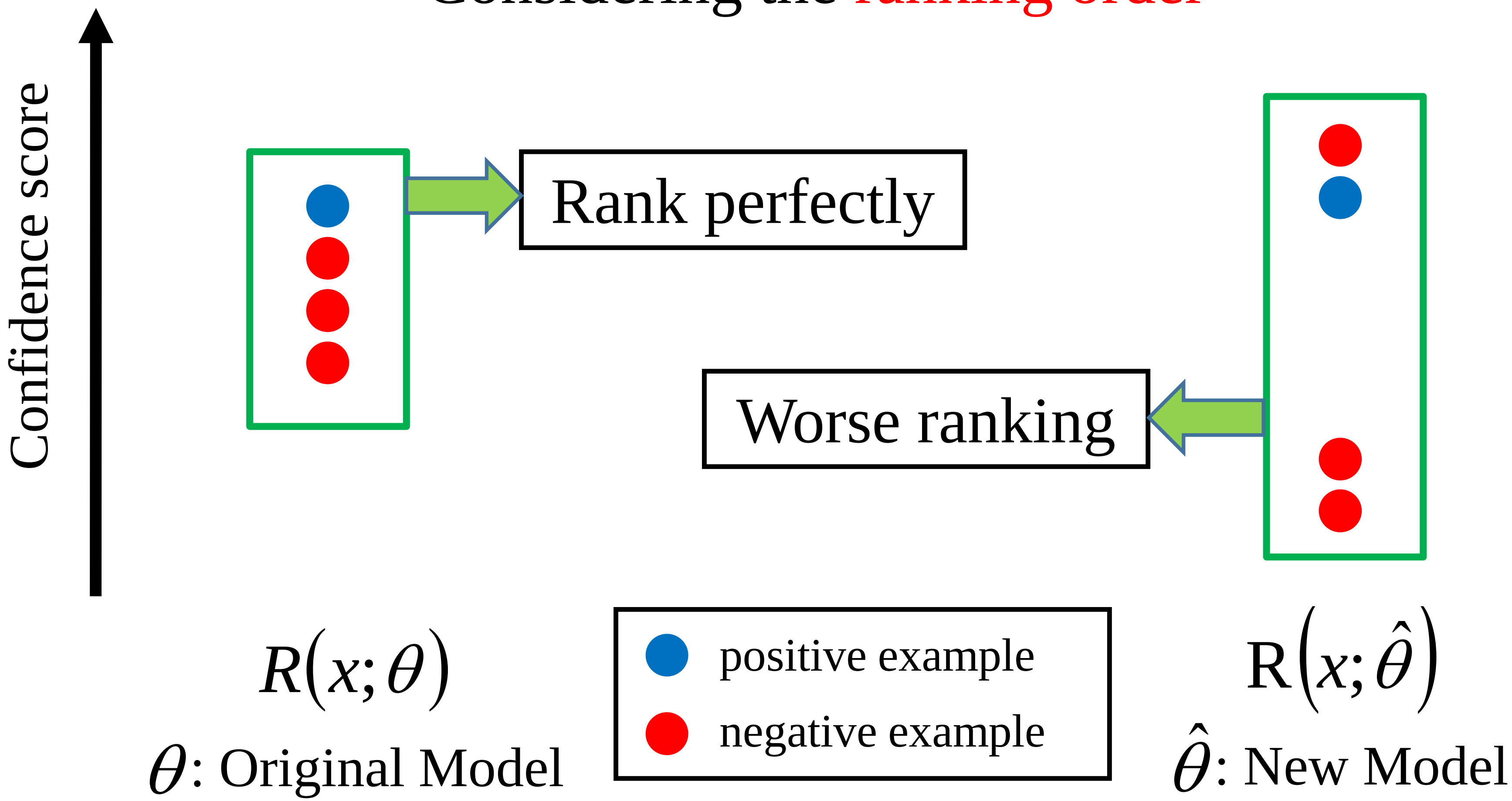
# Consider Ranking

$$F(\theta)=\sum_{x_+} R(x_+;\theta)-\sum_{x_-} R(x_-;\theta)$$



# Consider Ranking

Considering the ranking order



# Consider Ranking

$$F(\theta) = \sum_{x_+, x_-} \delta(x_+, x_-)$$

$$\delta(x_+, x_-) = \begin{cases} \boxed{1} & \boxed{\mathrm{R}(x_+;\theta) > \mathrm{R}(x_-;\theta)} \\ 0 & otherwise \end{cases}$$

- If the confidence score for a positive example exceed that for a negative example
  - the objective function adds 1.

# Consider Ranking

$$F(\theta) = \sum_{x_+, x_-} \delta(x_+, x_-)$$

$$\delta(x_+, x_-) = \begin{cases} 1 & \mathrm{R}(x_+;\theta) > \mathrm{R}(x_-;\theta) \\ 0 & otherwise \end{cases}$$

- $\delta(x_+,x_-)$ approximated by a sigmoid function during optimization.

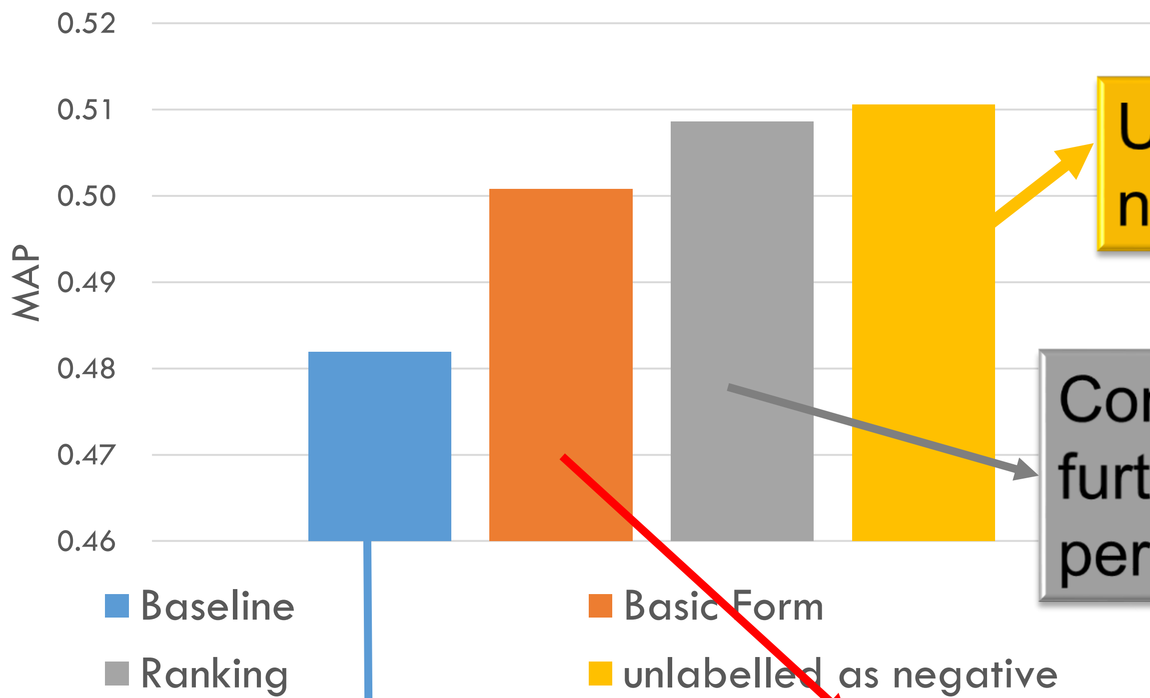
Little feedback data?

⇨ The unlabeled examples as negative examples

# Acoustic Models - Experiments

- Lecture recording (80 queries, each has 5 clicks)

[Lee & Lee, IEEE T. ASL 12]




Unlabeled data as negative examples

Considering the ranking further improved the performance

Baseline: before model re-estimation

Re-estimating acoustic models improved the performance (basic form)

# New Direction 1-2:

## Modified ASR
## for Retrieval Purposes
## Besides Acoustic Modeling

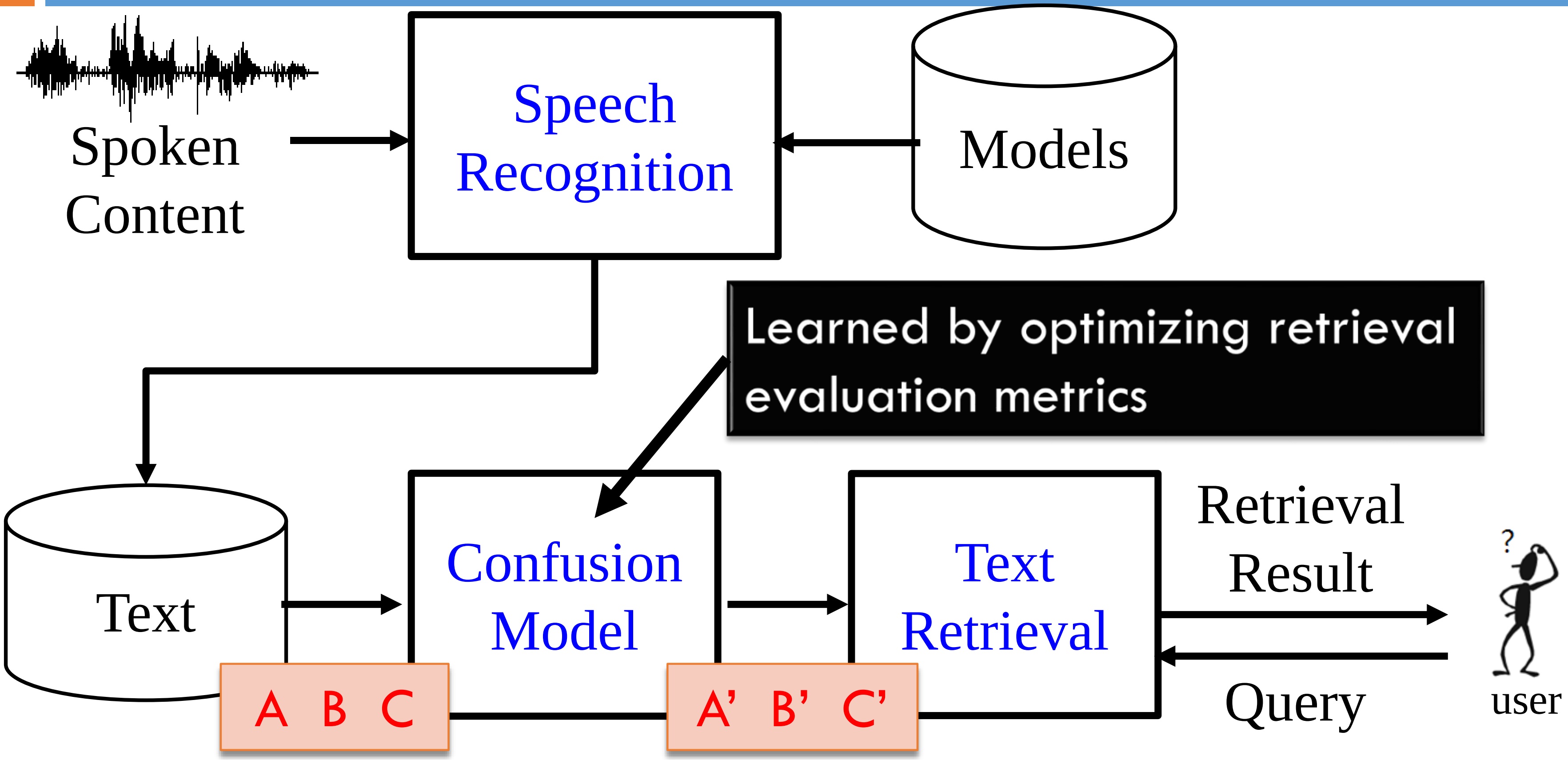
# Language Modeling

- The query terms are usually very specific. Their probabilities are underestimated.
- Boosting the probabilities of n-grams including query terms
  - By repeating the sentences including the query terms in training corpora
  - Helpful in DARPA's RATS program [Mandal, Interspeech 13] and NIST OpenKWS13 evaluation [Chen, ISCSLP 14]
- NN-based LM: Modifying training criterion, so the key terms are weighted more during training
  - Helpful in NIST OpenKWS13 evaluation [Gandhe, ICASSP 14]

# Decoding

- Give different words different pruning thresholds during decoding
  - The keywords given lower pruning thresholds than normal terms
  - Called white listing [Zhang, Interspeech 12] or keyword-aware pruning [Mandal, Interspeech 13]
- OOV words never correctly recognized
  - Two stage approach [Shao, Interspeech 08]
    - Identify the lattices probably containing OOV (by subword-based approach)
    - Insert the word arcs of OOV words into lattices and rescore

# Confusion Models

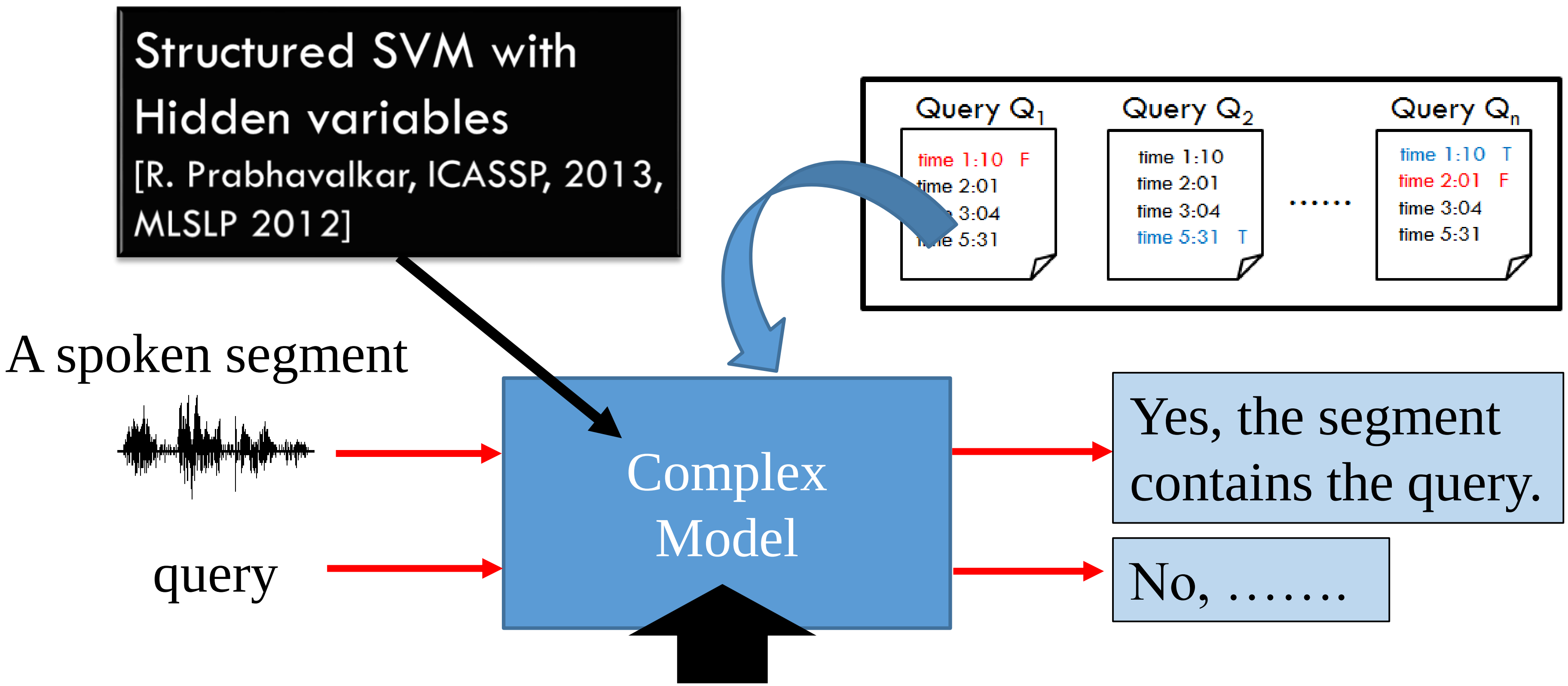



The ASR produces systematic errors, so it is possible to learn a confusion model to offer better retrieval results

[Karanasou, Interspeech 12][Wallace, ICASSP 10]

# Jointly Optimizing Speech Recognition and Retrieval Modules

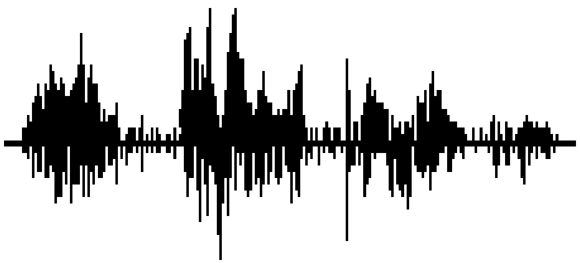
Sounds crazy?




End-to-end model performing speech recognition and retrieval jointly (learned jointly) in one step

# Much information lost during ASR

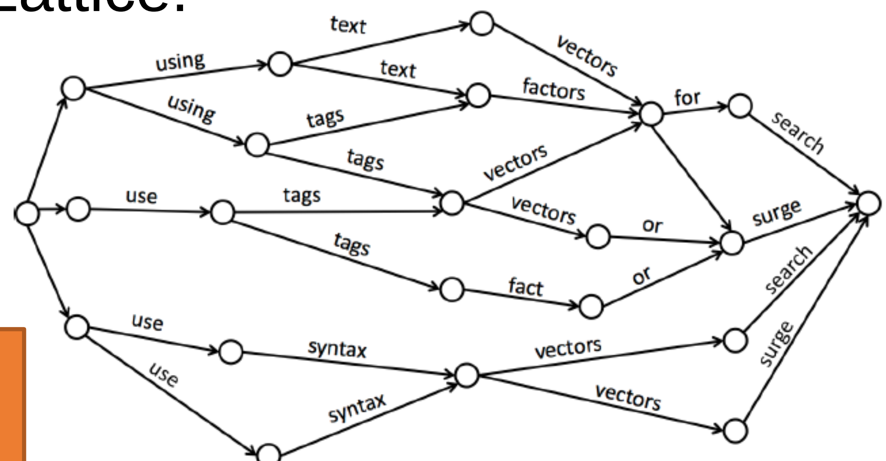
Spoken Content



ASR

Transcriptions:
using syntax vectors surge ……

Lattice:

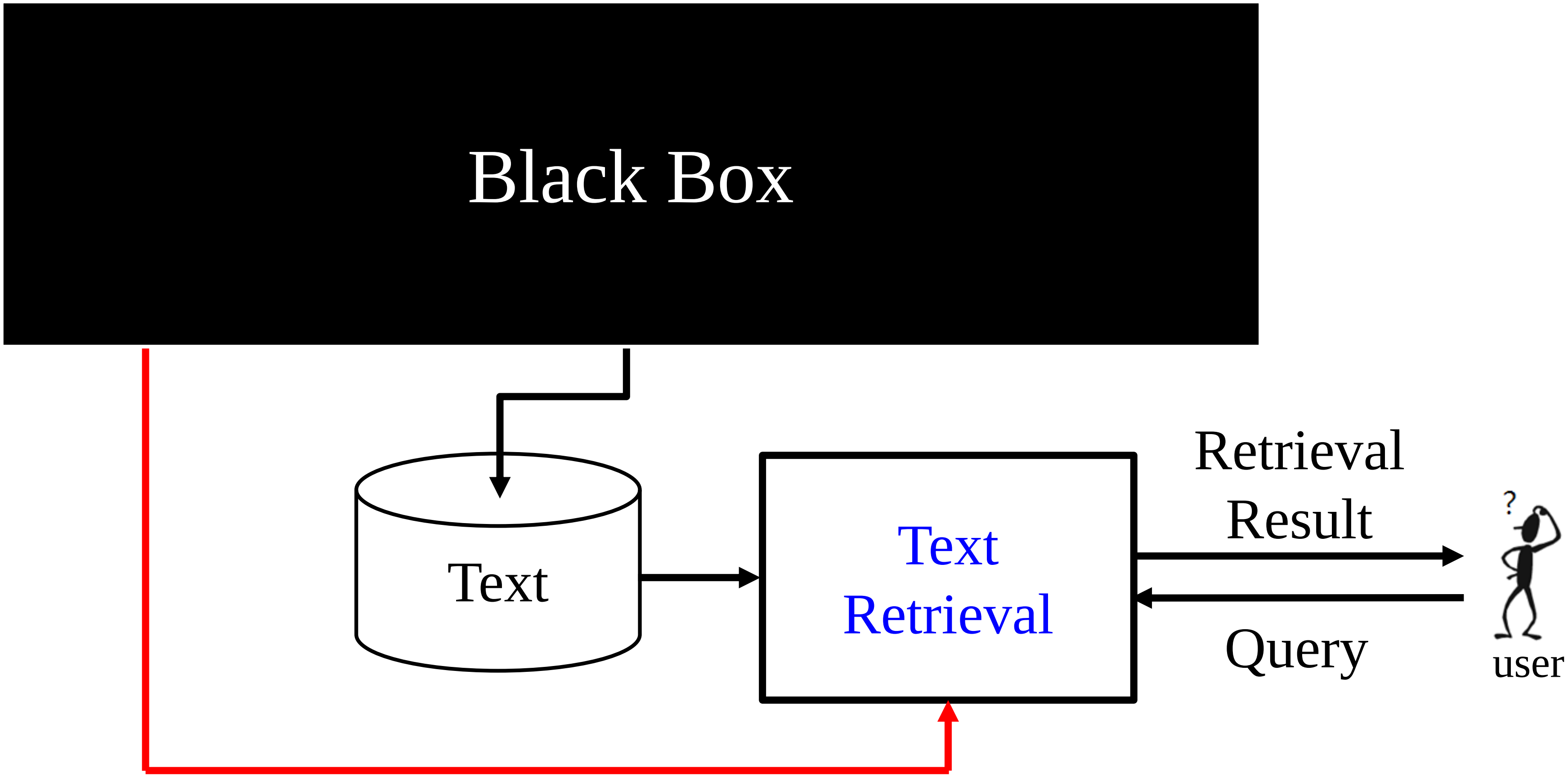



Much information lost during ASR

# New Direction 2:

## Incorporating
## Those Information Lost in ASR

# Information beyond Speech Recognition Output




Incorporating information lost in
ASR to help retrieval

# New Direction 2-1:

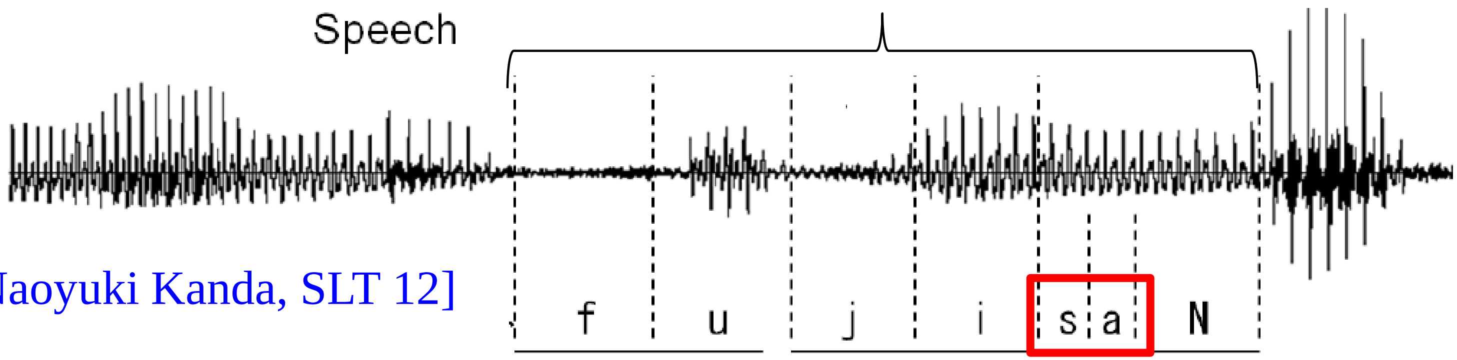
## Incorporating
## Those Information Lost in ASR

What kind of information can be helpful?

# Information beyond Speech Recognition Output

- Phoneme or syllable duration [Wollmer, ICASSP 09][Naoyuki Kanda, SLT 12][Teppei Ohno, SLT 12]

Query is Japanese word "fu-ji-sa-N"




[Naoyuki Kanda, SLT 12]

very short! False alarm!

- Pitch & Energy [Tejedor, Interspeech 10]
- Landmark and attribute detection with prosodic cues includes can reduce the false alarm [Ma, Interspeech 2007]

# Query-specific Information

- **"Jack of all trades, master of none“**

Speech Recognition

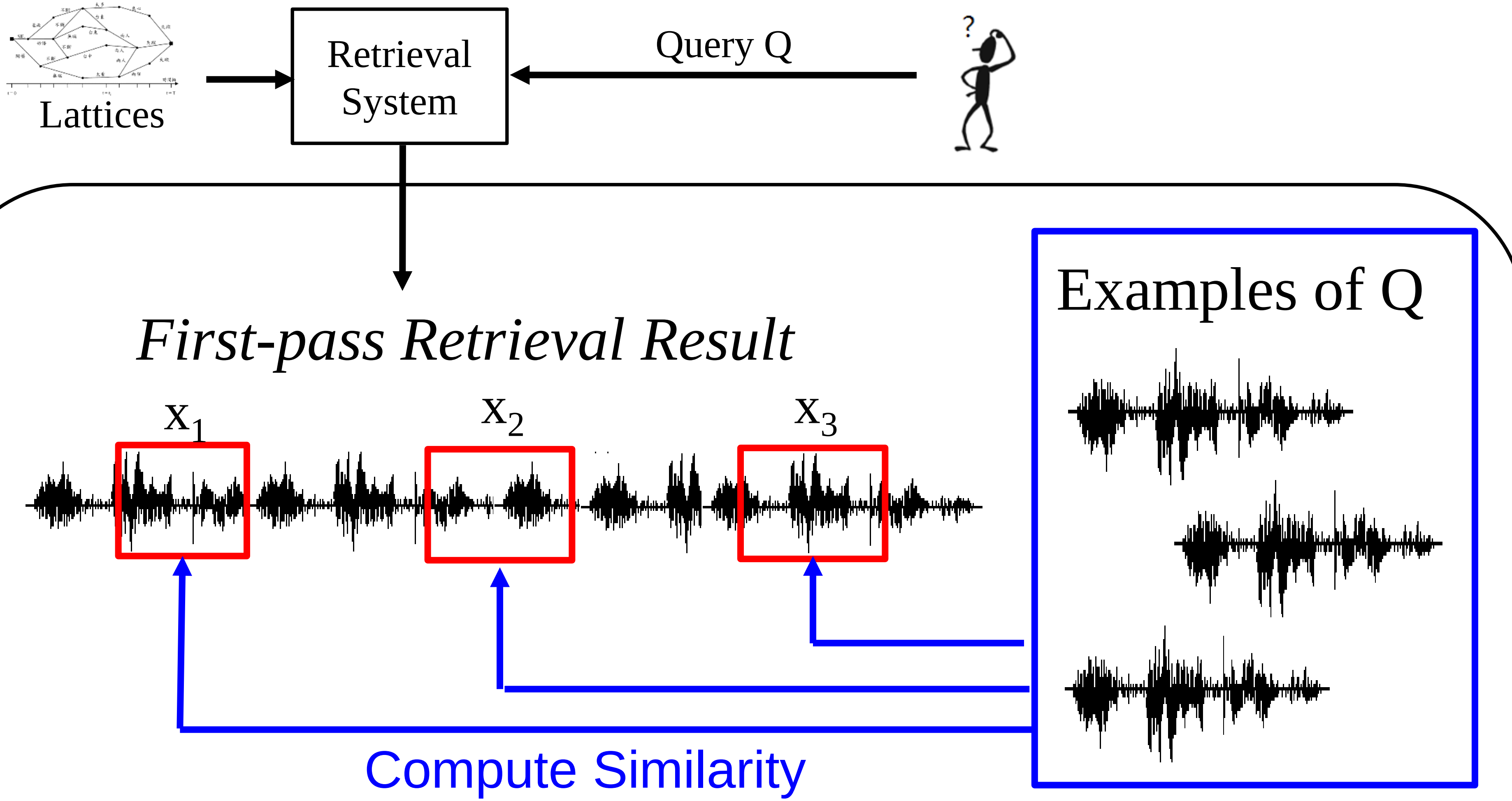

Spoken Term Detection

Query-specific detector

Correctly recognized all the words

higher detector accuracy on specific query

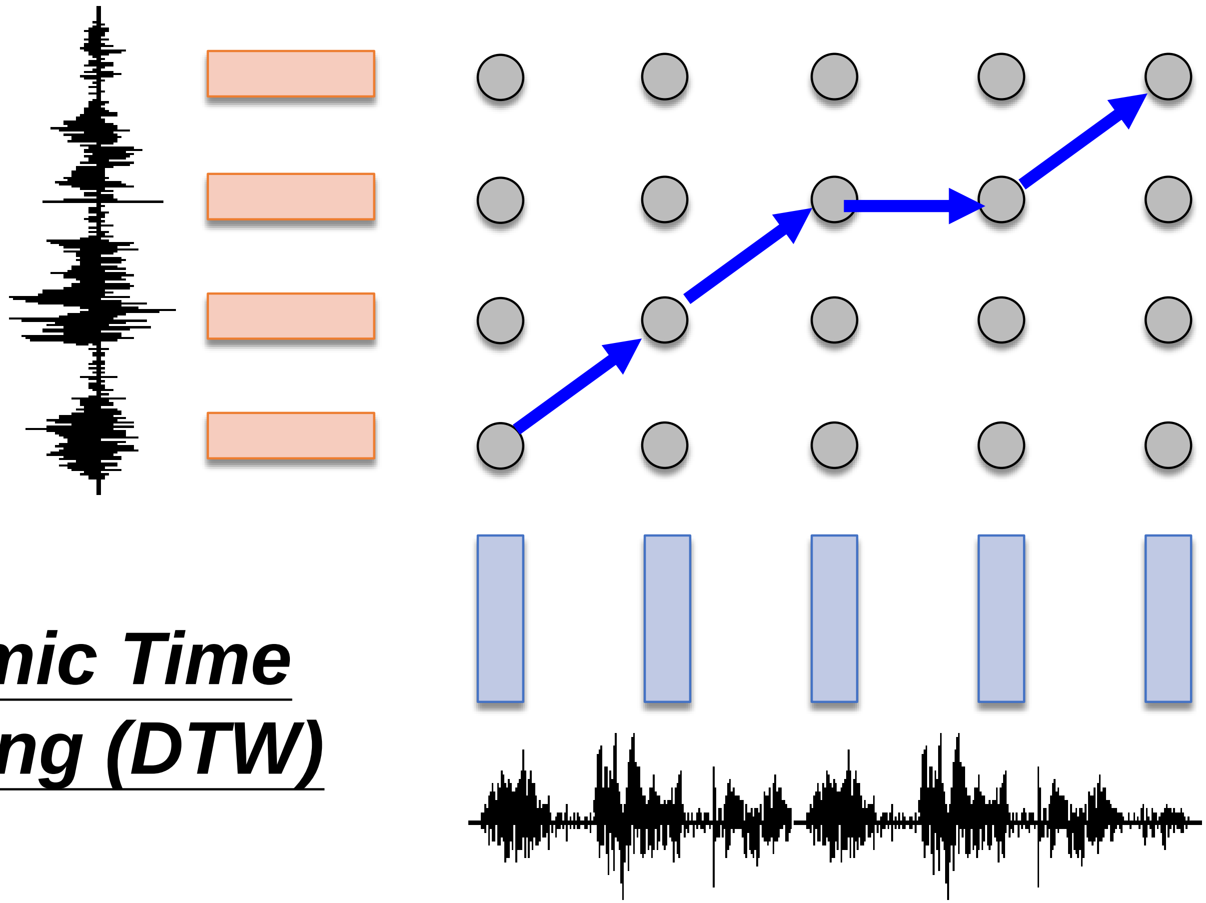
# Query-specific Detector




Exemplar-based approach also used in speech recognition

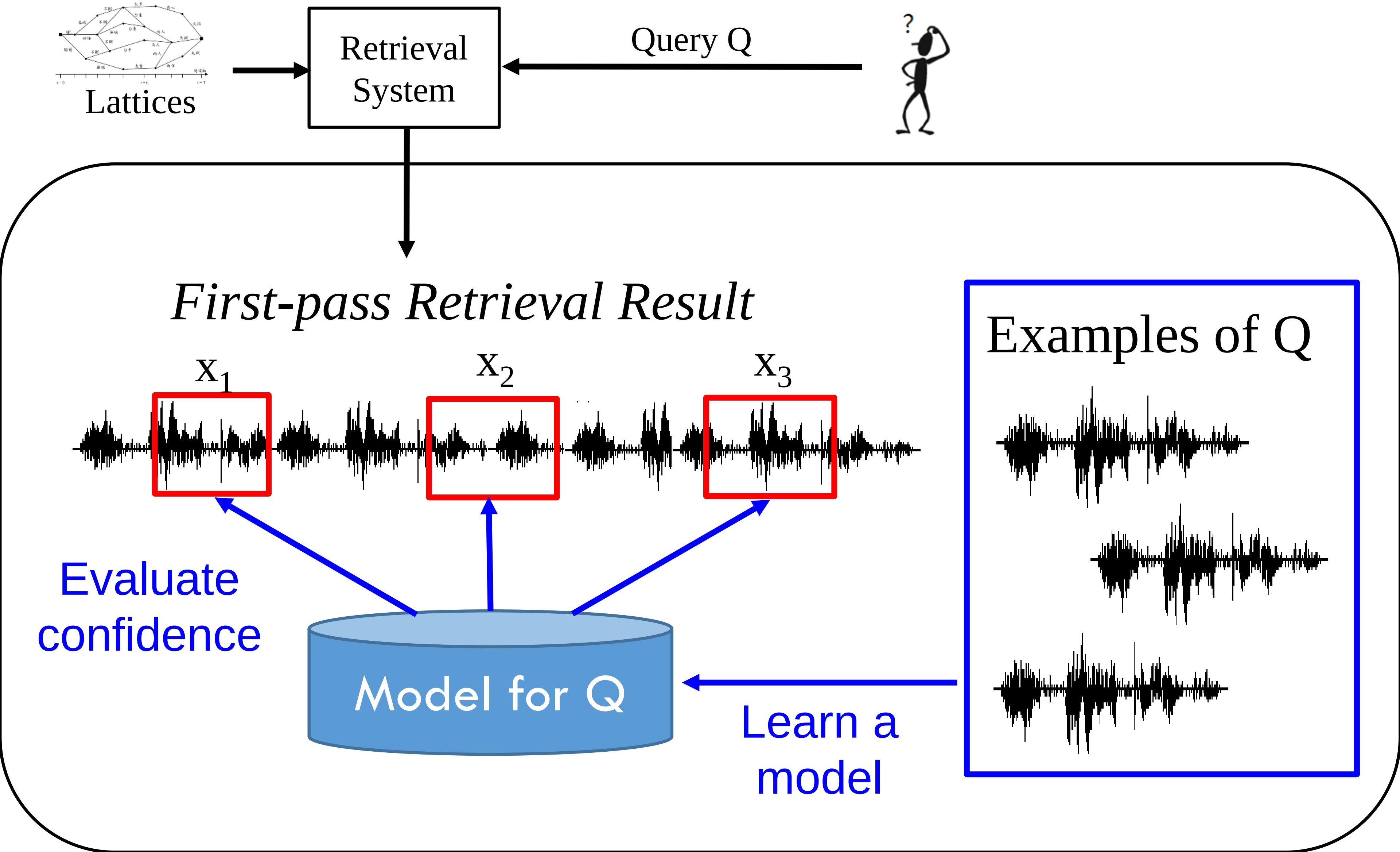
[Demuynck, ICASSP 2011][Heigold, ICASSP 2012][Nancy Chen, ICASSP 2016]

# Similarities between Audio Segments

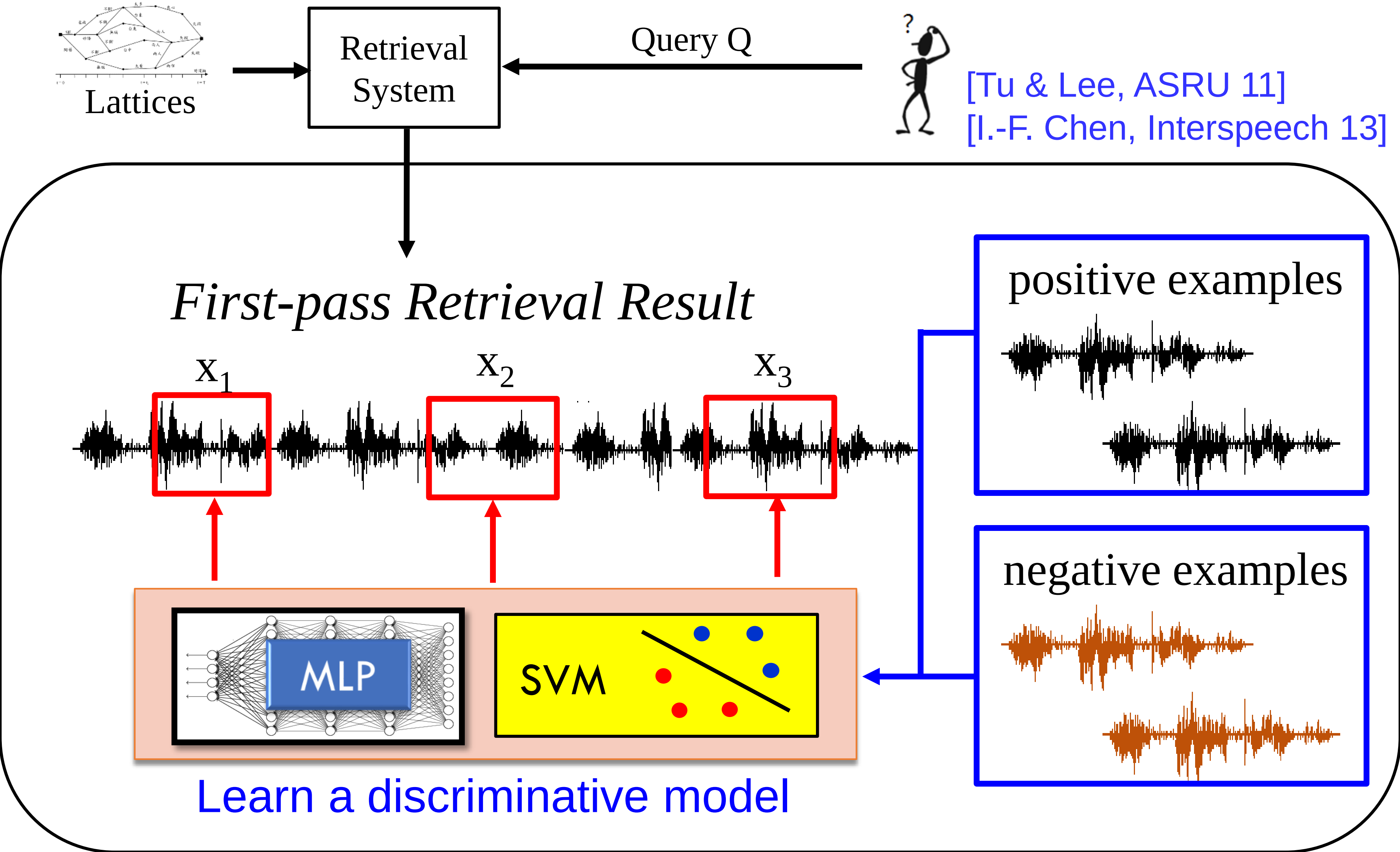


***Dynamic Time Warping (DTW)***

# Query-specific Detector



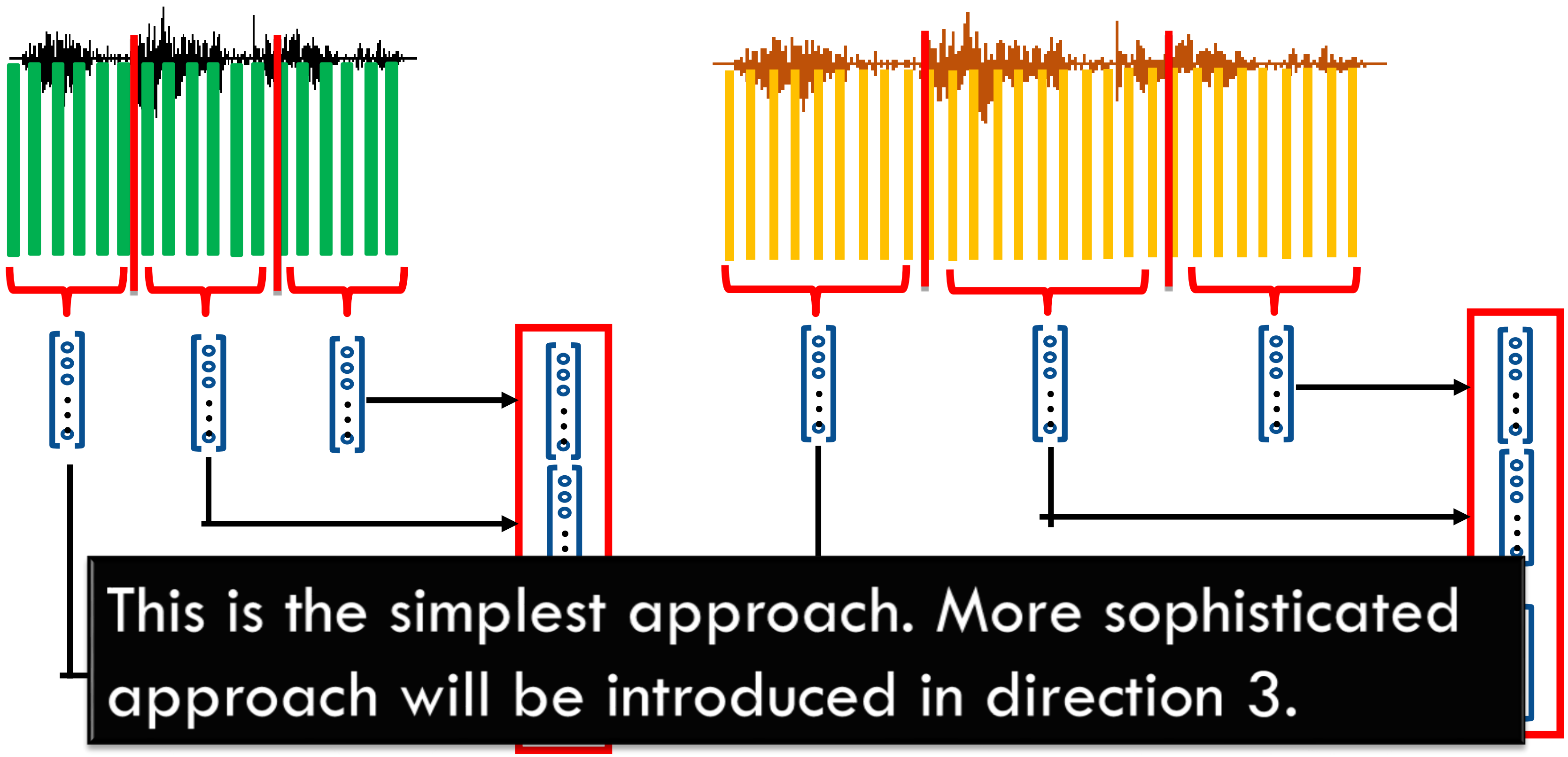
# Query-specific Detector



# Query-specific Detector

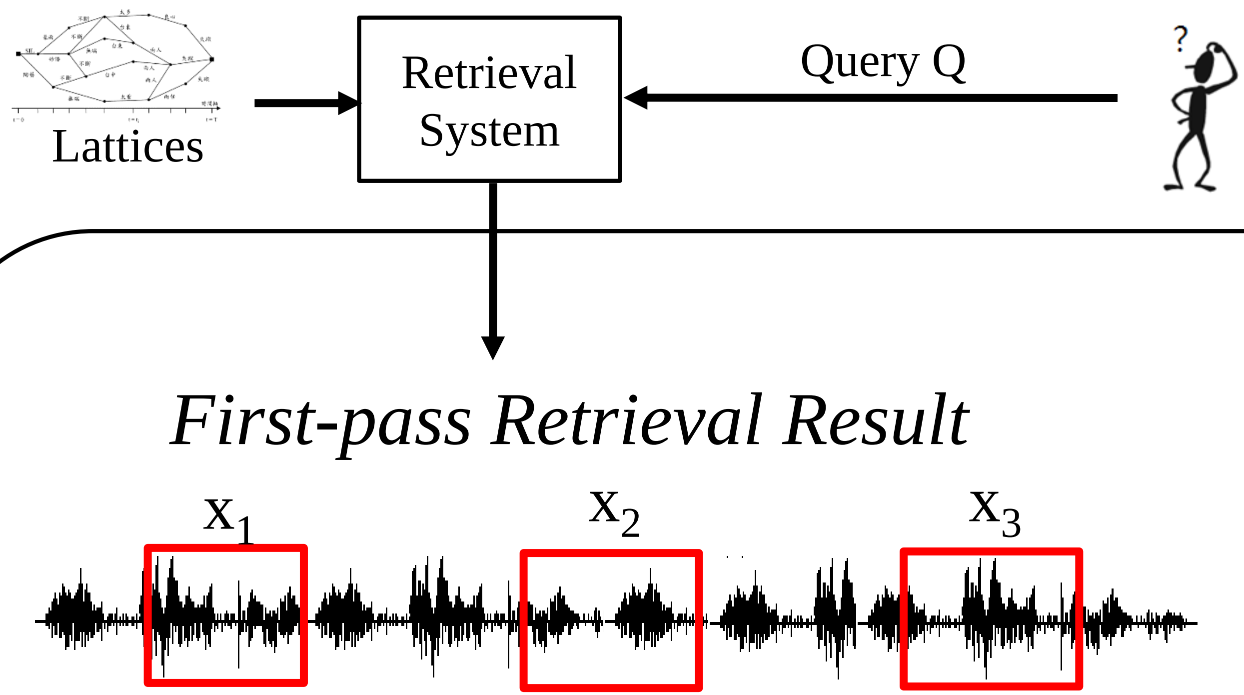
[Tu & Lee, ASRU 11]
[I.-F. Chen, Interspeech 13]

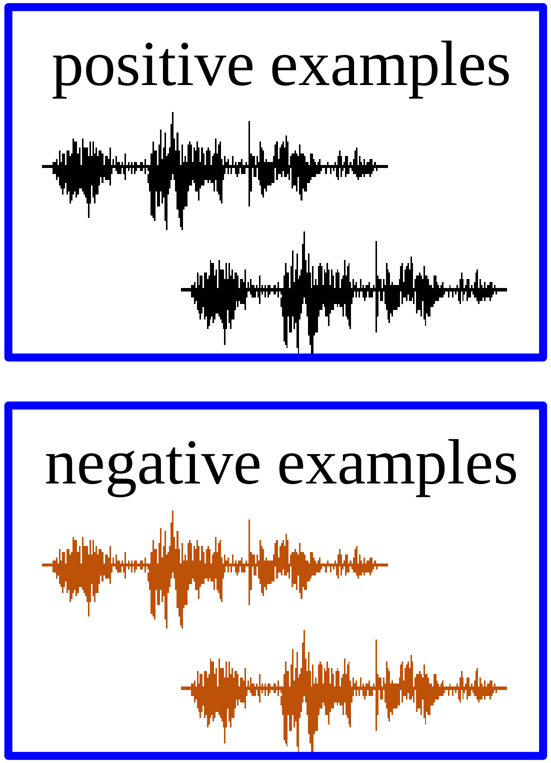
- The input of SVM or MLP has to be a fixed-length vector
- Representing an audio segment with different length into a fixed-length vector



This is the simplest approach. More sophisticated approach will be introduced in direction 3.

