Spoken Document Retrieval and Browsing

Ciprian Chelba

Overview

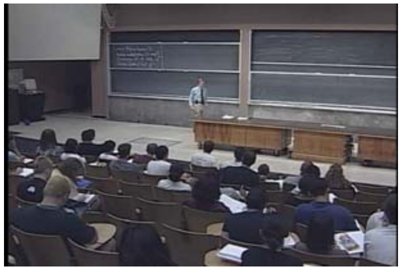
- Introduction
- Speech Recognition for Spoken Documents
- Spoken Document Retrieval & Browsing
- Summary and Questions

Motivation

- In the past decade there has been a dramatic increase in the availability of on-line audio-visual material…
  – More than 50% percent of IP traffic is video
- …and this trend will only continue as cost of producing audio-visual content continues to drop

Broadcast News

Podcasts



Academic Lectures

- Raw audio-visual material is difficult to search and browse
- Keyword driven Spoken Document Retrieval (SDR):
  – User provides a set of relevant *query terms*
  – Search engine needs to *return relevant spoken documents and provide an easy way to navigate them*

Spoken Document Processing

- **The goal is to enable users to:**
  - **Search for spoken documents as easily as they search for text**
  - **Accurately retrieve relevant spoken documents**
  - **Efficiently browse through returned *hits***
  - **Quickly find segments of spoken documents they would most like to listen to or watch**

- **Information (or meta-data) to enable search and retrieval:**
  - **Transcription of speech**
  - **Text summary of audio-visual material**
  - **Other relevant information:**
    - * speakers, time-aligned outline, etc.
    - * slides, other relevant text meta-data: title, author, etc.
    - * links pointing to spoken document from the www
    - * collaborative filtering (who else watched it?)

When Does Automatic Annotation Make Sense?

- Scale: Some repositories are too large to manually annotate
  - Collections of lectures collected over many years (Microsoft)
  - WWW video stores (Apple, Google, MSN, Yahoo, YouTube)
  - TV: all “new” English language programming is required by the FCC to be closed captioned
    http://www.fcc.gov/cgb/consumerfacts/closedcaption.html
- Cost: A basic text-transcription of a one hour lecture costs >$100
  - Some users have monetary restrictions
  - Amateur podcasters
  - Academic or non-profit organizations
- Privacy: Some data needs to remain secure
  - corporate customer service telephone conversations
  - business and personal voice-mails
  - VoIP chats

TREC SDR: "A Success Story"

- **The Text Retrieval Conference (TREC)**
  - **Pioneering work in spoken document retrieval (SDR)**
  - **SDR evaluations from 1997-2000 (TREC-6 toTREC-9)**
- **TREC-8 evaluation:**
  - **Focused on broadcast news data**
  - **22,000 stories from 500 hours of audio**
  - **Even fairly high ASR error rates produced document retrieval performance close to human generated transcripts**
  - **Key contributions:**
    - * Recognizer expansion using N-best lists
    - * query expansion, and document expansion
  - **Conclusion: SDR is "A success story" (Garofolo *et al*, 2000)**
- **Why don't ASR errors hurt performance?**
  - **Content words are often repeated providing redundancy**
  - **Semantically related words can offer support (Allan, 2003)**

Broadcast News: SDR Best-case Scenario

- **Broadcast news SDR is a best-case scenario for ASR:**
  - **Primarily prepared speech read by professional speakers**
  - **Spontaneous speech artifacts are largely absent**
  - **Language usage is similar to written materials**
  - **New vocabulary can be learned from daily text news articles**

State-of-the-art recognizers have word error rates ~10%

* comparable to the closed captioning WER (used as reference)

- **TREC queries were fairly long (10 words) and have low out-of-vocabulary (OOV) rate**
  - **Impact of query OOV rate on retrieval performance is high (Woodland et al., 2000)**