# Query-specific Detector



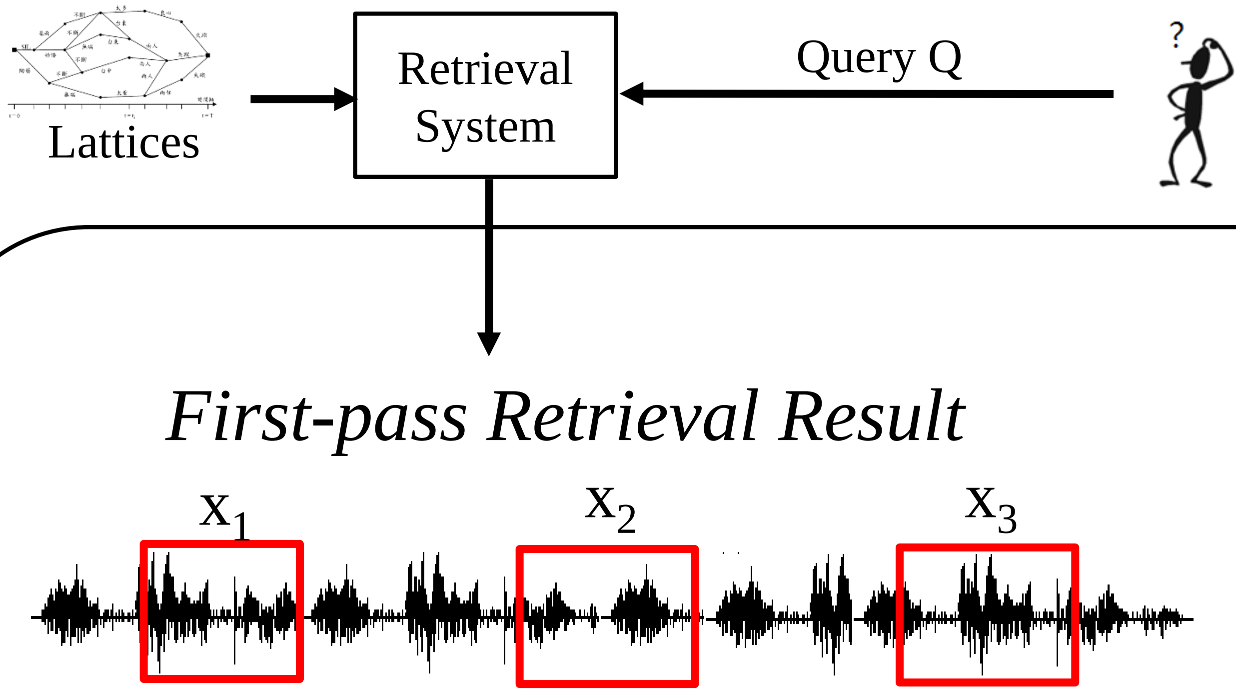

- Is it realistic to have those examples?
- Data collected from users (direction 1)
- Pseudo-relevance Feedback (PRF)

# New Direction 2-2:

## Incorporating Those Information Lost in ASR

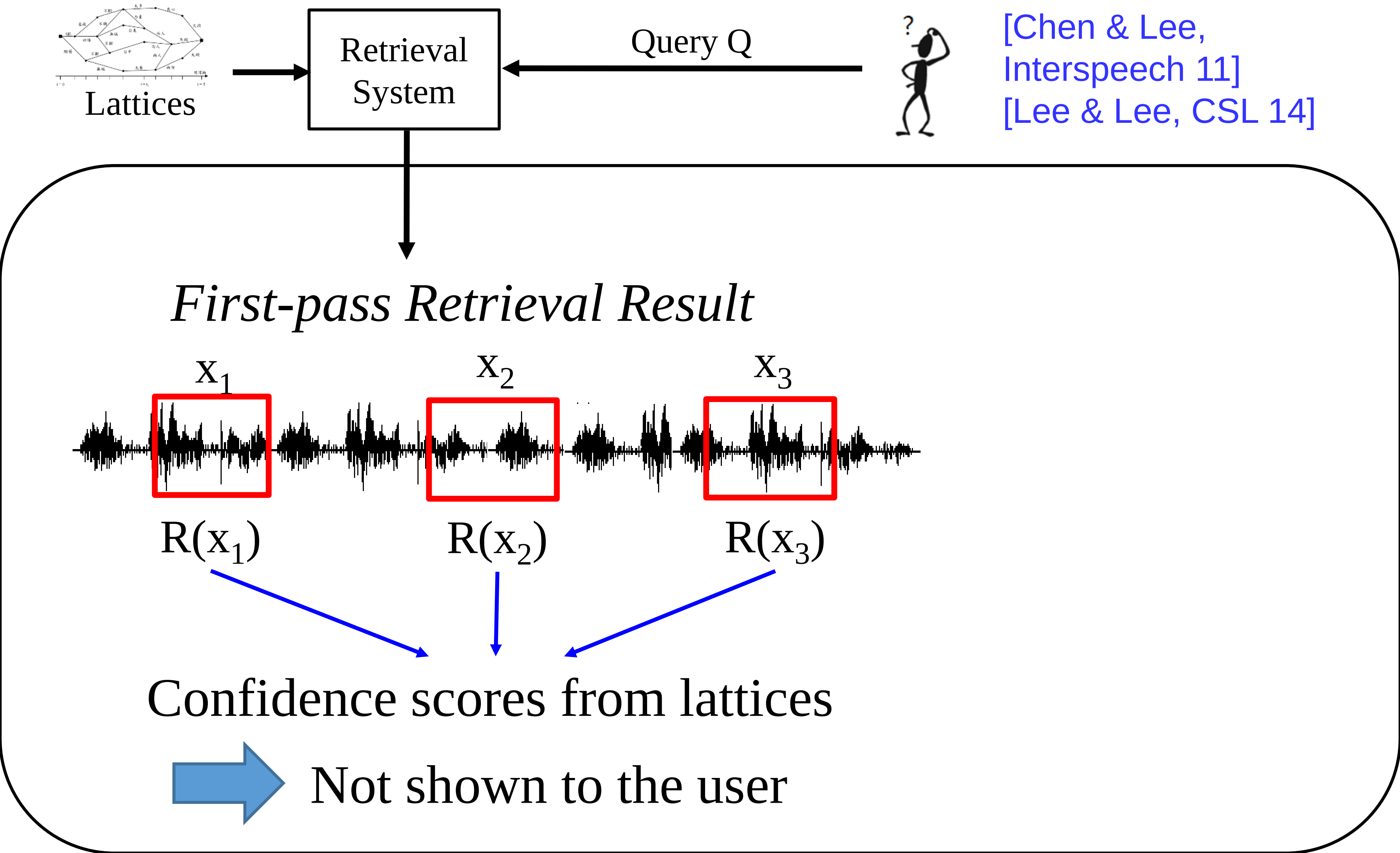
## Pseudo Relevance Feedback

# Pseudo Relevance Feedback (PRF)



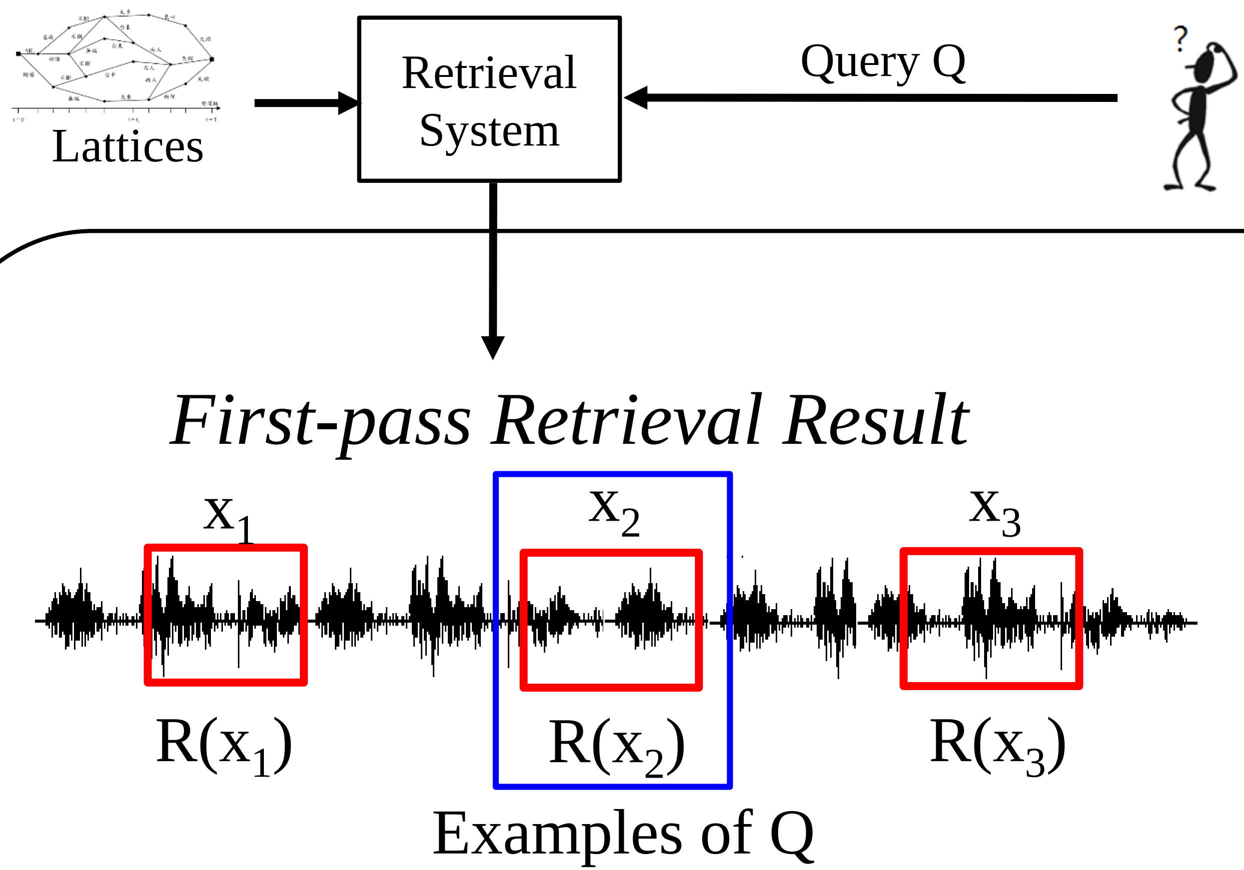

[Chen & Lee, Interspeech 11]
[Lee & Lee, CSL 14]

# Pseudo Relevance Feedback (PRF)




[Chen & Lee, Interspeech 11]
[Lee & Lee, CSL 14]

# Pseudo Relevance Feedback (PRF)

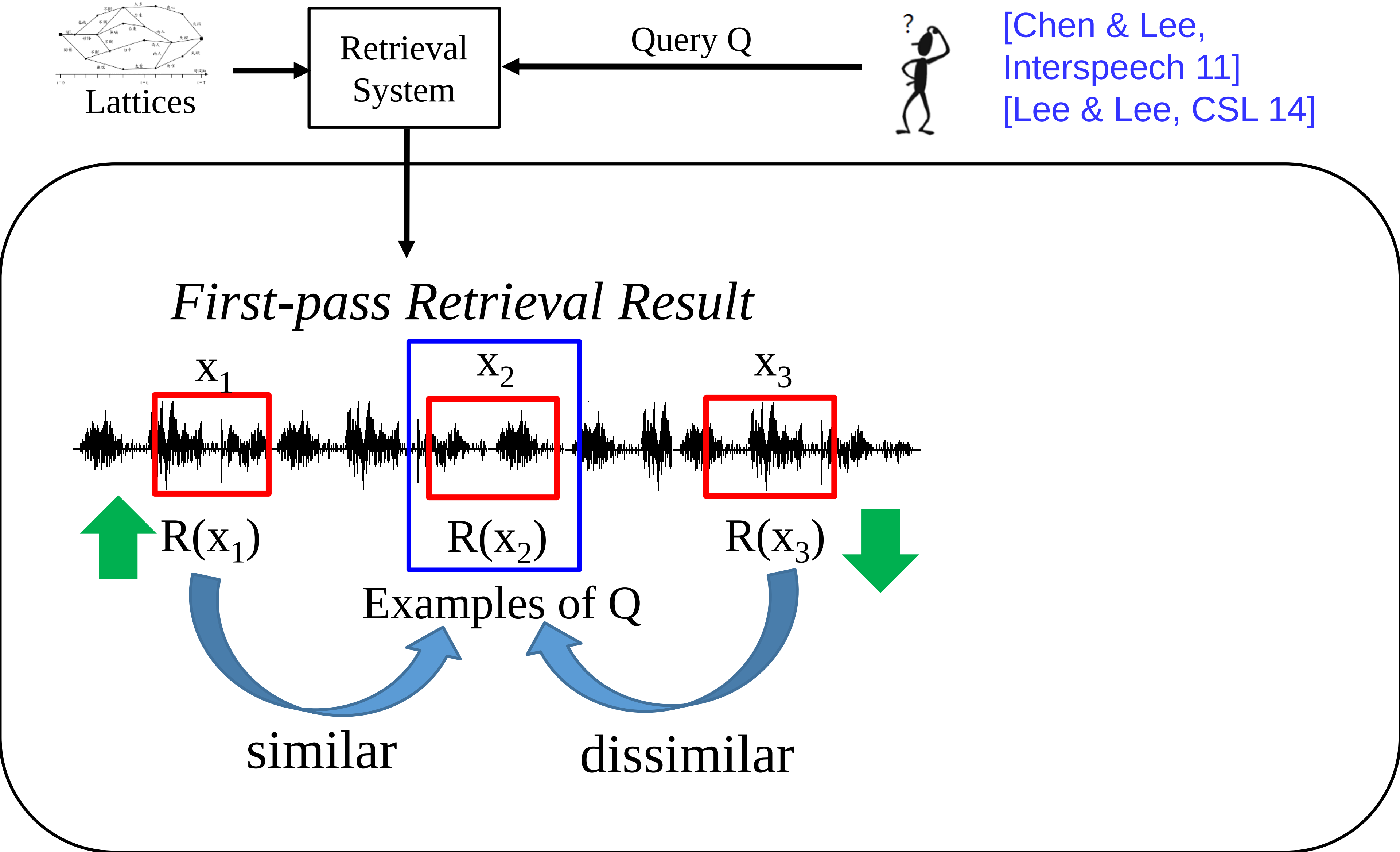



[Chen & Lee, Interspeech 11]
[Lee & Lee, CSL 14]

***Assume*** the results with high confidence scores as correct

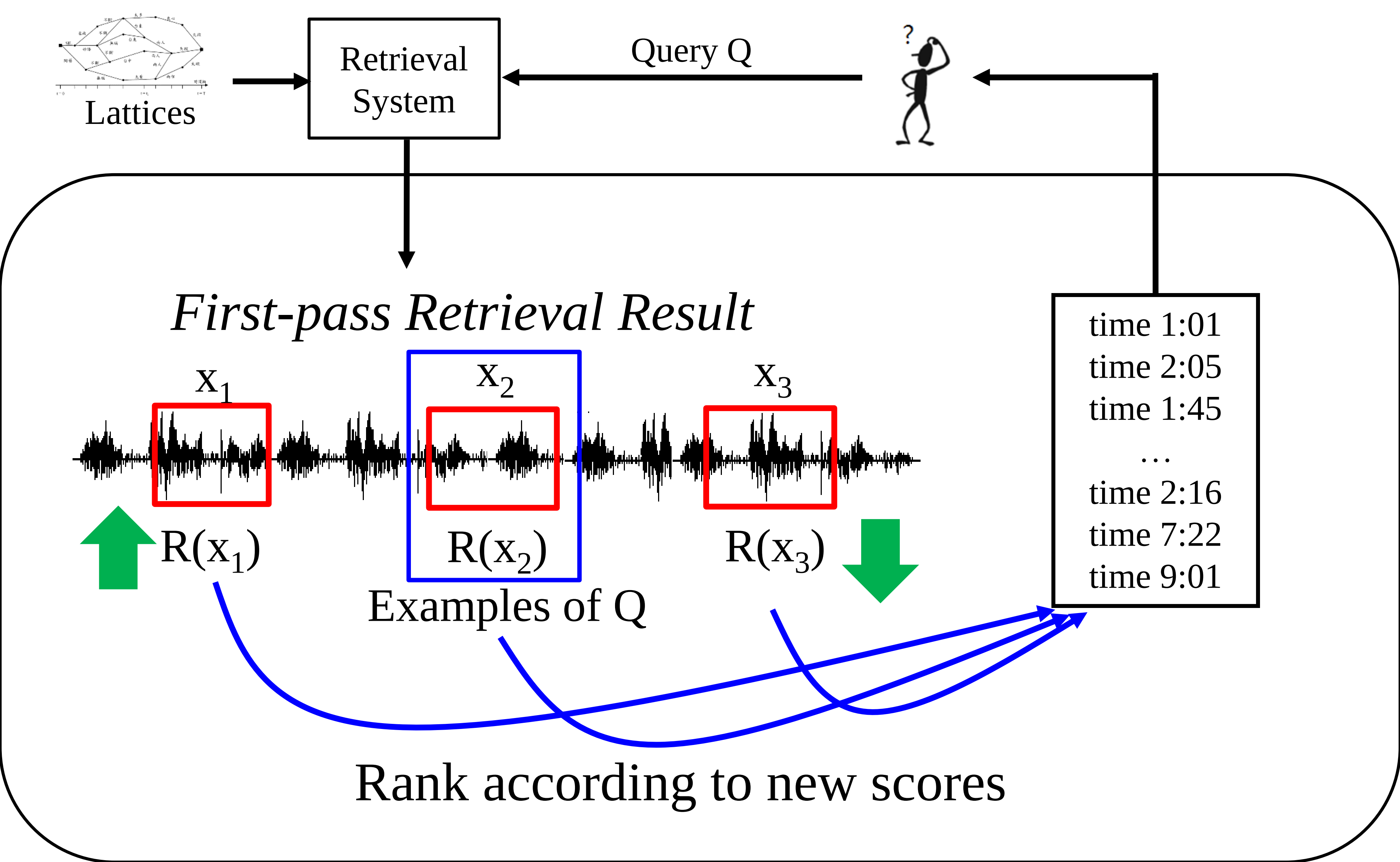
➡ Considered as examples of Q

# Pseudo Relevance Feedback (PRF)




[Chen & Lee, Interspeech 11]
[Lee & Lee, CSL 14]

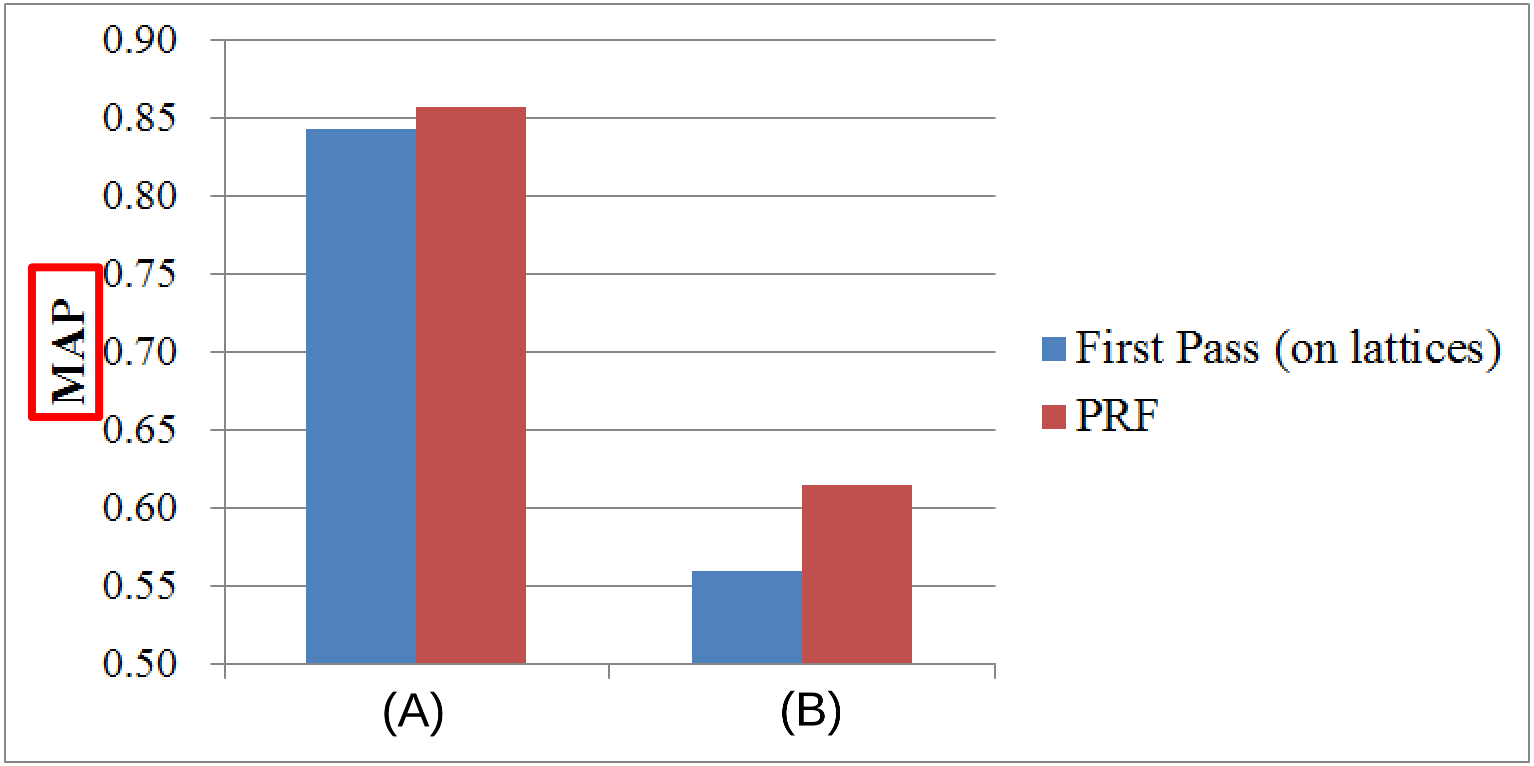
# Pseudo Relevance Feedback (PRF)



# Pseudo Relevance Feedback (PRF) - Experiments

- Lecture recording [Lee & Lee, CSL 14]

Evaluation Measure: MAP (Mean Average Precision)

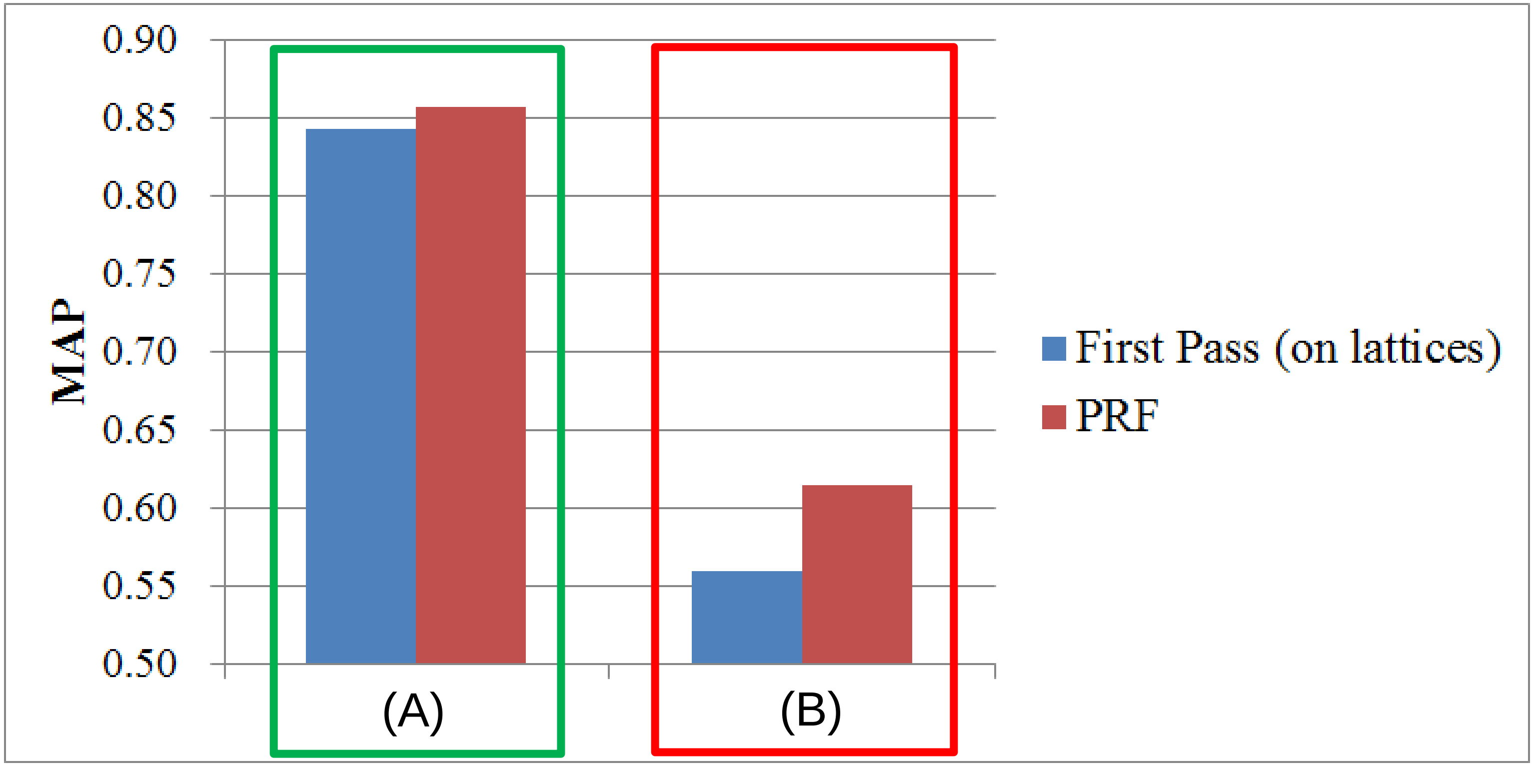


# Pseudo Relevance Feedback (PRF) - Experiments

(A) and (B) use different speech recognition systems

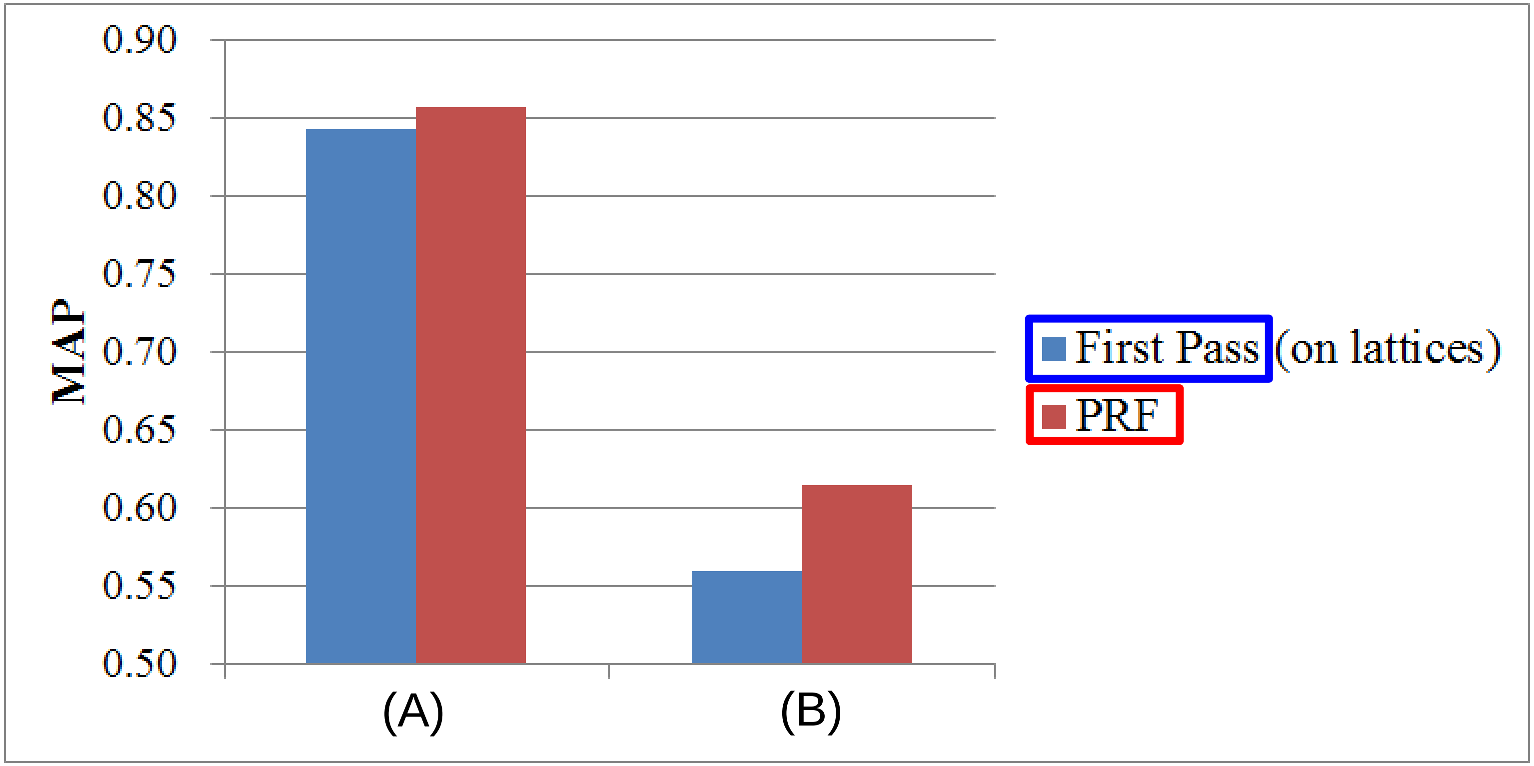
(A): speaker dependent (84% recognition accuracy)

(B): speaker independent (50% recognition accuracy)



# Pseudo Relevance Feedback (PRF) - Experiments

- PRF (red bars) improved the first-pass retrieval results with lattices (blue bars)



# New Direction 2-3:

## Incorporating
## Those Information Lost in ASR

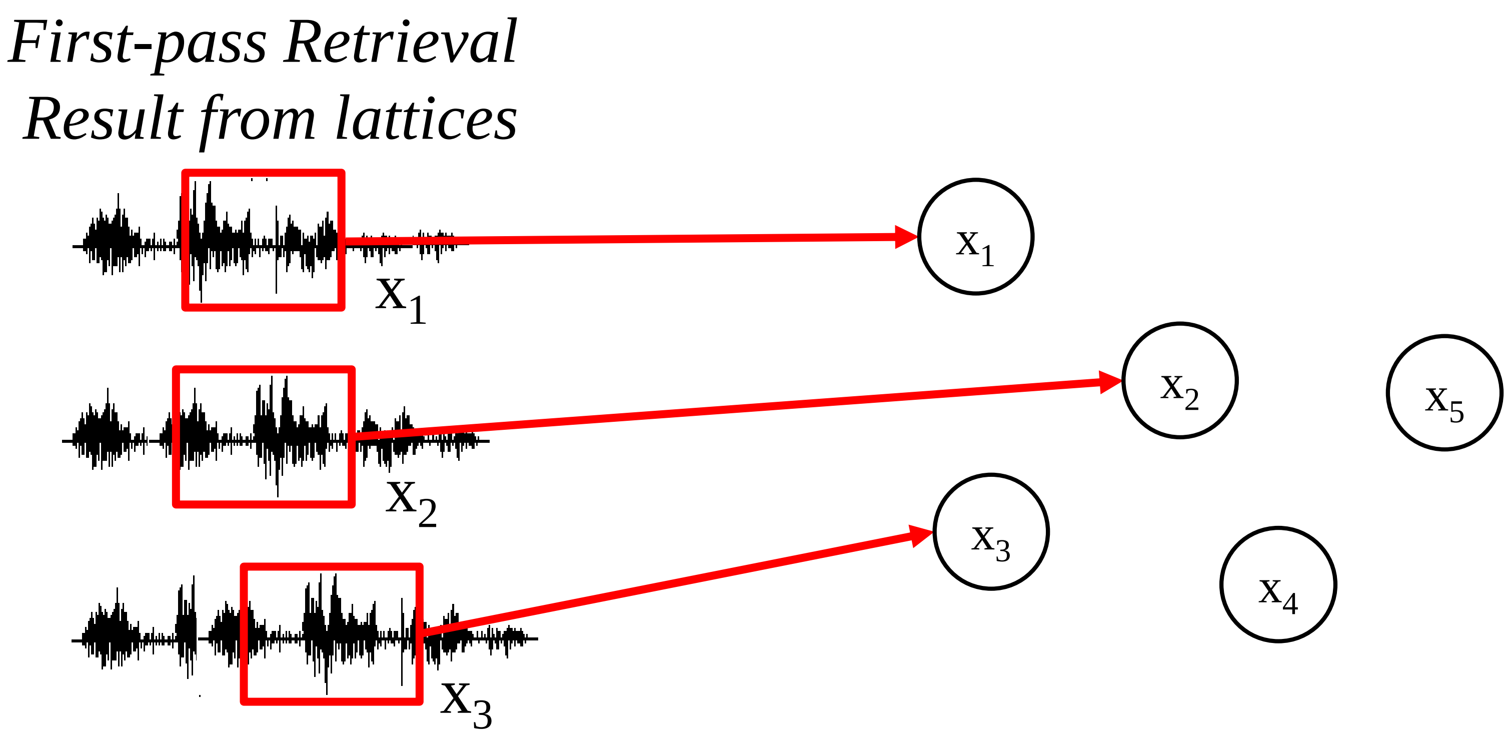
## Graph-based Approach

# Graph-based Approach

- PRF
  - Make some assumption to find the examples
  - Each result considers the similarity to the audio examples
- Graph-based approach [Chen & Lee, ICASSP 11][Lee & Lee, APSIPA 11][Lee & Lee, CSL 14]
  - Not assume some results are correct
  - Consider the similarity between all results

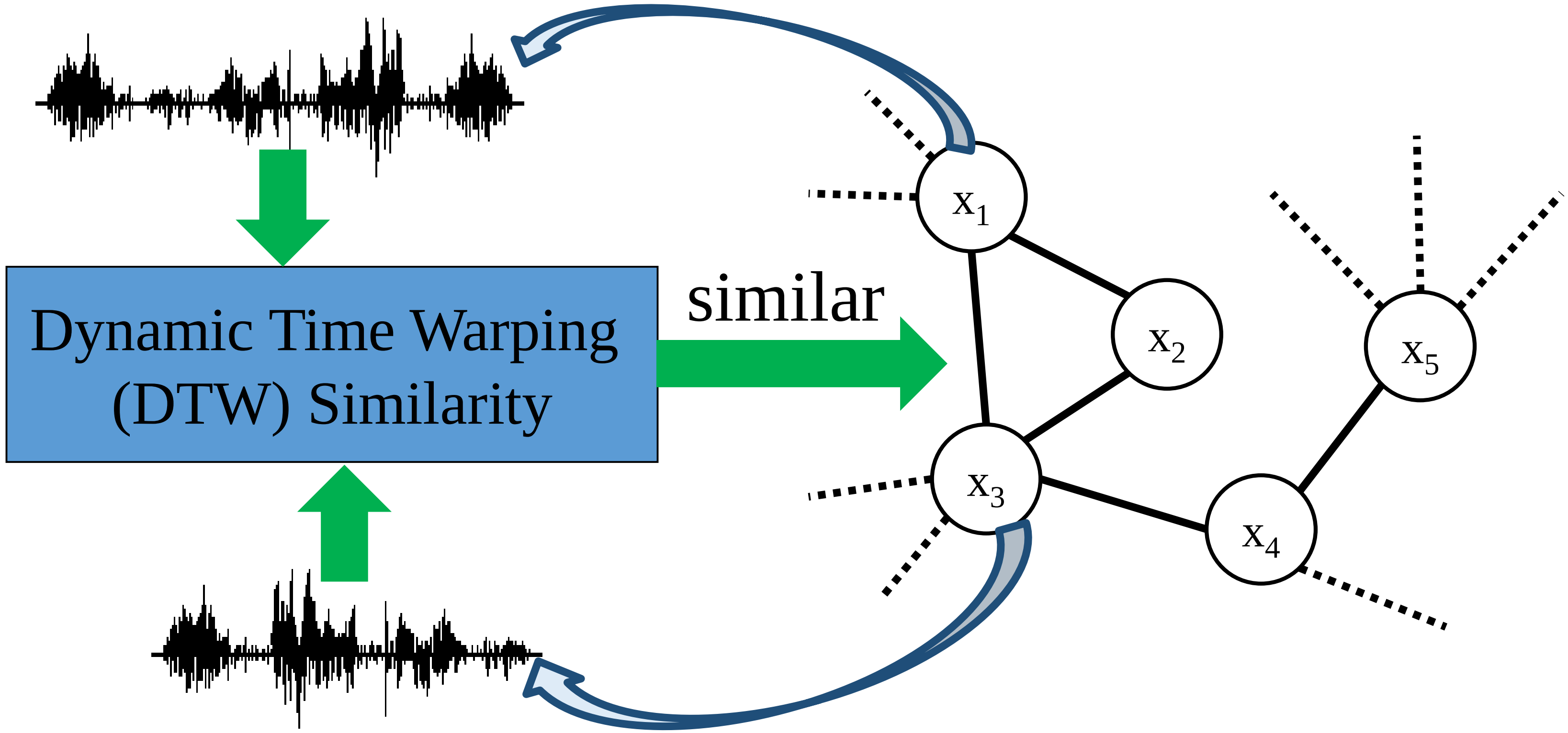
# Graph Construction

- The first-pass results is considered as a graph.
  - Each retrieval result is a node



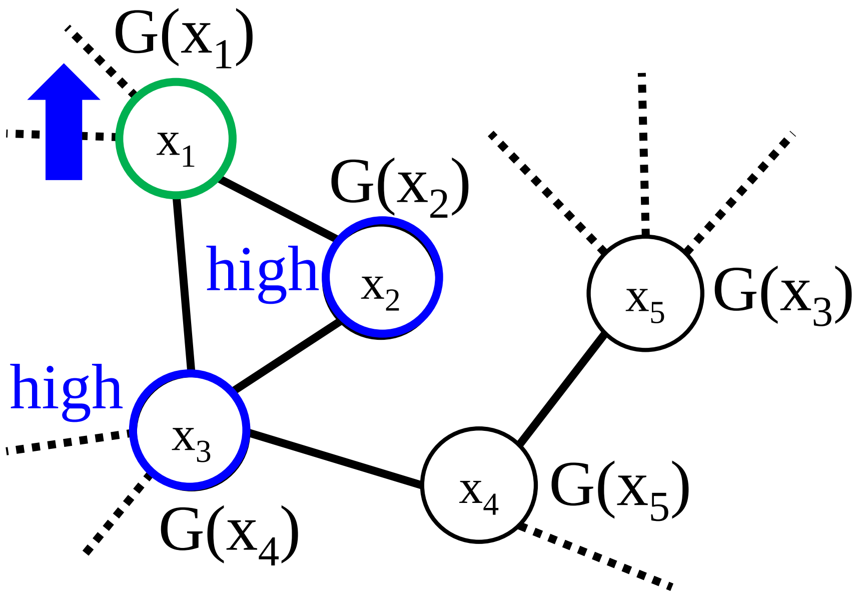
# Graph Construction

- The first-pass results is considered as a graph.
  - Nodes are connected if their retrieval results are similar.
    - DTW similarities are considered as edge weights



# Changing Confidence Scores by Graph

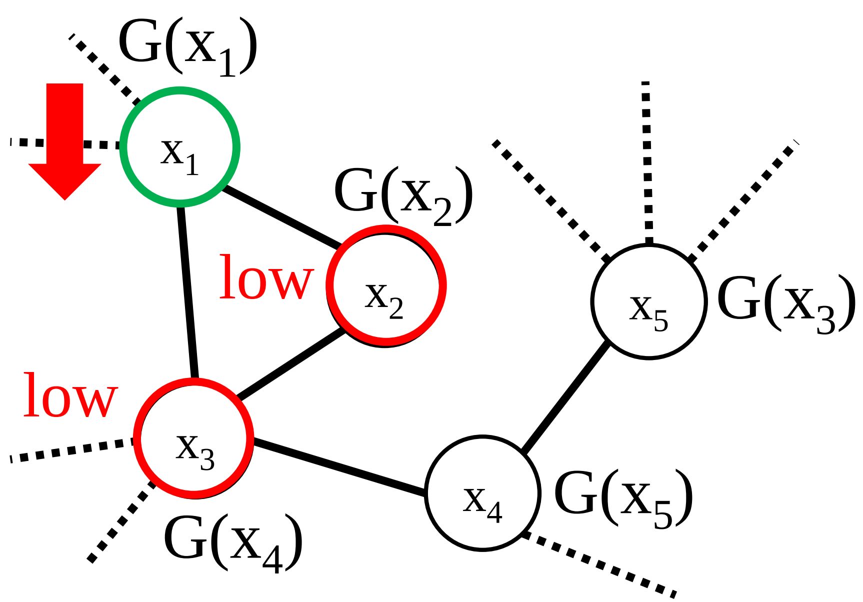
- New scores $G(x_i)$ for each node based on the graph structure → ranked according to new scores




*"You are known by the company you keep"*

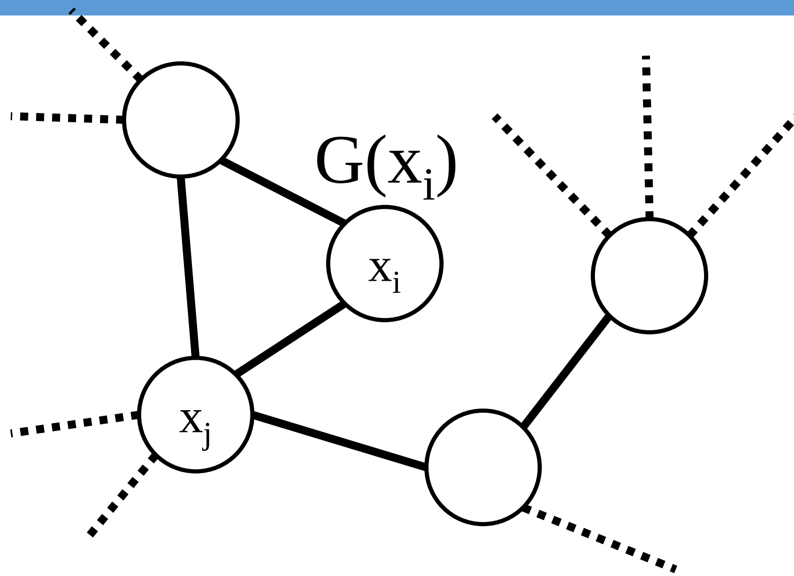
# Changing Confidence Scores by Graph

- New scores $G(x_i)$ for each node based on the graph structure ➡ ranked according to new scores



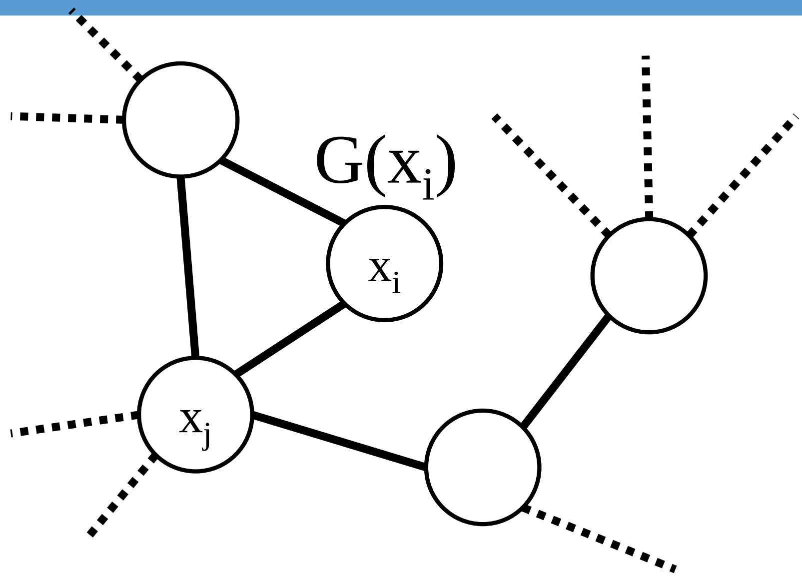

*"You are known by the company you keep"*

# Graph-based Re-ranking - Formulation




$$\boxed{G(x_i)} = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\hat{\mathrm{W}}(x_j, x_i)$$

# Graph-based Re-ranking - Formulation

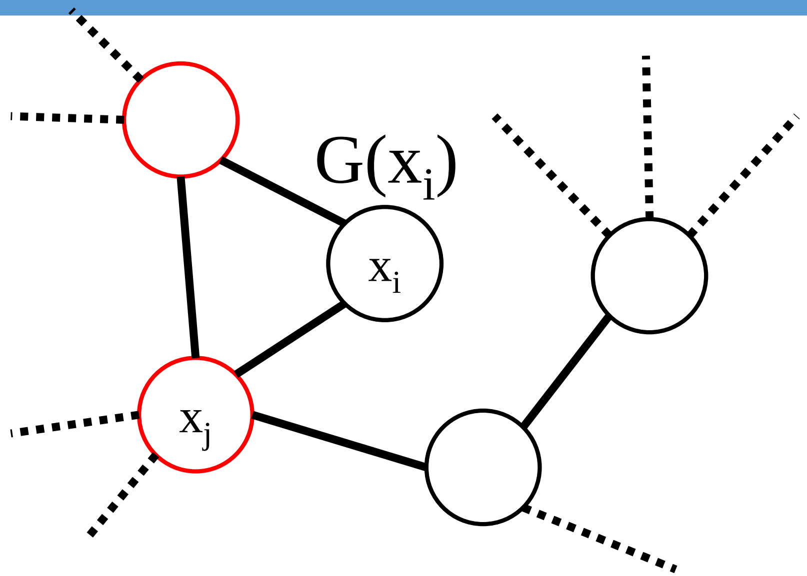



$$G(x_i) = \boxed{(1-\alpha)}\,\boxed{\mathrm{R}(x_i)} + \boxed{\alpha}\,\boxed{\sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\,\hat{\mathrm{W}}(x_j, x_i)}$$

original score
(from lattices)

considering graph structure

# Graph-based Re-ranking - Formulation

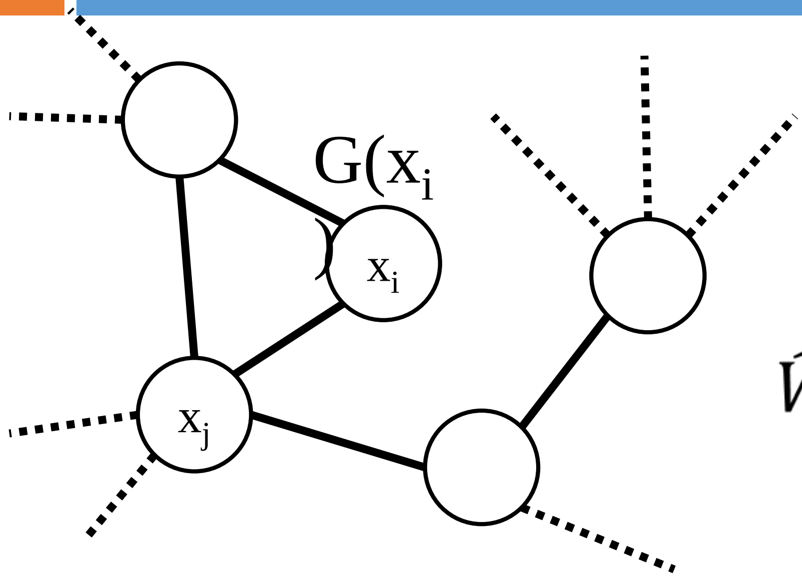



$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{\boxed{x_j} \in \boxed{\mathrm{N}(x_i)}} \boxed{G(x_j)}\boxed{\hat{\mathrm{W}}(x_j, x_i)}$$

$N(x_i)$: neighbors of $x_i$ (nodes connected to $x_i$)

$x_j$: neighbors of $x_i$ (nodes connected to $x_i$)

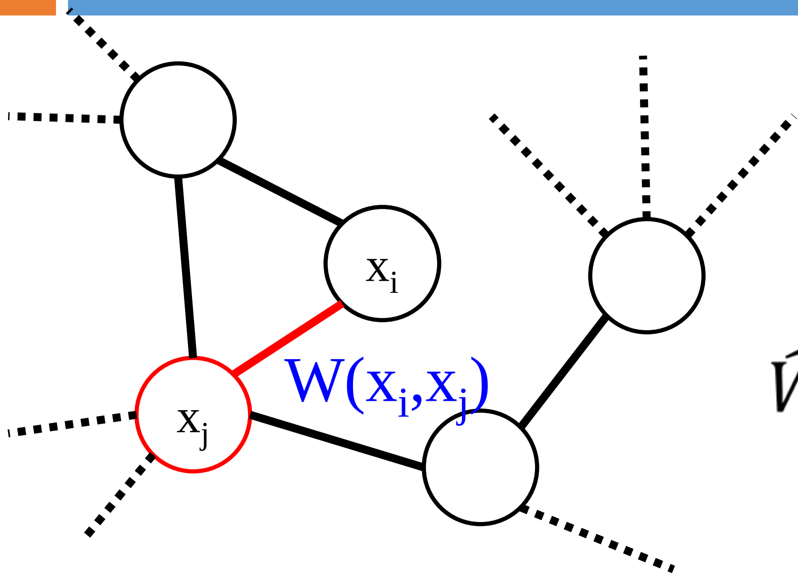
# Graph-based Re-ranking - Formulation




$$\hat{W}(x_j, x_i) = \frac{W(x_i, x_j)}{\sum_{x_k \in N(x_j)} W(x_k, x_j)}$$

$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\boxed{\hat{\mathrm{W}}(x_j, x_i)}$$

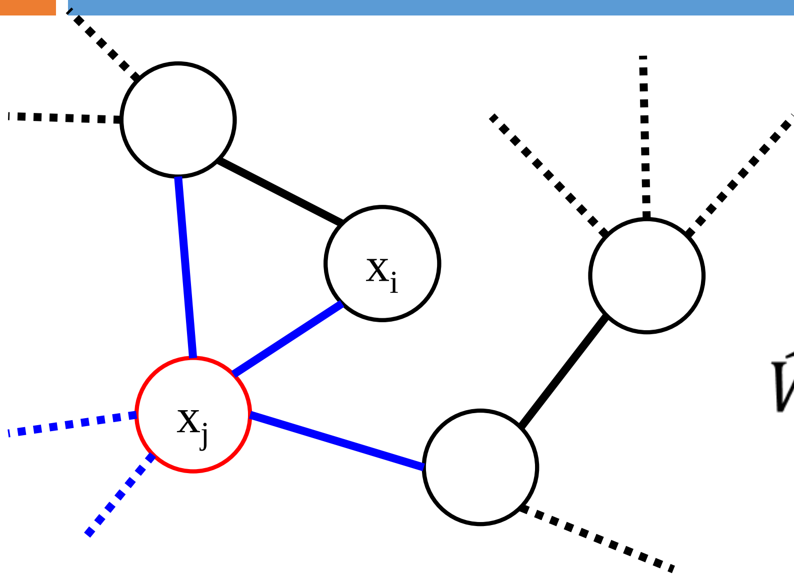
# Graph-based Re-ranking - Formulation




$$\hat{W}(x_j, x_i) = \frac{\boxed{W(x_i, x_j)}}{\sum_{x_k \in N(x_j)} W(x_k, x_j)}$$

$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\boxed{\hat{\mathrm{W}}(x_j, x_i)}$$

# Graph-based Re-ranking - Formulation

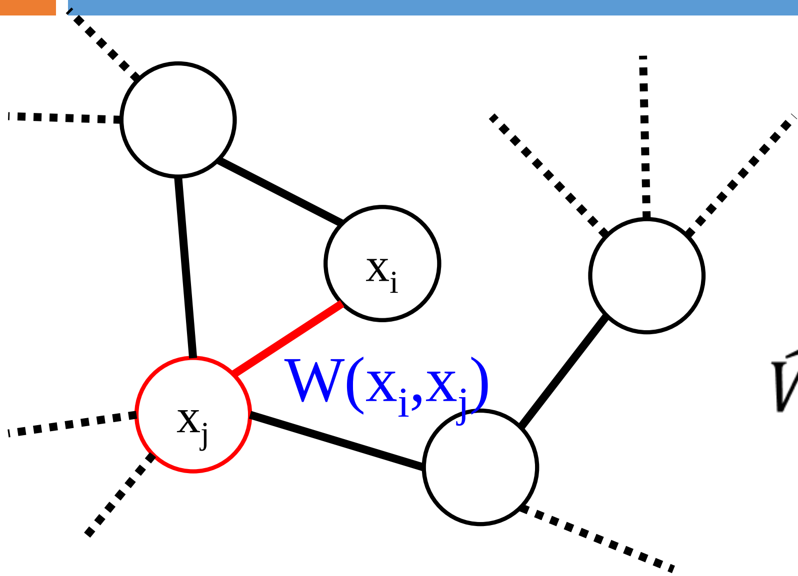



Normalized by the weights of all the edges connected to $x_j$

$$\hat{W}(x_j, x_i) = \frac{W(x_i, x_j)}{\boxed{\sum_{x_k \in N(x_j)} W(x_k, x_j)}}$$

$$G(x_i) = (1-\alpha) R(x_i) + \alpha \sum_{x_j \in N(x_i)} G(x_j) \boxed{\hat{W}(x_j, x_i)}$$

# Graph-based Re-ranking - Formulation



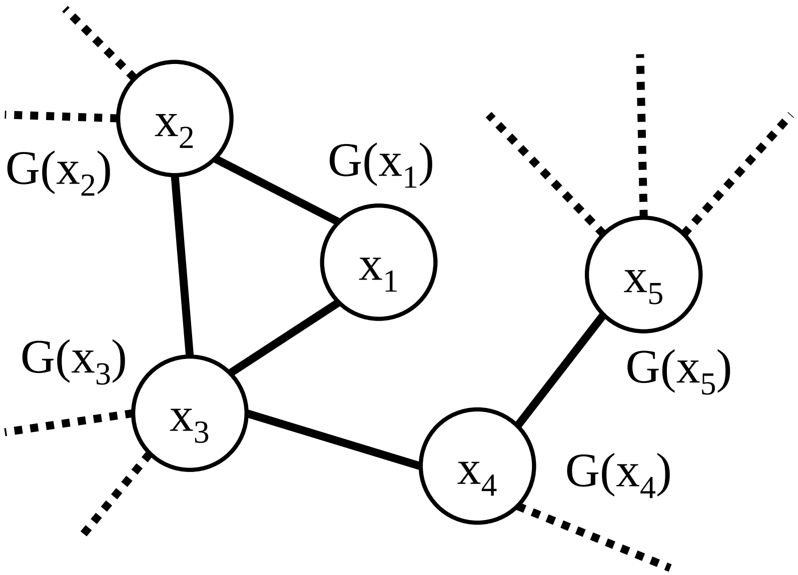

$$\hat{W}(x_j, x_i) = \frac{\boxed{W(x_i, x_j)}}{\sum_{x_k \in N(x_j)} W(x_k, x_j)}$$

$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\boxed{\hat{\mathrm{W}}(x_j, x_i)}$$

The score of $x_i$ would be more close to the nodes $x_j$ with larger edge weights.

# Graph-based Re-ranking - Formulation

- Assign score G(x) for each hit region based on the graph structure

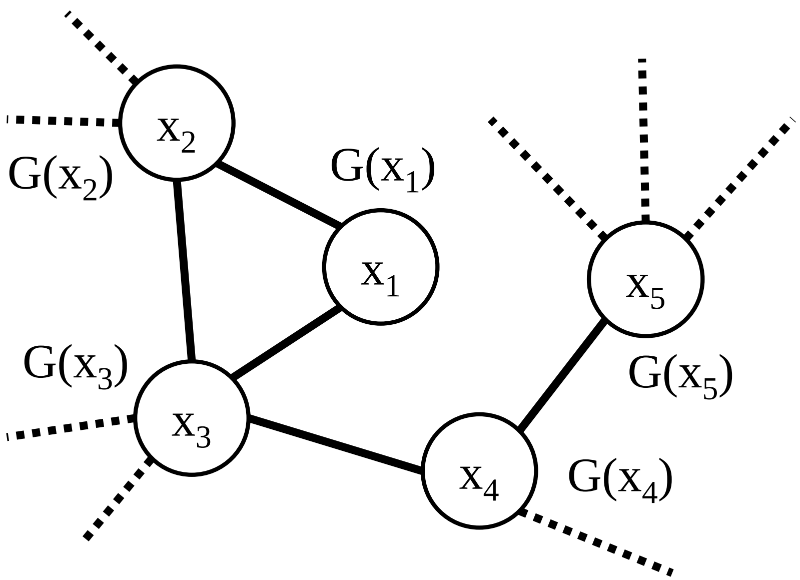



- $G(x_1)$ depends on $G(x_2)$ and $G(x_3)$

$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\hat{\mathrm{W}}(x_j, x_i)$$

# Graph-based Re-ranking - Formulation

- Assign score G(x) for each hit region based on the graph structure

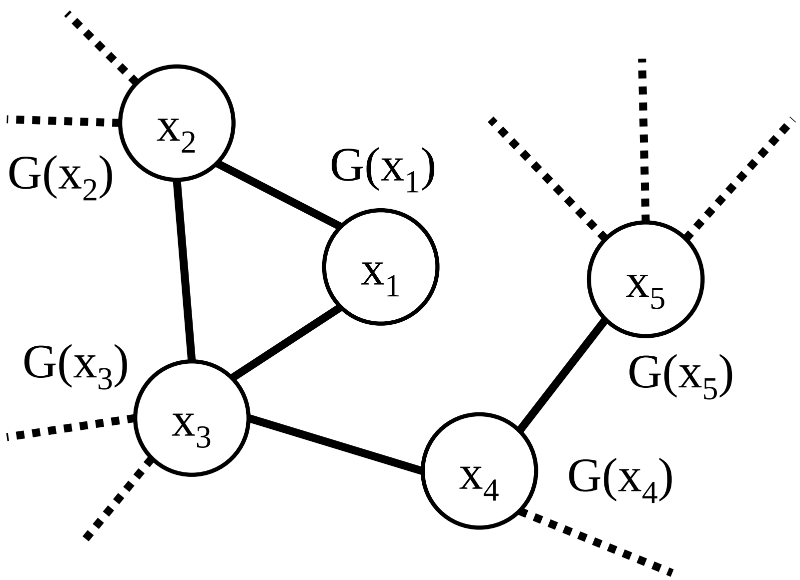



- $G(x_1)$ depends on $G(x_2)$ and $G(x_3)$
- $G(x_2)$ depends on $G(x_1)$ and $G(x_3)$ ……

$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\hat{\mathrm{W}}(x_j, x_i)$$

# Graph-based Re-ranking - Formulation

- Assign score G(x) for each hit region based on the graph structure

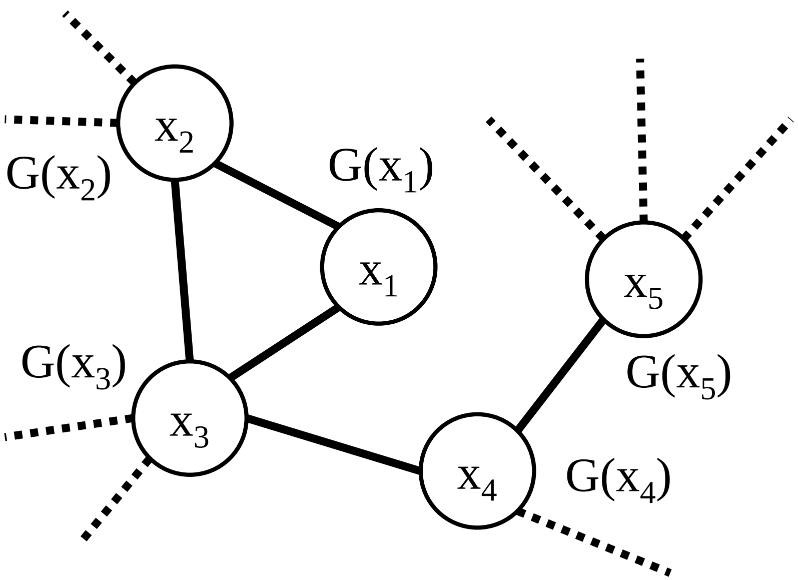



- $G(x_1)$ depends on $G(x_2)$ and $G(x_3)$
- $G(x_2)$ depends on $G(x_1)$ and $G(x_3)$ ……
- ……

$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\hat{\mathrm{W}}(x_j, x_i)$$

# Graph-based Re-ranking - Formulation

- Assign score G(x) for each hit region based on the graph structure




- How to find $G(x_1)$, $G(x_2)$, $G(x_3)$ …… satisfying the following equation?
- This is random walk.

$G(x_i)$ is uniquely and efficiently obtainable

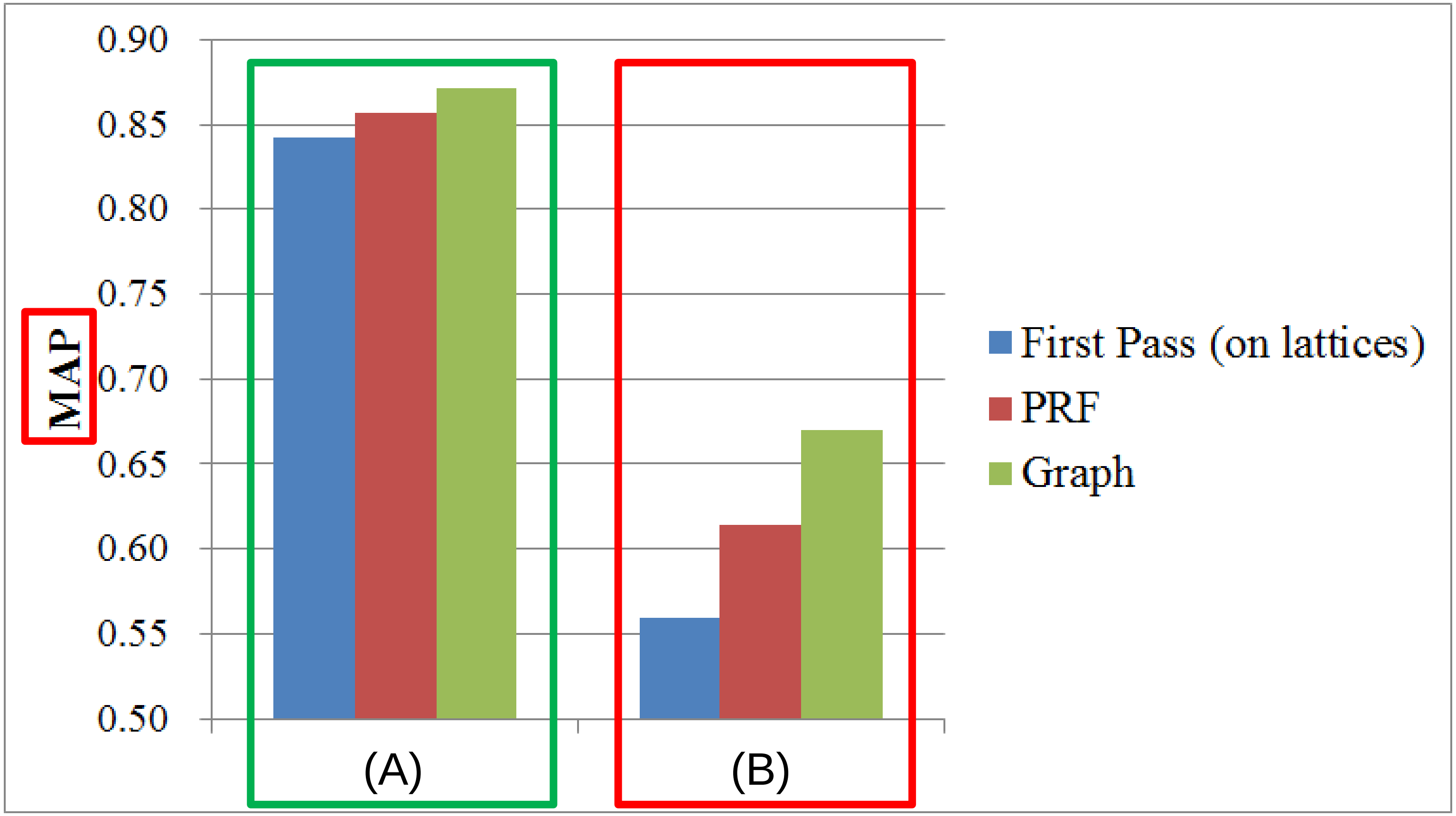
$$G(x_i) = (1-\alpha)\mathrm{R}(x_i) + \alpha \sum_{x_j \in \mathrm{N}(x_i)} G(x_j)\hat{\mathrm{W}}(x_j, x_i)$$

# Graph-based Approach - Experiments

- Lecture recording [Lee & Lee, CSL 14]

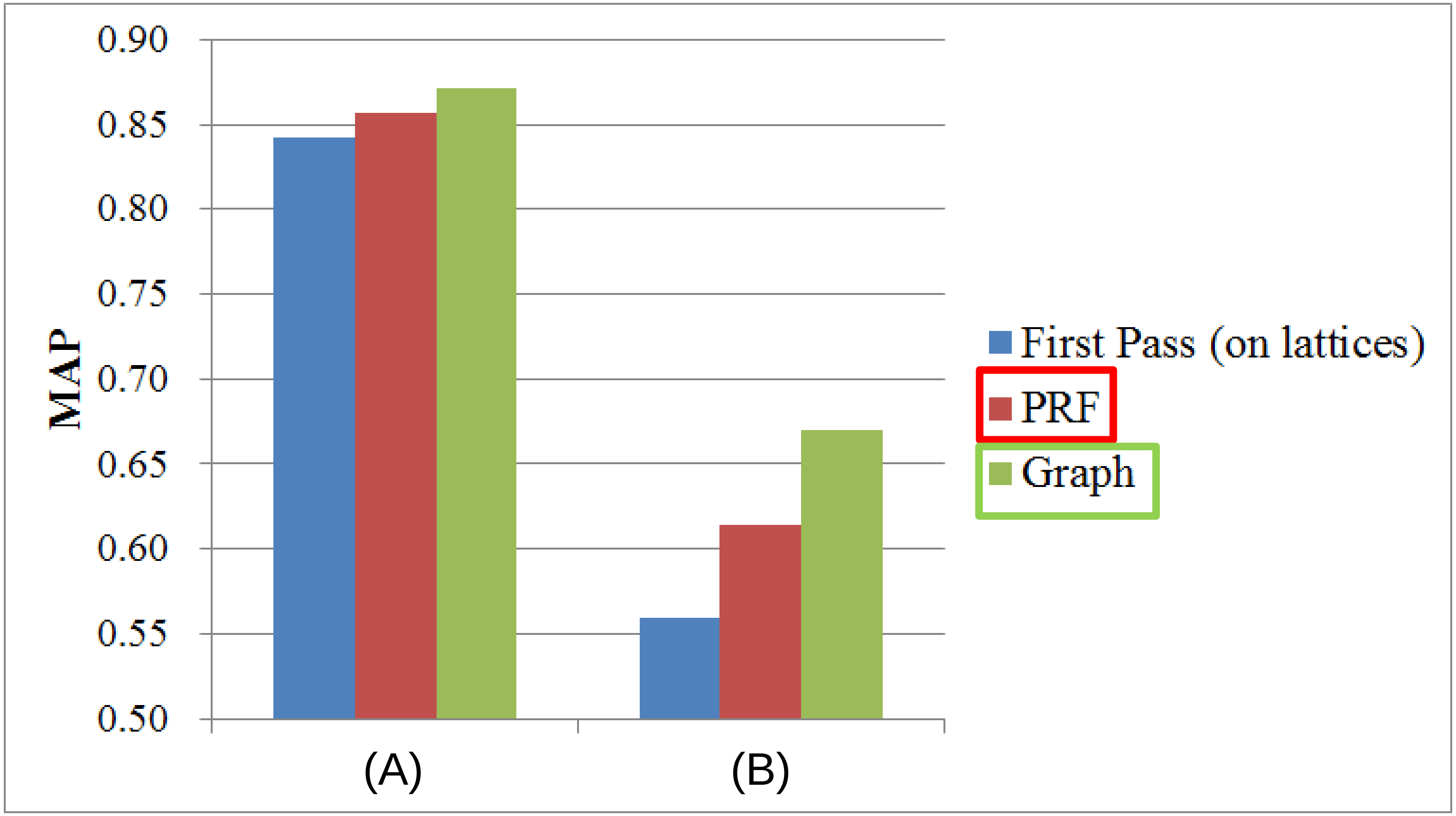
  (A): speaker dependent (high recognition accuracy)

  (B): speaker independent (low recognition accuracy)



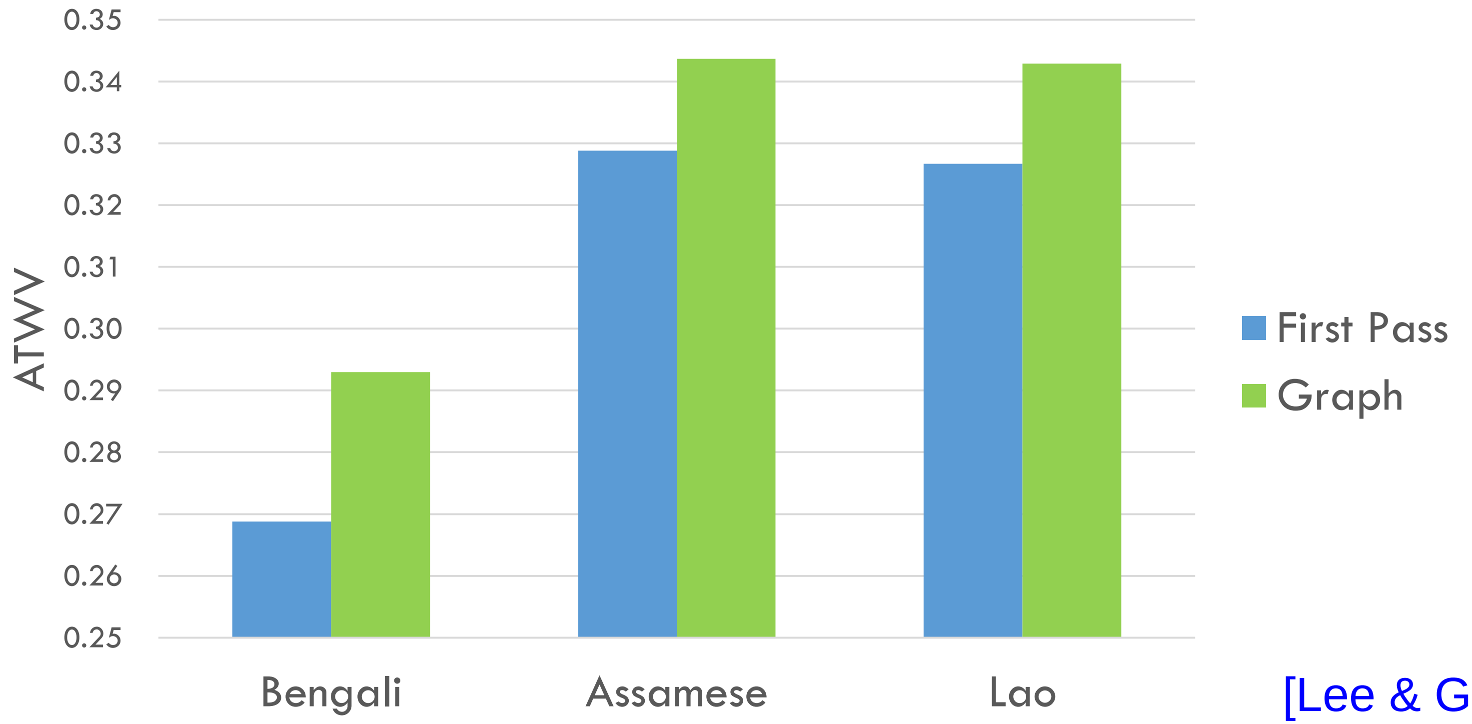
# Graph-based Approach - Experiments

- Graph-based re-ranking (green bars) outperformed PRF (red bars)



# Graph-based Approach – Experiments

- Graph-based approach on limited language data from the IARPA Babel program




[Lee & Glass, Interspeech 14]

# Graph-based Approach – Experiments

- 13% relative improvement on OOV queries on lecture recording (several speakers) [Jansen, ICASSP 13][Norouzian, ICASSP 13]
- 14% relative improvement on AMI Meeting Corpus [Norouzian, Interspeech 13]
  - Graph Spectral Clustering
- Optimizing evaluation measure and considering the graph structure at the same time [Audhkhasi, ICASSP 2014]
- 11% relative improvement with subword-based system on OpenKWS15 (Swahili) [Van Tung Pham, ICASSP, 2016]

# New Direction 3:

## No Speech Recognition!

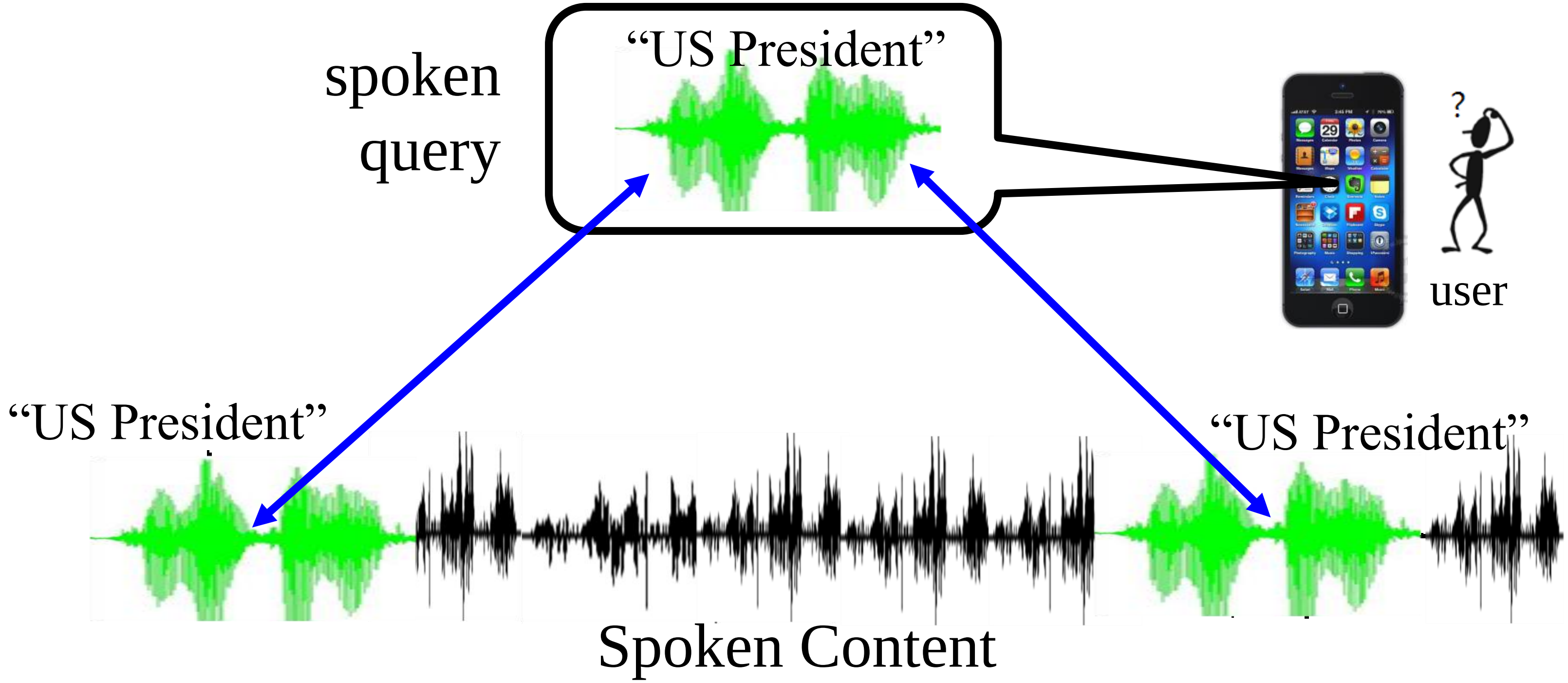
# Why Spoken Content Retrieval without Speech Recognition?

- Bypassing ASR to avoid information loss and all problems with ASR (errors, OOV words, background noise, etc. )
- Just to identify the query, no need to find out which words the query includes
- Audio files on the Internet in hundreds of different languages
  - Too limited annotated data for training reliable speech recognition systems for most languages
  - Written form even doesn't exist for some languages
- Many audio files are code-switched across several different languages

# Spoken Content Retrieval without Speech Recognition

Is it possible?

**This task is called "Query-by-Example".**




Compute similarity between spoken queries and audio files on acoustic level, and find the query term

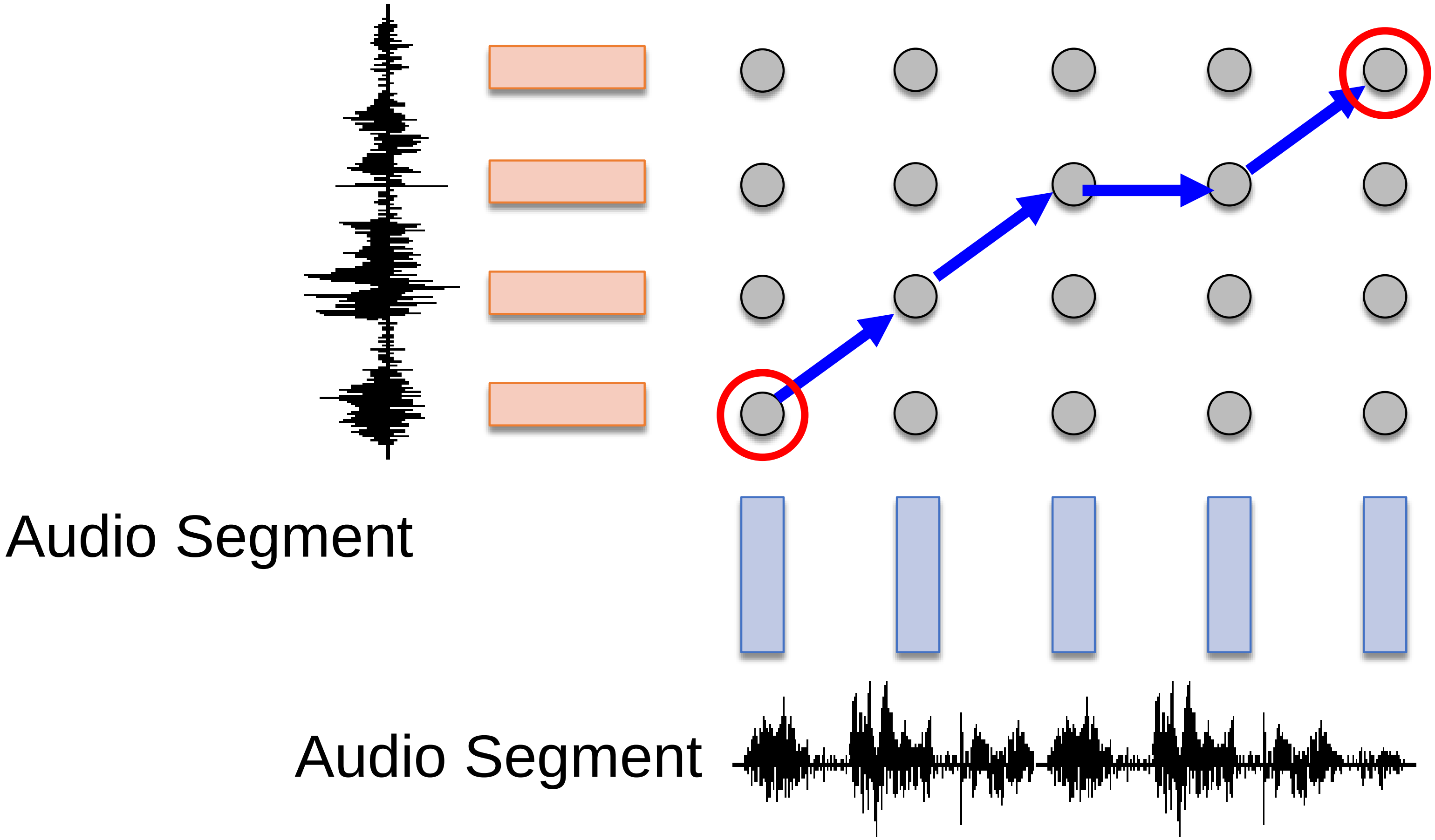
# Approach Categories

- DTW-based Approaches
  - Matching sequences with DTW
- Audio Segment Representation
  - Representing audio segments by fixed length vector representations
- Unsupervised ASR (or model-based approach)
  - Training word- or subword-like acoustic patterns (or tokens) from target audio archive
  - Transcribing both the audio archive and the query into word- or subword-like token sequences
  - Matching based on the tokens, just like text retrieval

# New Direction 3-1:
# No Speech Recognition!
## DTW-based Approaches

# DTW-based Approach

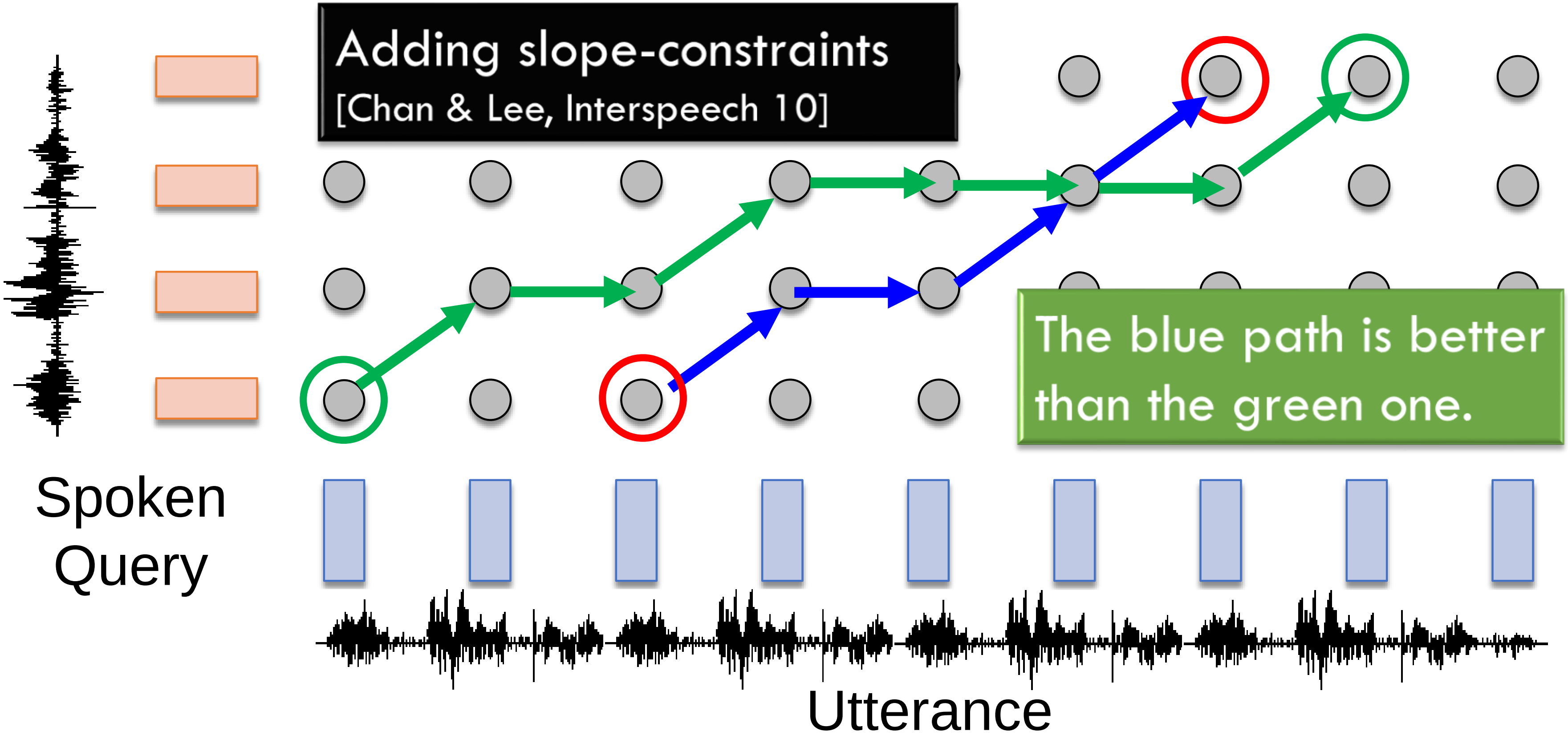
- Conventional DTW



# DTW-based Approach

Segmental DTW [Zhang, ICASSP 10], Subsequence DTW [Anguera, ICME 13][Calvo, MediaEval 14]

- DTW for query-by-example
  - Whether a spoken query is in an utterance



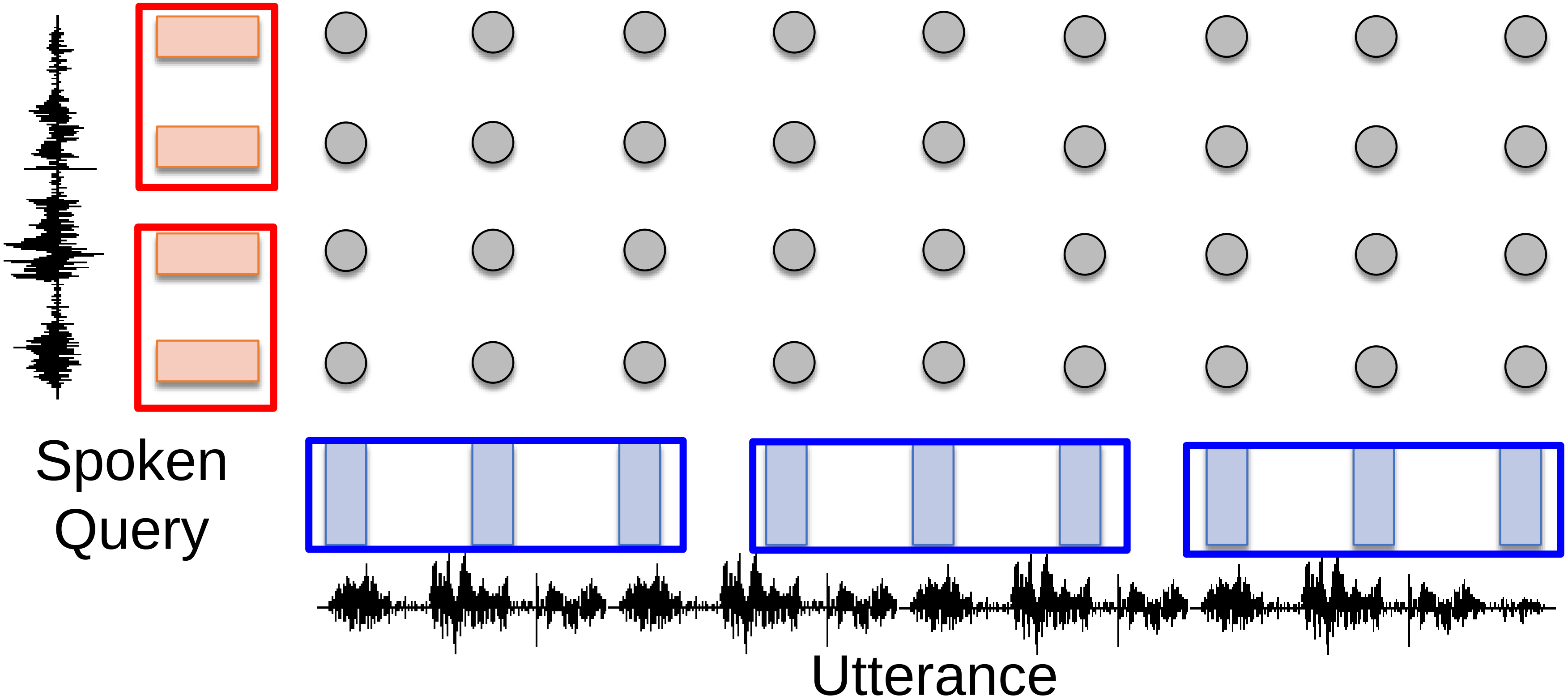
# Acoustic Feature Vectors

- Gaussian posteriorgram [Zhang, ICASSP 10][Wang, MediaEval 14]
- Phonetic posteriors [Hazen, ASRU 09]
  - MLP trained from another corpus (probably in a different language)
- Bottle-neck feature generated from MLP [Kesiraju, MediaEval 14]
- RBM posteriorgram [Zhang, ICASSP 12]
- Performance comparison [Carlin, Interspeech 11]

# Speed-up Approaches for DTW

- Segment-based matching [Chan & Lee, Interspeech 10][Chan & Lee, ICASSP 11]

Group consecutive acoustically similar feature vectors into a segment



# Speed-up Approaches for DTW

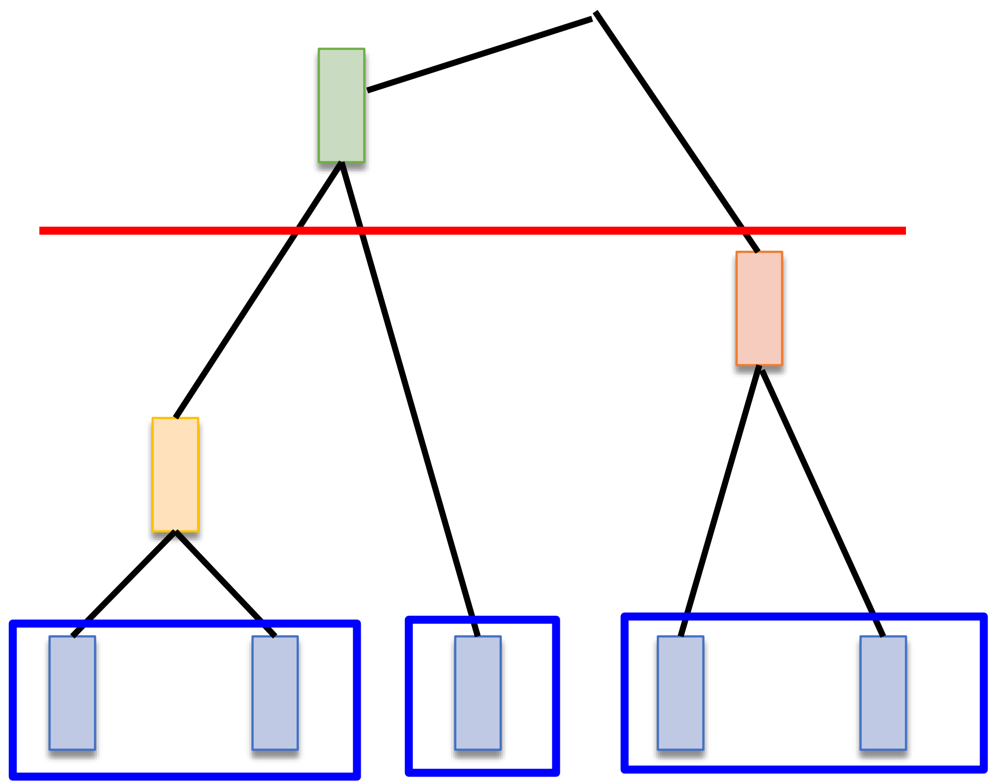
- Segment-based matching

Group consecutive acoustically similar feature vectors into a segment

Hierarchical Agglomerative Clustering (HAC)

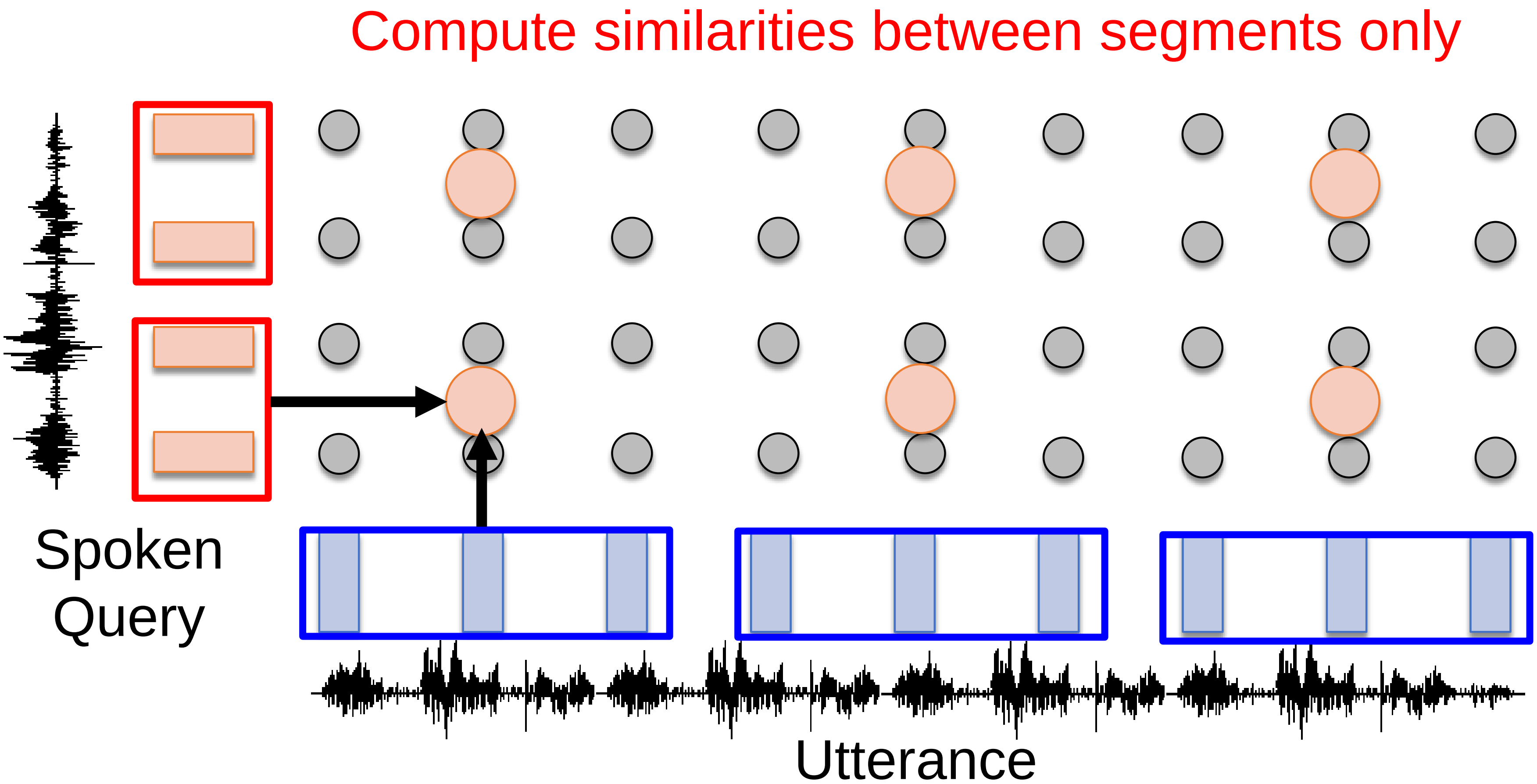
Step 1: build a tree

Step 2: pick a threshold



# Speed-up Approaches for DTW

- Segment-based matching [Chan & Lee, Interspeech 10][Chan & Lee, ICASSP 11]



# Speed-up Approaches for DTW

- Segment-based matching [Chan & Lee, Interspeech 10][Chan & Lee, ICASSP 11]
- Lower bound estimation [Zhang, ICASSP 11][Zhang, Interspeech 11]
- Indexing the frames in the target audio file [Jansen, ASRU 11][Jansen, Interspeech 12]
- Information Retrieval based DTW [Anguera, Interspeech 13]

# New Direction 3-2:
# No Speech Recognition!

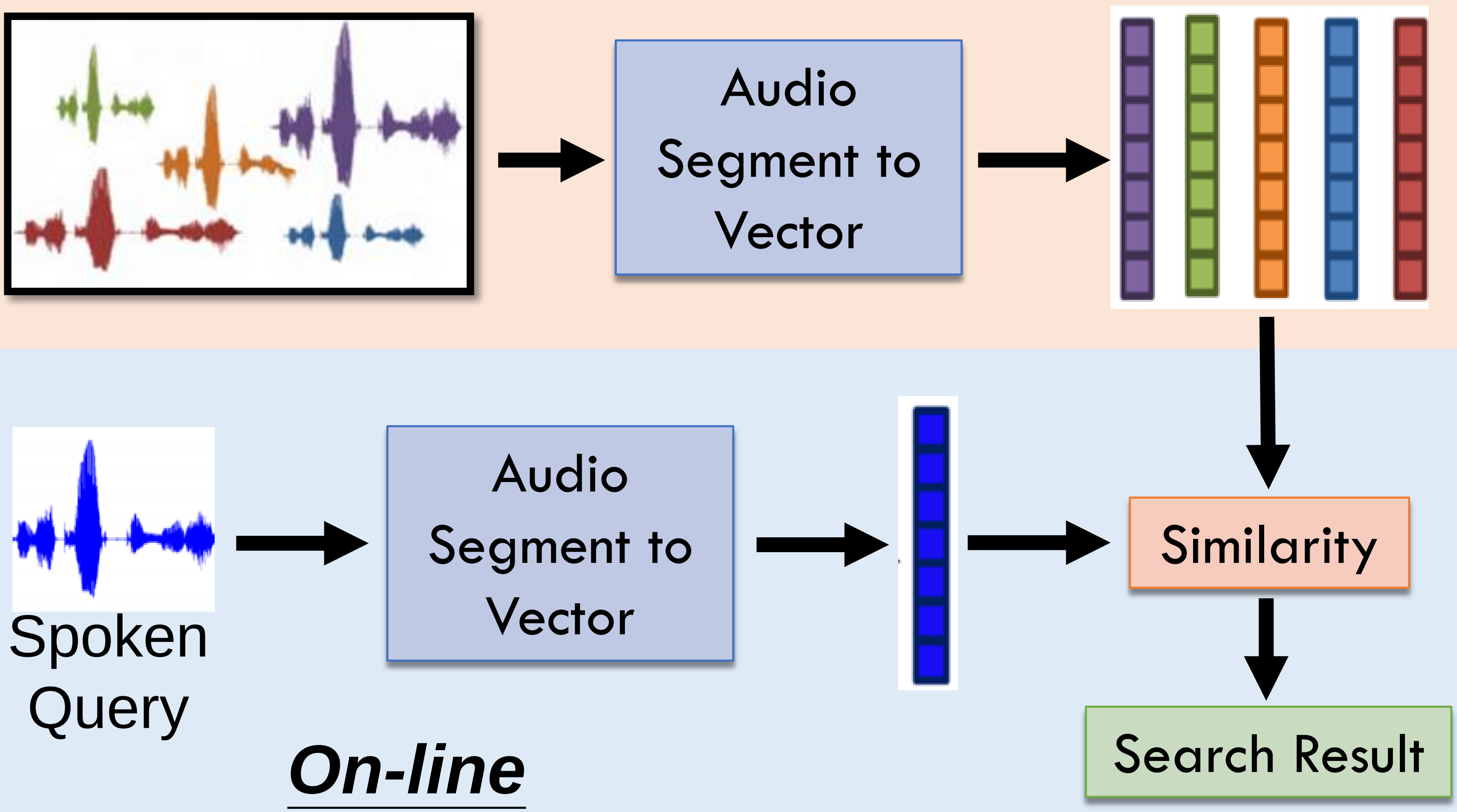
## Audio Segment Representation

# Framework

[Chung & Lee, Interspeech 16][Chen, ICASSP 15]
[Levin, ICASSP 15][Levin, ASRU 13]

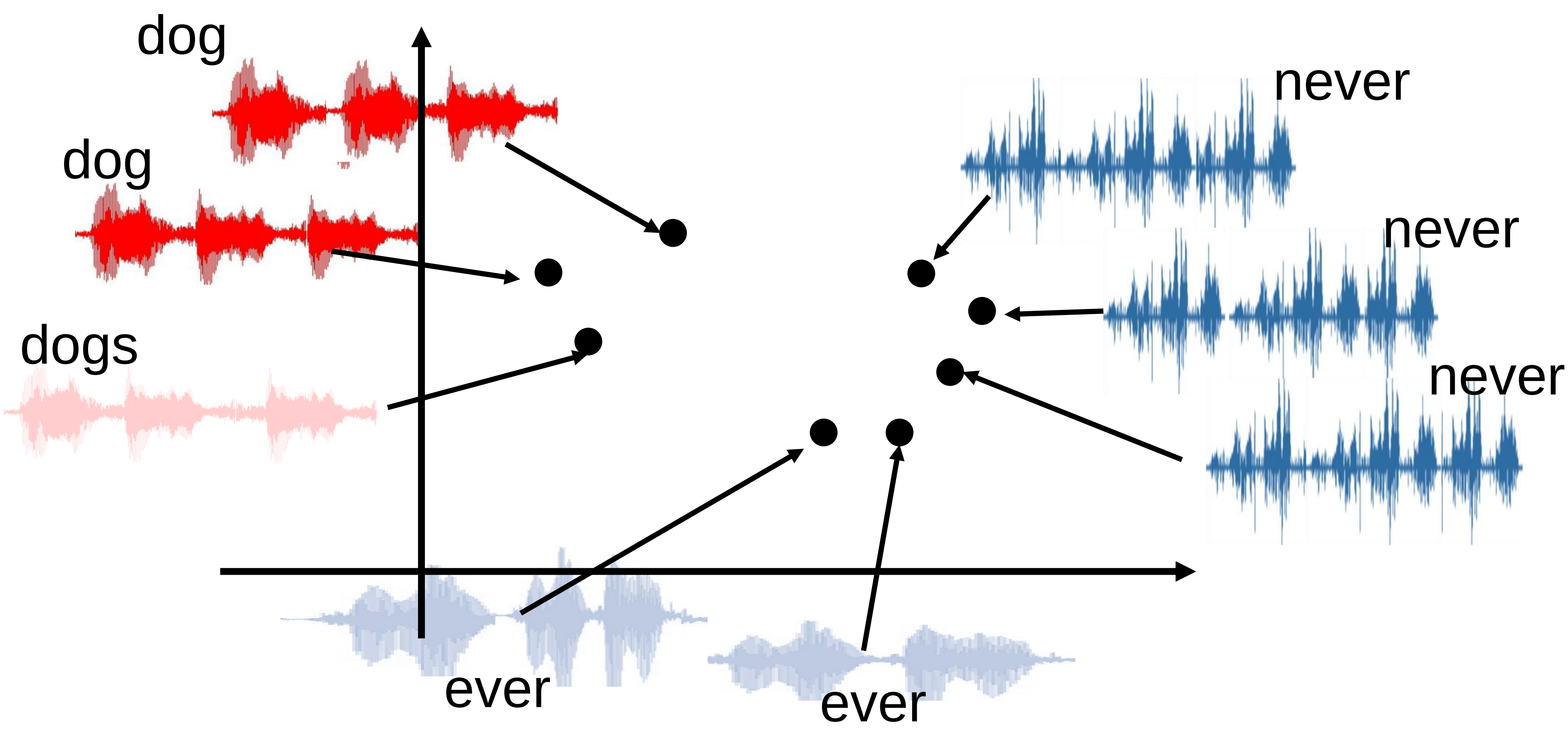
Audio archive divided into variable-length audio segments

***Off-line***



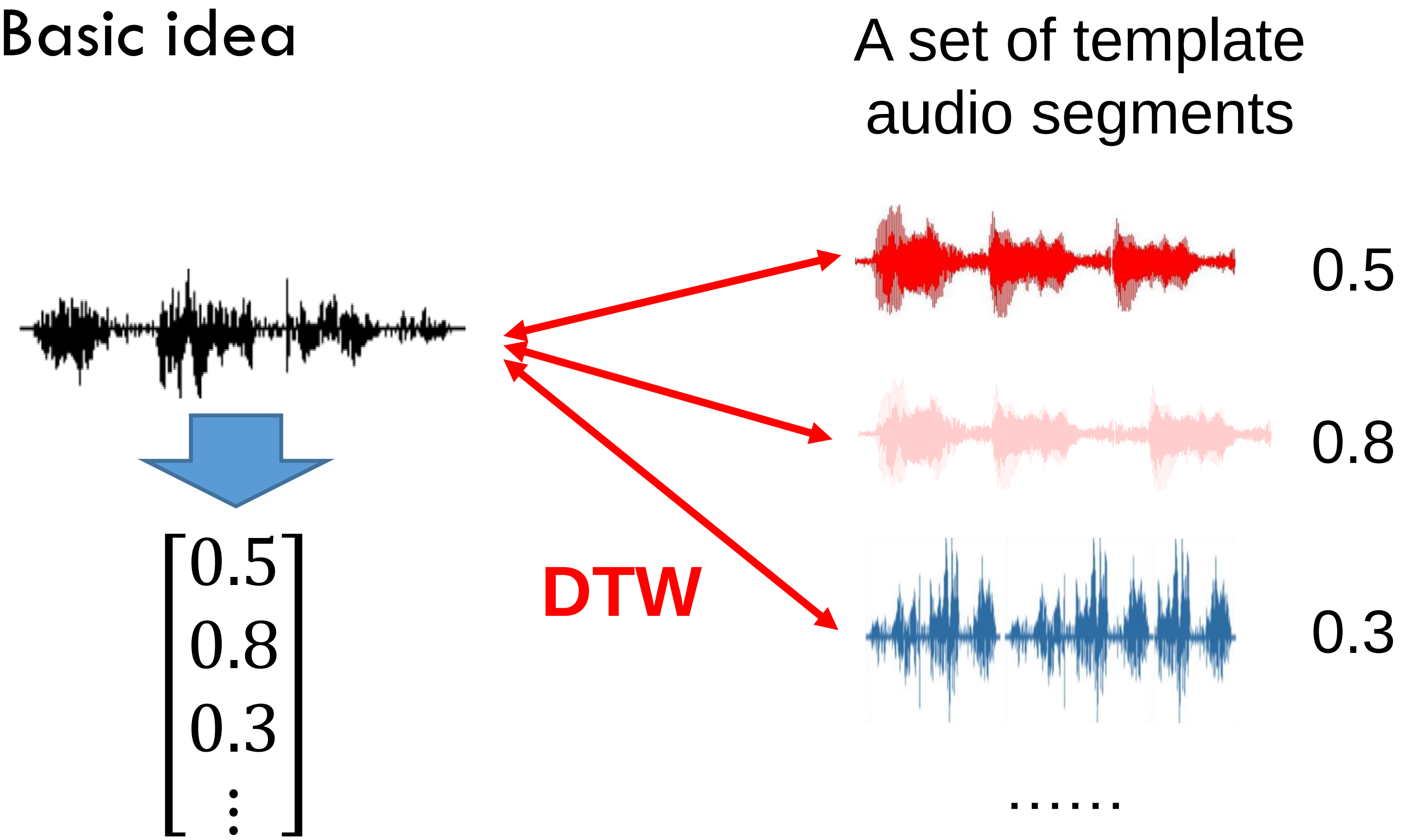
# Audio Word to Vector

- The audio segments corresponding to words with similar pronunciations are close to each other.



# Audio Word to Vector - Segmental Acoustic Indexing

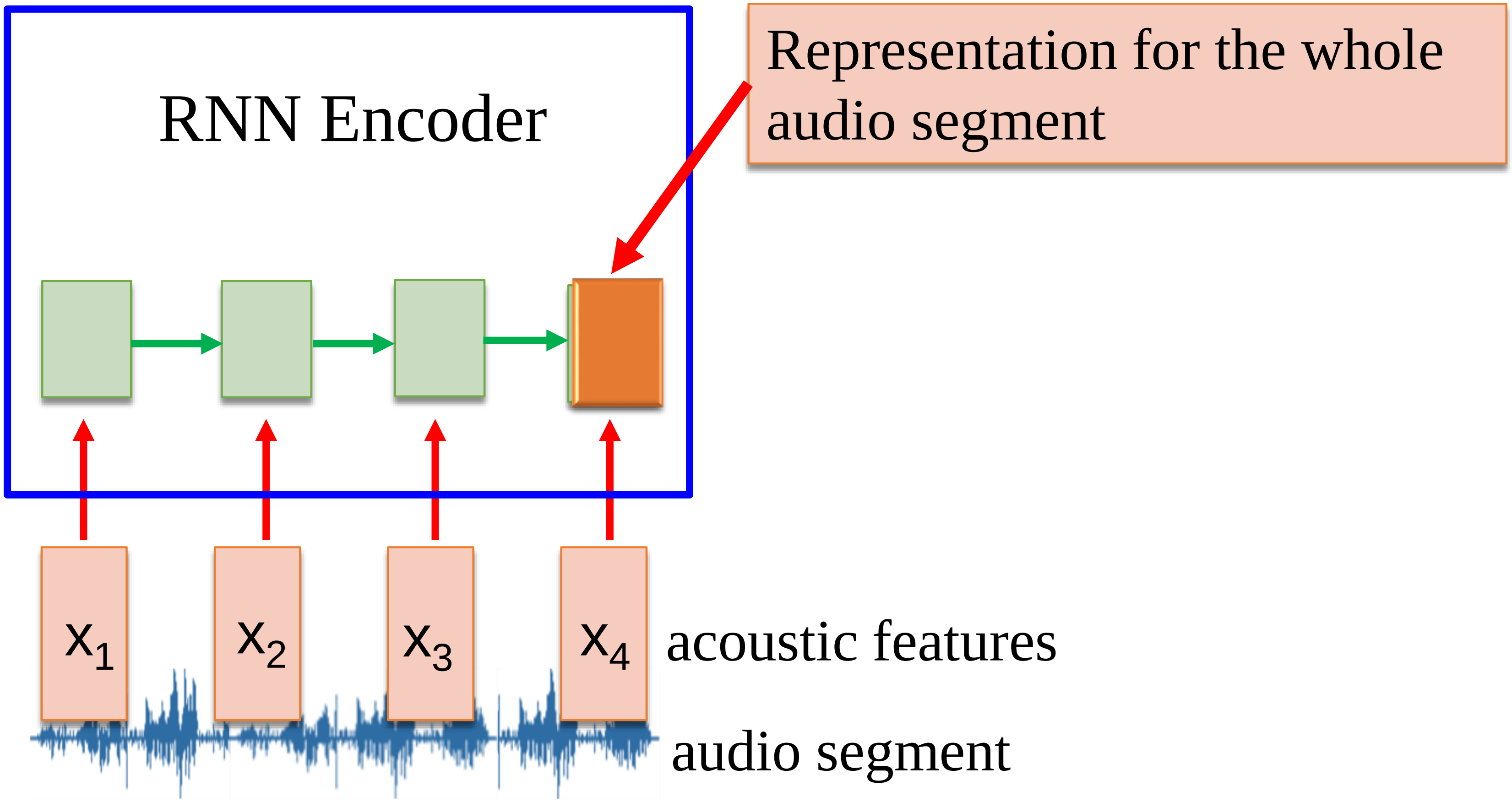
- Basic idea




[Levin, ICASSP 15][Levin, ASRU 13]

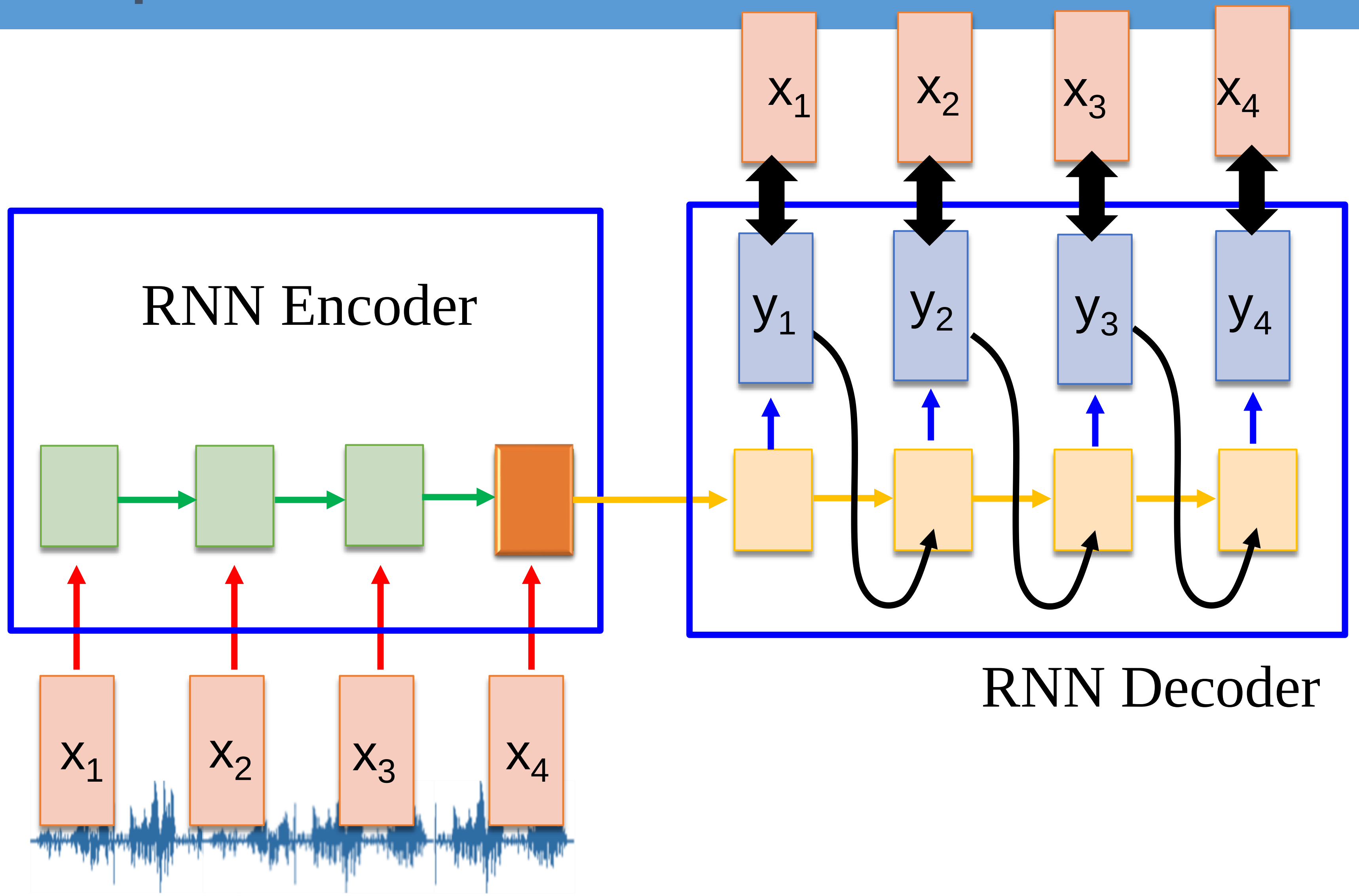
# Audio Word to Vector – Sequence Auto-encoder

[Chung & Lee, Interspeech 16]

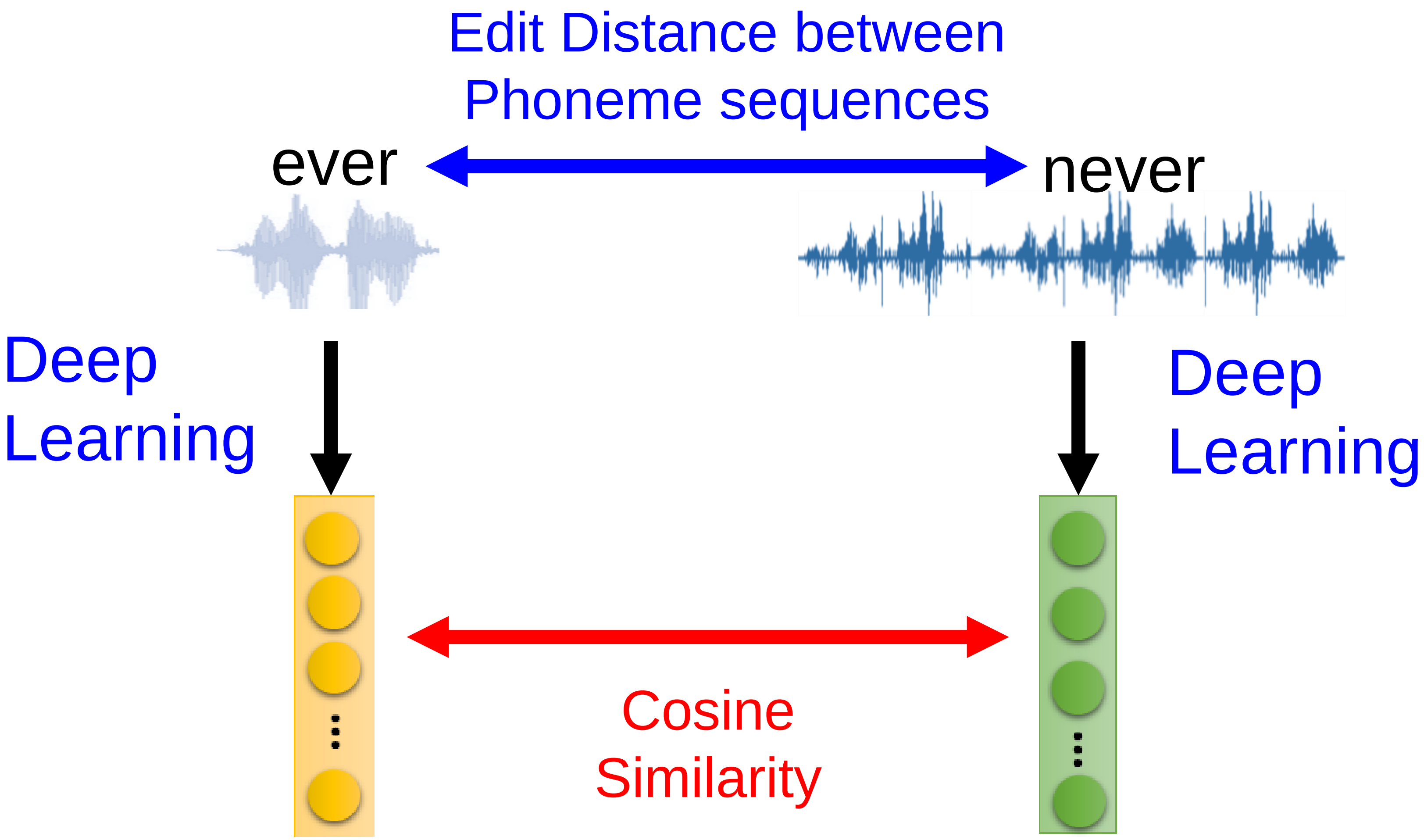


# Audio Word to Vector – Sequence Auto-encoder

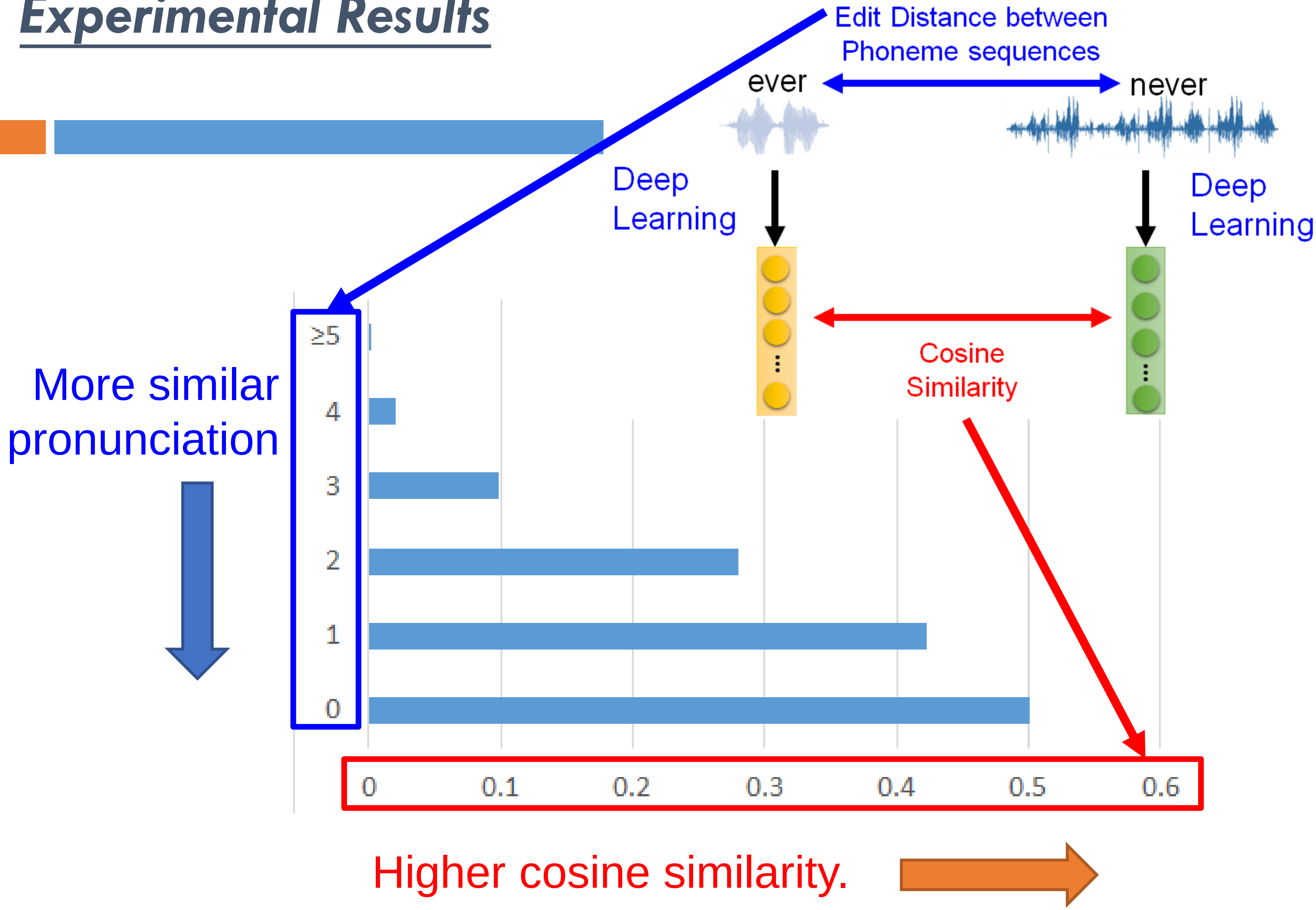
[Chung & Lee, Interspeech 16]



# Sequence Auto-encoder – Experimental Results

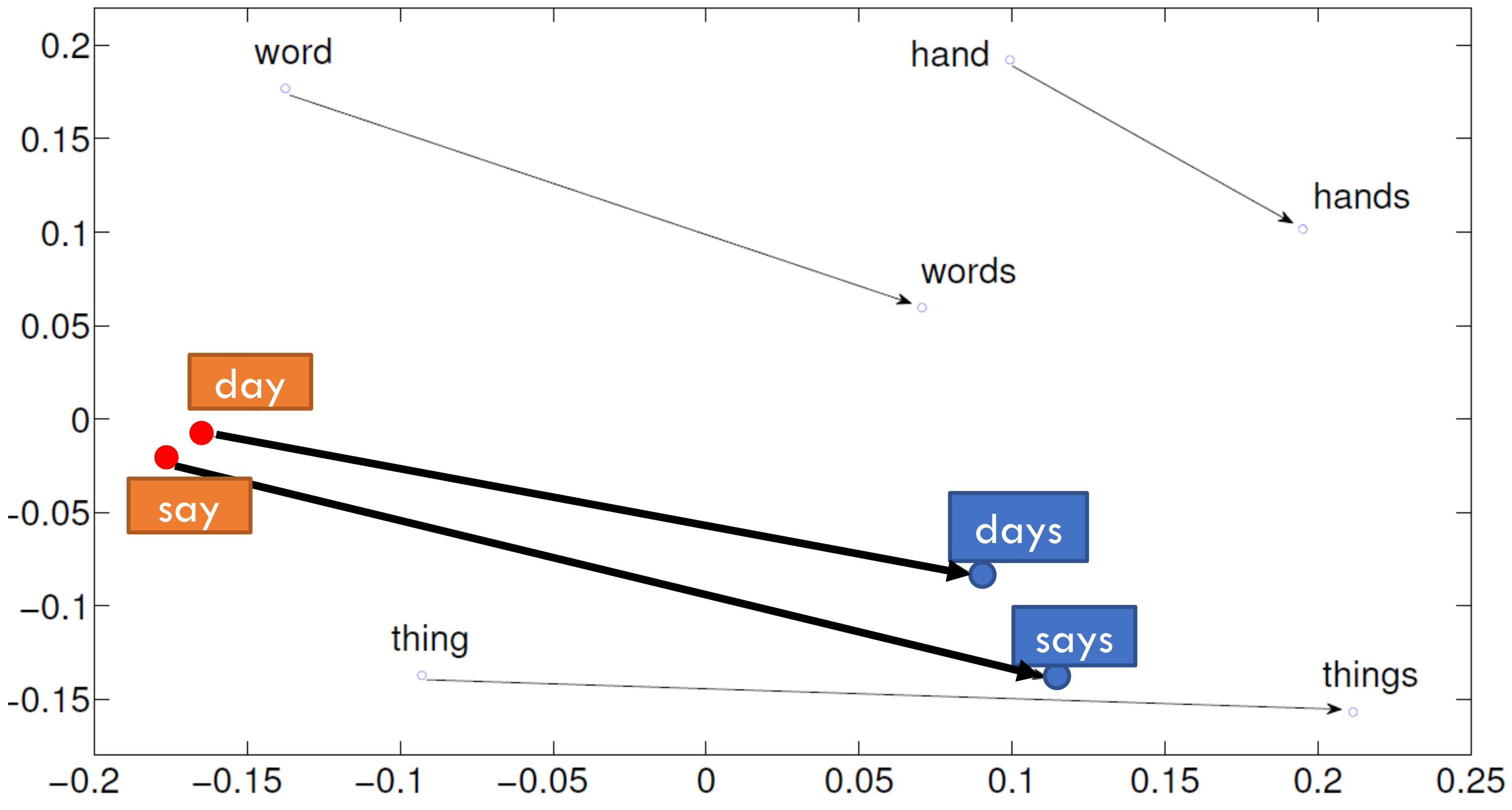


# *Experimental Results*



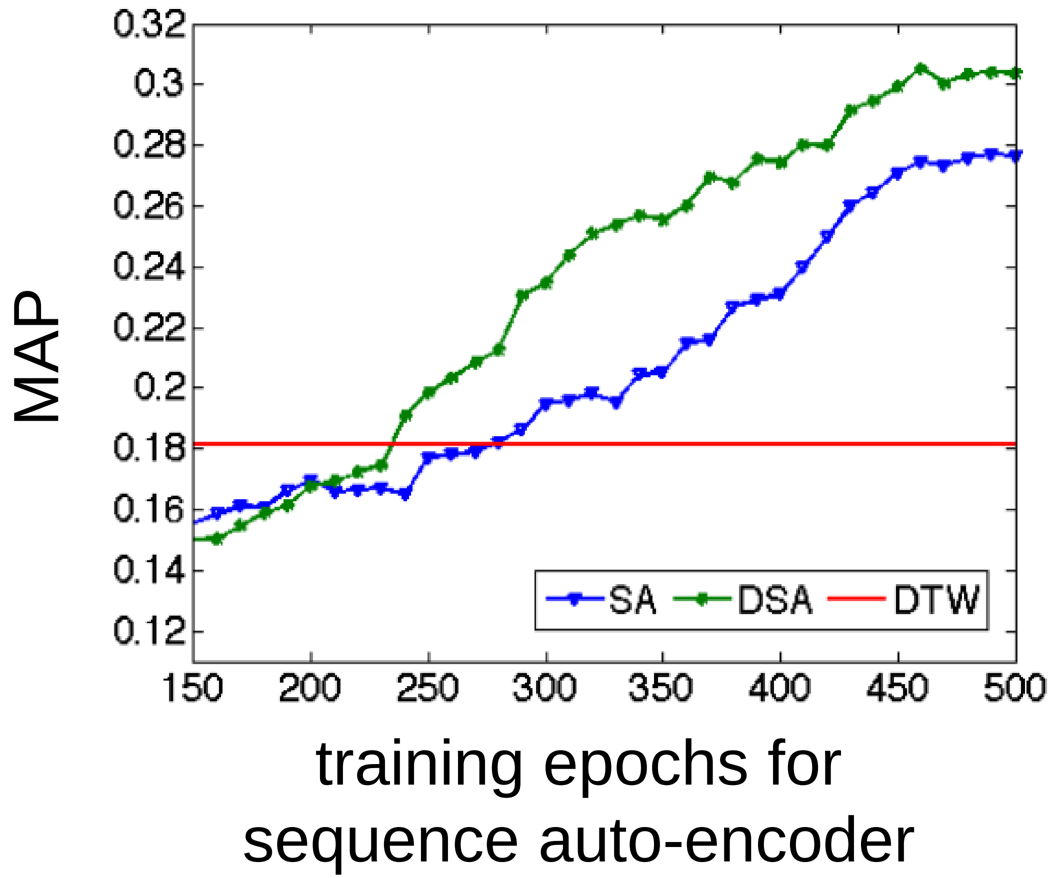
# Sequence Auto-encoder – Experimental Results

- Projecting the embedding vectors to 2-D



# Sequence Auto-encoder – Experimental Results

- Audio story (LibriSpeech corpus)




SA: sequence auto-encoder

DSA: de-noising sequence auto-encoder

**Input**: clean speech + noise

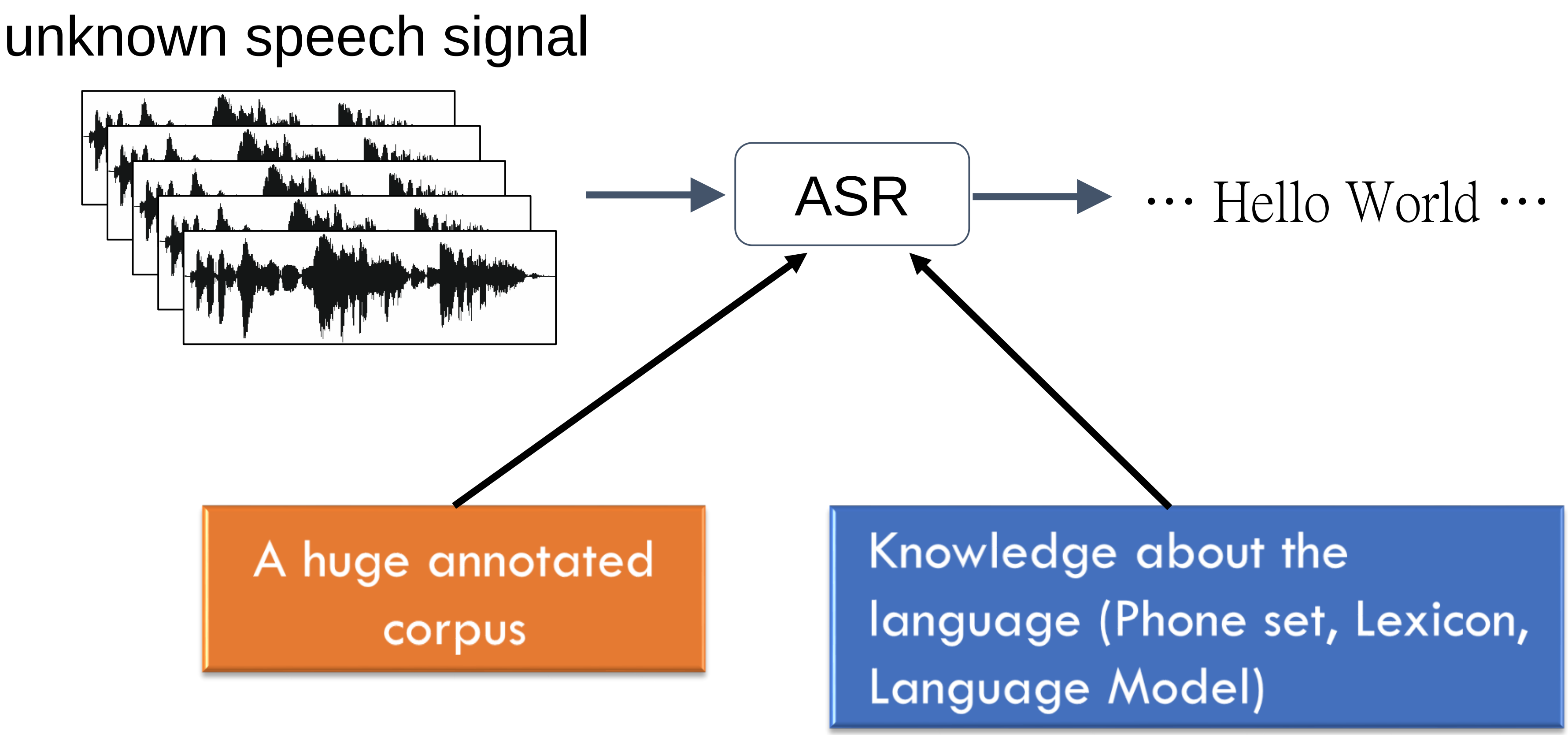
**output**: clean speech

# New Direction 3-3:

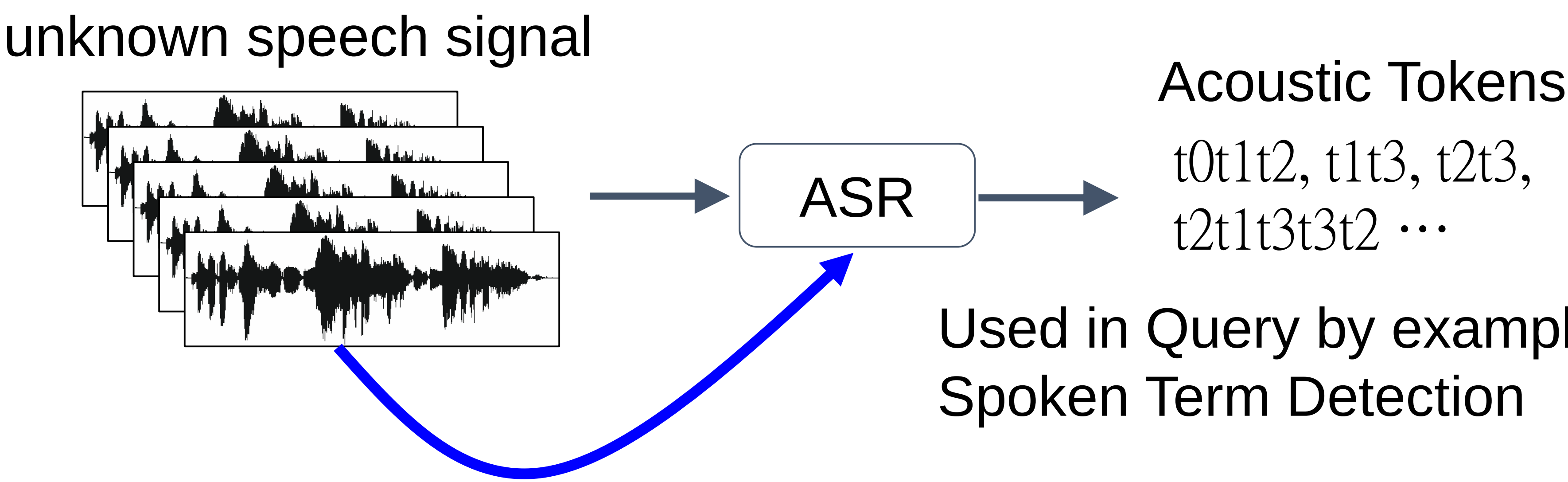
## No Speech Recognition!

### Unsupervised ASR

# Conventional ASR



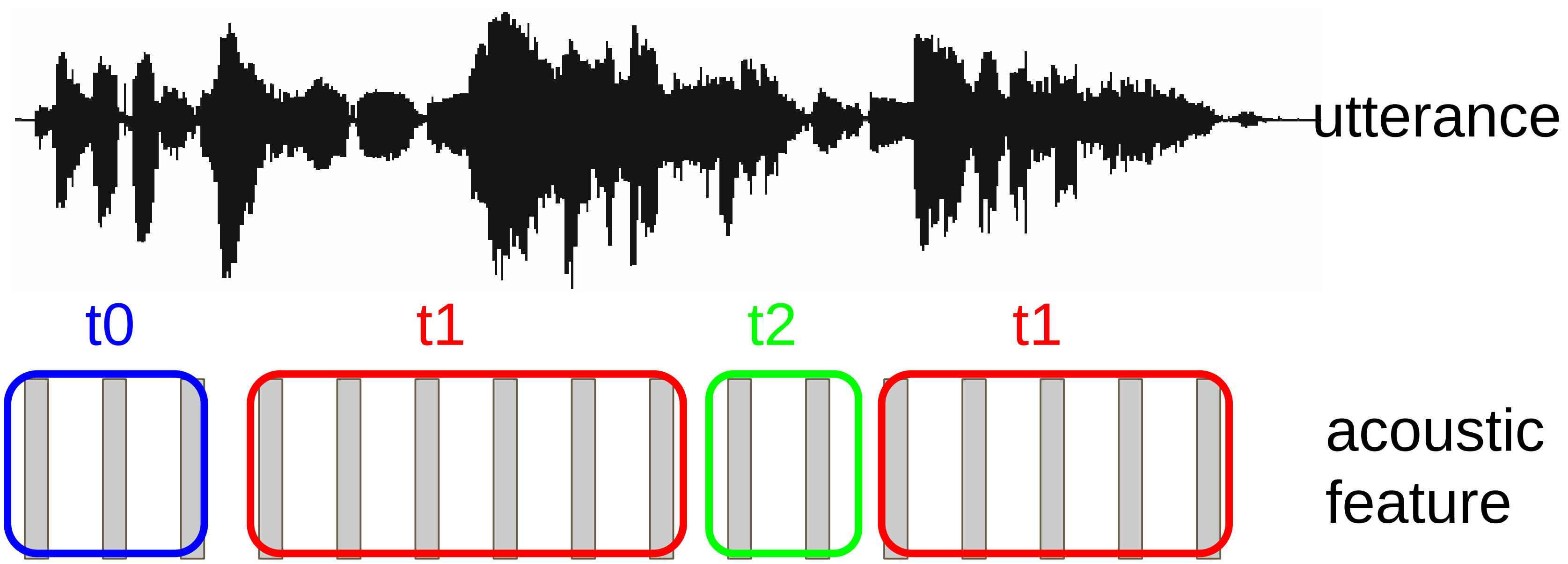
# Unsupervised ASR




Used in Query by example
Spoken Term Detection

Unsupervised ASR:
Learn the models for a set of acoustic patterns (tokens) directly from the corpus (target spoken archive)

# Unsupervised ASR - Acoustic Token




acoustic tokens: chunks of acoustically similar feature vectors with token ids

[Zhang & Glass, ASRU 09]
[Huijbregts, ICASSP 11]
[Chan & Lee, Interspeech 11]

# Unsupervised ASR
## - Overall Framework

$$\boxed{\omega_0 = \text{initialization}(X)}$$

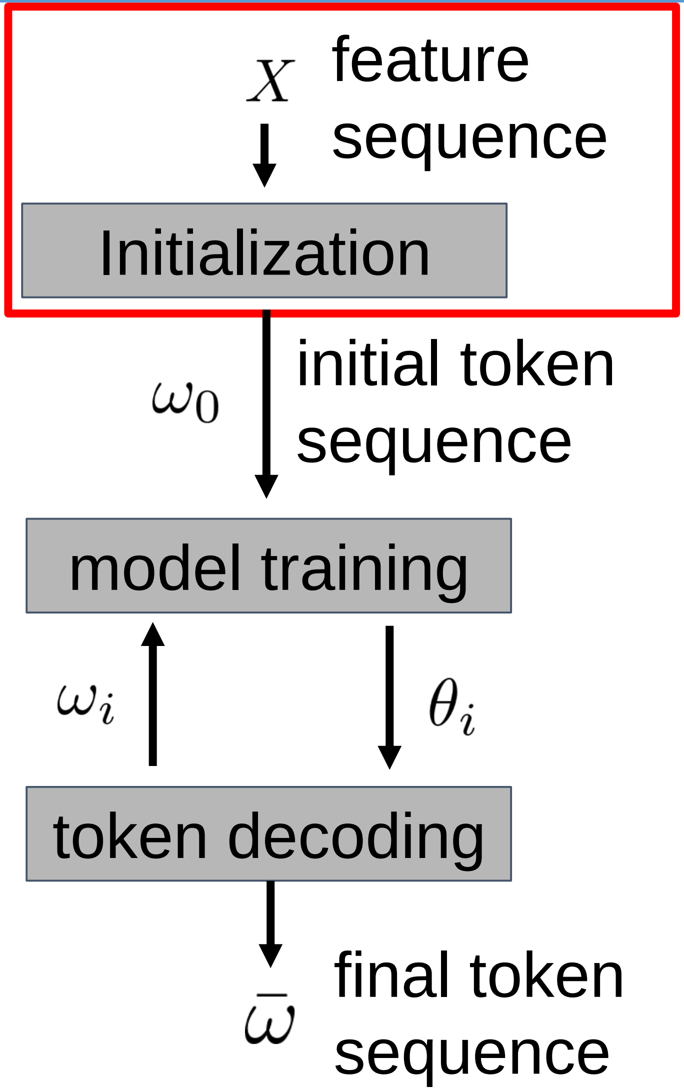
**simple segmentation and clustering**

$$\theta_i = \arg\max_{\theta} P(X|\omega_{i-1}, \theta)$$

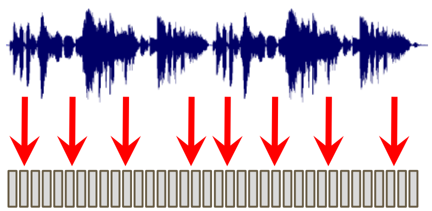
$$\omega_i = \arg\max_{\omega} P(X|\omega, \theta_{i-1})$$

$X$ : feature sequence for the whole corpus

$\omega$ : token sequences for X



# Unsupervised ASR
## - Initialization

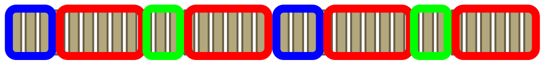


Extract acoustic features for every utterance

Grouping consecutive acoustically similar feature vectors into segments

Extract mean of each segment and perform K-means clustering on the entire archive

Get Token ID



# Unsupervised ASR
## - Overall Framework

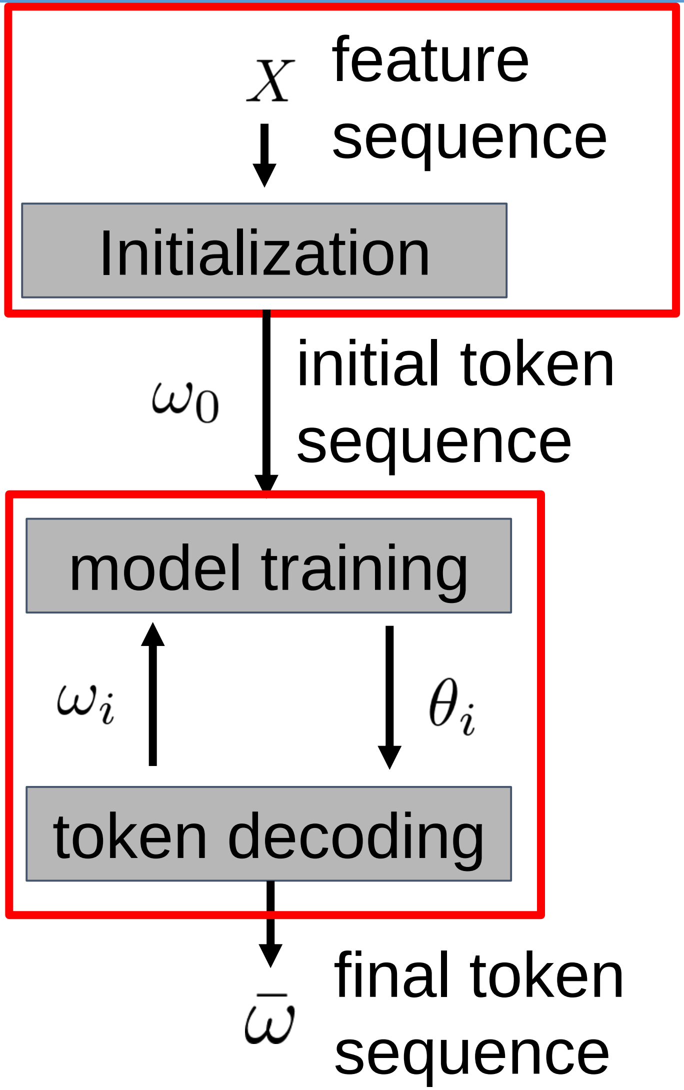
$$\boxed{\omega_0 = \text{initialization}(X)}$$

**simple segmentation and clustering**

$$\theta_i = \arg\max_{\theta} P(X|\omega_{i-1}, \theta)$$

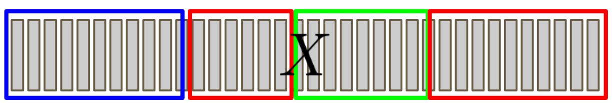
$$\omega_i = \arg\max_{\omega} P(X|\omega, \theta_{i-1})$$

$X$: feature sequence for the whole corpus
$\theta$ : Model (e.g. HMM) parameters
$\omega$ : token sequences for X
$i$ : training iteration



# Unsupervised ASR
## - Overall Framework

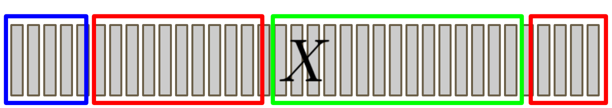
optimize HMM parameters using Baum–Welch algorithm on token sequence $\omega_{i-1}$ to get new models $\theta_i$



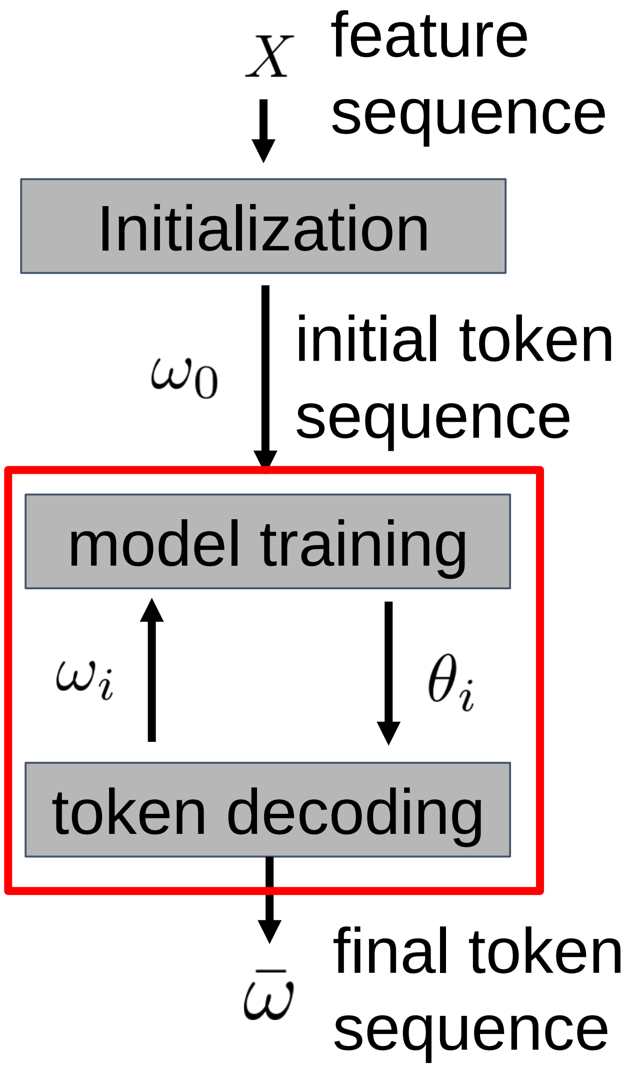

$\omega_{i-1}$

$$\theta_i = \arg\max_{\theta} P(X|\omega_{i-1}, \theta)$$

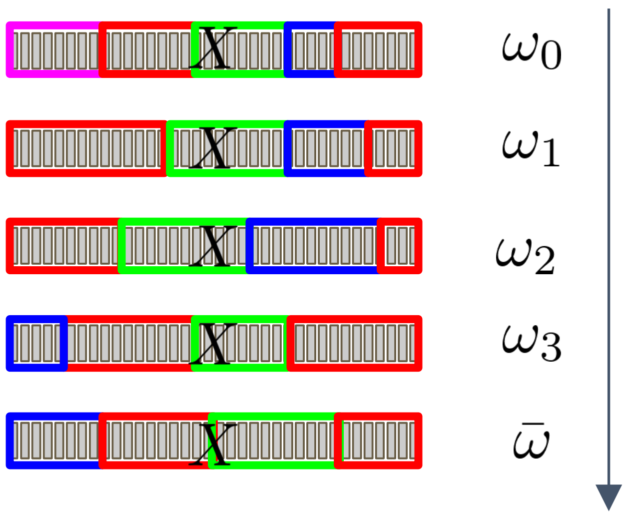
decode acoustic features into a new token sequence $\omega_i$ using Viterbi decoding



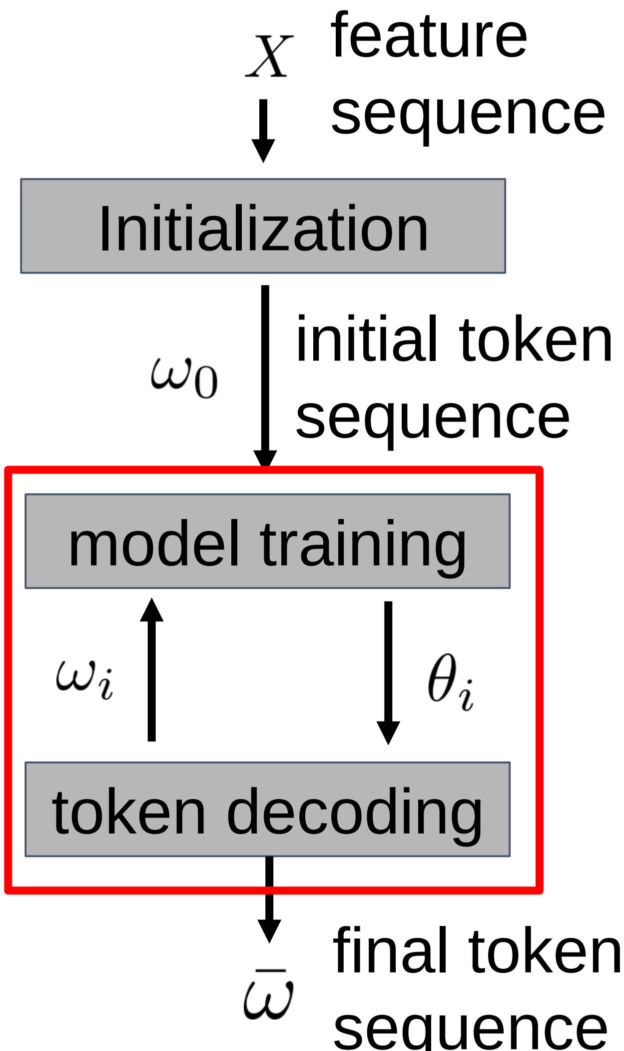

$$\omega_i = \arg\max_{\omega} P(X|\omega, \theta_{i-1})$$



# Unsupervised ASR
## - Overall Framework

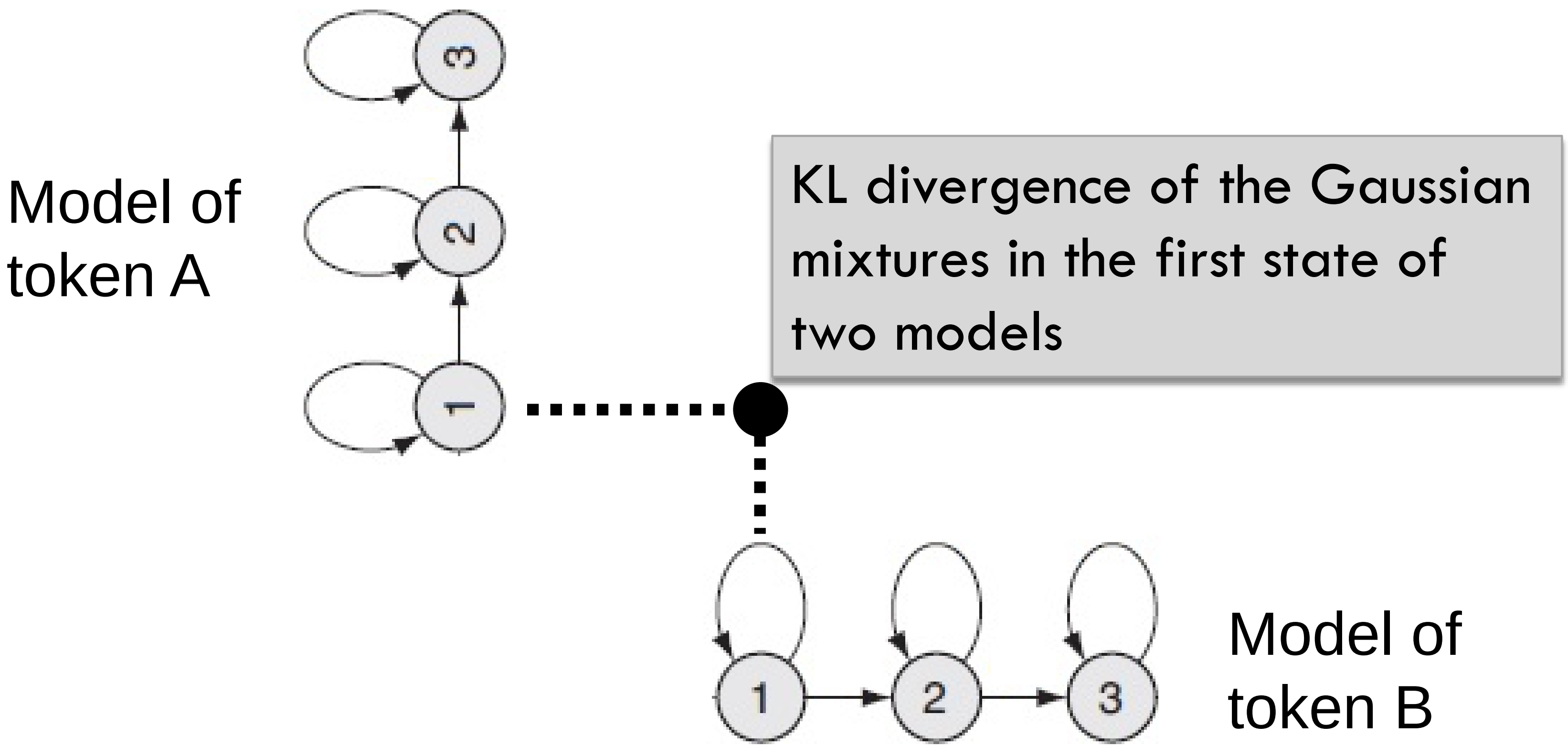



iterate until the token sequences (including token boudaries) converge



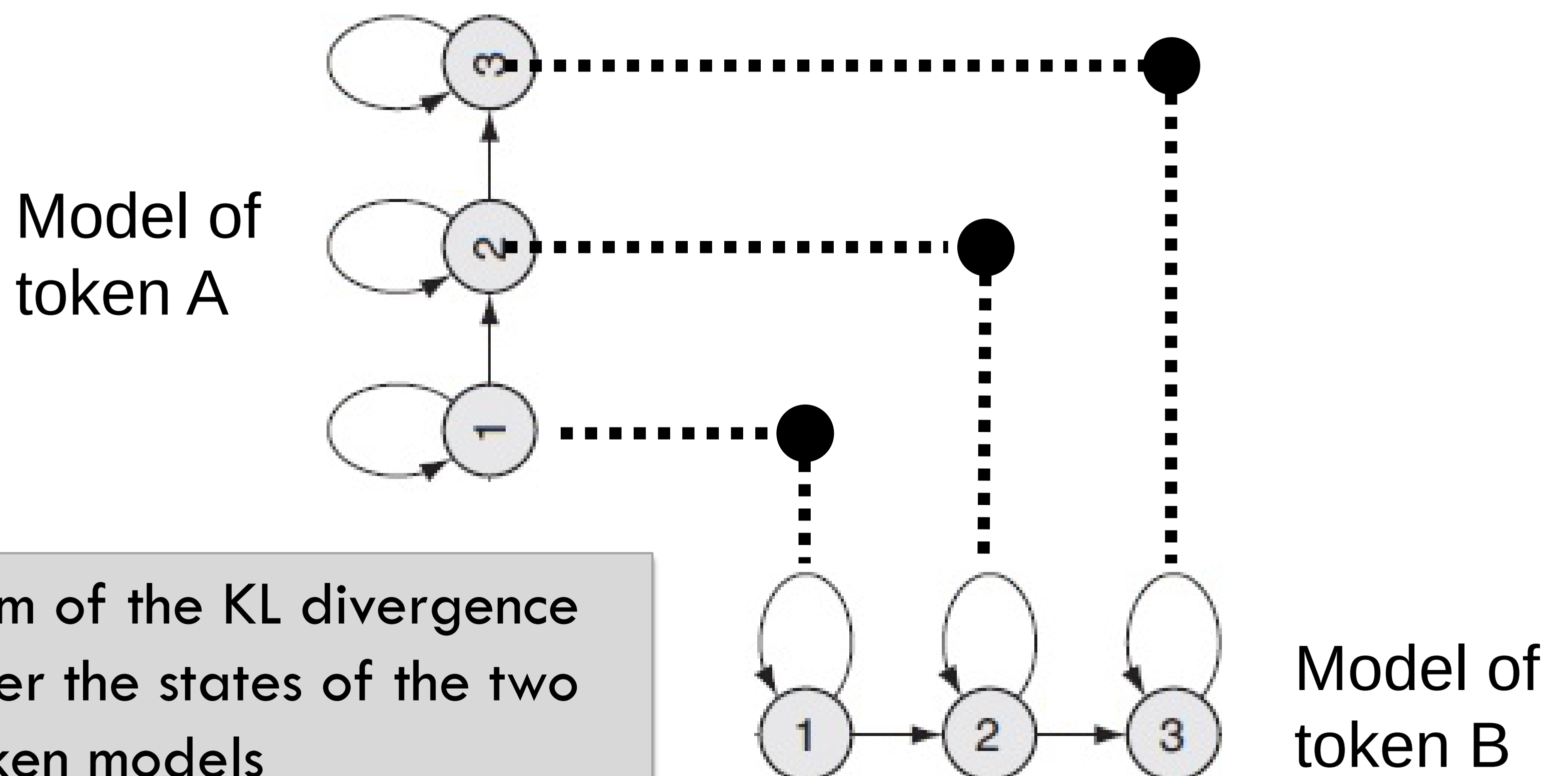
# Acoustic Token in Query by Example Spoken Term Detection

- Compute the similarity between the models of two tokens



# Acoustic Token in Query by Example Spoken Term Detection

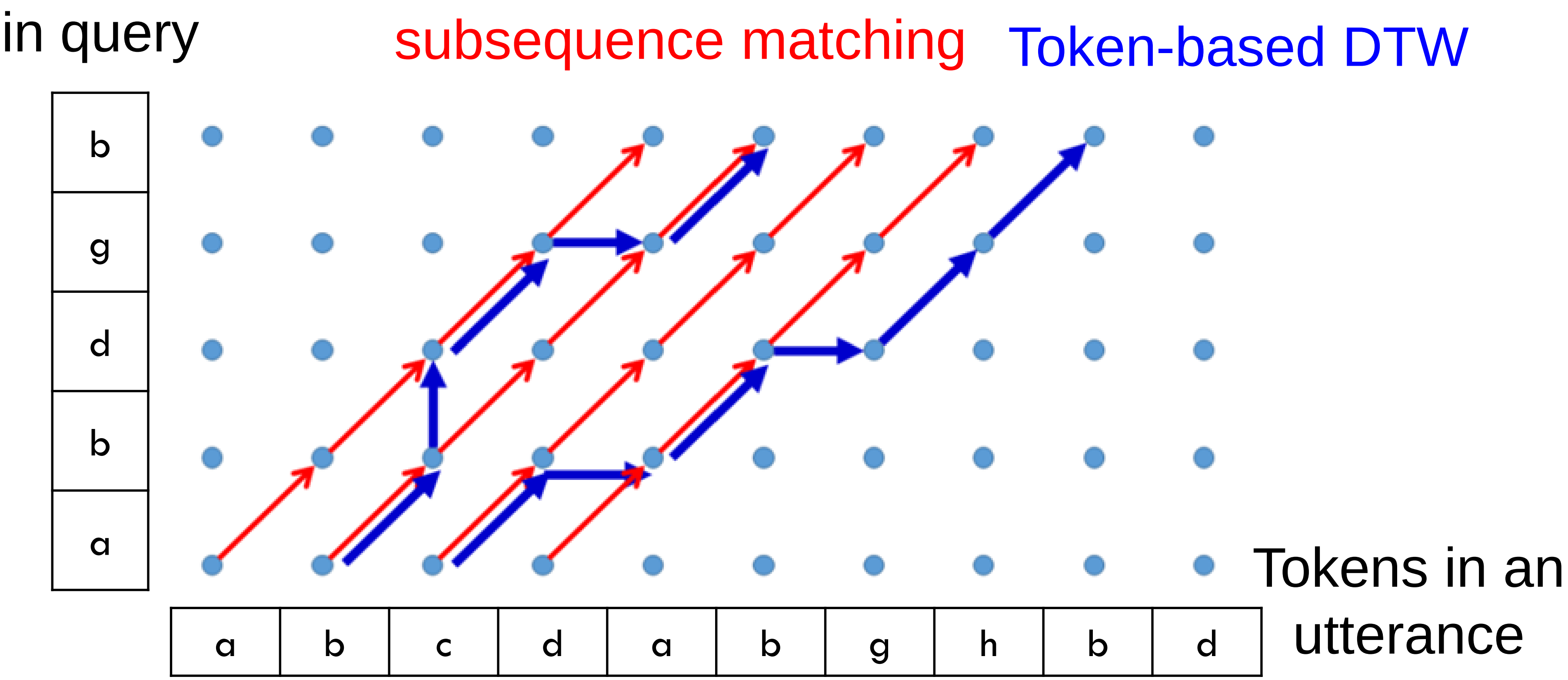
- Compute the similarity between the models of two tokens




Sum of the KL divergence over the states of the two token models

# Token-based DTW

- Tokens




- ➢ Signal-level DTW is more sensitive to signal variation (e.g. same phoneme across different speakers), while token models are able to cover better the distribution of signal variation
- ➢ Much lower on-line computation load

# Multi-granularity Space for Acousitc Tokens

- Unknown hyperparameters for the token models
  - Number of HMM states per token (m): token length
  - Number of distinct tokens (n)
- Multiple layers of intrinsic representations of speech

# Multi-granularity Space for Acousitc Tokens

- From short to long (Temporal Granularity)
  - phoneme
  - syllable
  - word
  - phrase

Number of states per HMM (m)

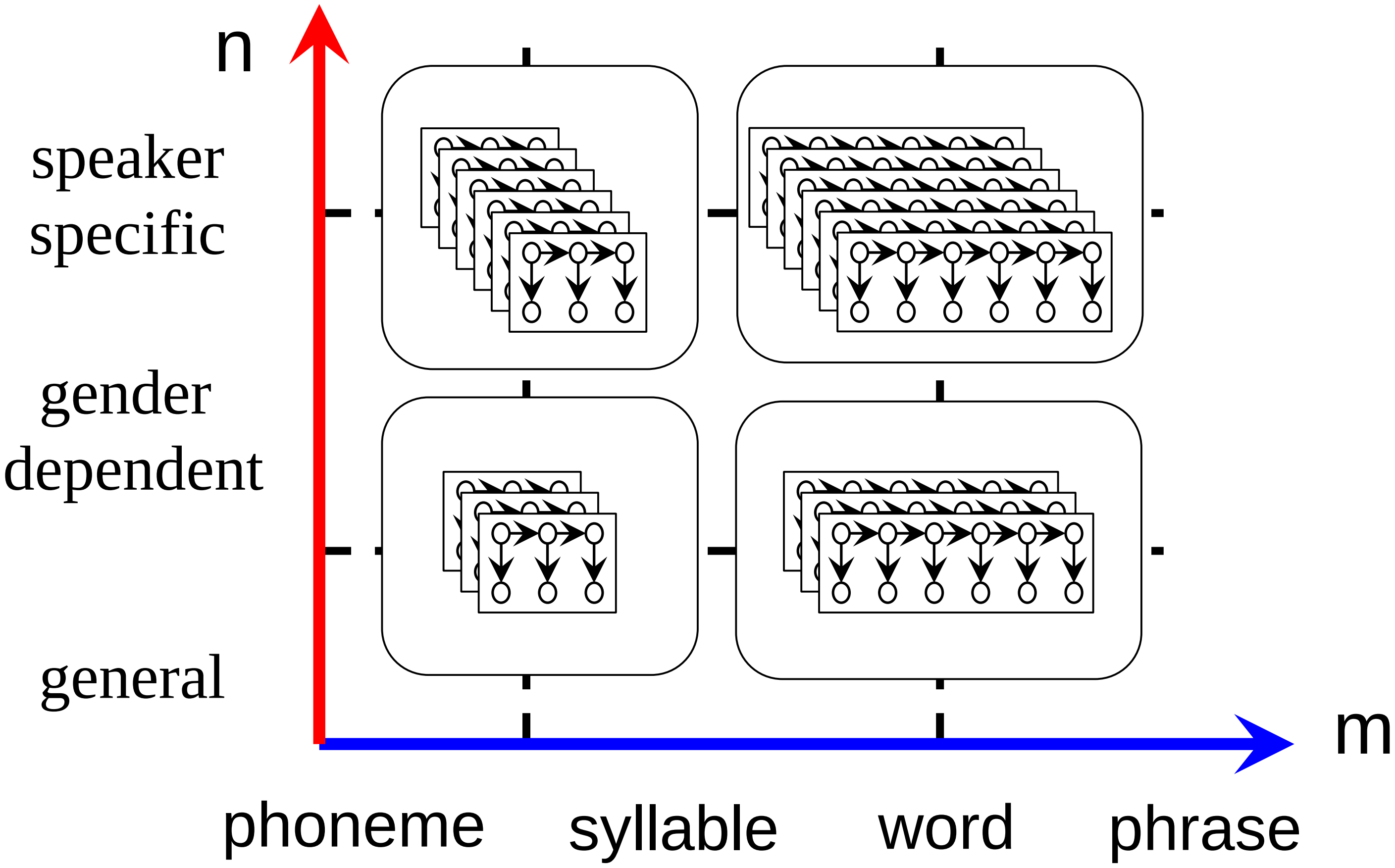
- From coarse units to fine units (Phonetic Granularity)
  - general phoneme set
  - gender dependent phoneme set
  - speaker specific phoneme set

Number of distinct HMMs (n)

# Multi-granularity Space for Acousitc Tokens

Training multiple sets of HMMs for with different granularity

[Chung & Lee, ICASSP 14 ]

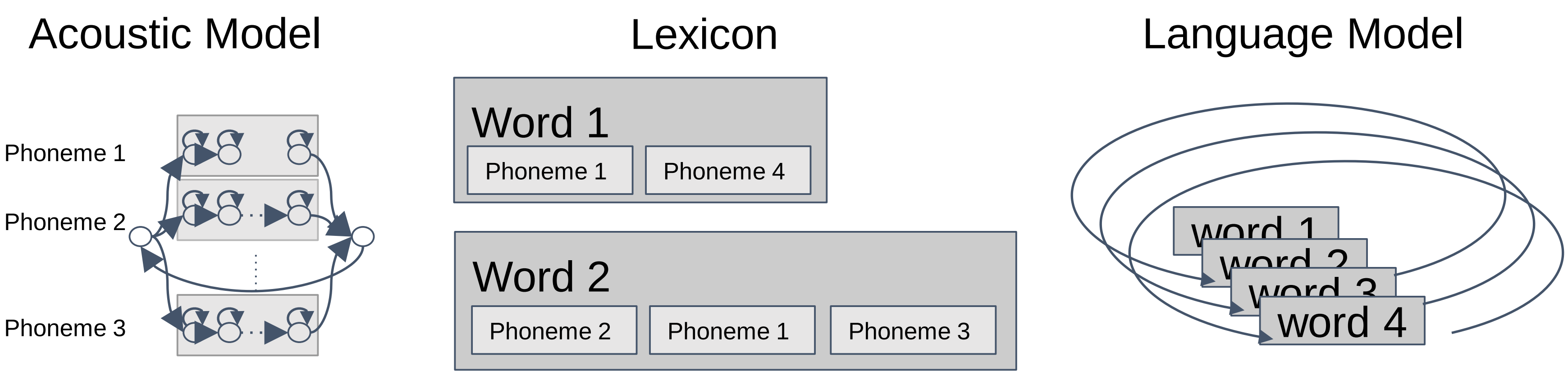


# Multi-granularity Space for Acousitc Tokens

- Token-based DTW using tokens with different granularity (m,n) averaged gave much better performance
- One example
  - Frame-level DTW: MAP = 10%
  - Using only the token set with the best performance: MAP = 11%
  - Using 20 sets of tokens (number of states per HMM m = 3,5,7,9,11, number of distinct HMMs n=50,100,200,300): MAP = 26%

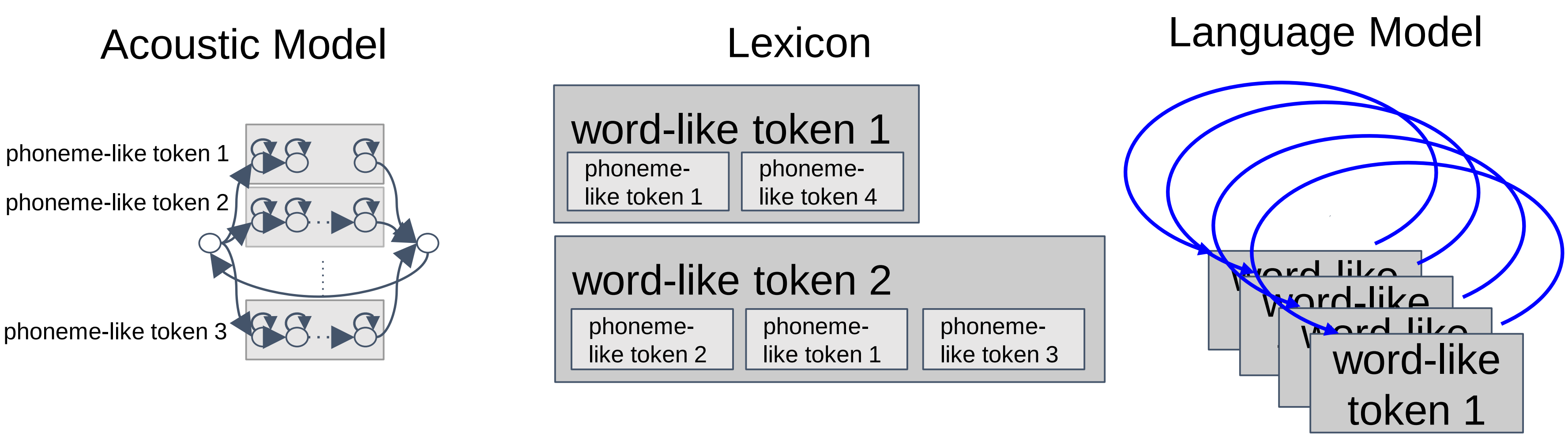
# Hierarchical Paradigm

- Typical ASR:
  - Acoustic Model: models for the phonemes
  - Lexicon: the pronunciation of every word as a phoneme sequence
  - Language Model: the transition between words



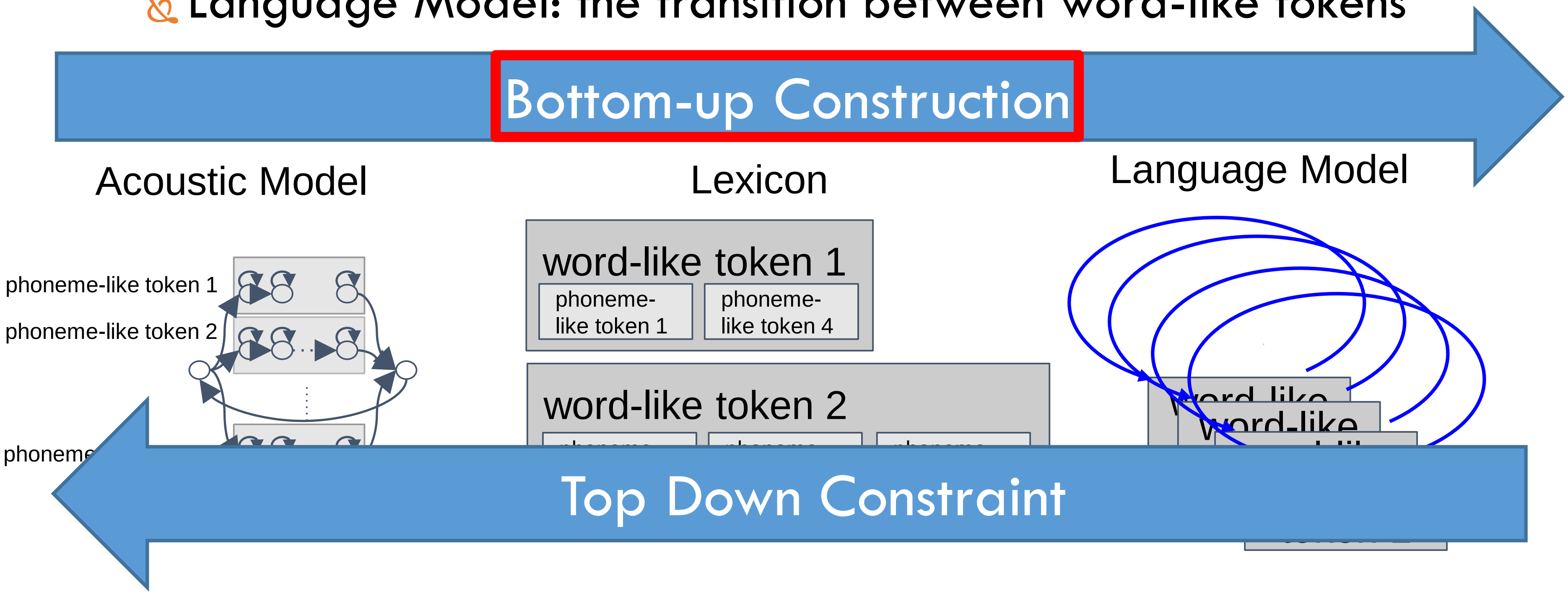
# Hierarchical Paradigm

- Similarly, in unsupervised ASR:
  - Acoustic Model: the phoneme-like token HMMs
  - Lexicon: the pronunciation of every word-like token as a sequence of phoneme-like tokens
  - Language Model: the transition between word-like tokens



# Hierarchical Paradigm

- Similarly, in unsupervised ASR:
  - Acoustic Model: the phoneme-like token HMMs
  - Lexicon: the pronunciation of every word-like token as a sequence of phoneme-like tokens
  - Language Model: the transition between word-like tokens



# Bottom Up Construction

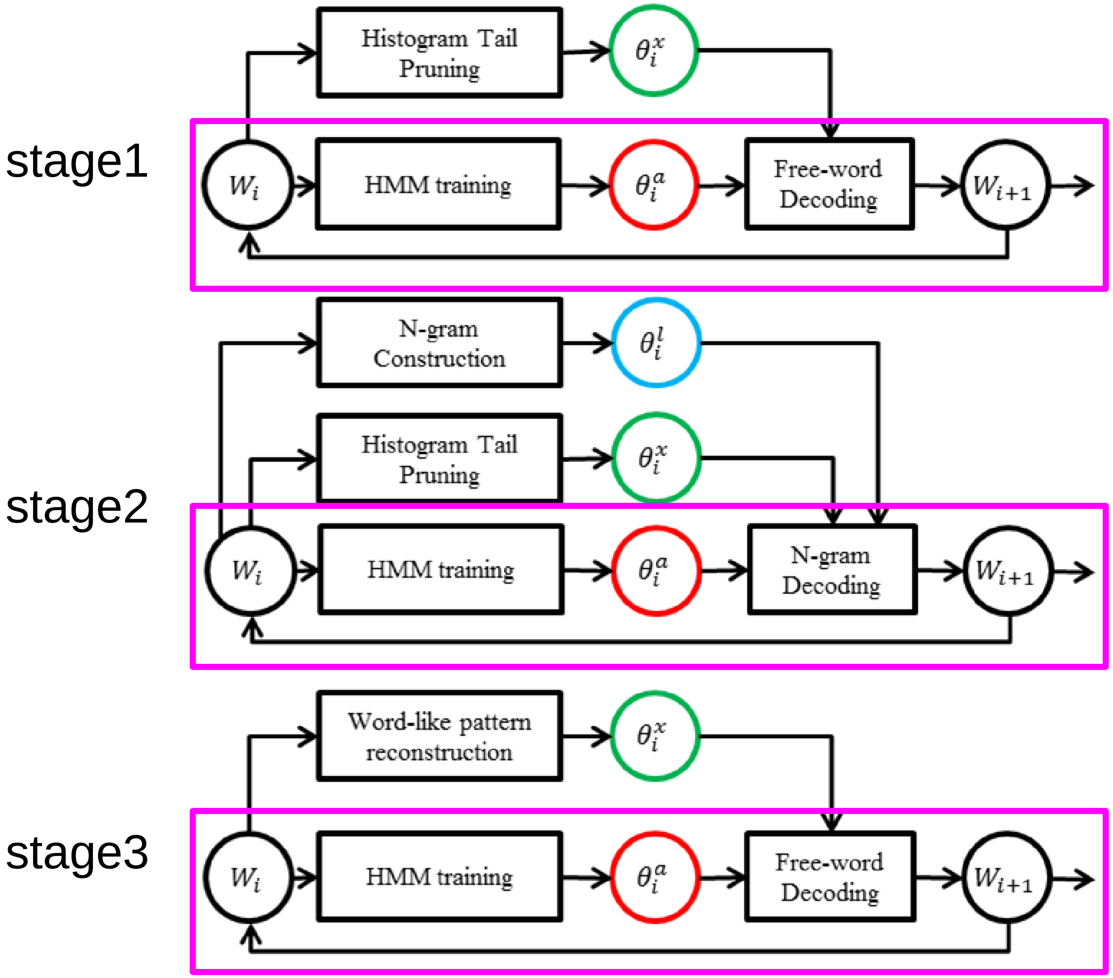
3 stages during training
focus on different constraints:

stage1: Acoustic Model $\theta_i^a$

stage2: Language Model $\theta_i^l$

stage3: Lexicon $\theta_i^x$

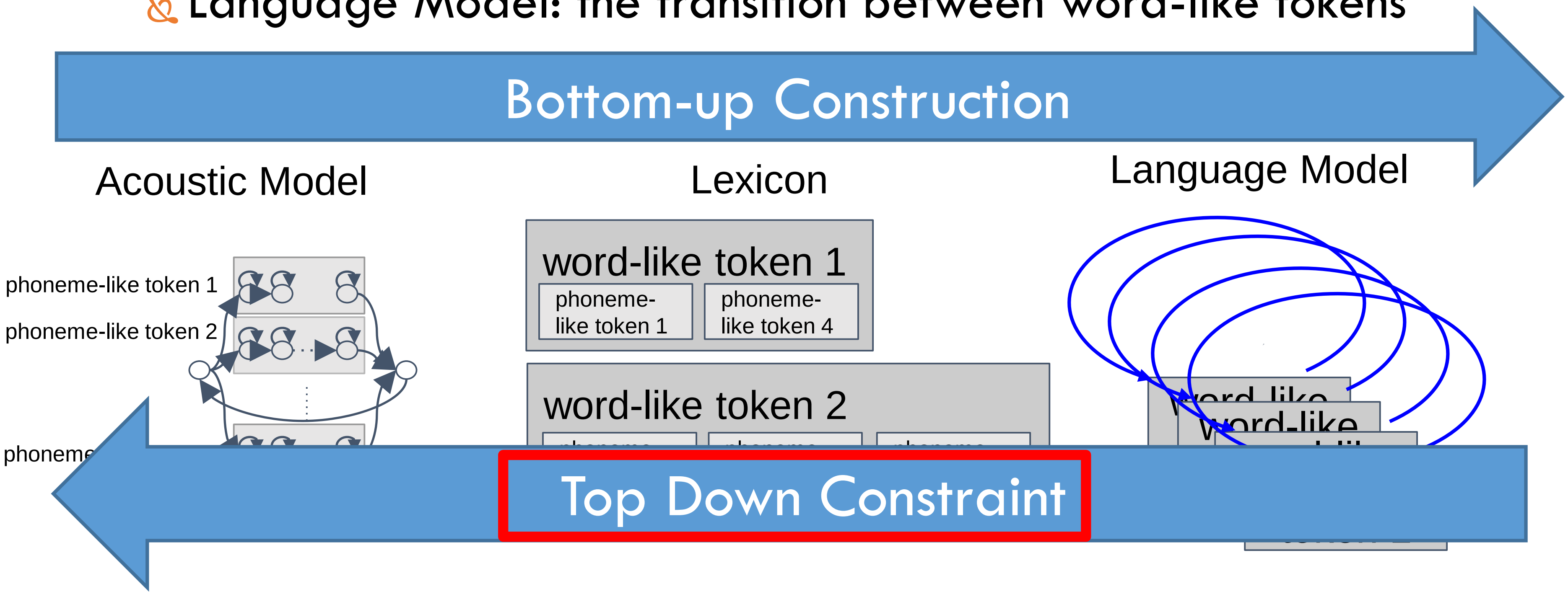
this part alone would be the HMM training we described earlier




[Chung & Lee, ICASSP 13]

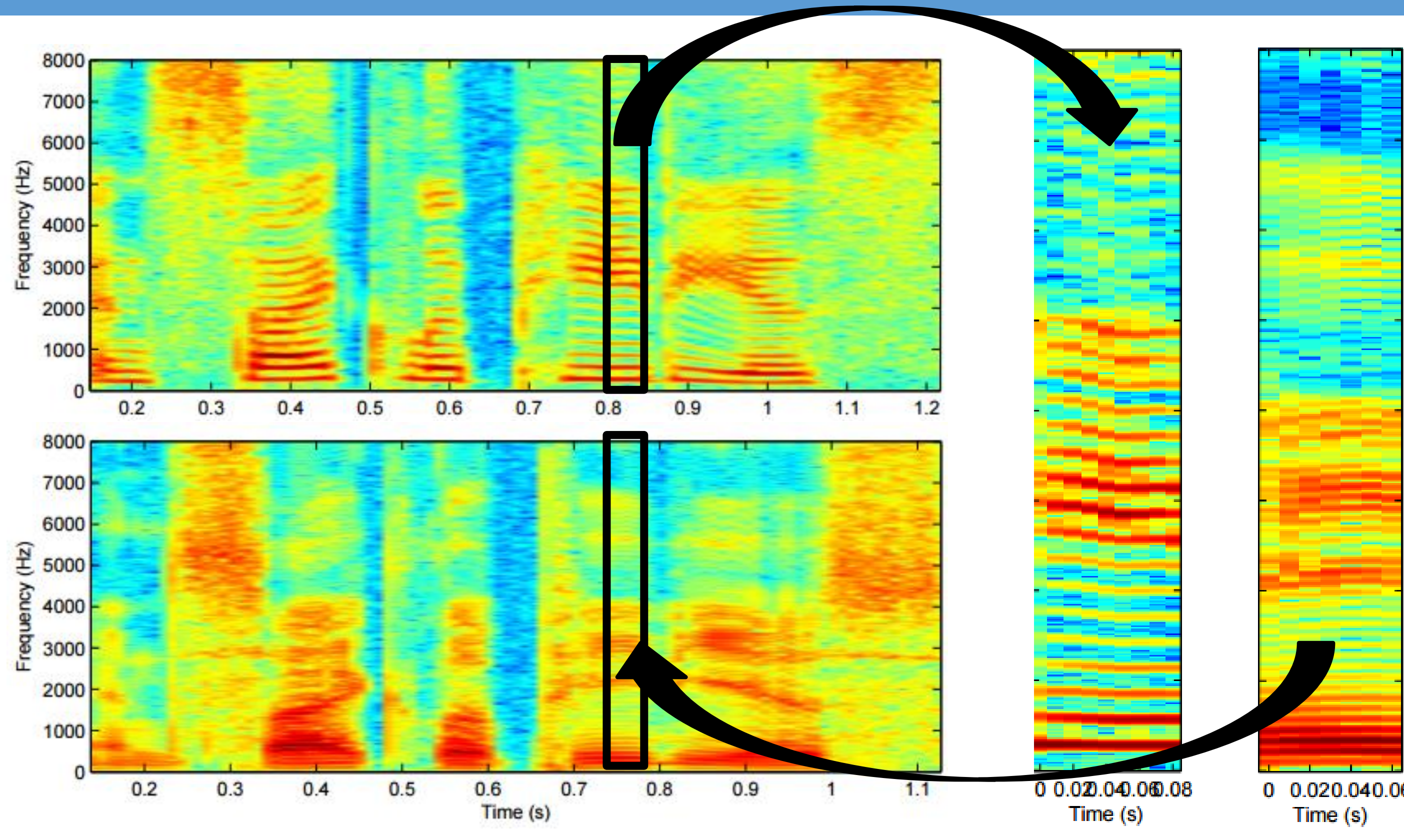
# Hierarchical Paradigm

- Similarly, in unsupervised ASR:
  - Acoustic Model: the phoneme-like token HMMs
  - Lexicon: the pronunciation of every word-like token as a sequence of phoneme-like tokens
  - Language Model: the transition between word-like tokens



# Top-down Constraints

This figure is from Aren Jansen's ICASSP paper. [Jansen, ICASSP 13]



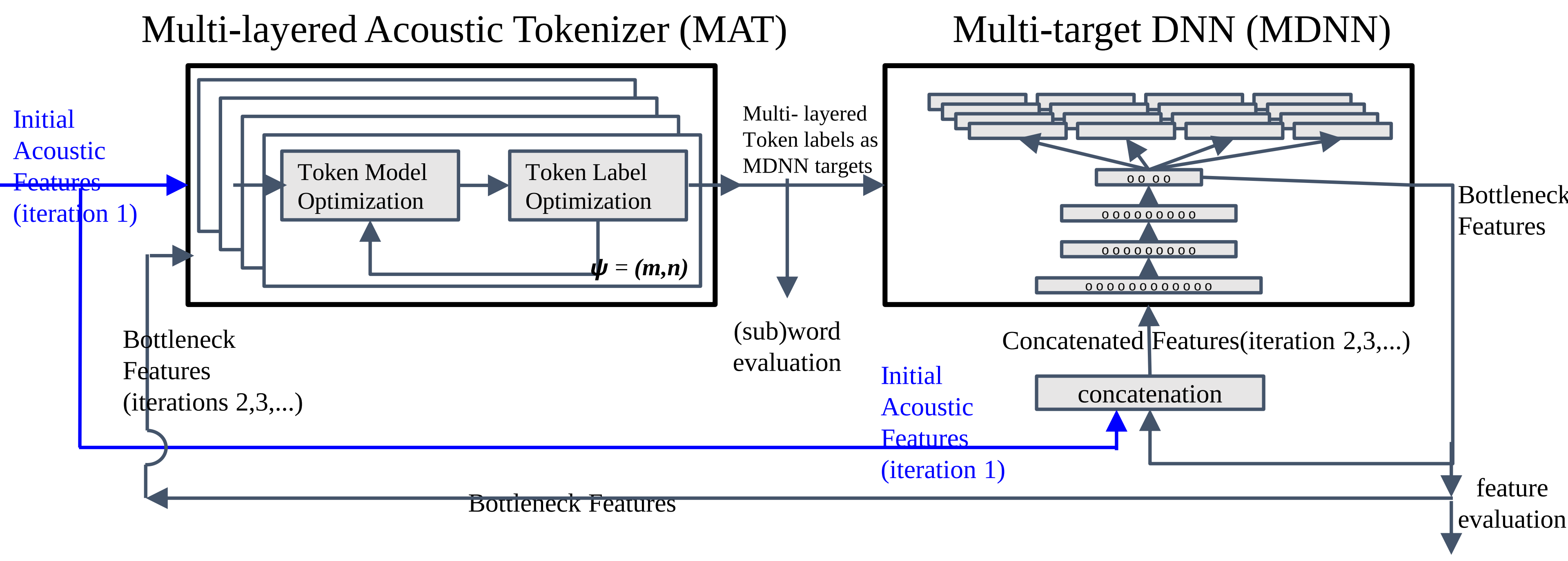

- Signals of the same phoneme may be very different on phoneme level, but the global structures of signals of the same word are very often very similar on word level
- Global structures help in building the hierarchical model

# Multi-layered Acoustic Tokenizing Deep Neural Networks (MAT-DNN)

[Chung & Lee, ASRU 15]

- Jointly learn high quality frame-level features (much better than MFCCs) and acoustic tokens in an unsupervised way
- Unsupervised training of multi-target DNN using unsupervised token labels as training target

In the first iteration, we use MFCC as the initial features
In the other iterations, we concatenate the bottleneck features with the MFCC

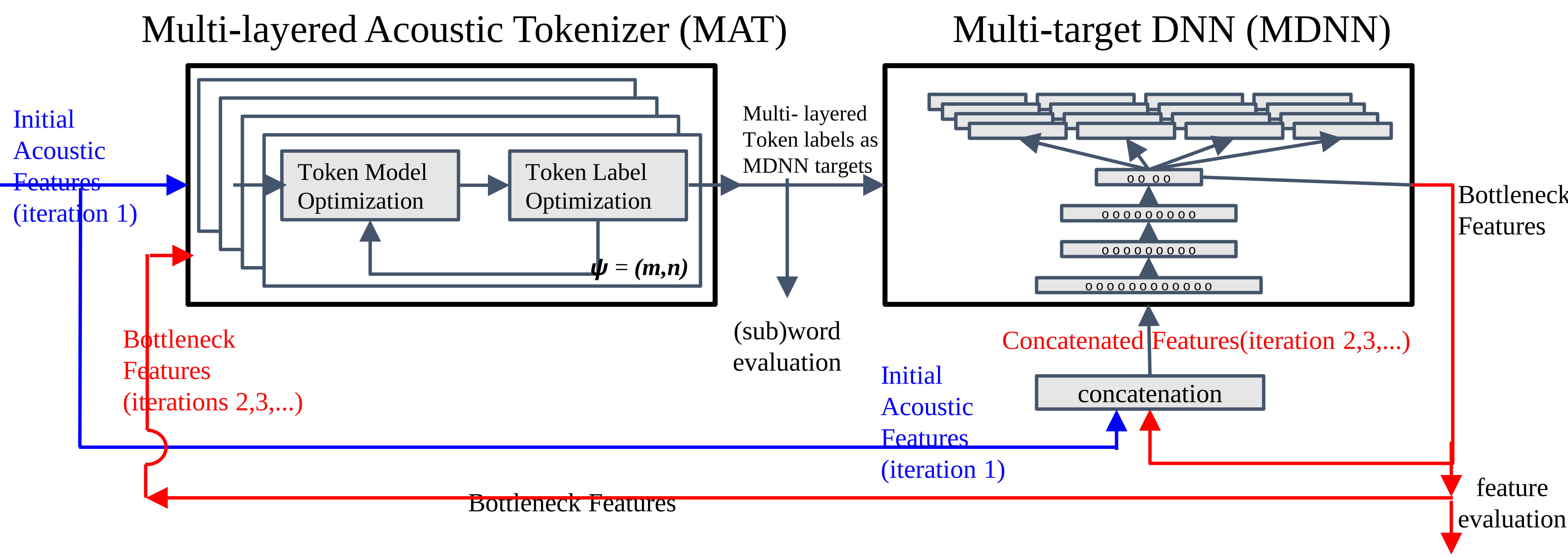


# Multi-layered Acoustic Tokenizing Deep Neural Networks (MAT-DNN)

[Chung & Lee, ASRU 15]

- Jointly learn high quality frame-level features (much better than MFCCs) and acoustic tokens in an unsupervised way
- Unsupervised training of multi-target DNN using unsupervised token labels as training target

In the first iteration, we use MFCC as the initial features
In the other iterations, we concatenate the bottleneck features with the MFCC



# Multi-layered Acoustic Tokenizing Deep Neural Networks (MAT-DNN)

[Chung & Lee, ASRU 13]

- Experimental Results
  - Query by Example Spoken Term Detection on Tsonga

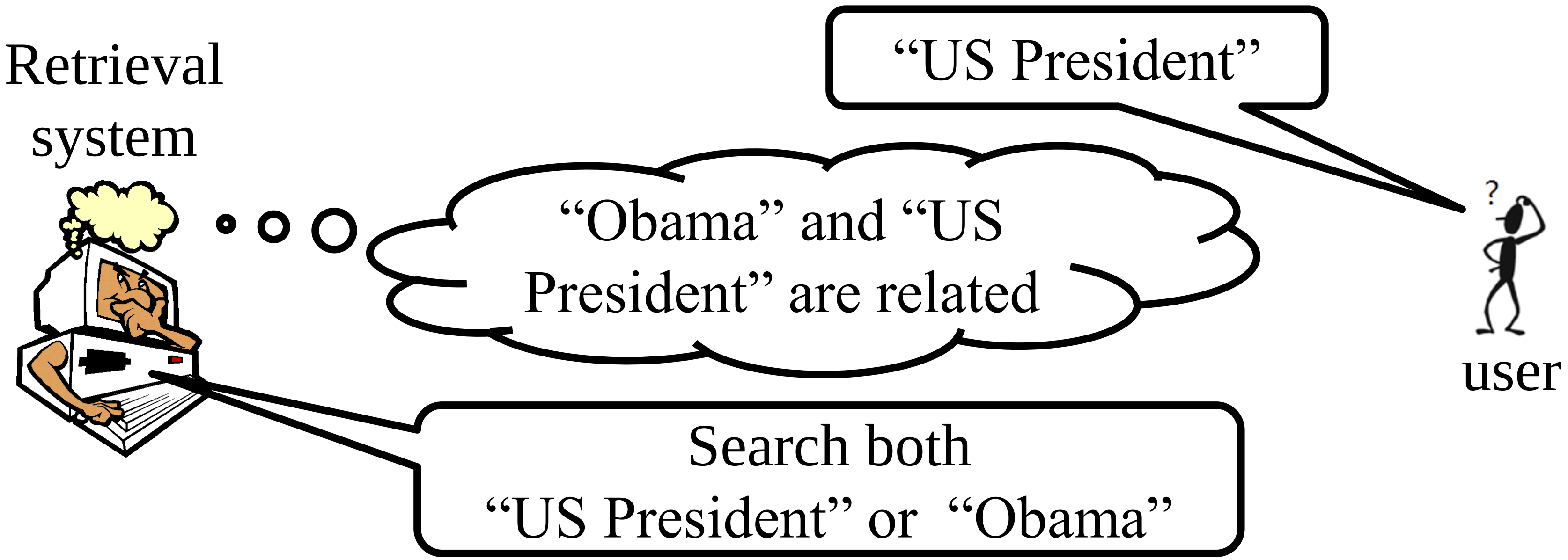
| Approach | | MAP |
|---|---|---|
| Frame-based DTW | MFCC | 9.0 |
| | New Feature | 28.7 |
| Token-based DTW | New Tokens | 26.2 |

# New Direction 4:

## Special Semantic Retrieval Techniques for Spoken Content

# Semantic Retrieval

- User expects semantic retrieval of spoken content.
  - User enters “US President”, system also finds “Obama”
- Widely studies on text retrieval
  - Take ***query expansion*** as example

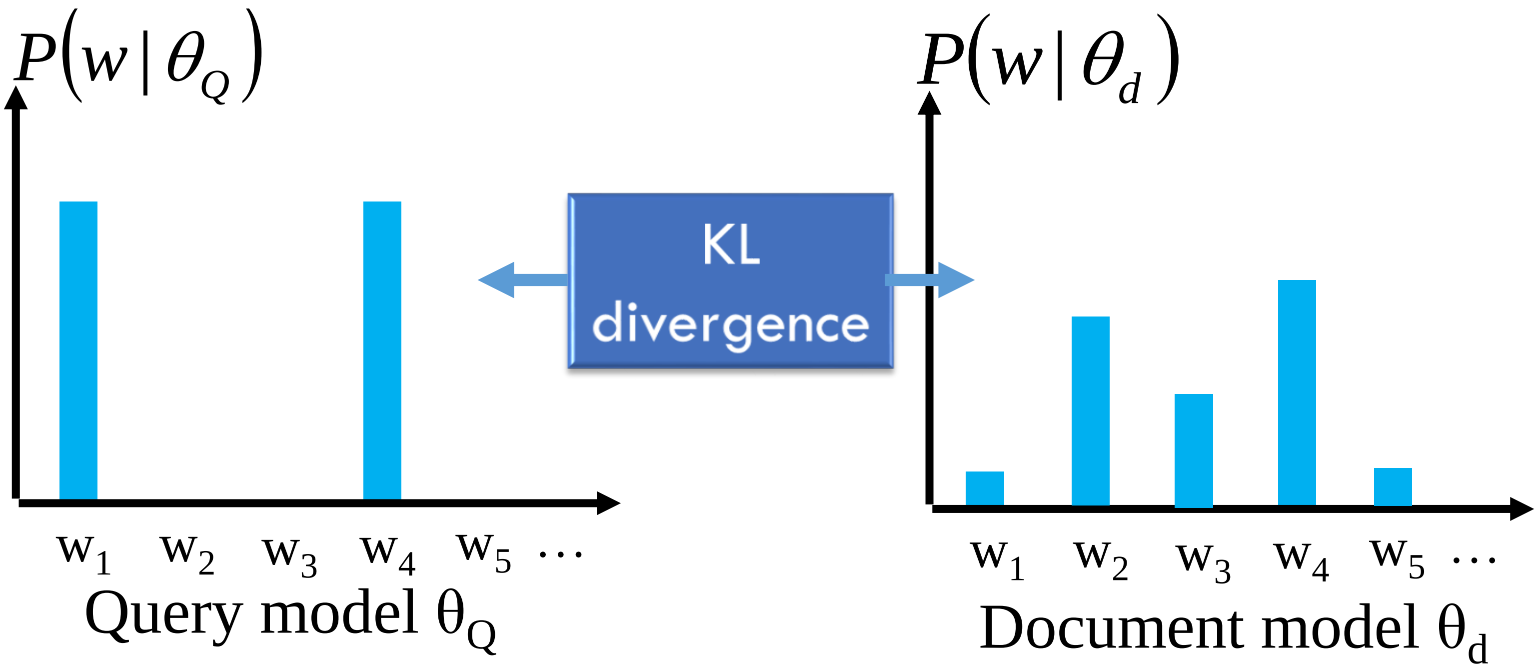


# Semantic Retrieval

- User expects semantic retrieval of spoken content.
  - User enters “US President”, system also finds “Obama”
- Widely studies on text retrieval
  - Take ***query expansion*** as example
- The semantic retrieval techniques developed for text can be directly applied on spoken content
- Query/document expansion based on language modeling retrieval approach as example

# Review: Language Modeling Retrieval Approach

- Both query Q and document d are represented as unigram language models $\theta_Q$ and $\theta_d$




KL divergence between the two models can be evaluated.

# Review: Language Modeling Retrieval Approach

- Given query Q, rank document d according to a relevance score function $S_{LM}(Q,d)$:

$$S_{LM}\left(Q,d\right) = -KL\left(\theta_Q \mid \theta_d\right)$$

  - Inverse of KL divergence between query model $\theta_Q$ and document model $\theta_d$
  - The documents with document models $\theta_d$ similar to query model $\theta_Q$ are more likely to be relevant.

# Review: Basic Query/Document Models in Text Retrieval

- Query model $\theta_Q$ for text:

$$P(w \mid \theta_Q) = \frac{\boxed{N(w,Q)}}{\boxed{\sum_{w'} N(w,Q)}}$$

N(w,Q): term frequency of word w in query Q
Normalize into probability

- Document model $\theta_d$ for text :

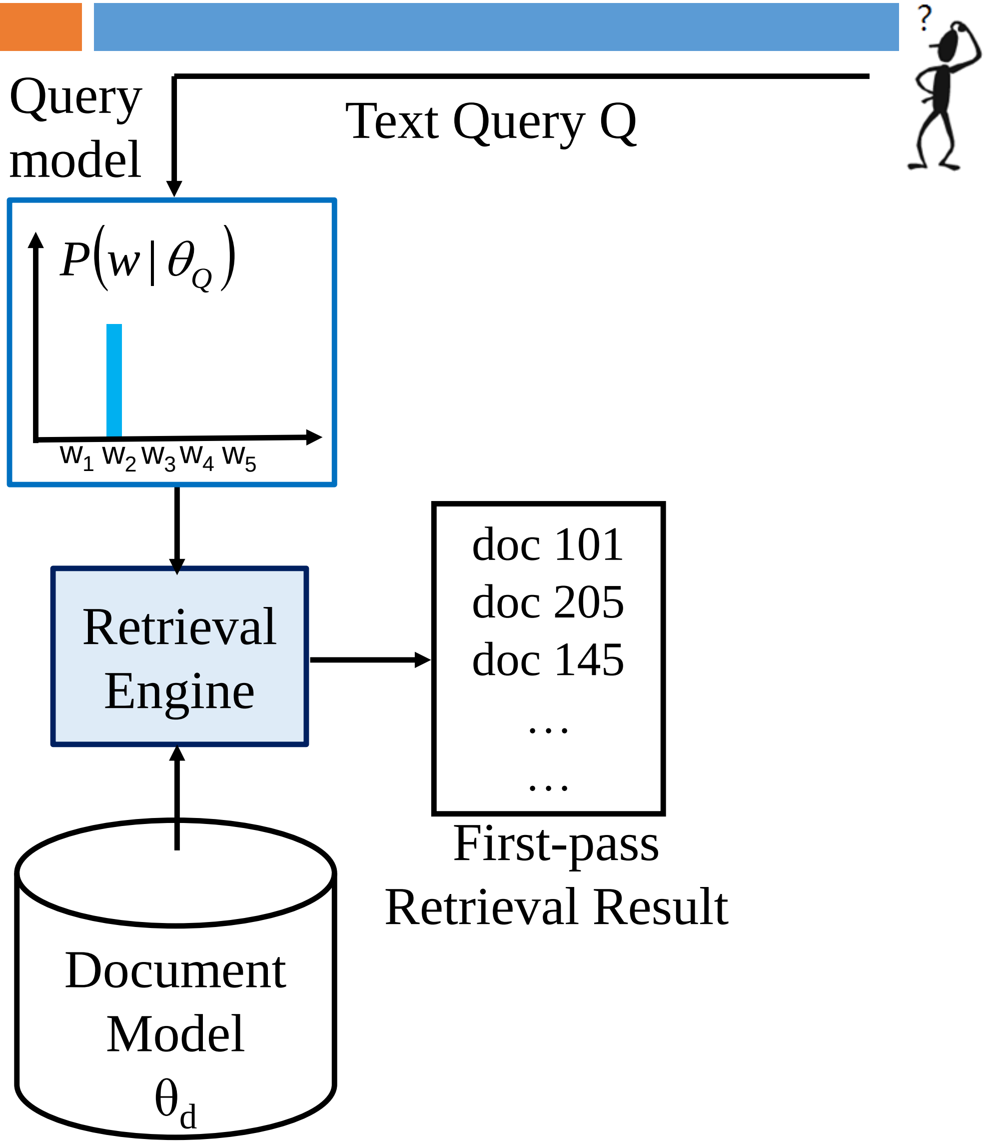
$$P(w \mid \theta_d) = \frac{\boxed{N(w,d)}}{\boxed{\sum_{w'} N(w,d)}}$$

N(w,d): term frequency of word w in document d
Normalize into probability

Those basic models can be enhanced by query/document expansion to handle the problem of semantic retrieval.

# Review: Query Expansion

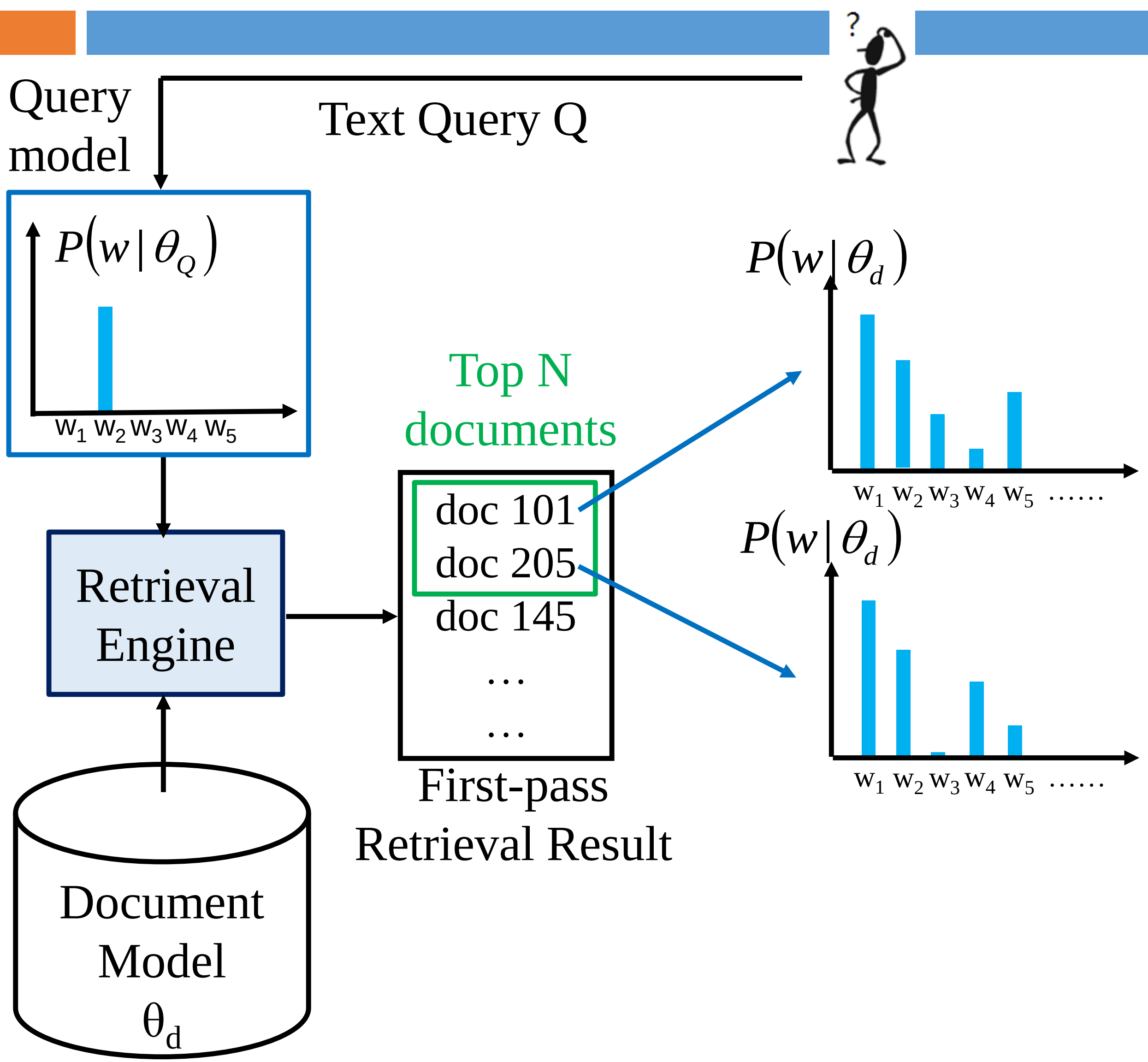
Parallel to PRF




[Tao, SIGIR 06]

# Review: Query Expansion

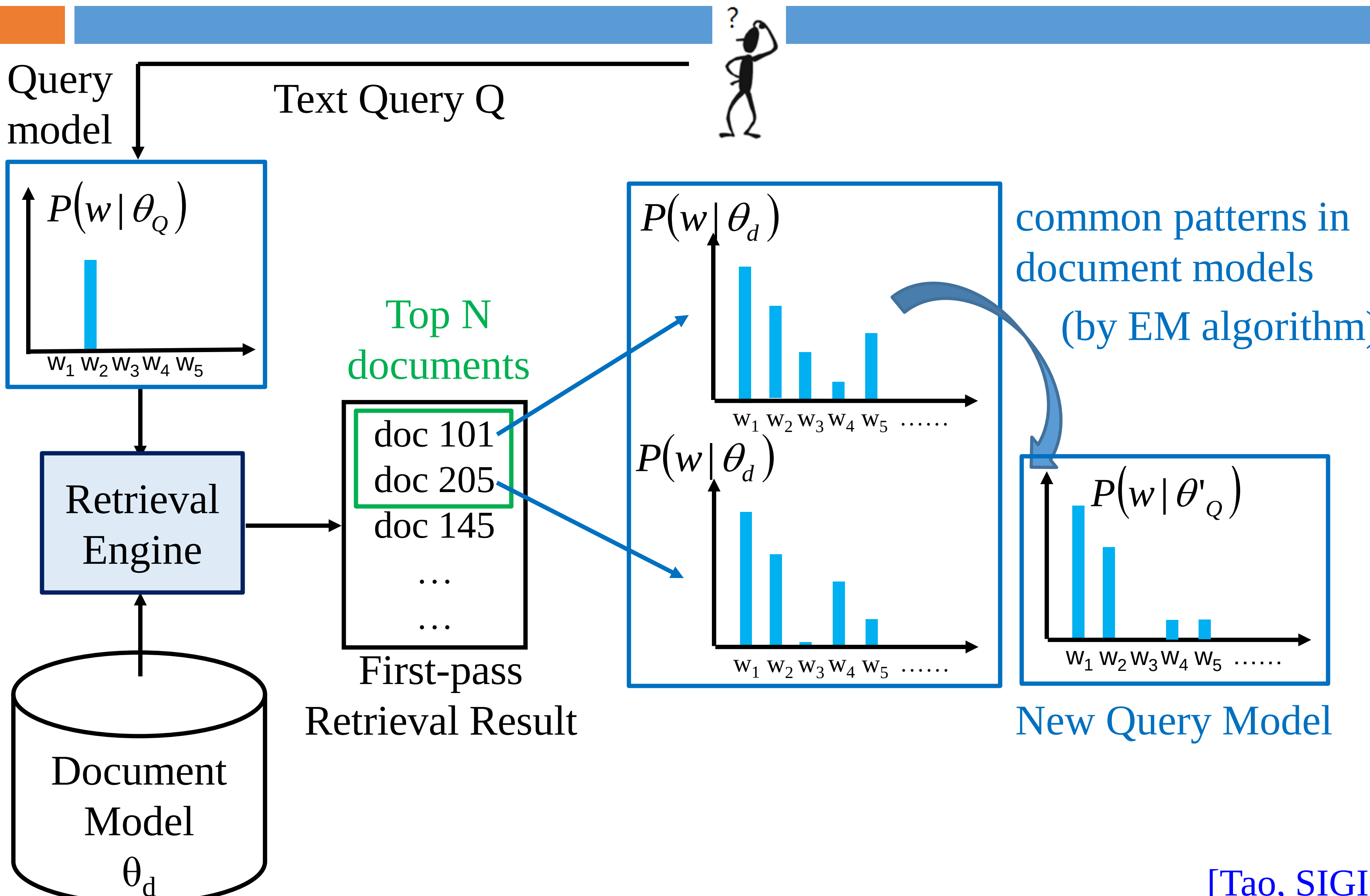
Parallel to PRF




[Tao, SIGIR 06]

# Review: Query Expansion

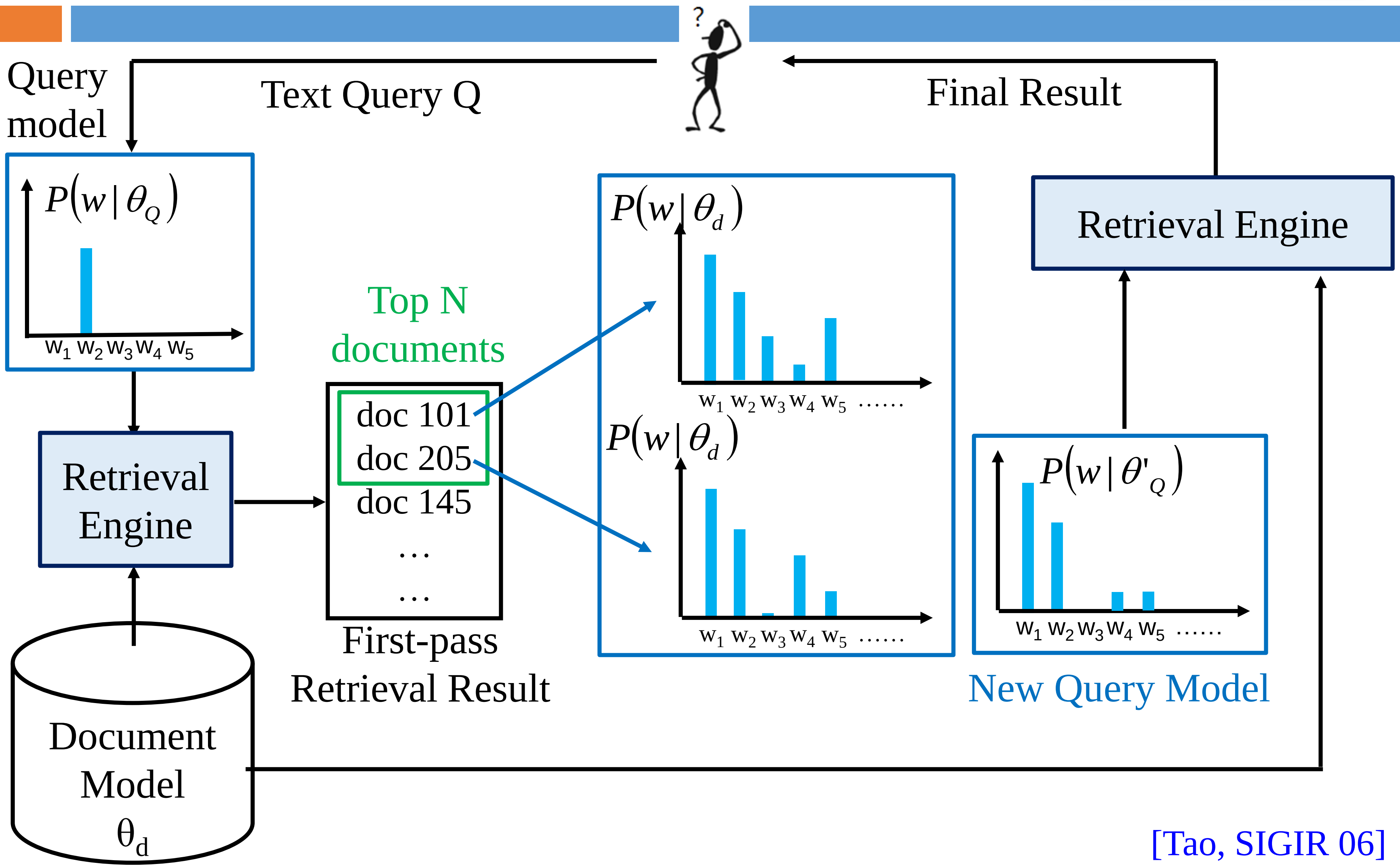
Parallel to PRF




[Tao, SIGIR 06]

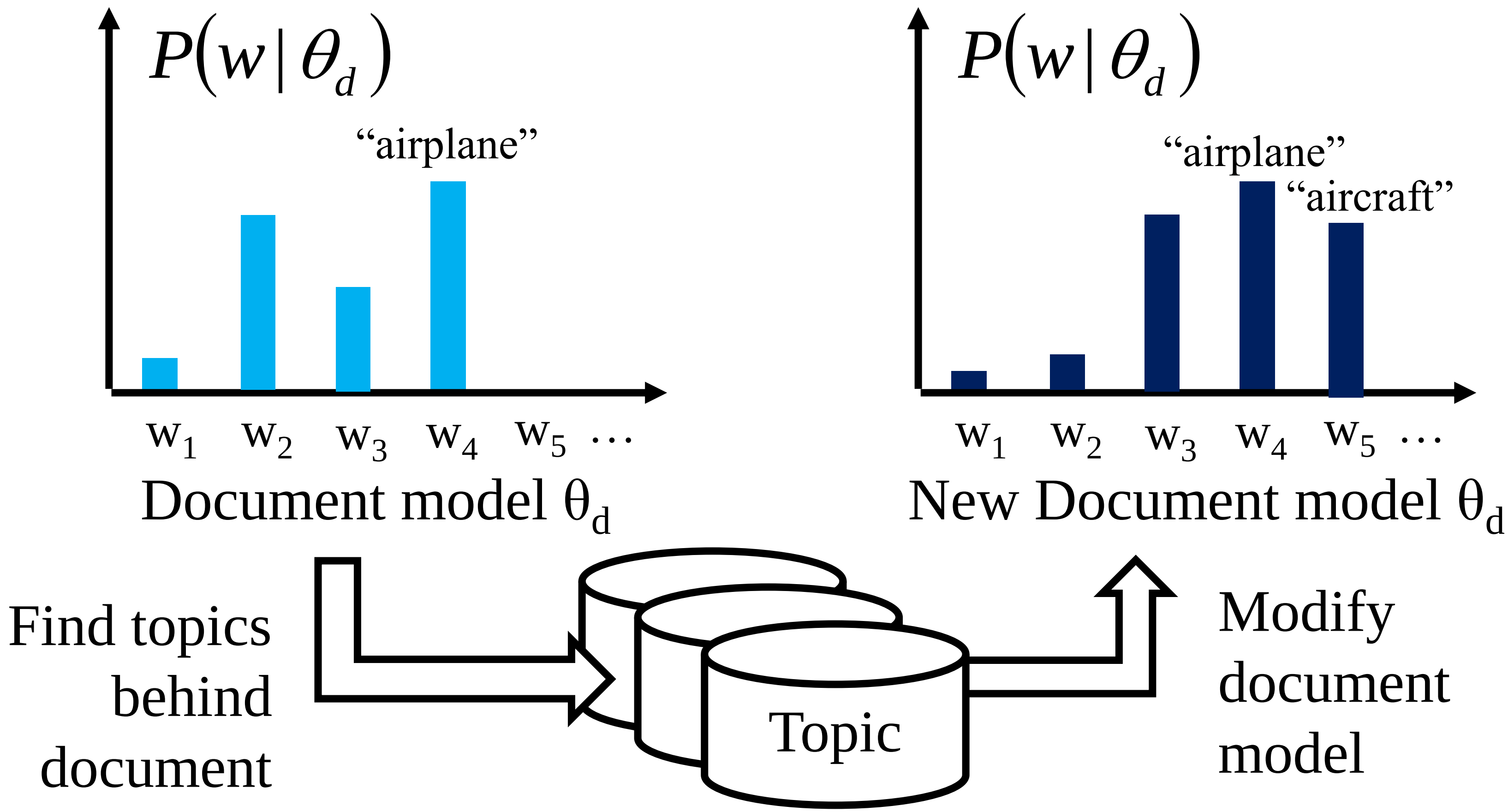
# Review: Query Expansion

Parallel to PRF




[Tao, SIGIR 06]

# Review: Document Expansion




This is realized by PLSA, LDA, etc.

[Wei, SIGIR 06]

# Semantic Retrieval on Lattices

- Modify retrieval model for ***lattices***:

| *Original Retrieval Model of Text* | *For Lattices* |
|---|---|
| Term Frequency | ***Expected*** Term Frequency |
| Document Length | ***Expected*** Document Length |
| …… | …… |

- Take the basic ***language modeling retrieval approach*** as example

# Document Model from Lattices

- Document model $\theta_d$ for text

$$P(w \mid \theta_d) = \frac{\boxed{N(w,d)}}{\sum_{w'} \boxed{N(w,d)}}$$

- (Spoken) Document model $\theta_d$ from lattice

$$P(w \mid \theta_d) = \frac{\boxed{E(w,d)}}{\sum_{w'} \boxed{E(w,d)}}$$

query/document expansion can be applied

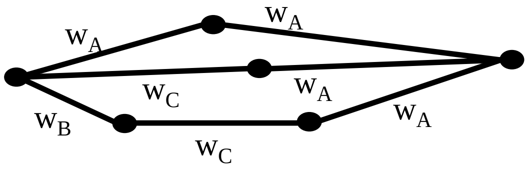
Replace term frequency N(w,d) with ***expected term frequency E(w,d) computed from lattices***

# Expected Term Frequency

- ***Expected*** term frequency E(w,d) for word w in spoken document d based on lattice

$$E(w,d) = \sum_{u \in L(d)} N(w,u) P(u \mid d)$$

lattice of spoken document d



# Expected Term Frequency

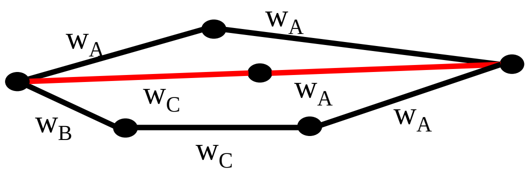
- ***Expected*** term frequency E(w,d) for word w in spoken document d based on lattice

$$E(w,d) = \sum_{u \in \boxed{L(d)}} \boxed{N(w,u)}\boxed{P(u \mid d)}$$

Can we have better estimation?

- ➢u: a word sequence in the lattice of d
- ➢P(u|d): posterior probability of word sequence u
- ➢N(w,u): the number of word w appearing in word sequence u
- ➢L(d): all the word sequences in the lattice of d

lattice of spoken document d



# New Direction 4-1:

## Special Semantic Retrieval Techniques for Spoken Content

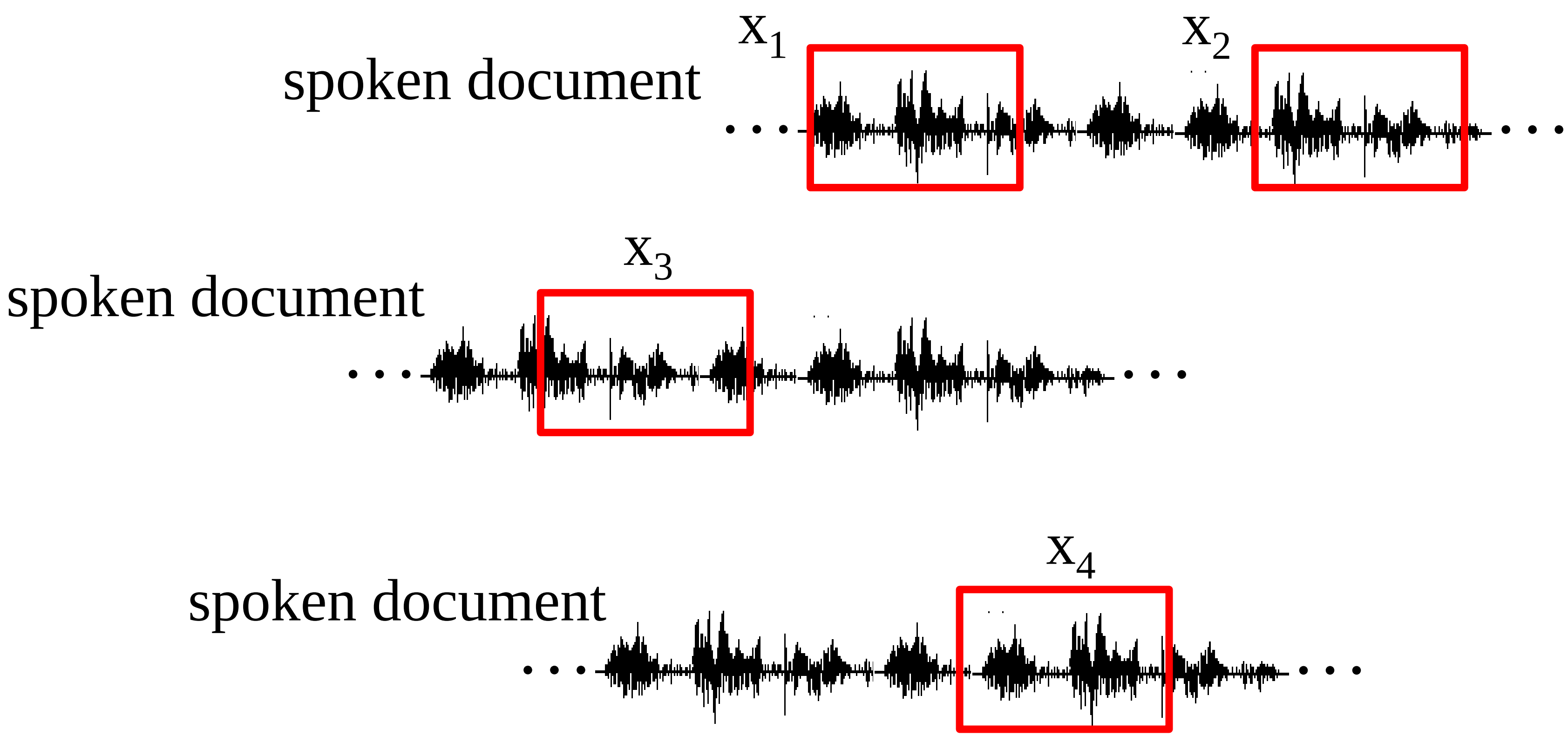
## Better Estimation of Term Frequencies

# Better Estimation of Term Frequencies

- Context of each term in the lattices [Tu & Lee, ICASSP 12]
  - The occurrences of a given term are usually characterized by similar context, while widely-varying contexts typically imply different terms [Schneider, Interspeech 10]
- Graph-based approach
  - Graph-based approach improved ***spoken term detection***
  - It can also improve ***semantic retrieval*** of spoken content
  - Idea: Replace expected term frequency E(w,d) with scores from graph-based approach [Lee & Lee, SLT 12] [Lee & Lee, IEEE/ACM T. ASL 14]

# Graph-based Approach for Semantic Retrieval

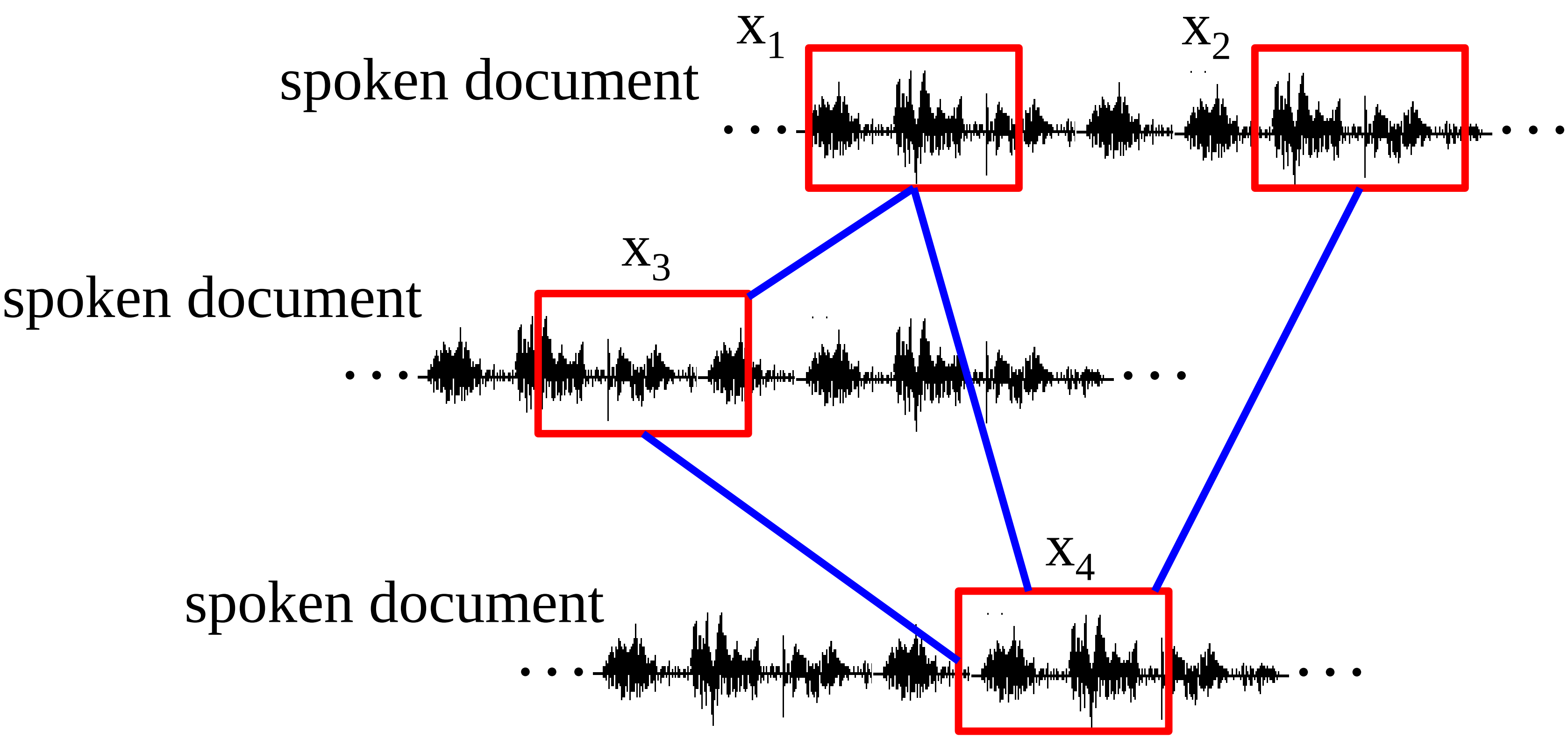
- For each word w in the lexicon




Find the occurrence regions of word w from lattices

# Graph-based Approach for Semantic Retrieval

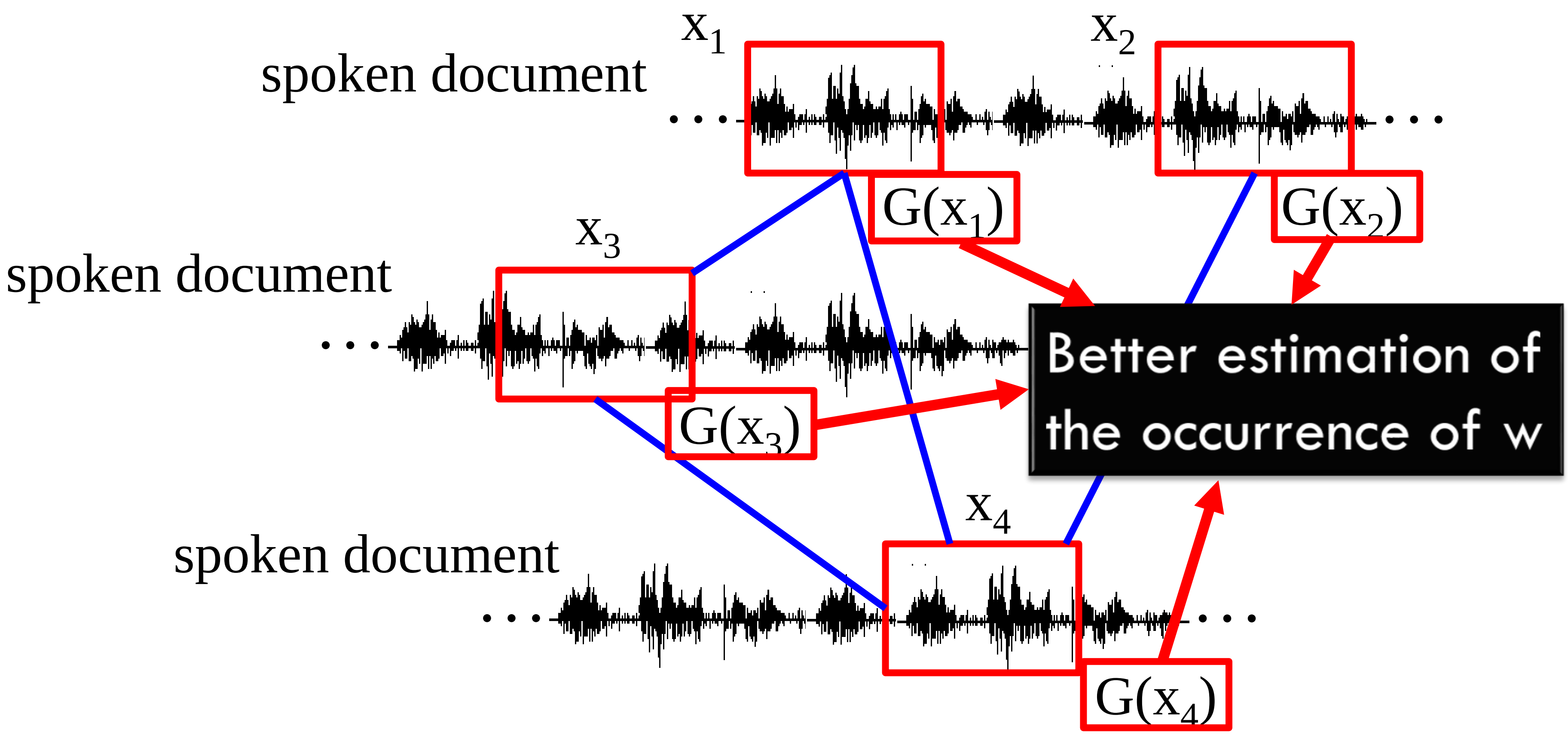
- For each word w in the lexicon




Connect the occurrence regions as a graph by similarities

# Graph-based Approach for Semantic Retrieval

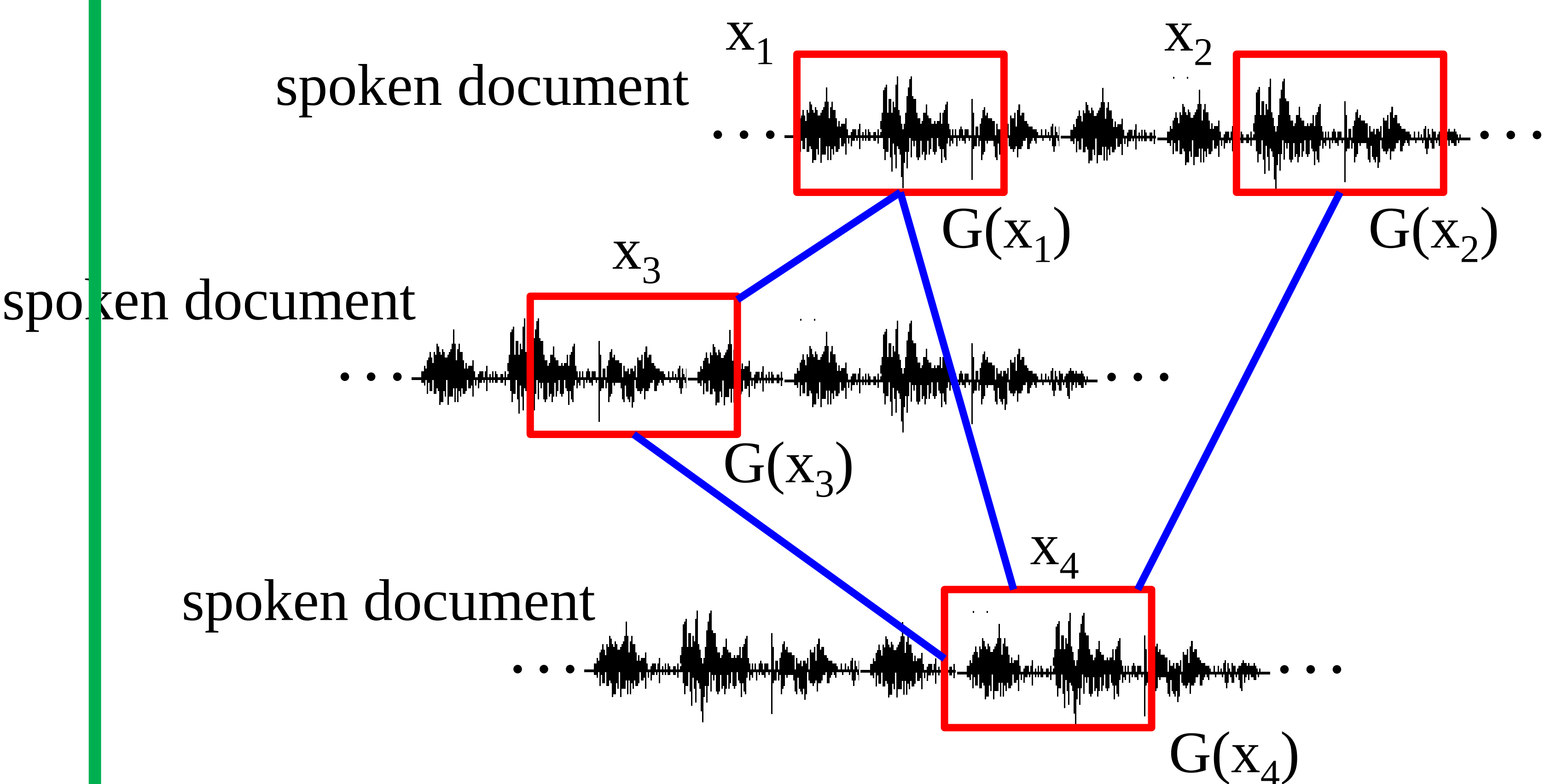
- For each word w in the lexicon




Obtain new score G(x) by random walk

# Graph-based Approach for Semantic Retrieval

- For each word w in the lexicon




Repeat this process for all the words w in the lexicon

# Graph-based Approach for Semantic Retrieval

Lattice-based Document Model:

$$P(w|\theta_d) = \frac{\boxed{E(w,d)}}{\sum_{w'} \boxed{E(w,d)}}$$

Graph-enhanced document model:

$$P(w|\theta'_d) = \frac{E'(w,d)}{\sum_{w'} E'(w,d)}$$

replace

query/document expansion can be applied

Better estimation of term frequencies for each word w in d

$$\boxed{E'(w,d)}$$

spoken document d

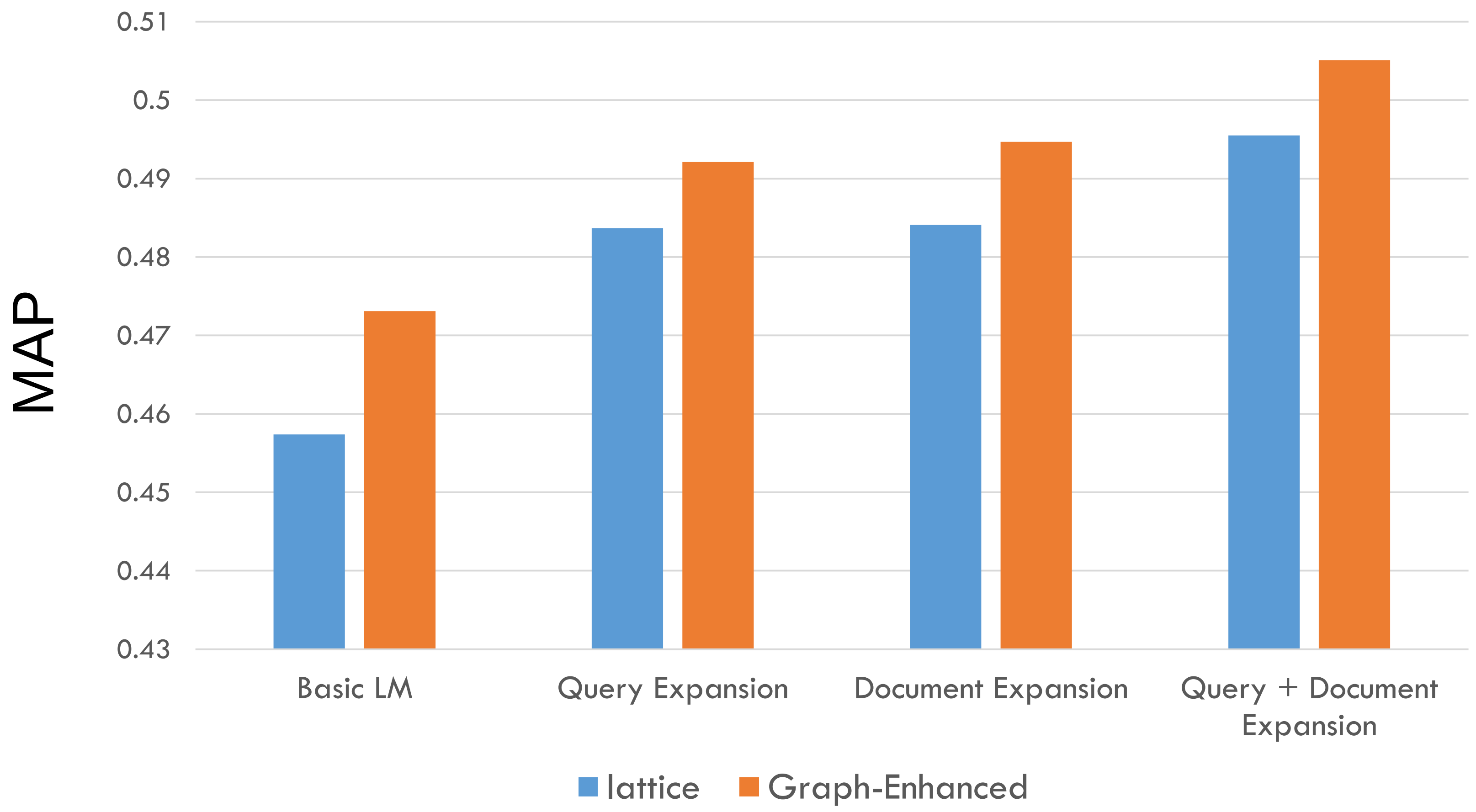
x1 x2

G(x1) G(x2)

Scores from graph

# Graph-based Approach for Semantic Retrieval - Experiments

- Experiments on TV News [Lee & Lee, IEEE/ACM T. ASL 14]



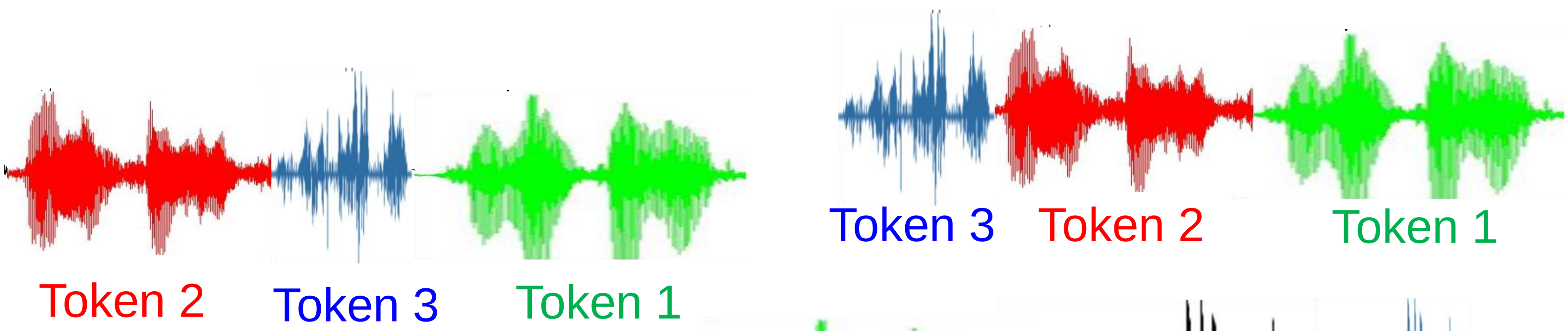
# New Direction 4-2:

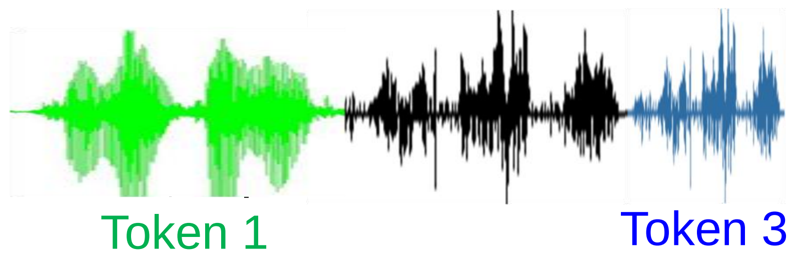
## Special Semantic Retrieval Techniques for Spoken Content

## Exploiting Acoustic Tokens

# Acoustic Tokens

- We can discover "***acoustic tokens***" in direction 3



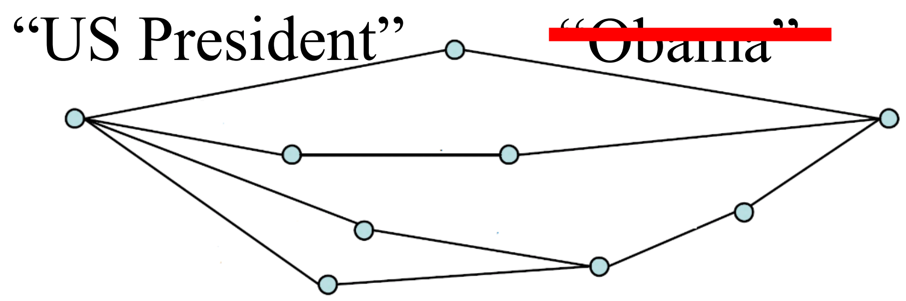

Can be useful in semantic retrieval of spoken content:

- ➢Query expansion with acoustic tokens
- ➢Unsupervised semantic retrieval of spoken content

# Query Expansion with Acoustic Tokens

Basic idea of query expansion

Related terms frequently co-occur in the same spoken document




If “Obama” is not in the lexicon

- “Obama” will never appear in lattices.
- We can never know “Obama” co-occur with “US President” in query expansion.
- Typical approach: using subwords
- Query expansion with acoustic tokens

Complementary to each other

# Query Expansion with Acoustic Tokens

[Lee & Lee, ICASSP 13]

Original Text Query:

"US President"

⇩

$d_{100}$: …… US President …

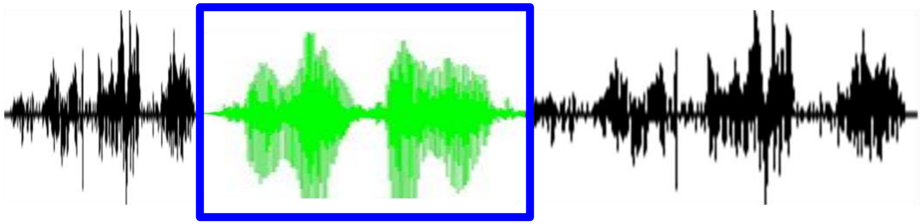
$d_{205}$: … US President ……

**First pass**: Retrieve spoken documents containing "US President" in the transcriptions

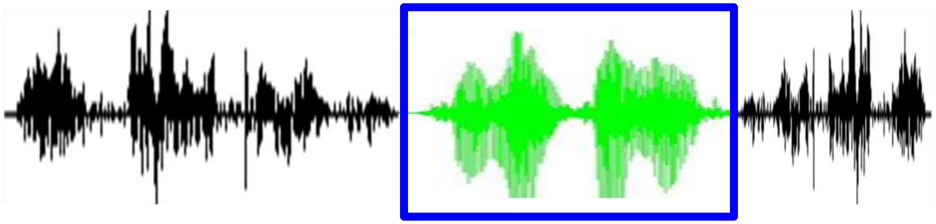
# Query Expansion with Acoustic Tokens

Original Text Query:
"US President"

$d_{100}$: …… US President …



$d_{205}$: … US President ……



Find acoustic tokens frequently appear in the signals of these retrieved documents

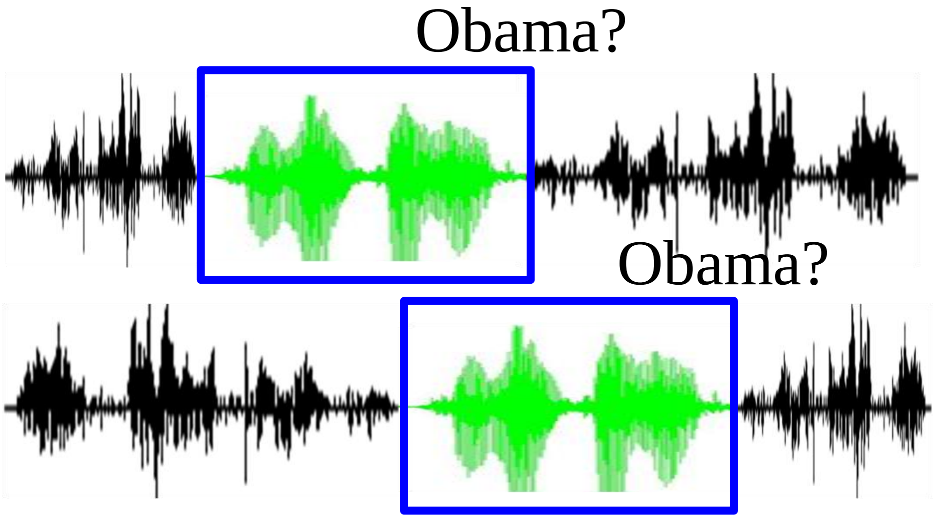
# Query Expansion with Acoustic Tokens

Original Text Query:

"US President"

$d_{100}$: …… US President …

$d_{205}$: … US President ……



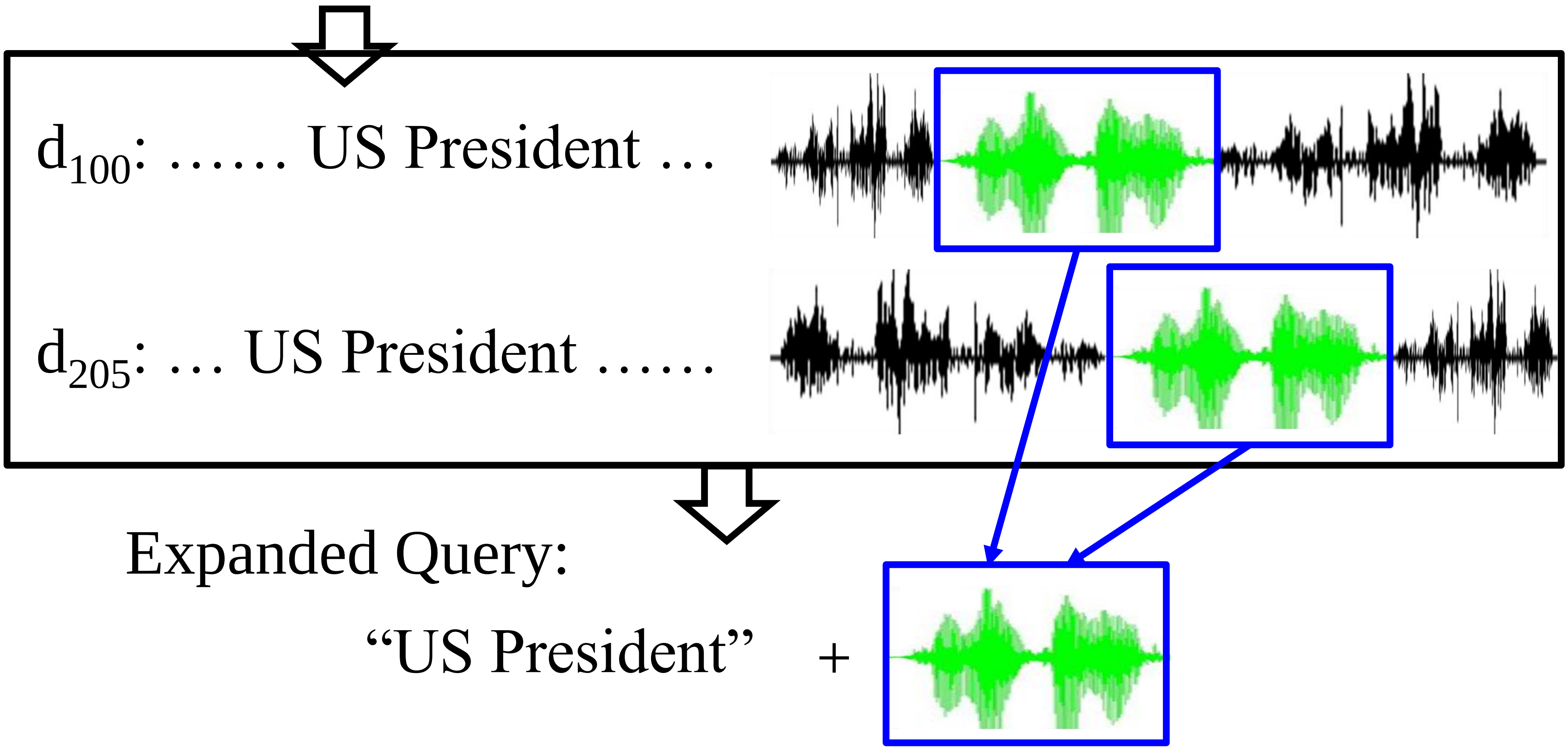

Even the terms related to the query is OOV

- If they co-occur with the query in speech signals
- Find acoustic tokens corresponding to these terms

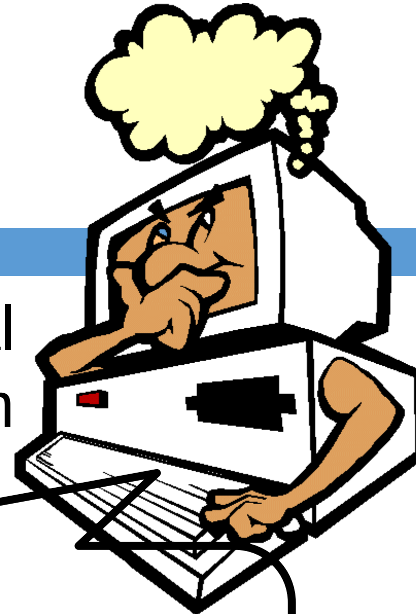
# Query Expansion with Acoustic Tokens

Original Text Query:

"US President"



# Query Expansion with Acoustic Tokens



Retrieval system

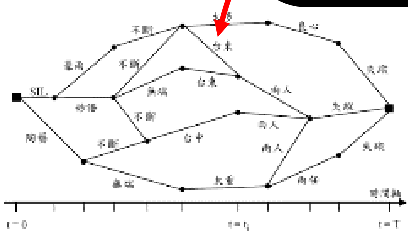
user

"US President"

***Expanded Query:***

"US President"
"White House"

and



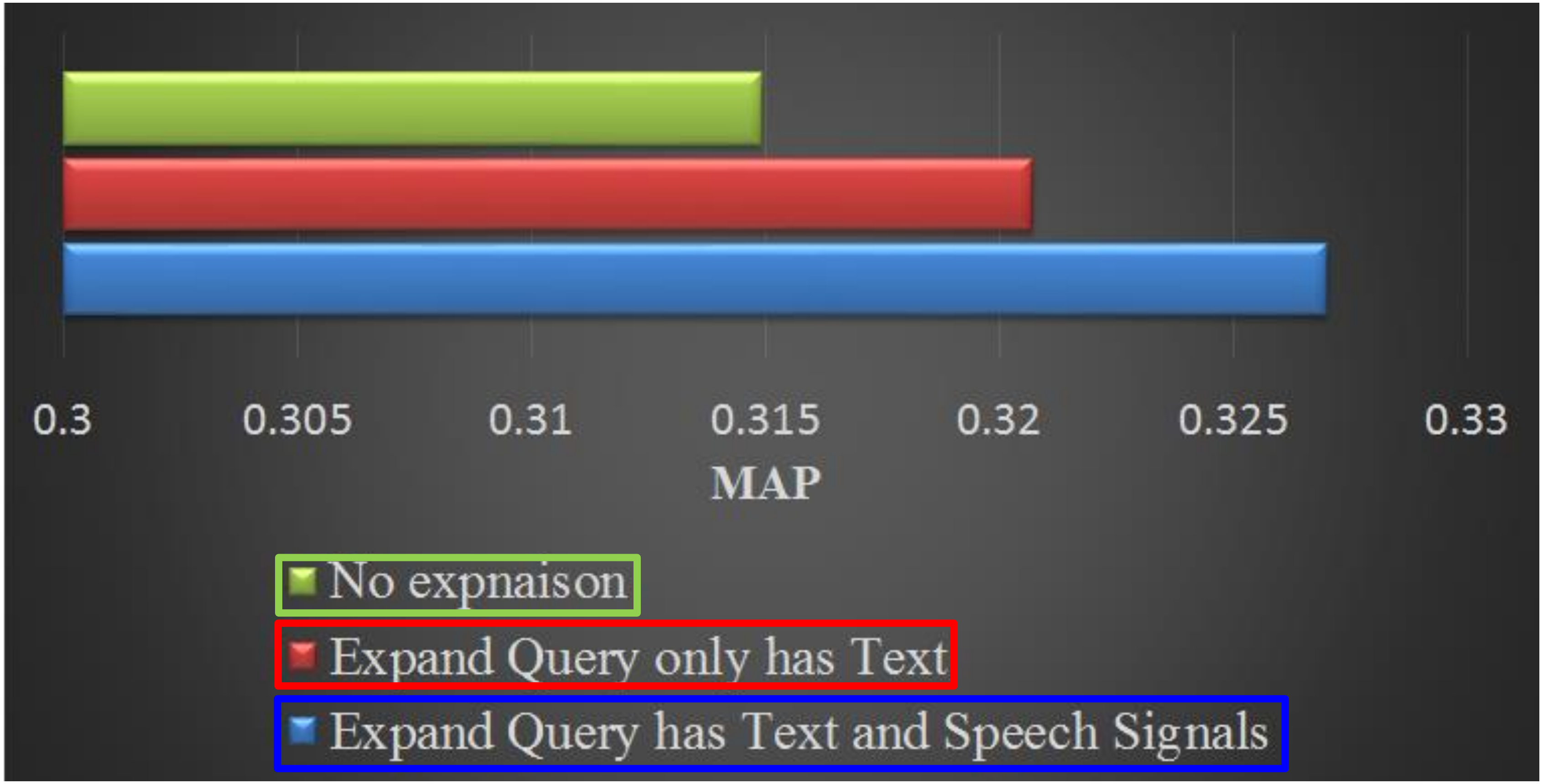
Lattices

Find the same tokens

By expanding the text query with acoustic tokens, more semantically related audio files can be retrieved.

# Query Expansion – Acoustic Patterns

- Experiments on TV News [Lee & Lee, ICASSP 13]



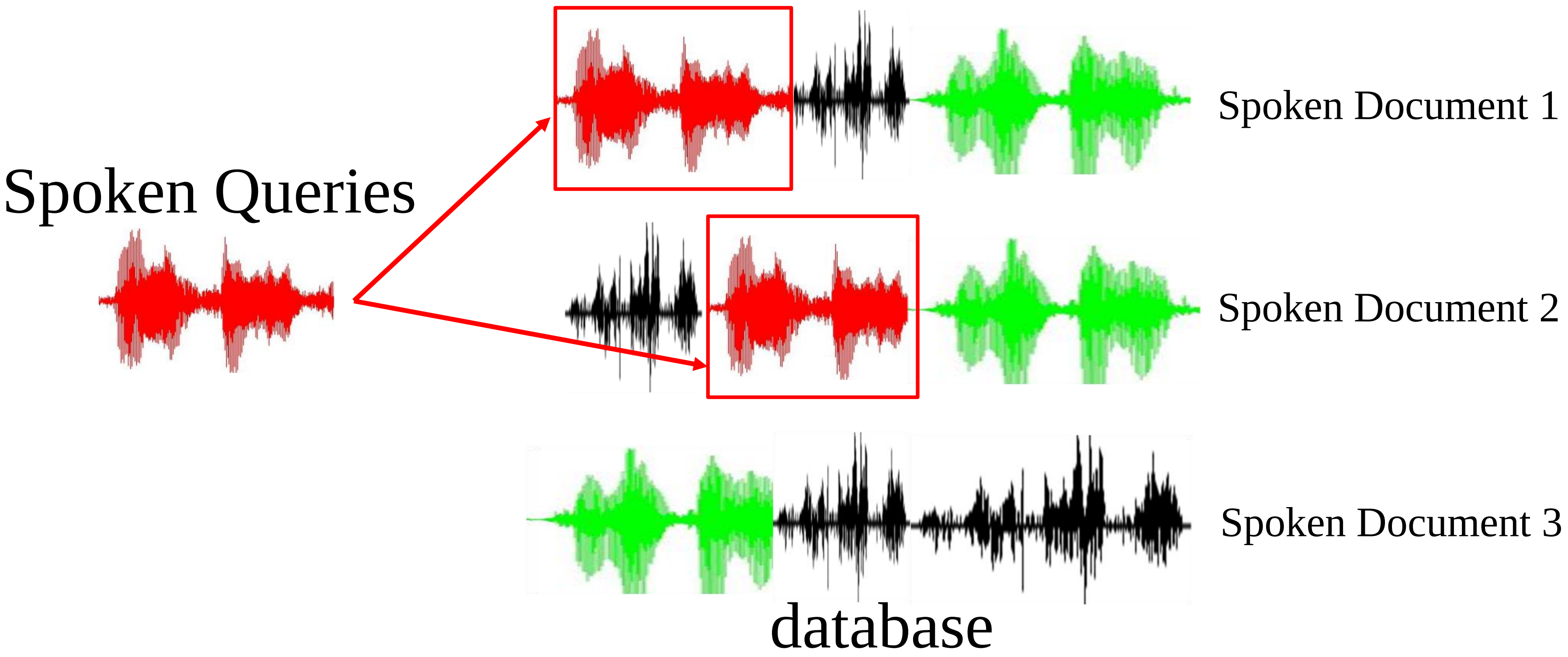
# Unsupervised Semantic Retrieval

- ***Unsupervised Semantic Retrieval*** [Li & Lee, ASRU 13][Oard, FIRE 13]
  - Find spoken documents ***semantically related*** to the ***spoken queries***
  - Without speech recognition
- New task, not too much previous work
  - Below is just a very preliminary study based on query expansion with acoustic tokens [Li & Lee, ASRU 13]

# Unsupervised Semantic Retrieval

1. Find spoken documents containing the spoken query

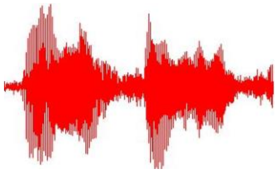
Done by the query-by-example spoken term detection approaches (e.g. DTW)

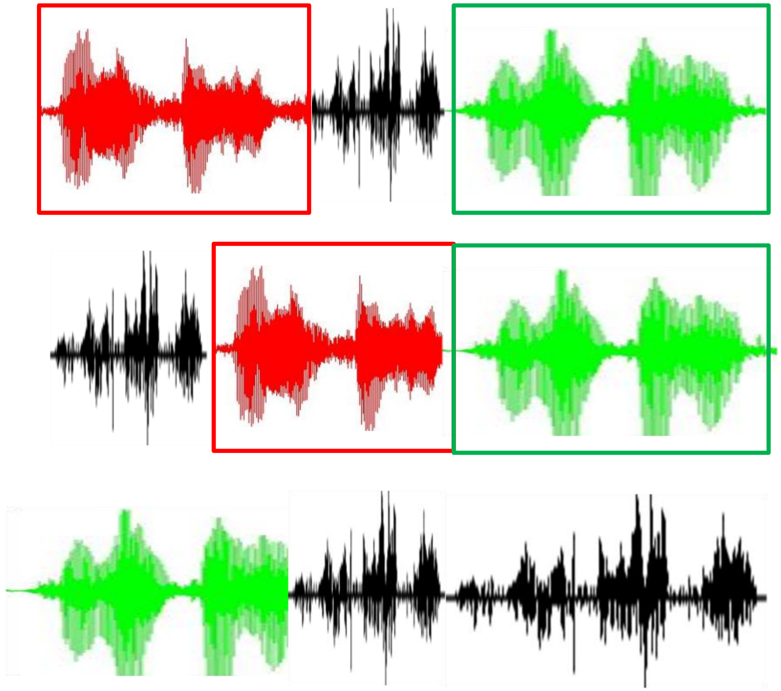


# Unsupervised Semantic Retrieval

2. Find acoustic tokens frequently co-occurring with the spoken queries in the same document

Spoken Queries

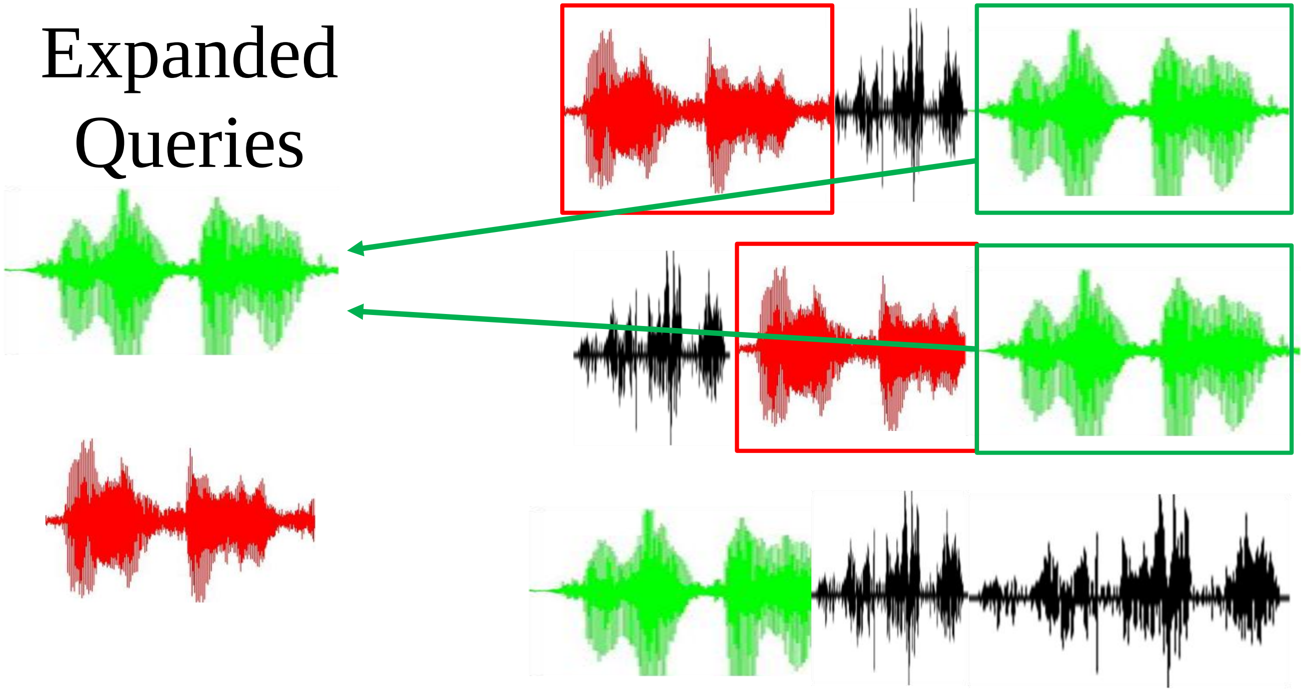




# Unsupervised Semantic Retrieval

3. Use the acoustic tokens to expand the original spoken query

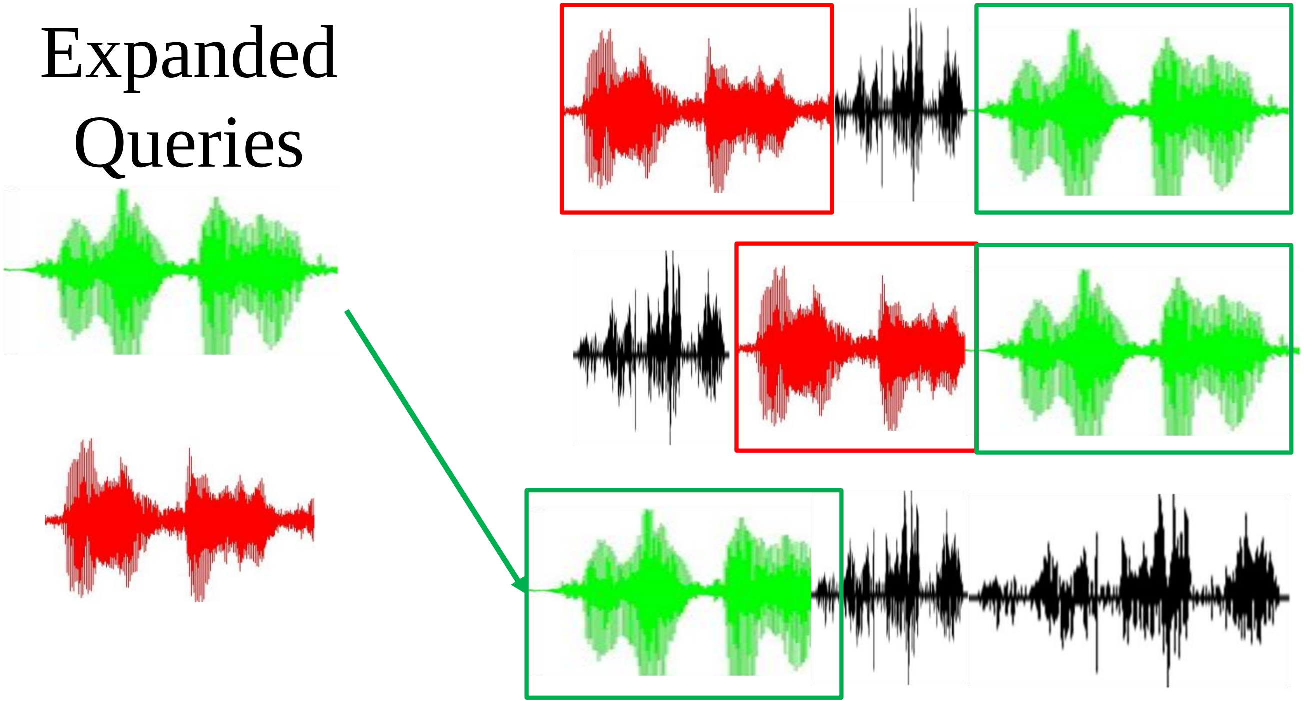
Expanded Queries



# Unsupervised Semantic Retrieval

4. Retrieve again by the expanded queries
   - Can retrieve spoken documents not containing the original spoken queries

Expanded Queries



# Unsupervised Semantic Retrieval - Experiments

- Broadcast news, MAP as evaluation measure
  - Using only DTW for unsupervised semantic retrieval: MAP= 8.76%
    - The semantically related documents without the query term cannot be retrieved by DTW.
  - Expanded by Acoustic Tokens: MAP= 9.70%
    - Unsupervised semantic retrieval has a long way to go

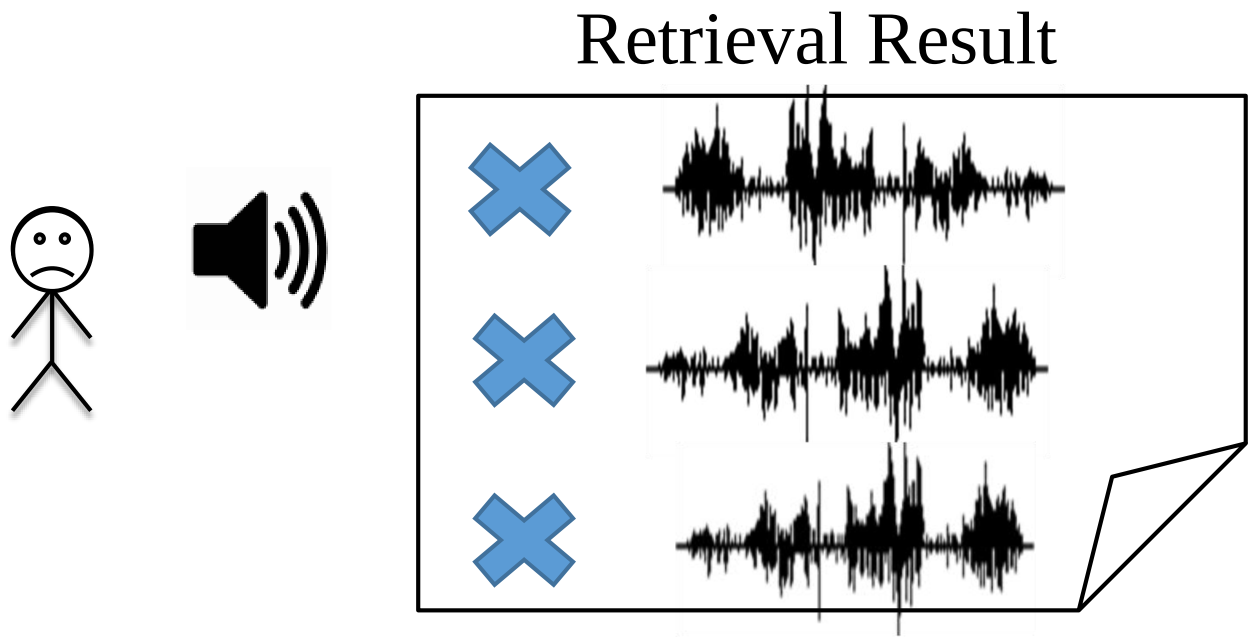
[Li & Lee, ASRU 13]

# New Direction 5:

## Speech Content is Difficult to Browse!

# Audio is hard to browse

- When the system returns the retrieval results, user doesn’t know what he/she get at the first glance



# Audio is hard to browse

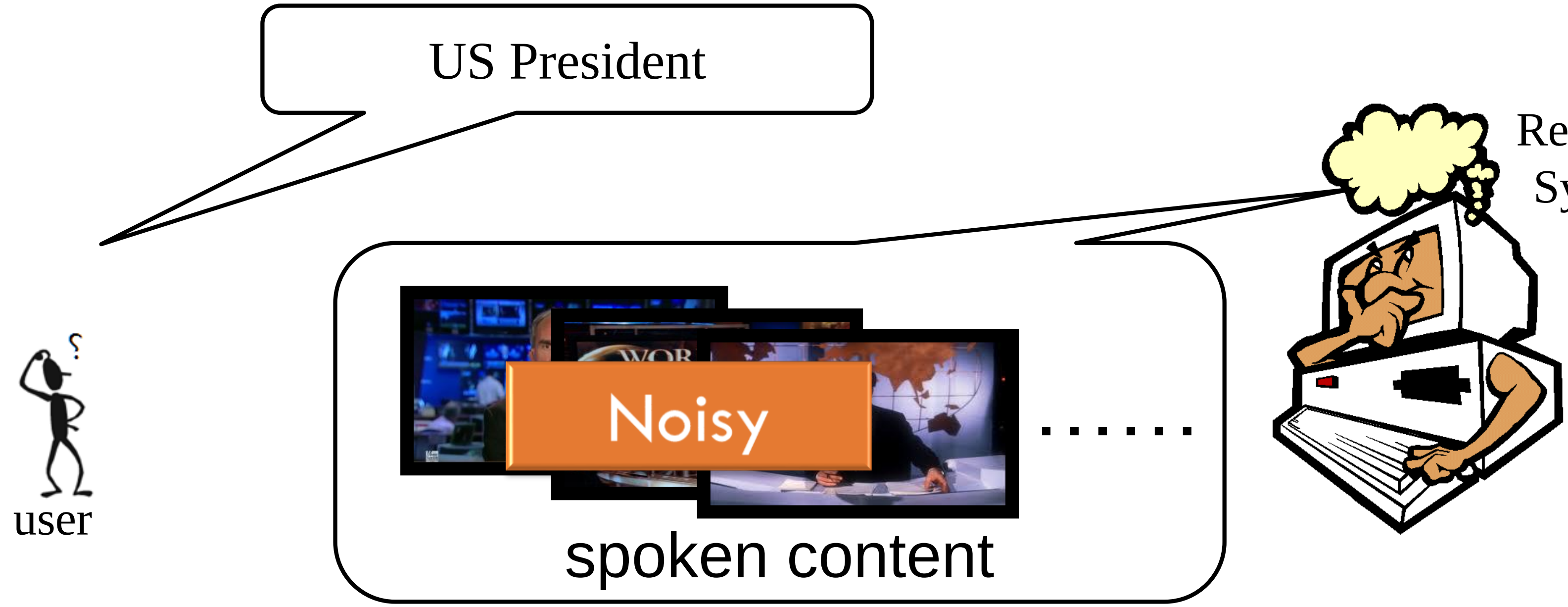
- Interactive spoken content retrieval
- Extracting Core Information
- Organizing Retrieved Results
- Spoken Question answering

# New Direction 5-1:

## Speech Content is Difficult to Browse!

## Interactive Spoken Content Retrieval

# Interactive spoken content retrieval




Input query is usually short ➡ Cannot describe the information need clearly

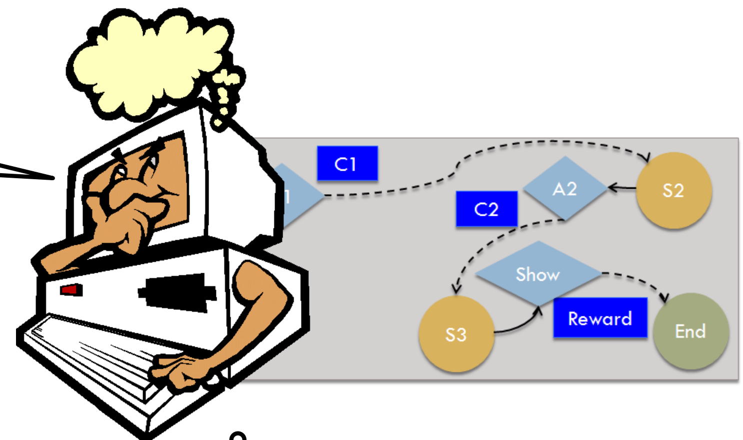
Speech recognition always produces errors.

# Interactive spoken content retrieval

US President

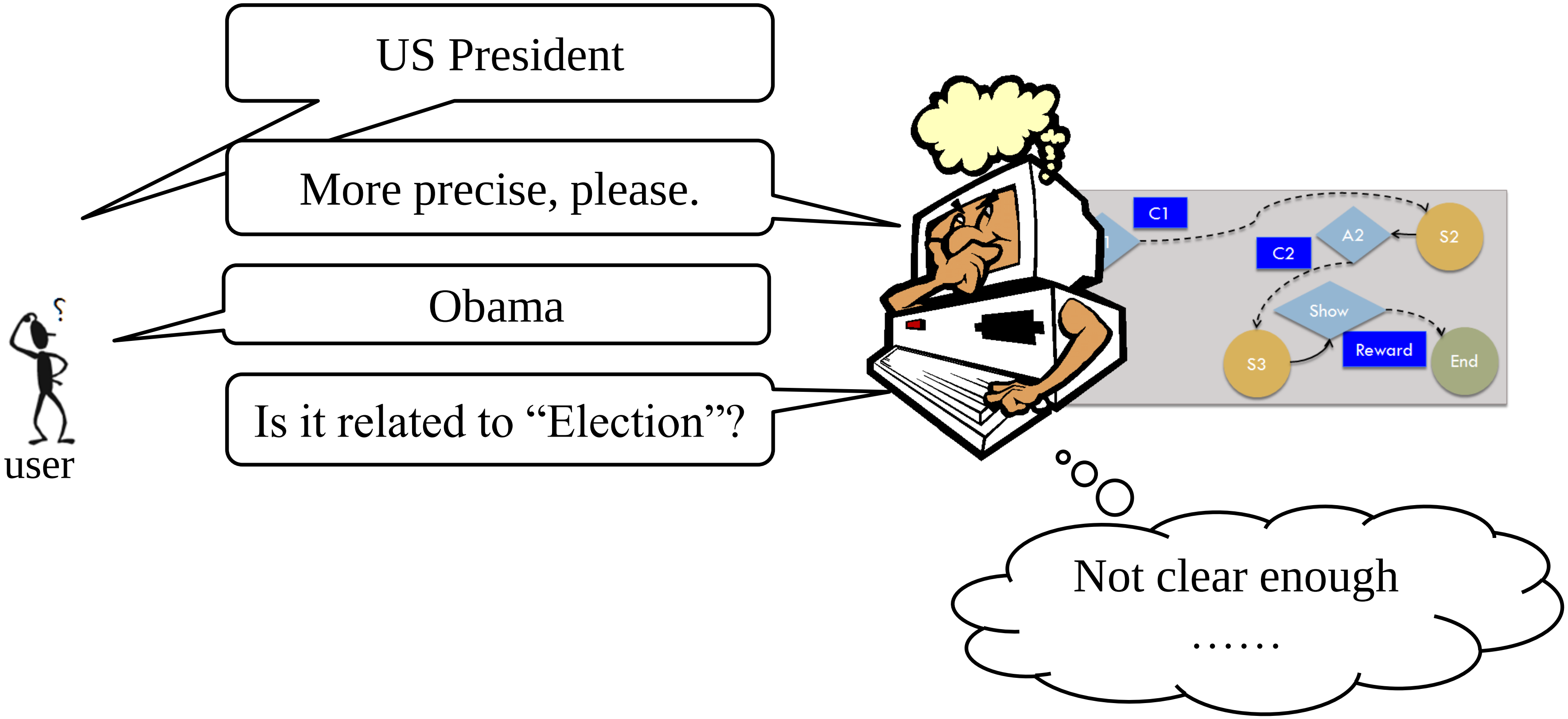
More precise, please.

user

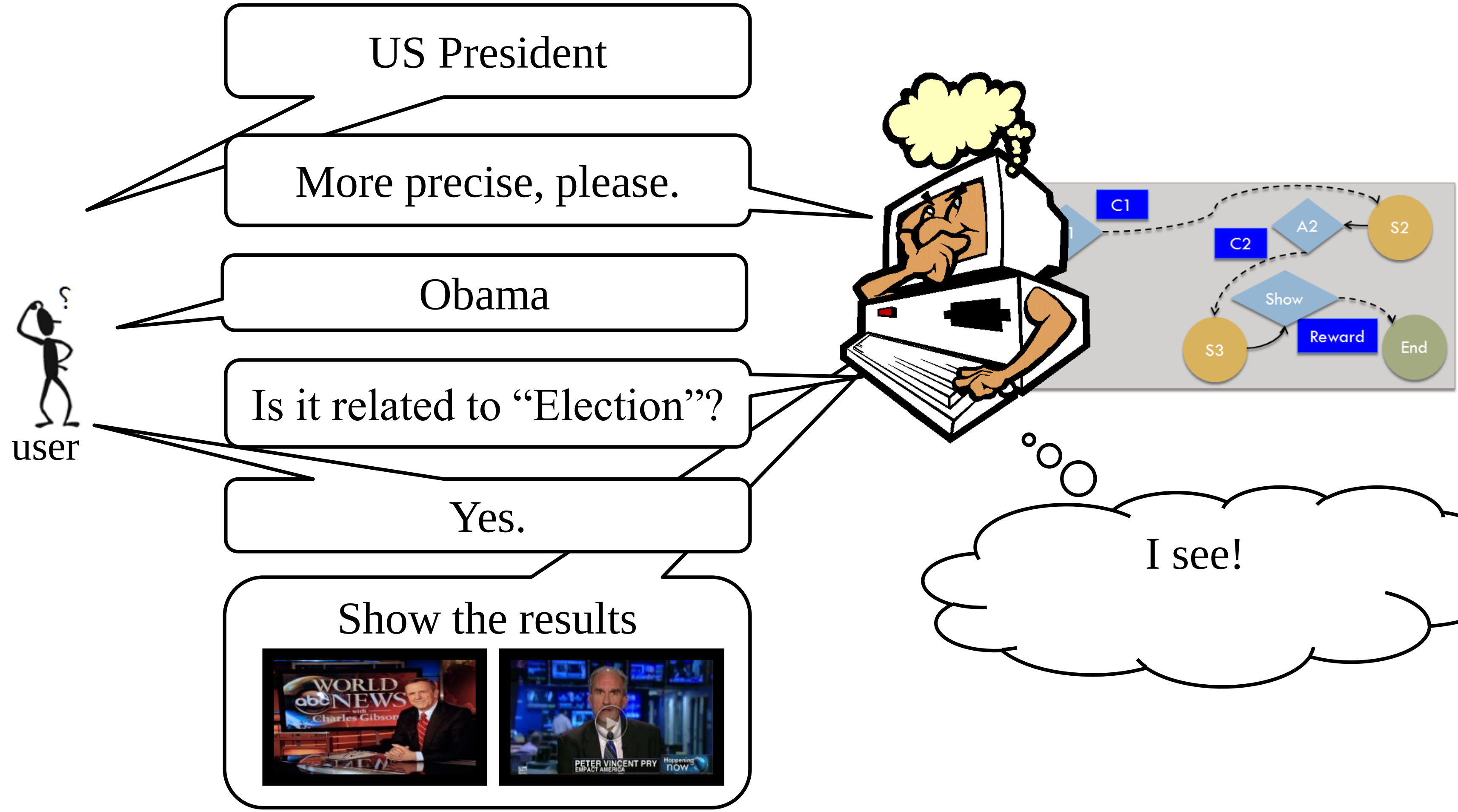



??????

# Interactive spoken content retrieval



# Interactive spoken content retrieval



# Challenges

- Given the information entered by the users, which action should be taken?

“More precise, please.”

Gain much information, but much user labor

“Is it relevant to XXX?”

Gain little information, but little user labor

“Give me an example.”

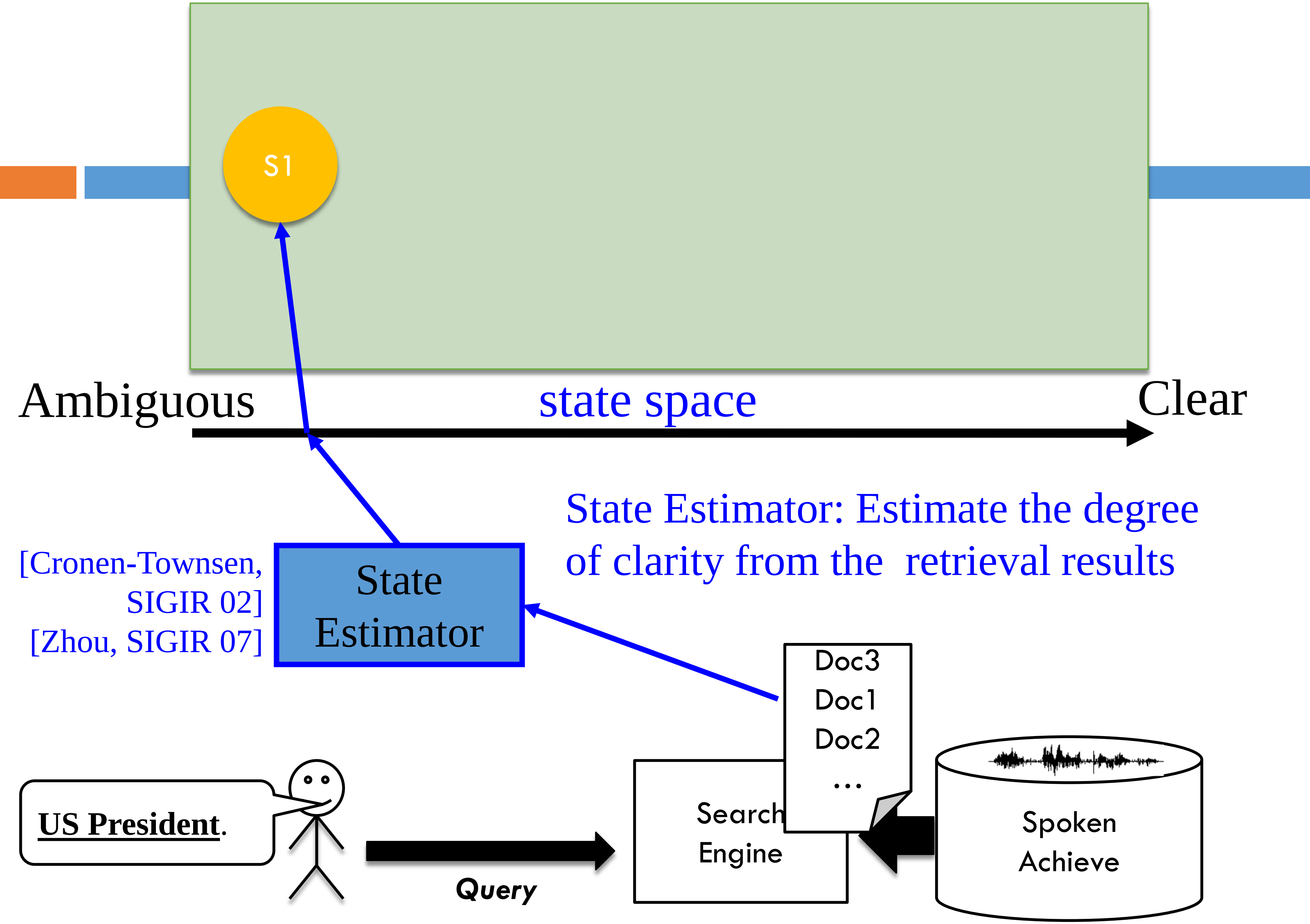
“Show the results.”

Borrowing the experiences from developing dialogue system (air ticket booking, city guides, personal assistant .., )

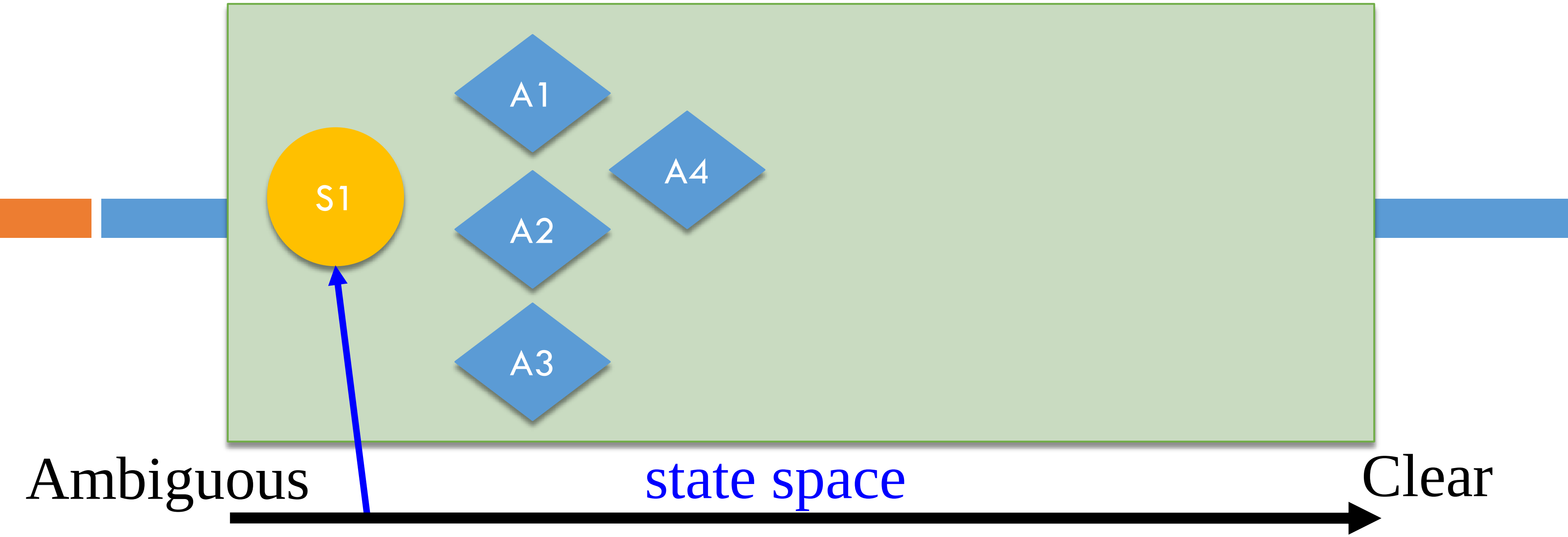
# MDP for Interactive Retrieval

- Markov Decision Process (MDP)
  - The system is in certain states.
  - Which action should be taken depends on the state the system is in.
- MDP for Interactive retrieval [Wen & Lee, Interspeech 12][Wen & Lee, ICASSP 13]
  - ***State***: the degree of clarity of the user's information need

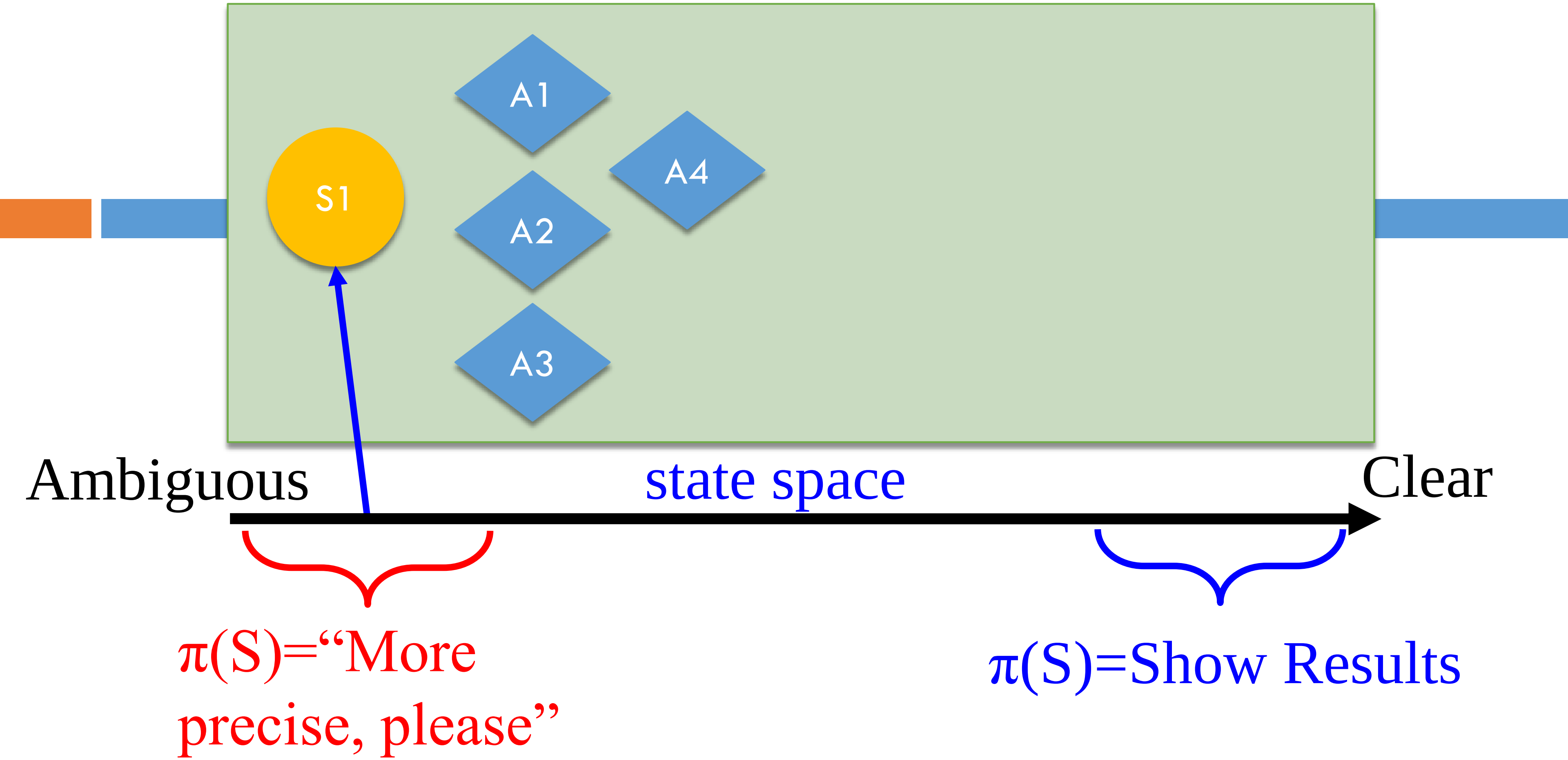
Ambiguous state space Clear



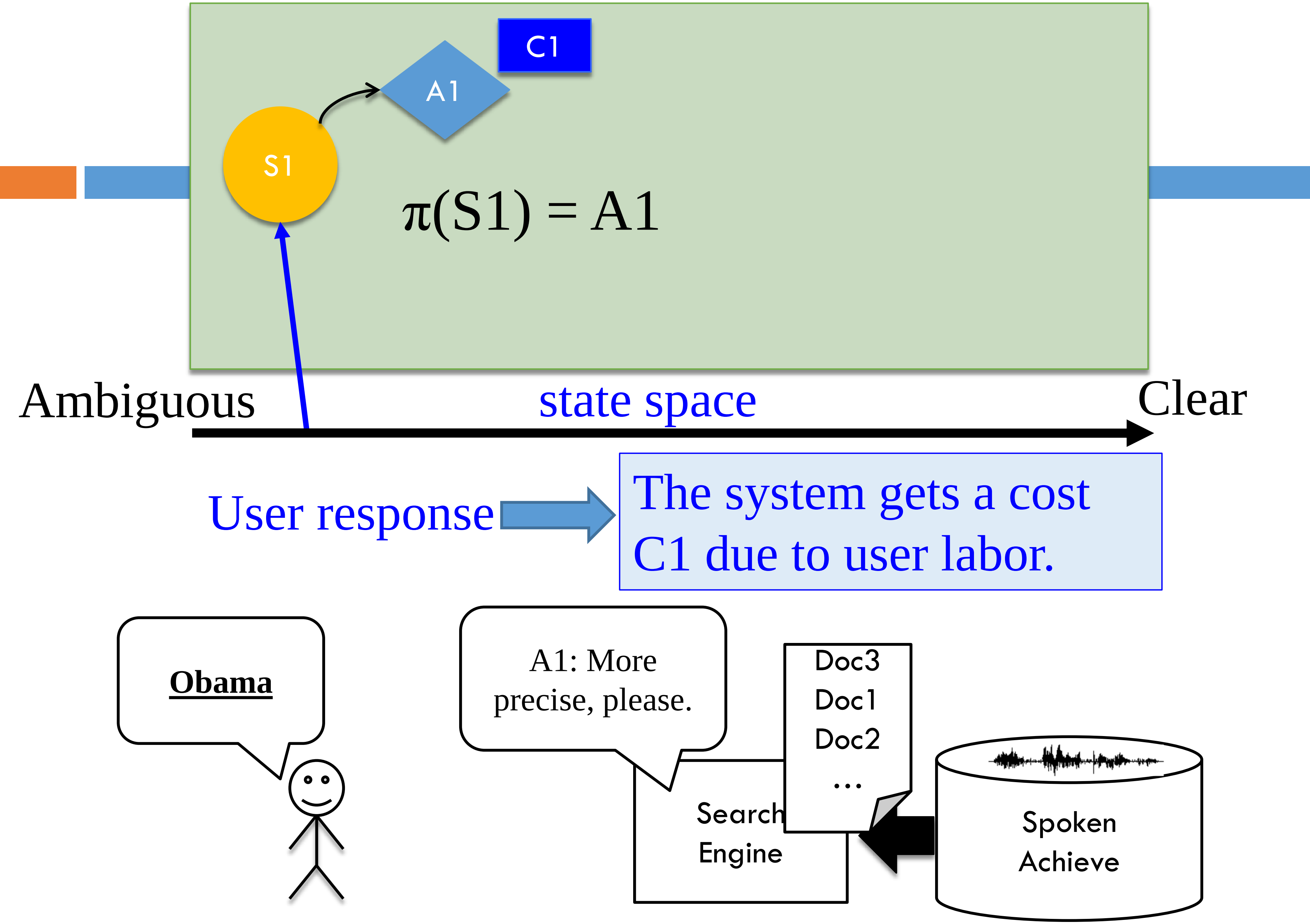
S1
Ambiguous
state space
Clear
State Estimator: Estimate the degree of clarity from the retrieval results
[Cronen-Townsen, SIGIR 02]
[Zhou, SIGIR 07]
State Estimator
Doc3
Doc1
Doc2
…
Search Engine
Spoken Achieve
US President.
Query

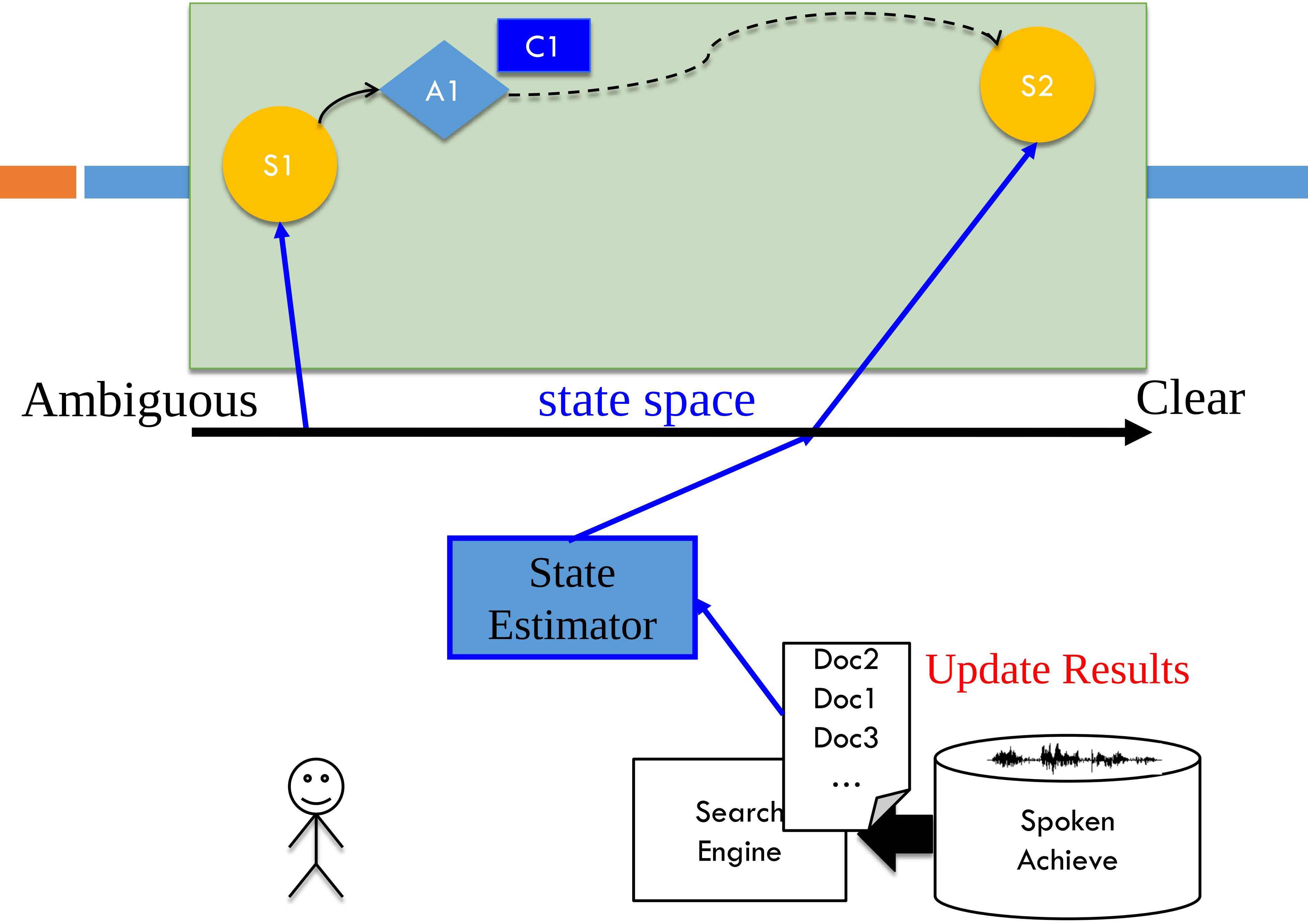



- A set of candidate actions
  - System: “More precisely, please.”
  - System: “Is it relevant to XXX?”
  - …..
- There is an action “show results”
  - When the system decides to show the results, the retrieval session is ended

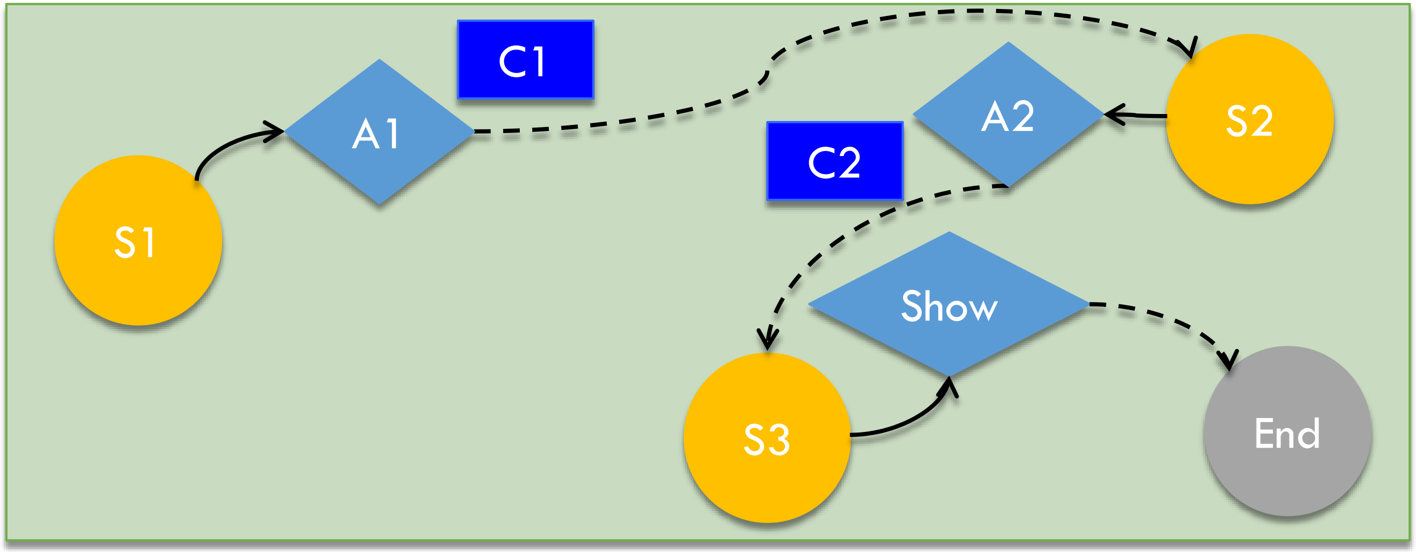



- Choose the actions by intrinsic policy π(S)
  - The policy is a function
  - Input: state S, output: action A


C1
A1
S1
$\pi(S1) = A1$
Ambiguous
state space
Clear
User response
The system gets a cost C1 due to user labor.
Obama
A1: More precise, please.
Doc3
Doc1
Doc2
...
Search Engine
Spoken Achieve

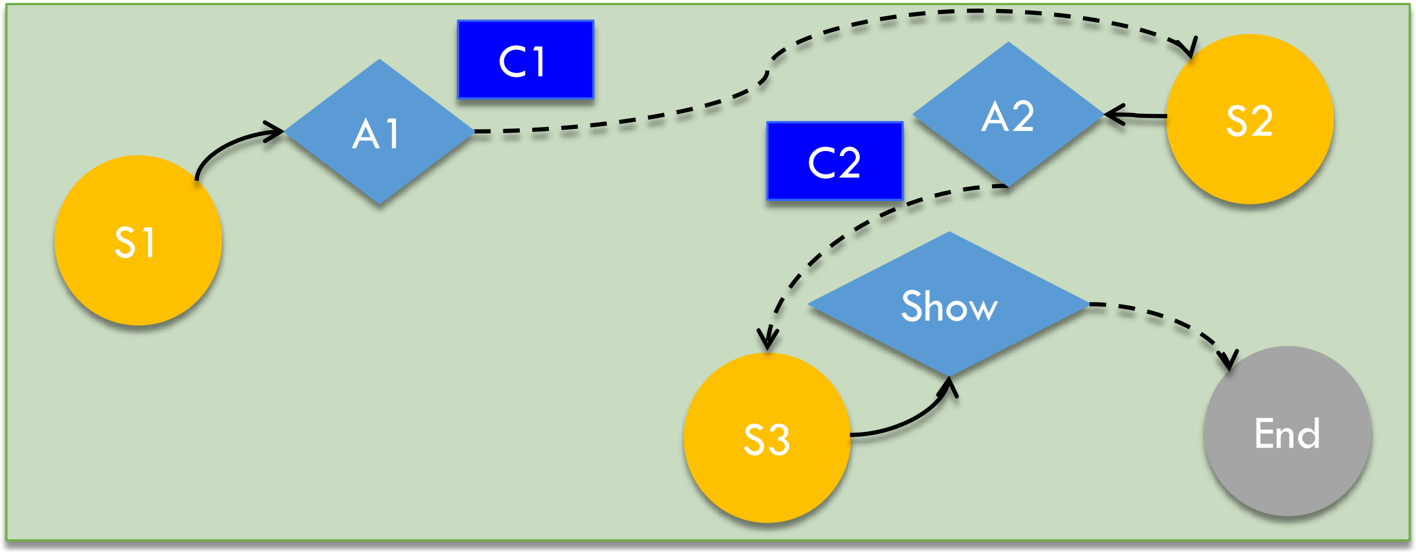

C1
A1
S1
S2
Ambiguous
state space
Clear
State Estimator
Doc2
Doc1
Doc3
...
Update Results
Search Engine
Spoken Achieve

# Interact with Users - MDP




- Good interaction:
  - The quality of final retrieval results shown to the users are as good as possible
  - The user labors (C1, C2) are as small as possible

# Interact with Users - MDP




- Find a polity π that

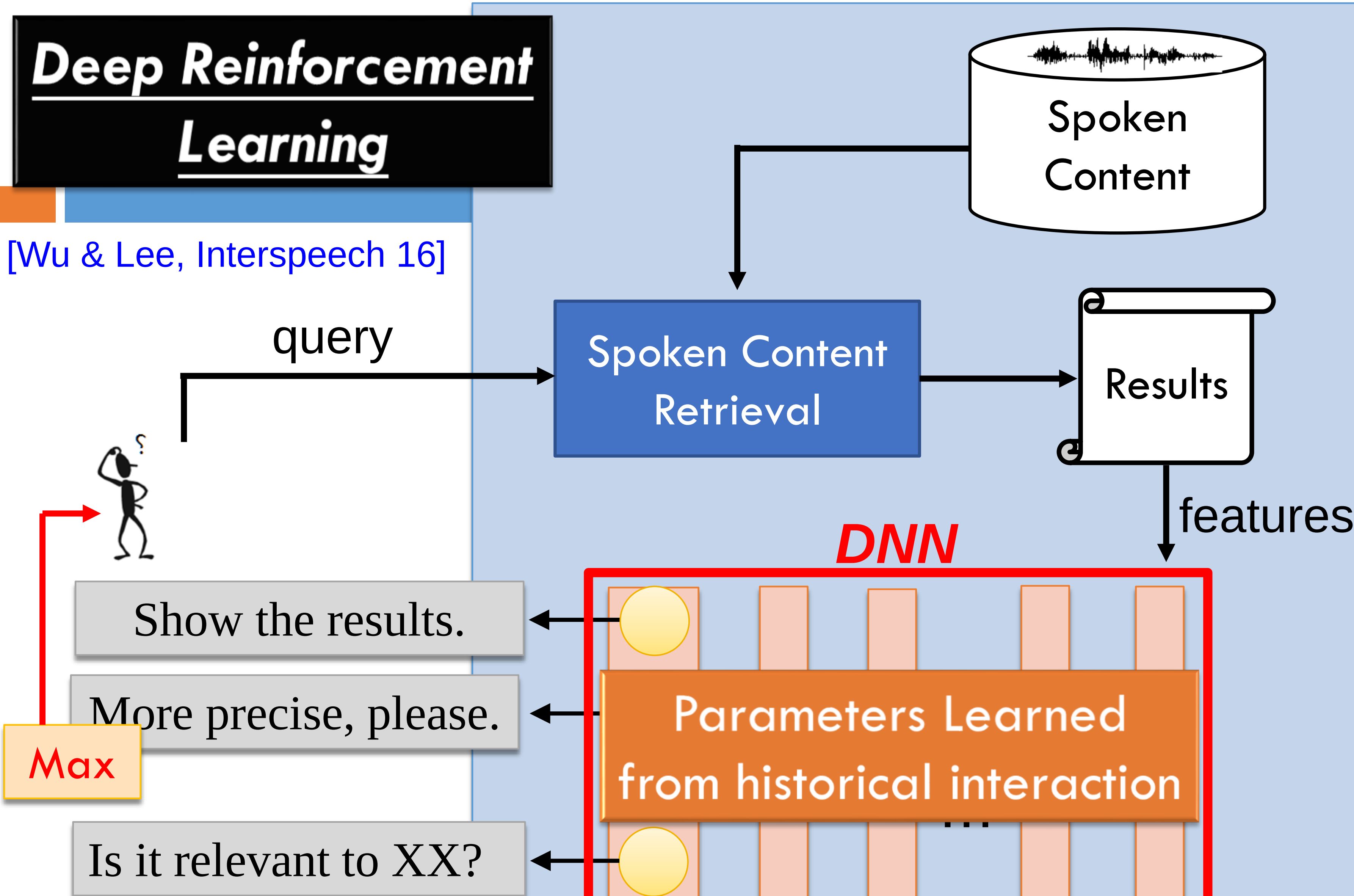
Maximizing Retrieval Quality, Minimizing User Labor

- The polity π can be learned from historical interaction [Wen & Lee, Interspeech 12][Wen & Lee, ICASSP 13]

# Deep Reinforcement Learning


# Deep Reinforcement Learning

[Wu & Lee, Interspeech 16]




Learning Target:
Maximizing Retrieval Quality, Minimizing User Labor

[Wu & Lee, Interspeech 16]

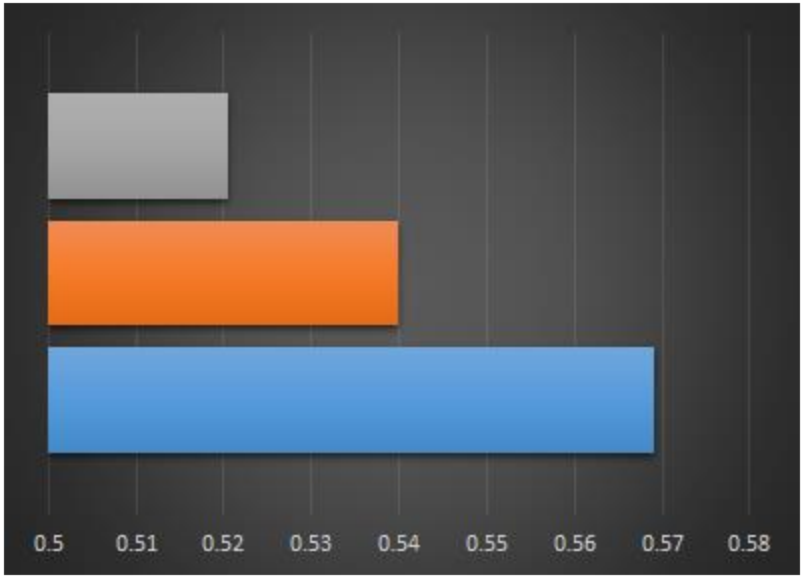
# Experimental Results

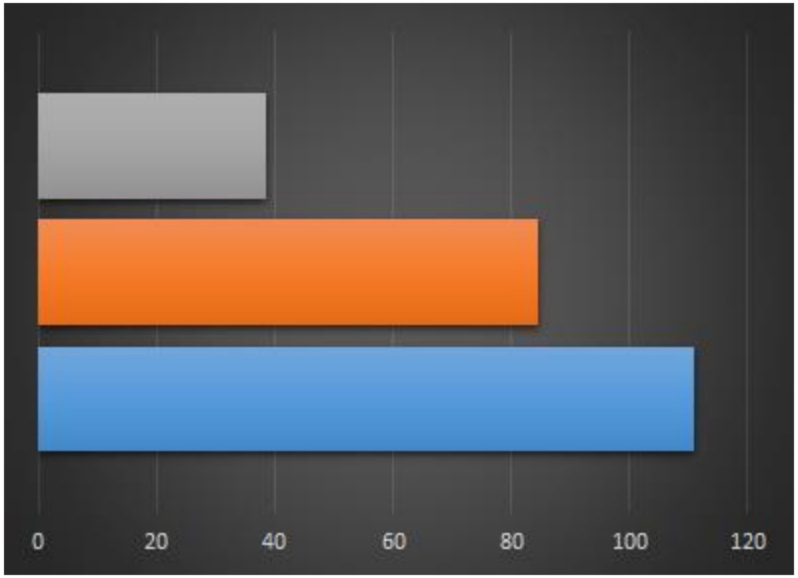
Oral, Friday, 6:40 p.m.,
Spoken Term Detection

- Broadcast news, semantic retrieval

Optimization Target:

***Retrieval Quality (MAP)*** ***Retrieval Quality - User labor***




Hand-crafted MDP Deep Reinforcement Learning

# Deep Reinforcement Learning v.s. MDP for interactive retrieval

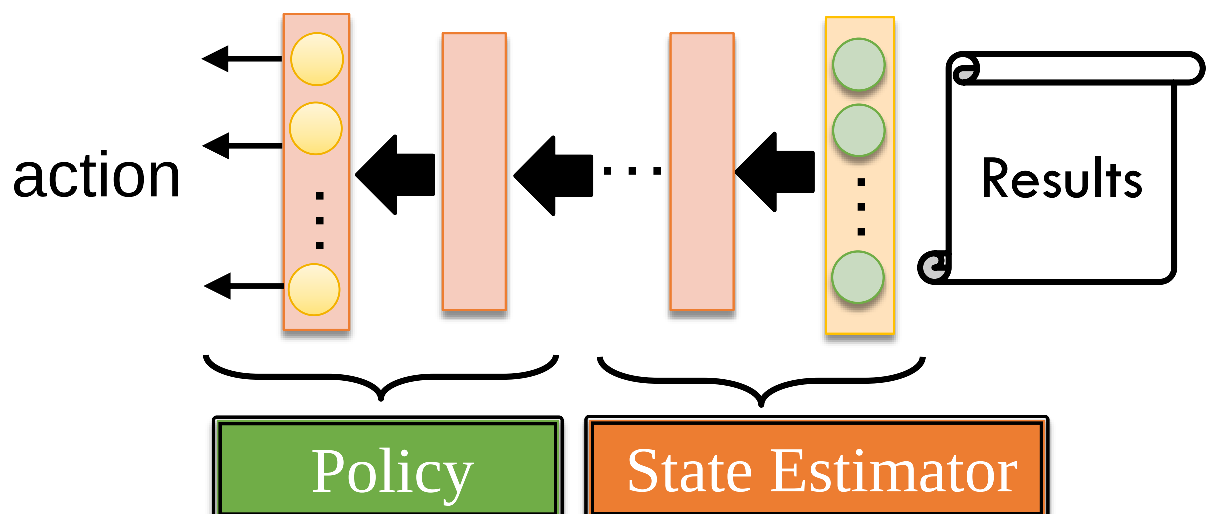
- MDP for interactive retrieval [Wen & Lee, Interspeech 12][Wen & Lee, ICASSP 13]


The two stages were learned separately.

- Deep

  End-to-end learning

  No hand-crafted states

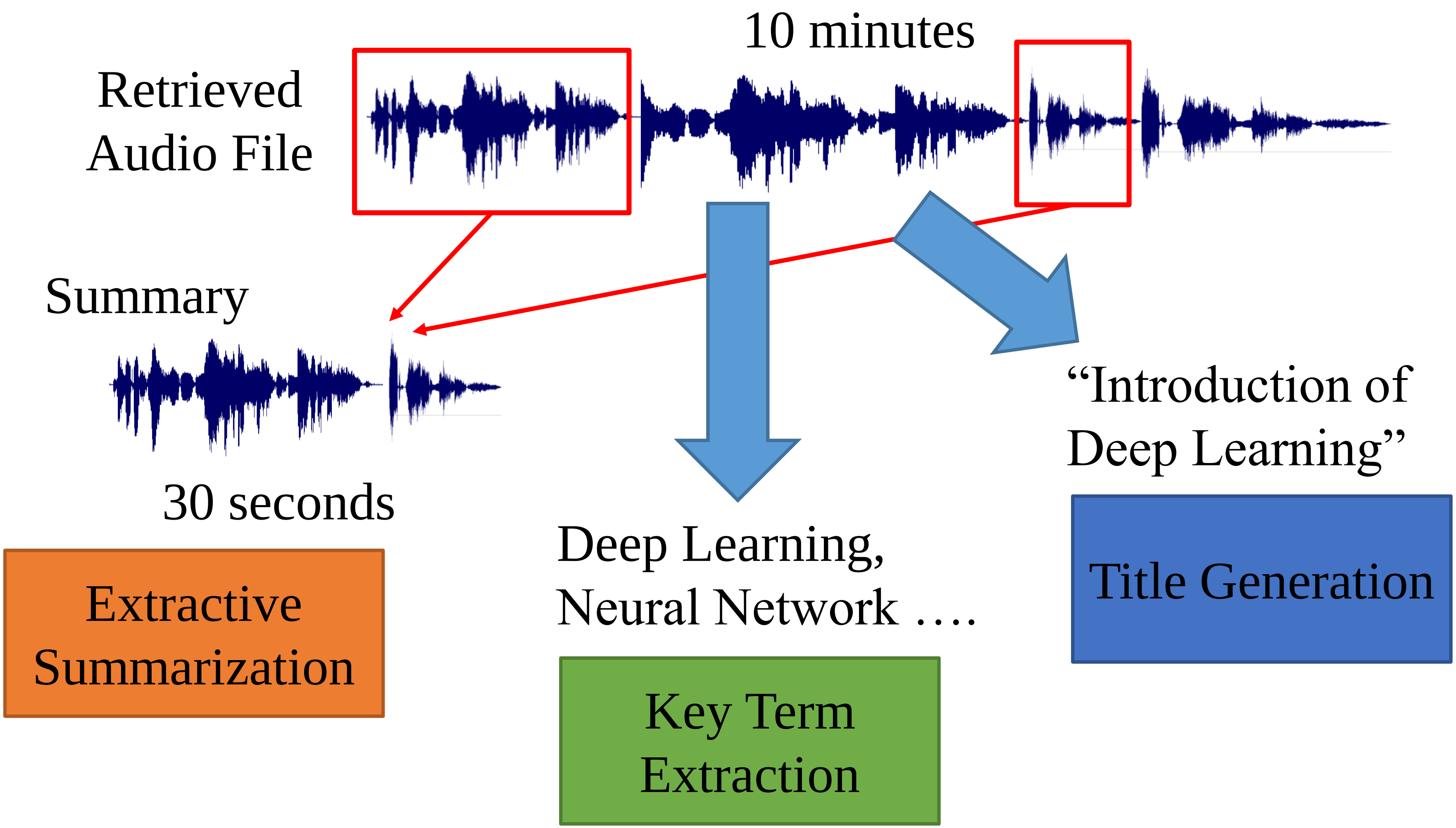


# New Direction 5-2:

## Speech Content is Difficult to Browse!

## Extracting Core Information

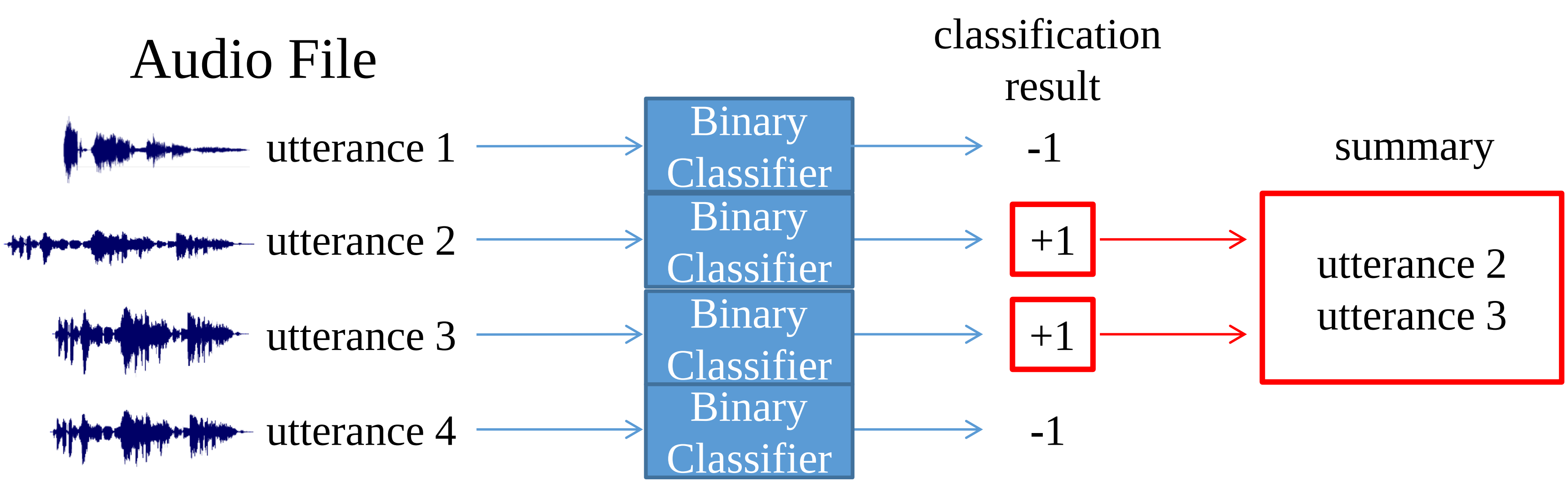
# Extracting Core Information



# Summarization

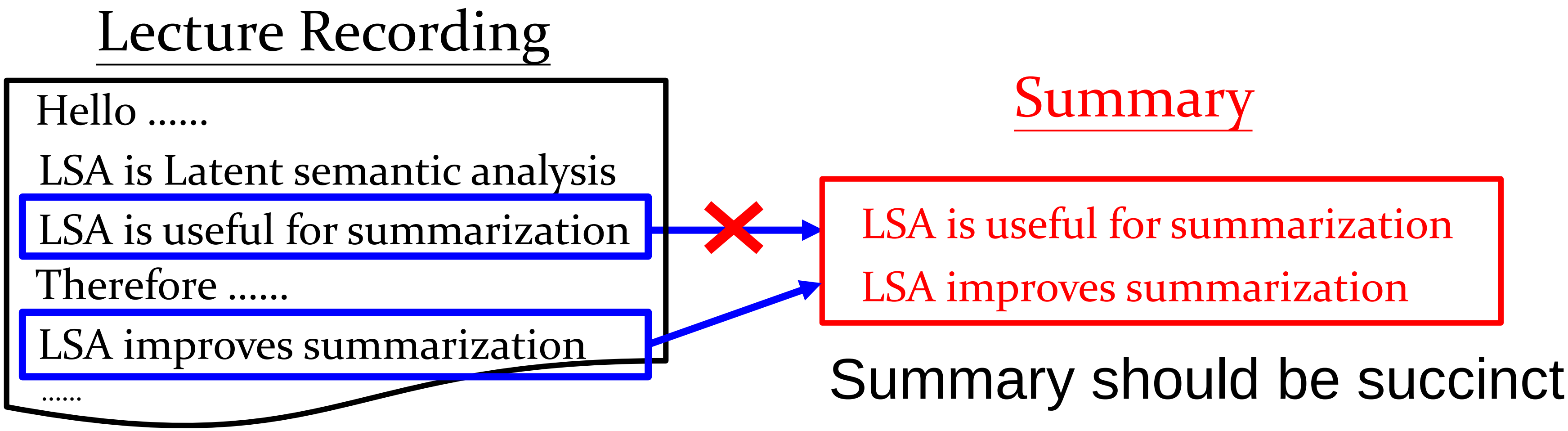
Reference: **13 Speech Summarization** (Gokhan Tur, Renato De Mori, Yang Liu, Dilek Hakkani-Tür). G. Tur and R. DeMori, Spoken Language Understanding: Systems for Extracting Semantic Information from Speech.

- **Unsupervised Approach**: Maximum Margin Relevance (MMR) and Graph-based Approach
- **Supervised approach**
  - Naïve approach: Summarization problem can be formulated as binary classification



# Summarization – Binary Classification

- Binary classifier individually considers each utterance
- Not sufficient
  - Example: summary should be concise



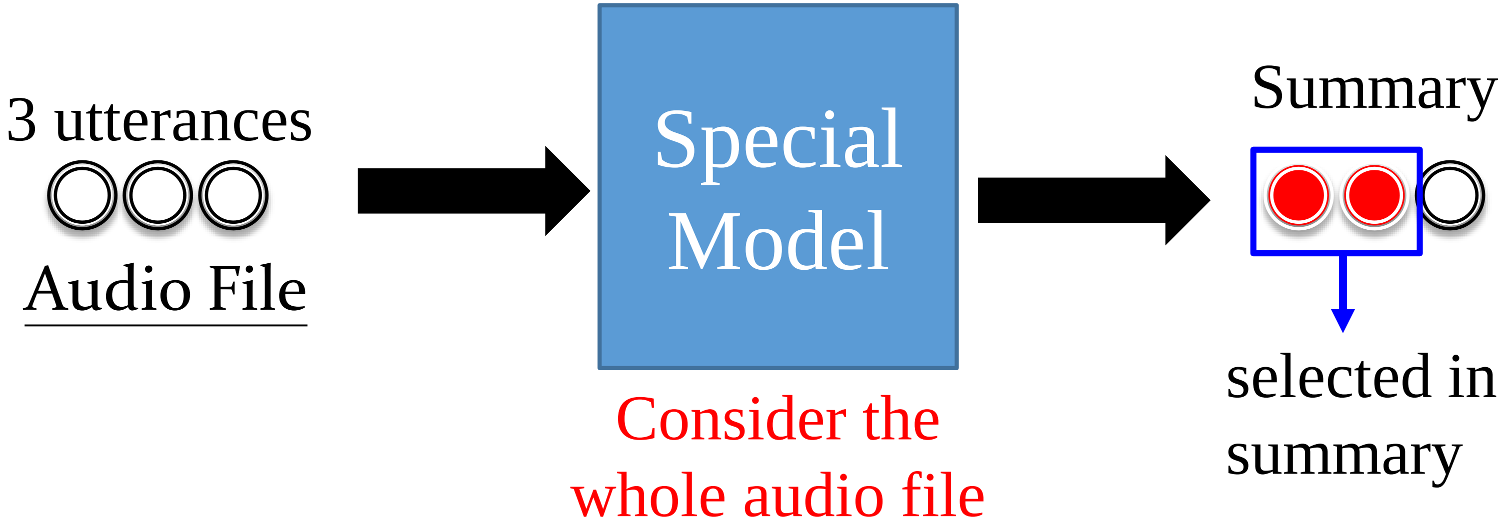

To generate a good summary, "***global information***" should be considered

More advanced machine learning techniques

# Summarization
# - Considering Global Information

- Learn a special model
  - Input: whole audio file
  - Output: summary

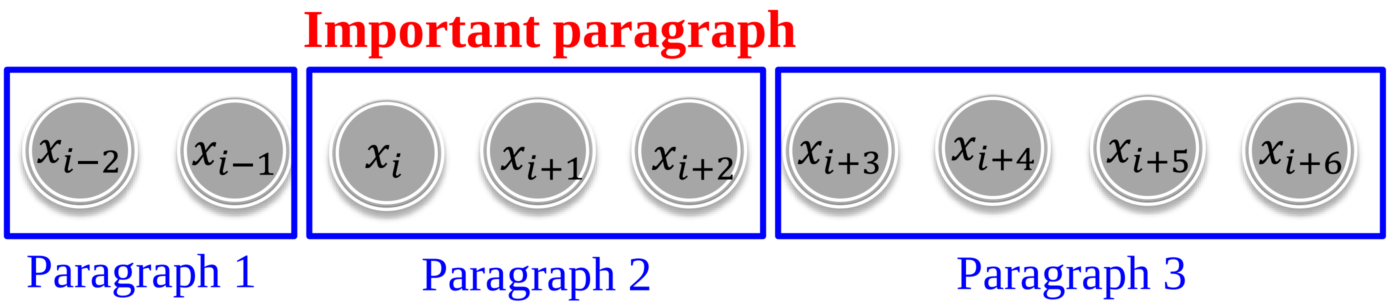
[Lee & Lee, ICASSP 13]
[Lee & Lee, Interspeech 12]



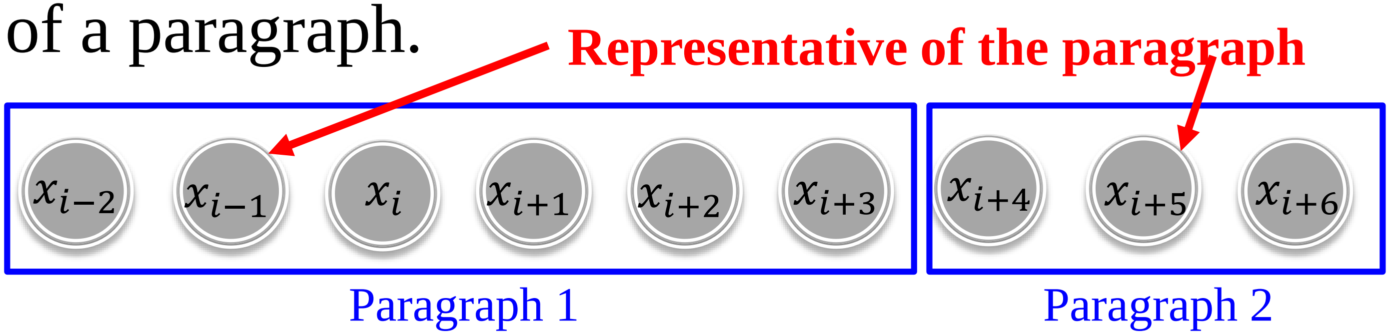

Structured SVM: I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support Vector Learning for Interdependent and Structured Output Spaces, ICML, 2004.

# Summarization
# - Structure in Spoken Content

- Temporal structure helps summarization
  - Long summary: consecutive utterances in a paragraph are more likely to be




  - Short summary: one utterance is selected on behalf of a paragraph.

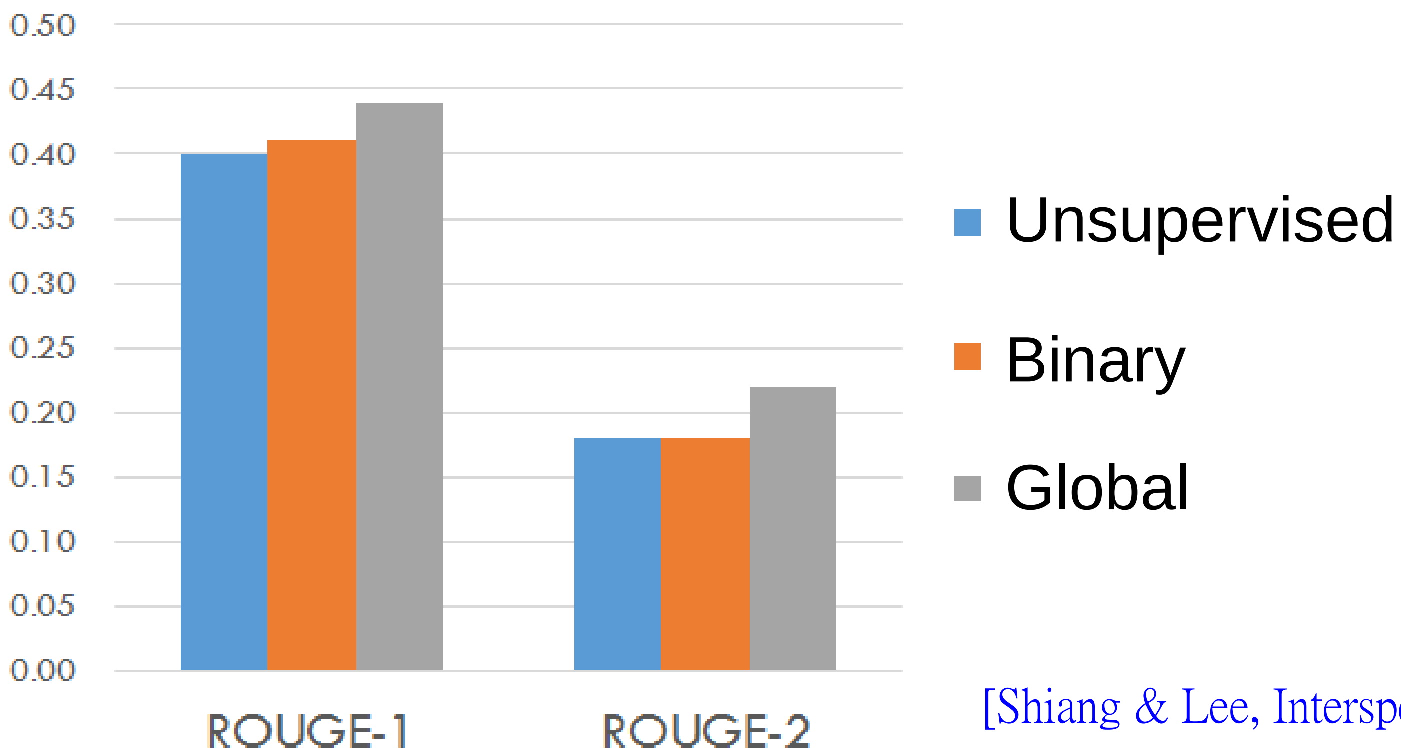


# Summarization
# - Structure in Spoken Content

- Structure in text are clear
  - Paragraph boundaries are directly known
- **For spoken content, there is no obvious structure**
  - The structure can be considered as "hidden variables"
  - Jointly learning structure of spoken document and summarization [Shiang & Lee, Interspeech 13]

# Summarization - Experiments

- Evaluation Measure: ROUGE-1 and ROUGE-2
  - Larger scores means the machine-generated summaries is more similar to human-generated summaries.




[Shiang & Lee, Interspeech 13]

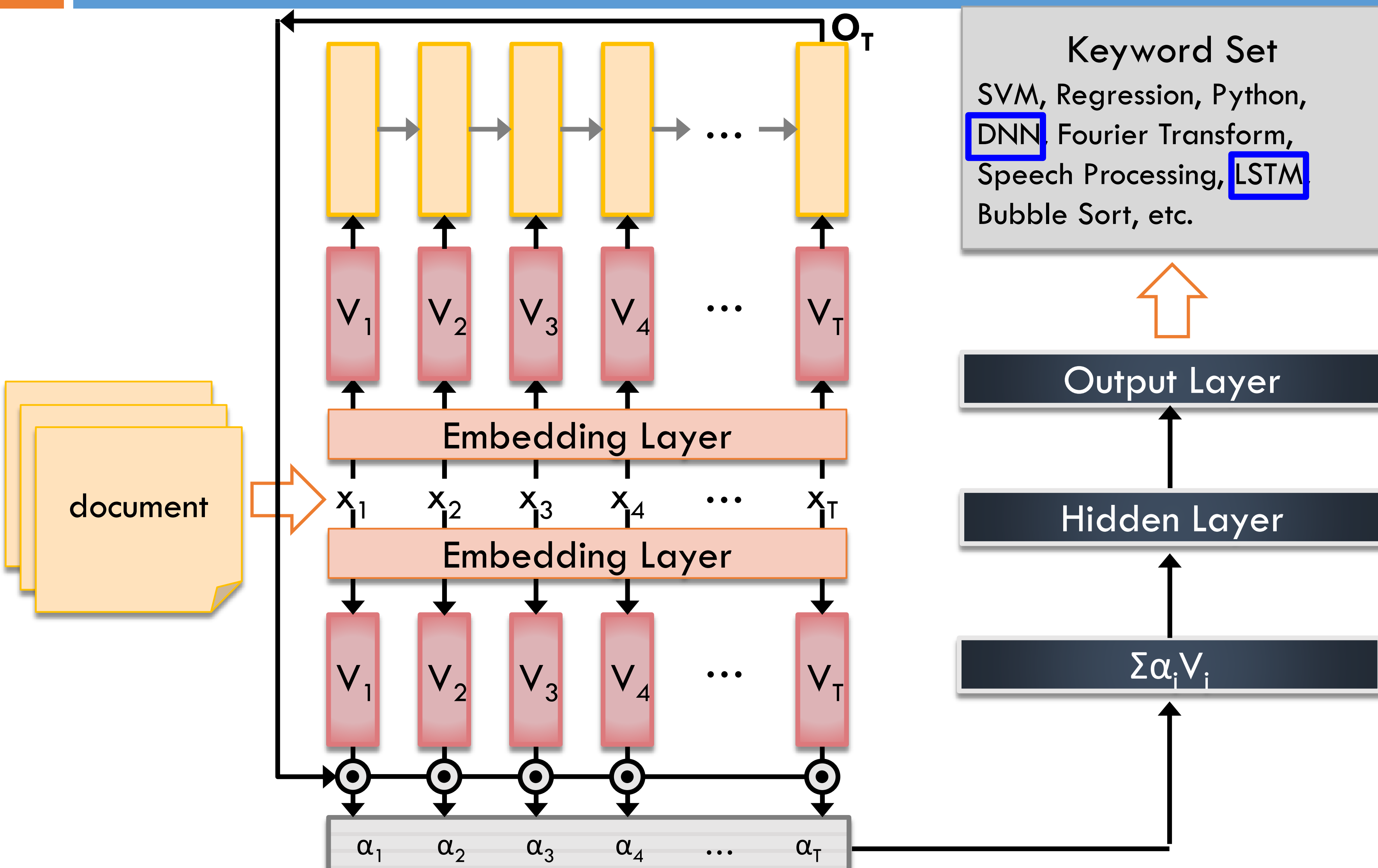
# Key Term Extraction

- TF-IDF is a good measure for identifying key terms [E. D'Avanzo, DUC 04][Jiang, SIGIR 09]
- Feature parameters from latent topic models [Hazen, Interspeech 11] [Chen & Lee, SLT 10]
  - Key terms are usually focused on small number of topics
- Prosodic Features [Chen & Lee, ICASSP 12]
  - slightly lower speed, higher energy, wider pitch range
- Machine Learning methods
  - Input: a term, output: key term or not [Liu, SLT 08][Chen & Lee, SLT 10]
  - Input: a document, output: key terms in the document [Kamal Sarkar, arXiv, 2010]

# Key Term Extraction – Deep Learning

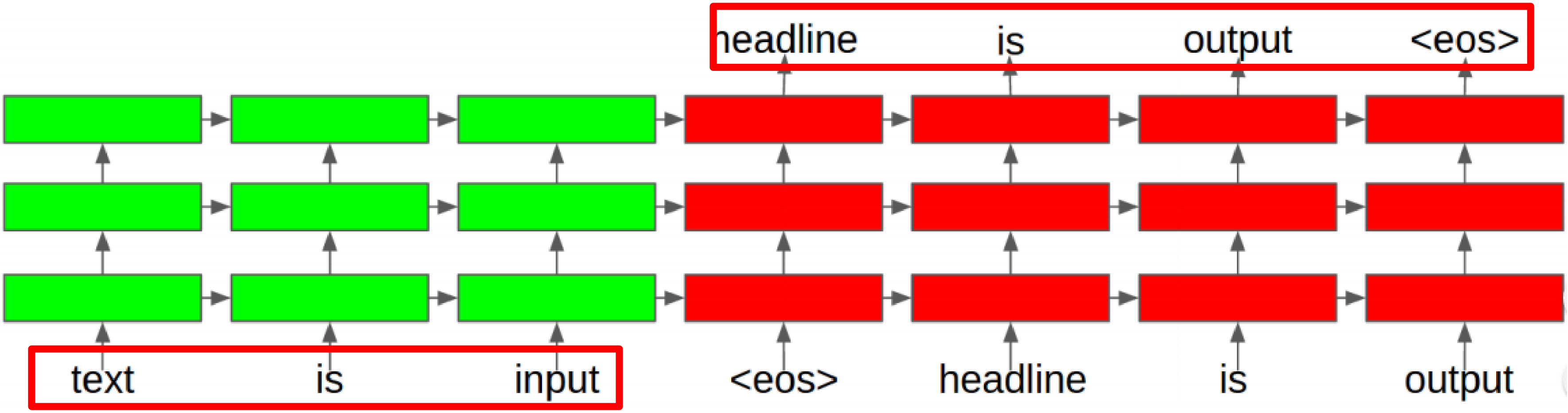
[Shen & Lee, Interspeech 16]

Poster, Sunday, 1:30 p.m., Dialogue Systems and Analysis of Dialogue



# Title Generation

- **Deep Learning based Approach** [Alexander M Rush, EMNLP 15][Chopra, NAACL 16][Lopyrev, arXiv 2015][Shen, arXiv 2016]
  - Based on Sequence-to-sequence learning
  - Input: a document (word sequence), output: its title (shorter word sequence)




https://arxiv.org/pdf/1512.01712v1.pdf

# New Direction 5-3:

## Speech Content is Difficult to Browse!

## Organizing Retrieved Results

# Introduction

- Organizing the retrieval results to help users know what is retrieved
- Taking retrieving on-line lectures as example
  - Searching spoken lectures is a very good application for spoken content retrieval
  - The speech of the instructors conveys most knowledge in the lectures

# Retrieving One Course

- NTU Virtual Instructor

[Kong & Lee, ICASSP 09]
[Lee & Lee, IEEE/ACM T. ASL 14]


Searching the course Digital Speech Processing of NTU

# Massive Open On-line Courses (MOOCs)

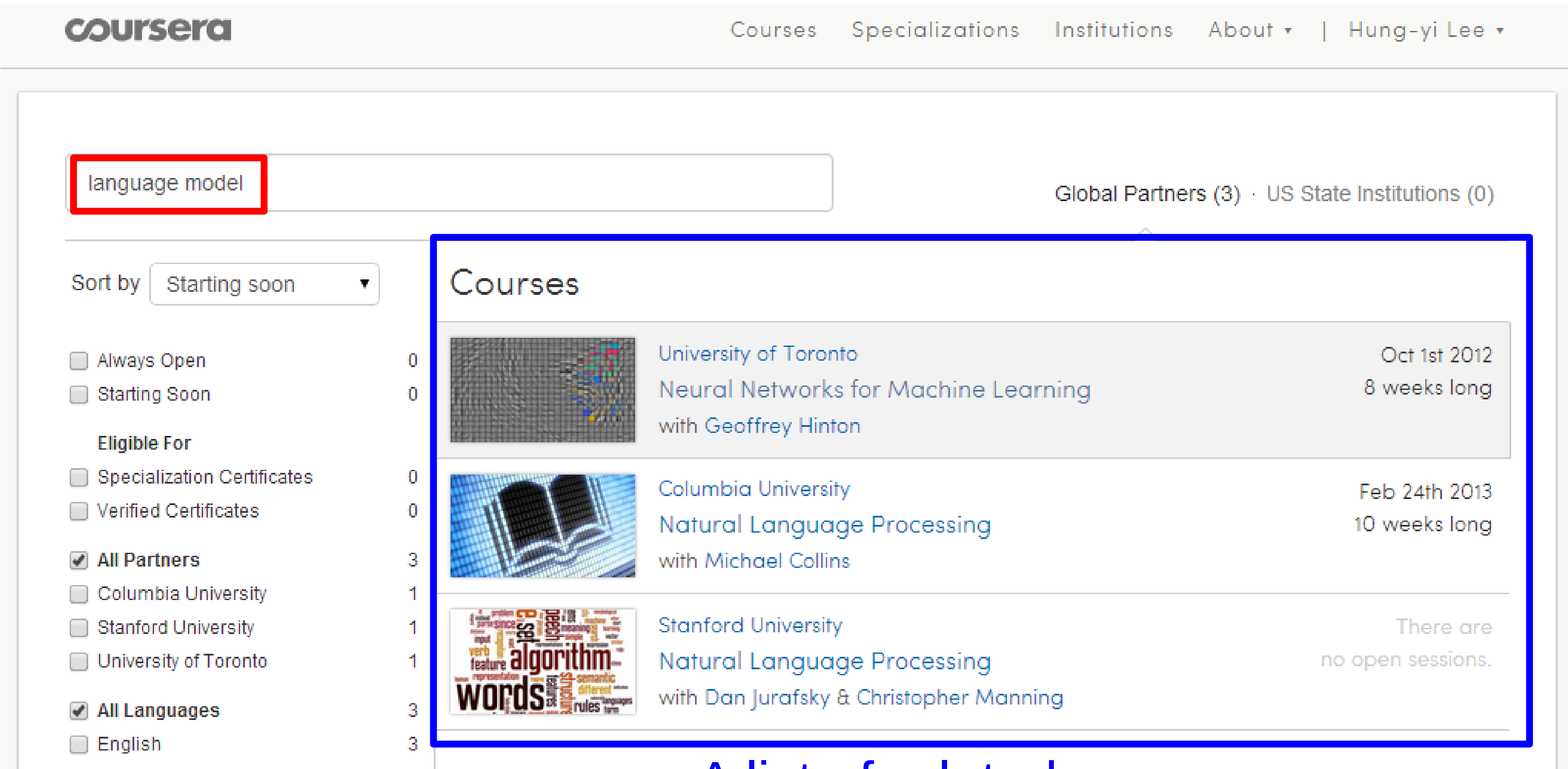
- Enormous on-line courses

coursera

edX

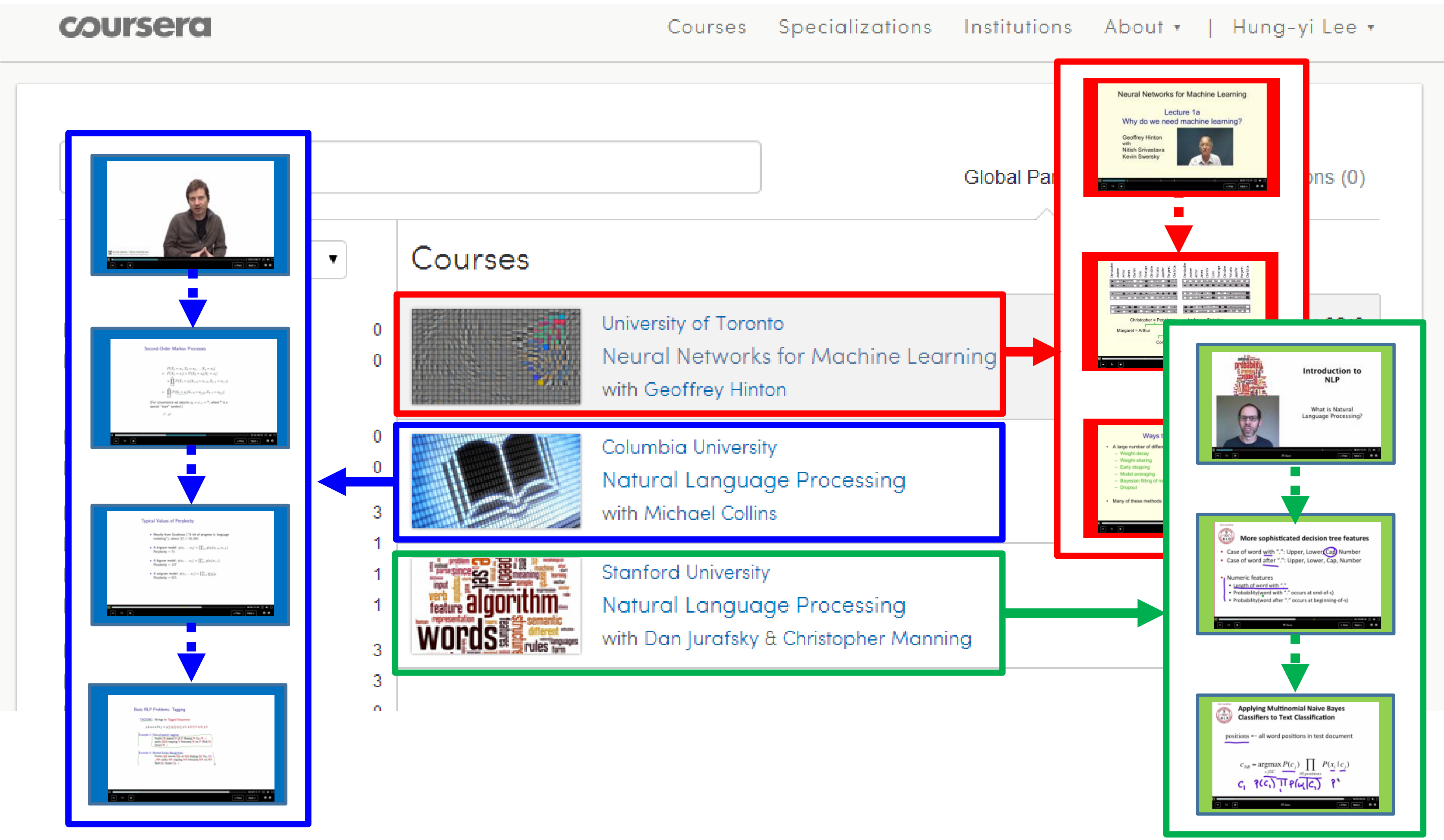
UDACITY

# Today's Retrieval Techniques



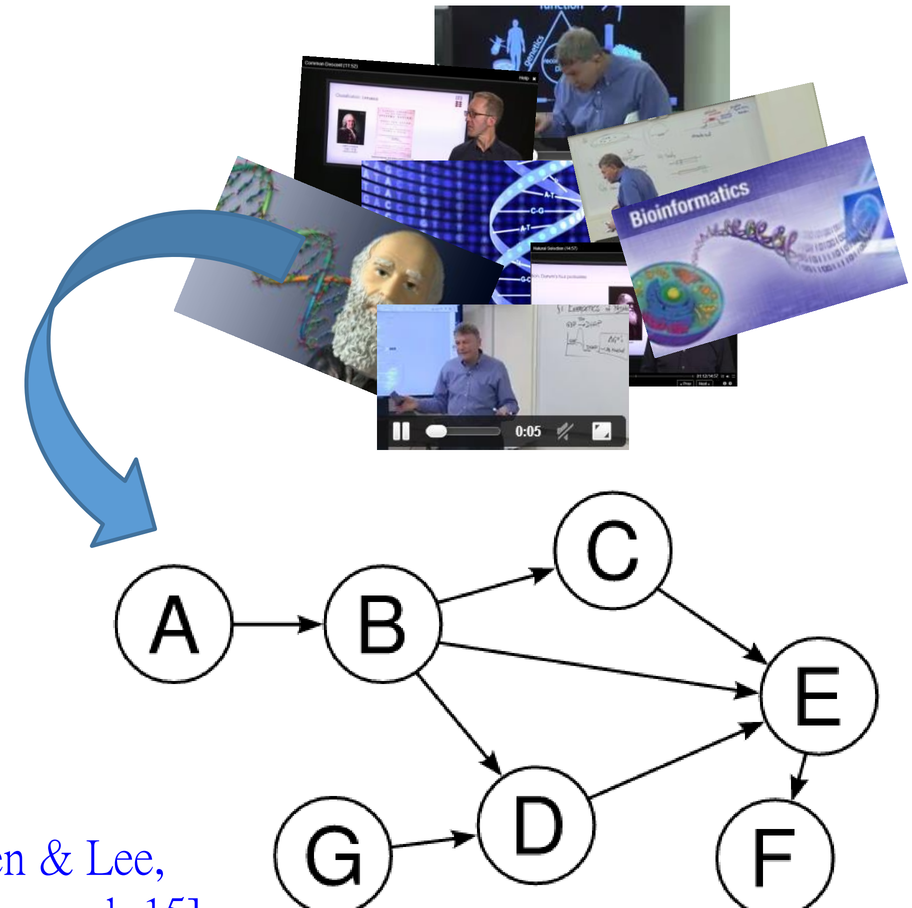

A list of related courses

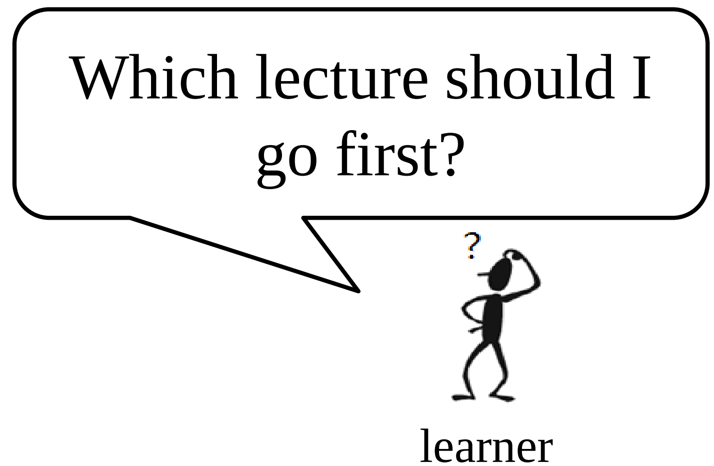
# Today's Retrieval Techniques



# More is less ……

- Given all the related lectures from different courses



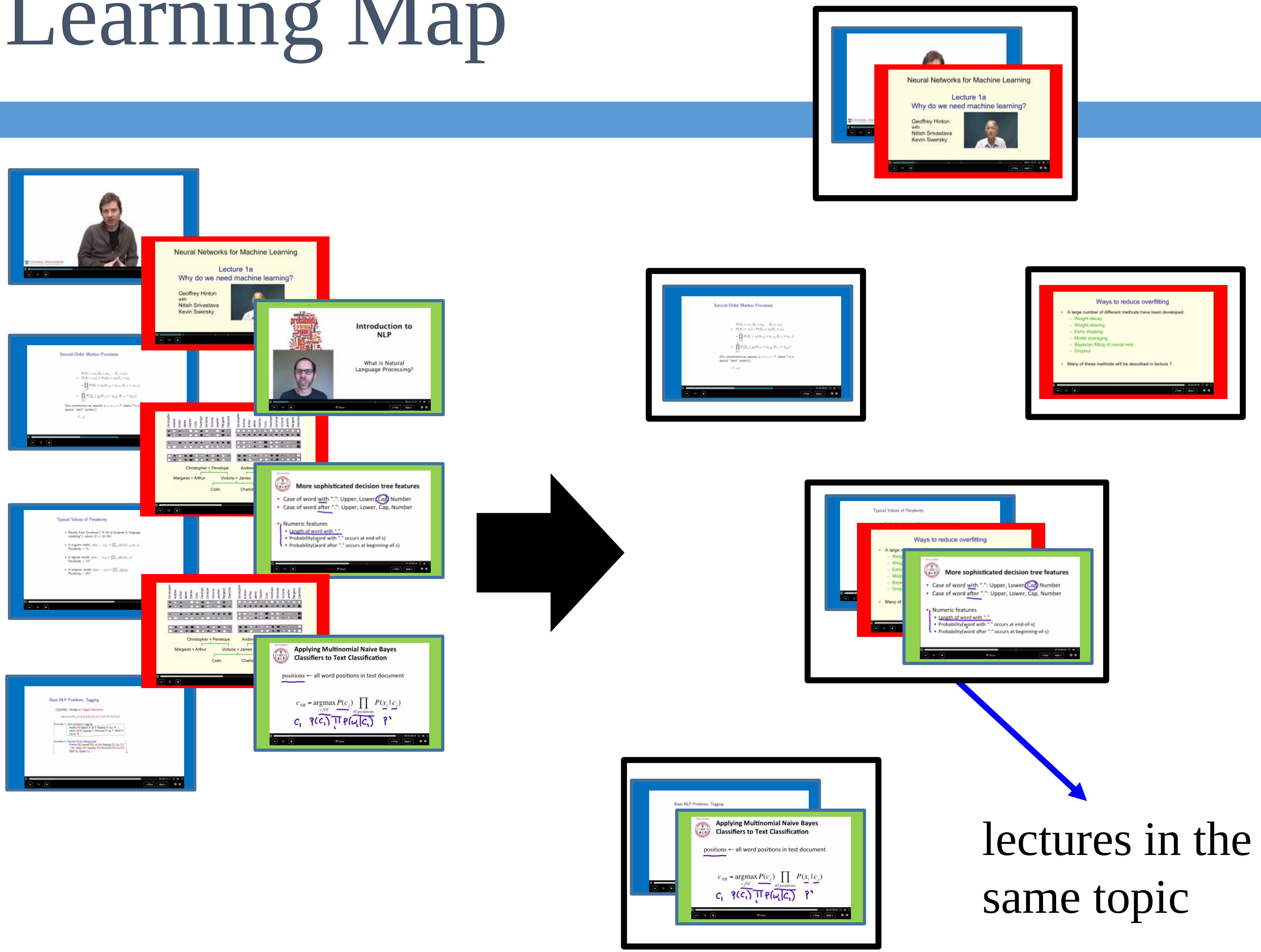




***<u>Learning Map</u>***

- Nodes: lectures in the same topics
- Edges: suggested learning order

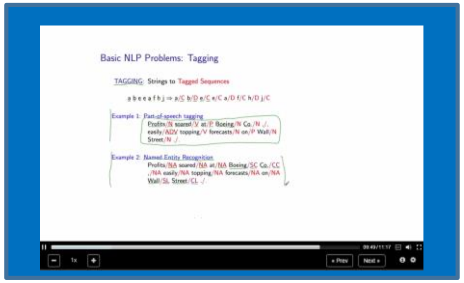
[Shen & Lee, Interspeech 15]

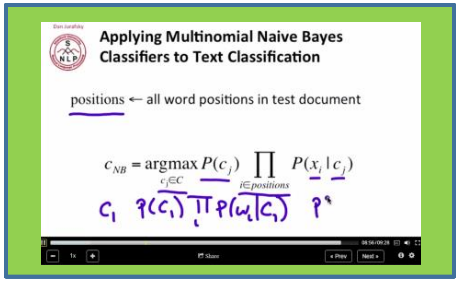
# Learning Map



# Lectures in the same topic

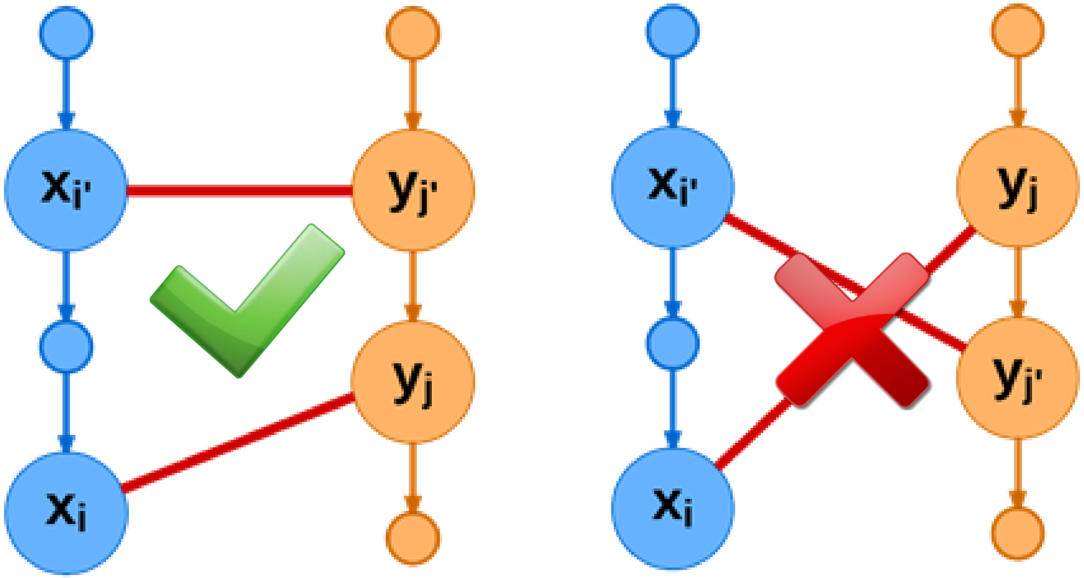
"Local" Information:



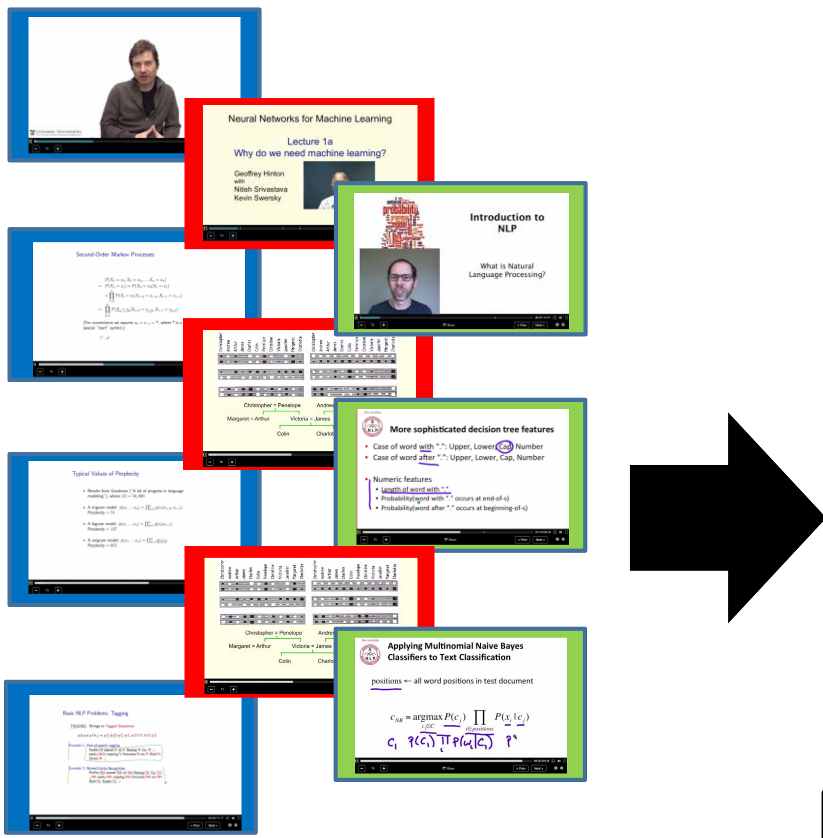



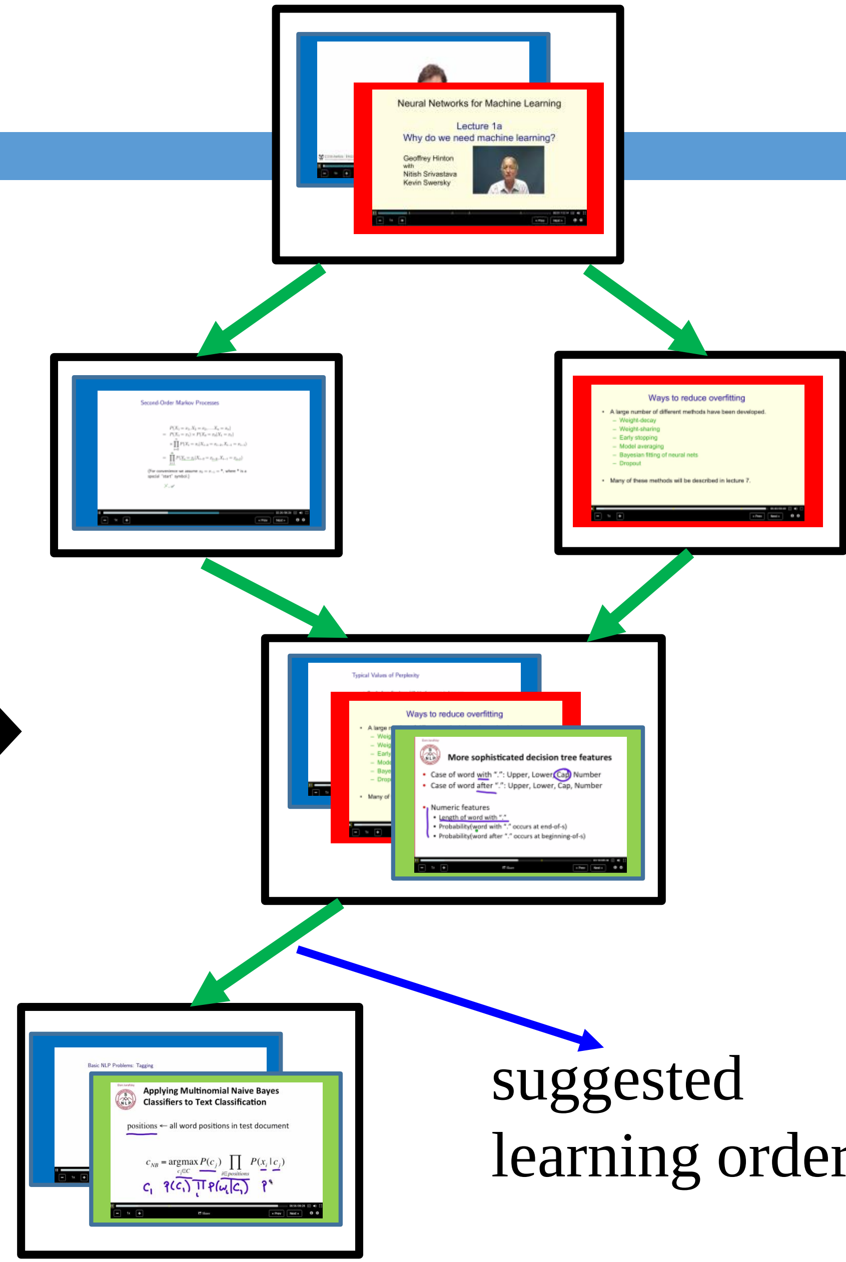

Similarity?

"Global" Information:



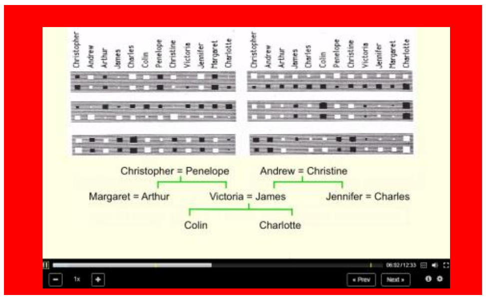
# Learning Map



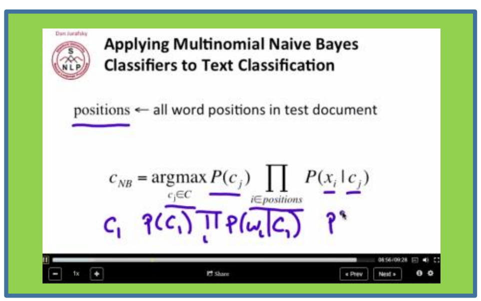



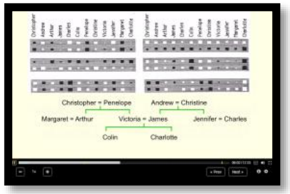
# Prerequisite

Lectures in
different courses

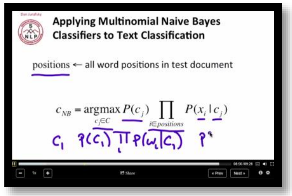



Prerequisite?



Content of
Lecture A



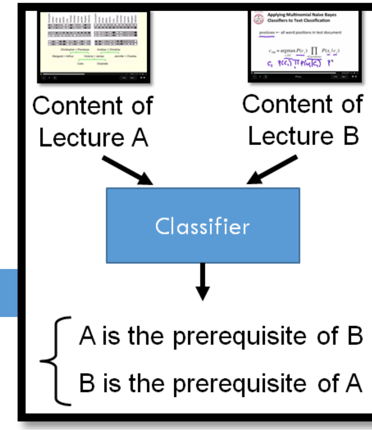

Content of
Lecture B

Classifier

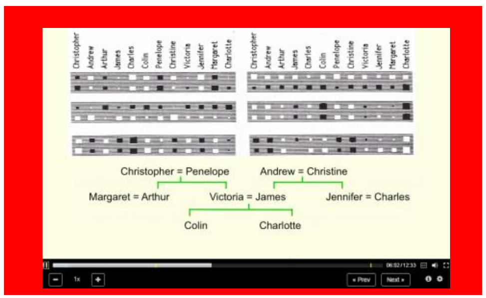
A is the prerequisite of B

B is the prerequisite of A

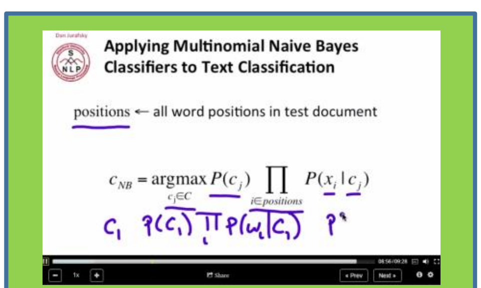
# Prerequisite




Lectures in different courses

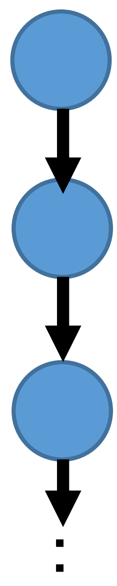



Prerequisite?




The existing courses on-line can be the training data

An existing course



Lecture 1

Lecture 2

Lecture 3

⋮

Lecture 1 is a prerequisite of lecture 2

Lecture 2 is a prerequisite of lecture 3

……

Training examples

# Demo

# Vision: Personalized Courses

on-line learning material

- I want to learn “deep learning”.
- I am a graduate student of computer science.
- I can spend 6 hours.

Learner


I open a course for you.

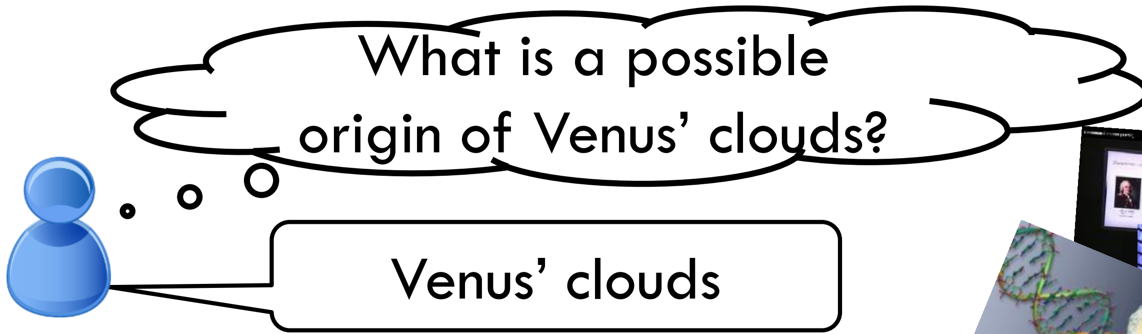
- With MOOCs and Spoken Language Processing techniques
  - It is possible to have a personalized course for each learning need.

# New Direction 5-4:

## Speech Content is Difficult to Browse!

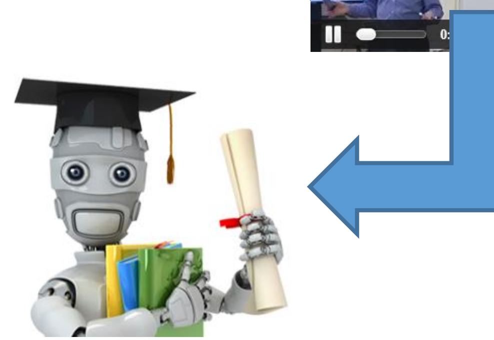
## Spoken Question Answering

# Spoken Question Answering





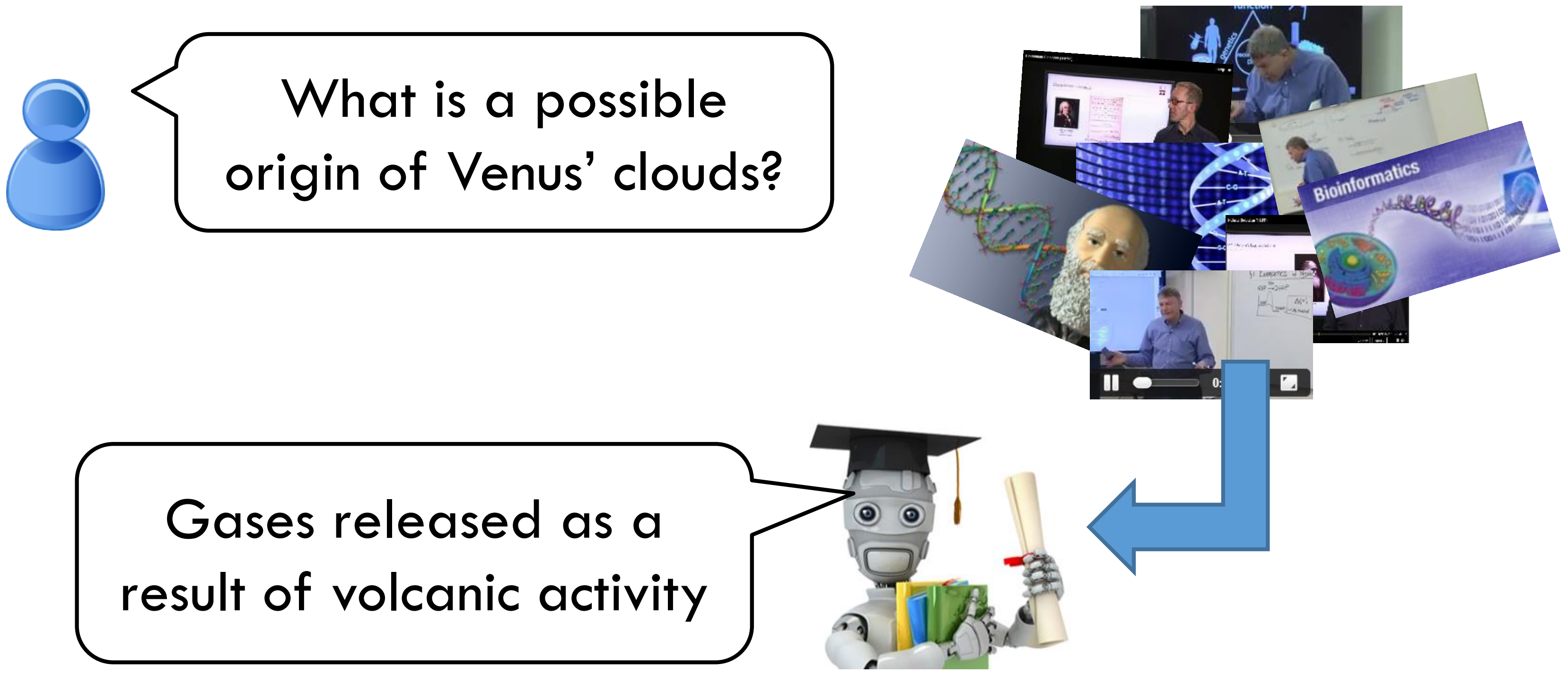
Taking some time to find the answer



Lectures
about Venus

***<u>Spoken Content Retrieval</u>***

# Spoken Question Answering




***<u>Spoken Question Answering</u>***: Machine answers questions based on the information in spoken content

# Spoken Question Answering

- Question Answering in Speech Transcripts (QAST) has been a well-known evaluation program of spoken question answering.
  - 2007, 2008, 2009

Reference: **6 Spoken Question Answering** (*Sophie Rosset, Olivier Galibert and Lori Lamel*). G. Tur and R. DeMori, Spoken Language Understanding: Systems for Extracting Semantic Information from Speech.

- Focused on factoid questions in the previous study
  - E.g. "What is name of the highest mountain in Taiwan?".
- To answer more difficult questions, machine has to understand questions and spoken documents.
  - How good can it achieve?

# New task for Machine Comprehension of Spoken Content

- **TOEFL Listening Comprehension Test by Machine**

[Tseng & Lee, Interspeech 16]

Audio Story: (The original story is 5 min long.)

Question: “ What is a possible origin of Venus’ clouds? ”

Choices:

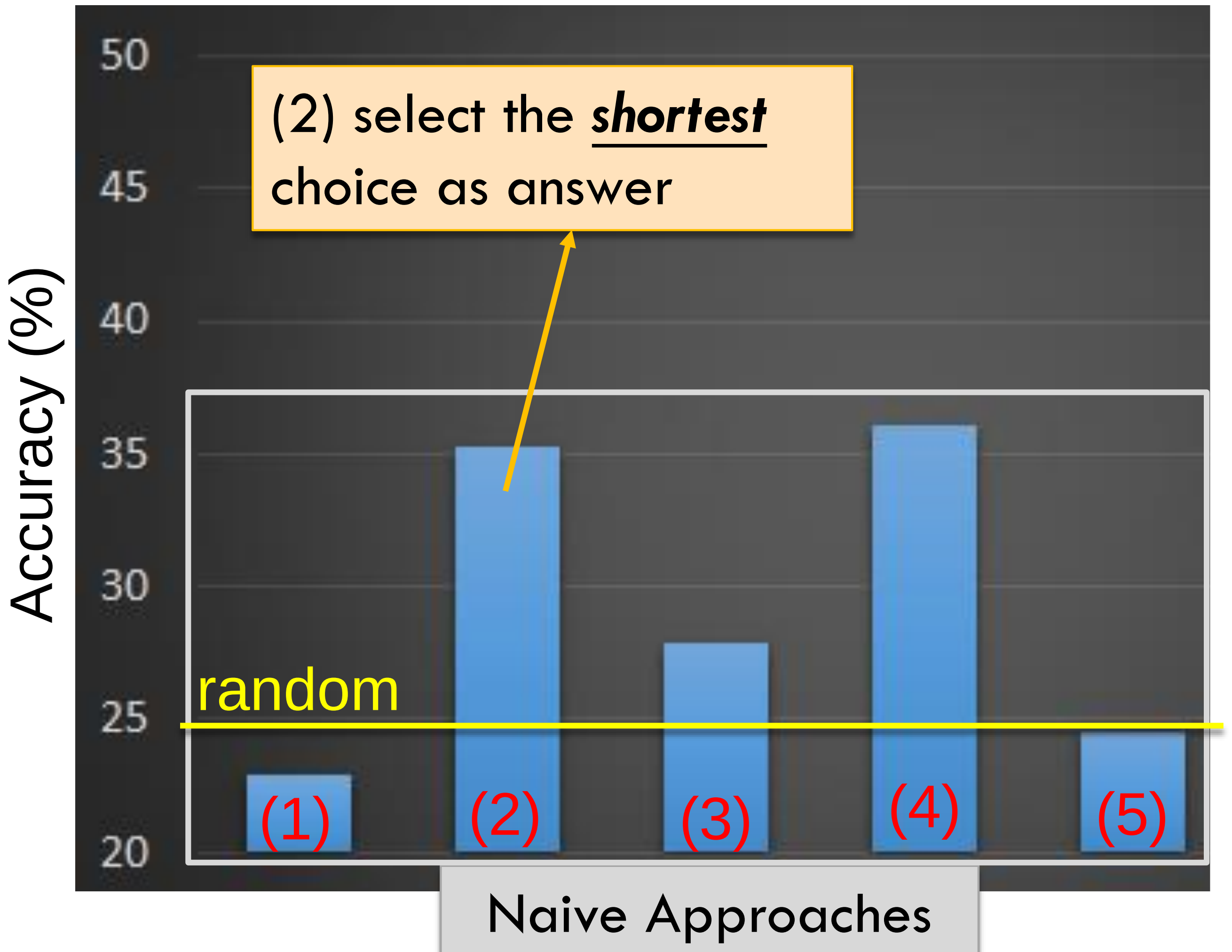
(A) gases released as a result of volcanic activity

(B) chemical reactions caused by high surface temperatures

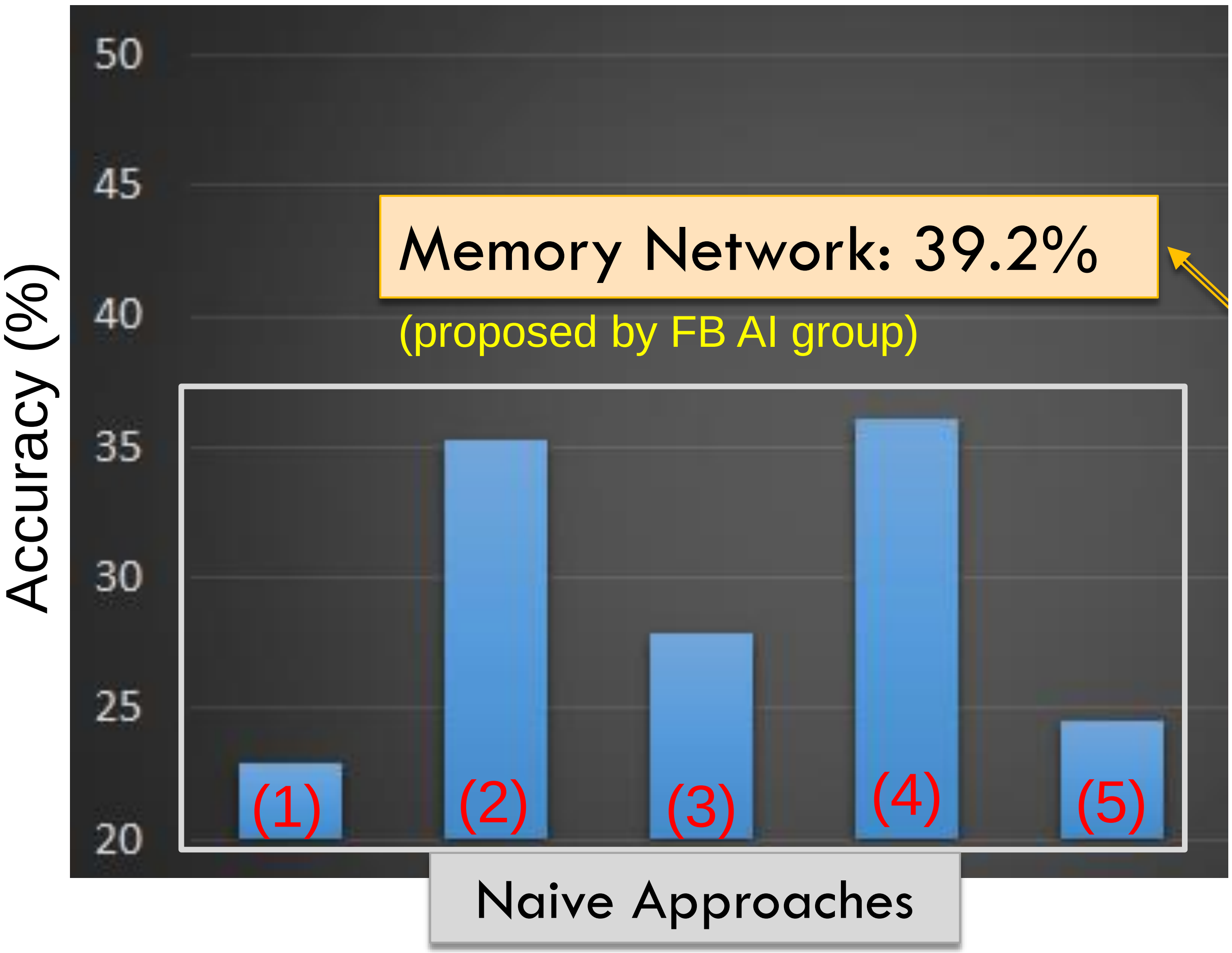
(C) bursts of radio energy from the plane's surface

(D) strong winds that blow dust into the atmosphere

# Simple Baselines



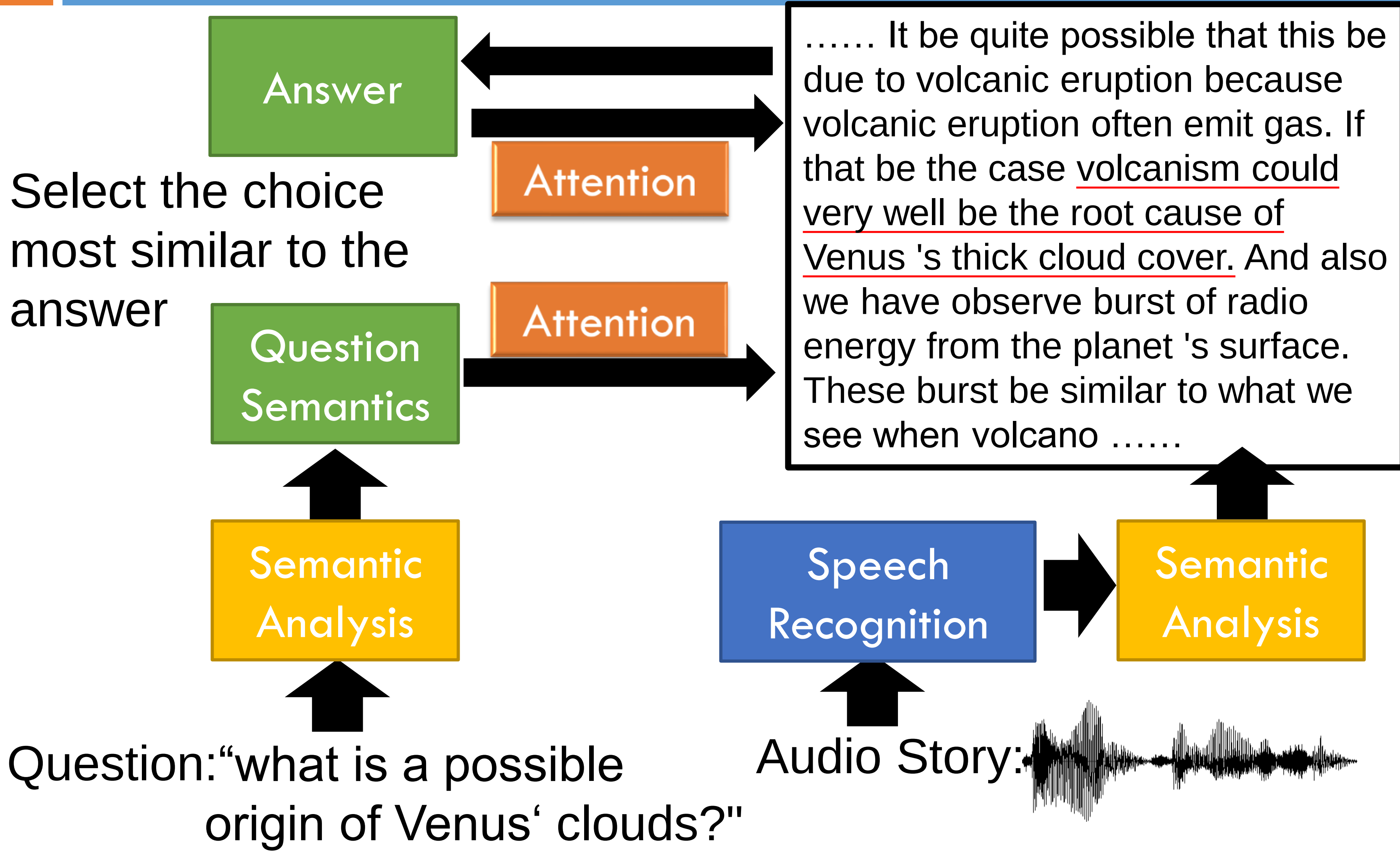
# Results



Accuracy (%)
50
45
40
35
30
25
20
Memory Network: 39.2%
(proposed by FB AI group)
(1)
(2)
(3)
(4)
(5)
Naive Approaches

# Model Architecture

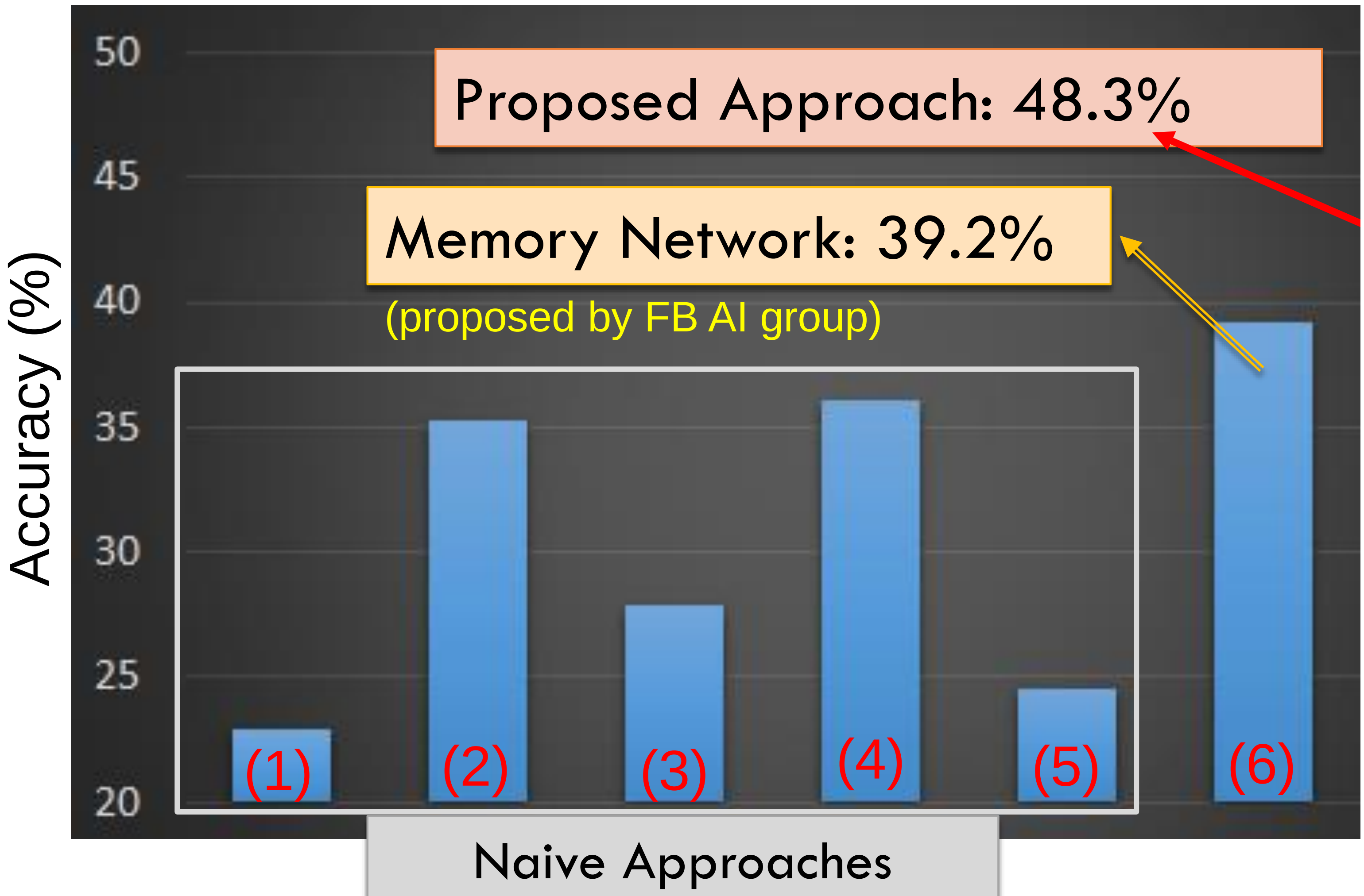
The model is learned end-to-end.



[Tseng & Lee, Interspeech 16]

# Results

Poster, Sunday, 1:30 p.m., Dialogue Systems and Analysis of Dialogue



# Concluding Remarks

# Conclusion Remarks

- New research directions for spoken content retrieval
  - Modified ASR for Retrieval Purposes
  - Incorporating Those Information Lost in ASR
  - No Speech Recognition!
  - Special Semantic Retrieval Techniques for Spoken Content
  - Spoken Content is Difficult to Browse!

# Take-Home Message

Spoken Content Retrieval

$\cancel{=}$

Speech Recognition

+

Text Retrieval

# Spoken Content Retrieval

300 hrs multimedia is uploaded per minute. (2015.01)

1874 courses on coursera (2016.04)

- Nobody is able to go through the data.
- In these multimedia, the spoken part carries very important information about the content
- Spoken content retrieval: Machine listens to the data, and extract the desired information for each individual user.
  - Just as Google does on text data

# Overview Paper

- Lin-shan Lee, James Glass, Hung-yi Lee, Chun-an Chan, "Spoken Content Retrieval —Beyond Cascading Speech Recognition with Text Retrieval," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol.23, no.9, pp.1389-1420, Sept. 2015
- http://speech.ee.ntu.edu.tw/~tlkagk/paper/Overview.pdf
- This tutorial includes updated information after this paper is published.

*Thank You for Your Attention*