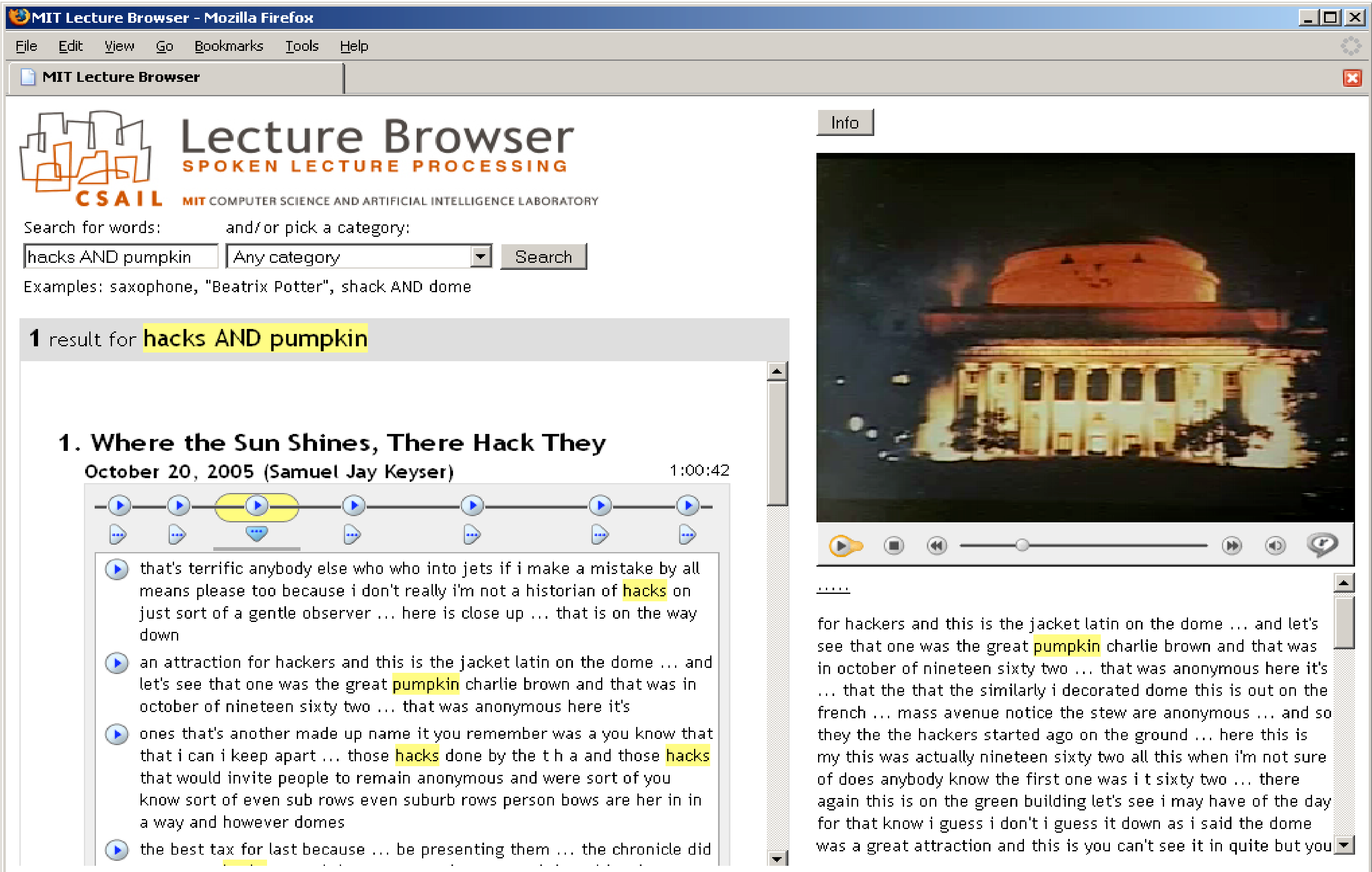
- **Vast amount of content is closed captioned**

Beyond Broadcast News

- **Many useful tasks are more difficult than broadcast news**
  - **Meeting annotation (e.g., Waibel *et al*, 2001)**
  - **Voice mail (e.g., SCANMail, Bacchiani *et al*, 2001))**
  - **Podcasts (e.g., Podzinger, www.podzinger.com)**
  - **Academic lectures (e.g., MIT iCampus)**
- **Primary difficulties due to limitations of ASR technology:**
  - **Highly spontaneous, unprepared speech**
  - **Topic-specific or person-specific vocabulary & language usage**
  - **Unknown content and topics potentially lacking support in general language model**
  - **Wide variety of accents and speaking styles**
  - **OOVs in queries: ASR vocabulary is not designed to recognize infrequent query terms, which are most useful for retrieval**
- **General SDR still has many challenges to solve**

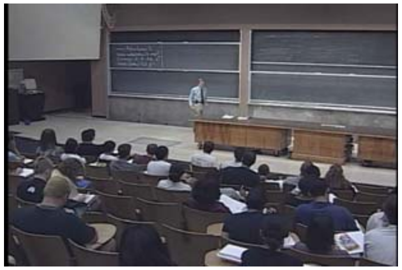
Demonstration of MIT Lecture Browser

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)



The Research Challenge

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)



1) I've been talking -- I've been multiplying matrices already, but certainly time for me to discuss the rules for matrix multiplication.
2) And the interesting part is the many ways you can do it, and they all give the same answer.
3) So it's -- and they're all important.
4) So matrix multiplication, and then, uh, come inverses.
5) So we're -- uh, we -- mentioned the inverse of a matrix, but there's -- that's a big deal.
6) Lots to do about inverses and how to find them.
7) Okay, so I'll begin with how to multiply two matrices.
8) First way, okay, so suppose I have a matrix A multiplying a matrix B and -- giving me a result -- well, I could call it C.
9) A times B. Okay.
10) Uh, so, I- let me just review the rule for w- for this entry.

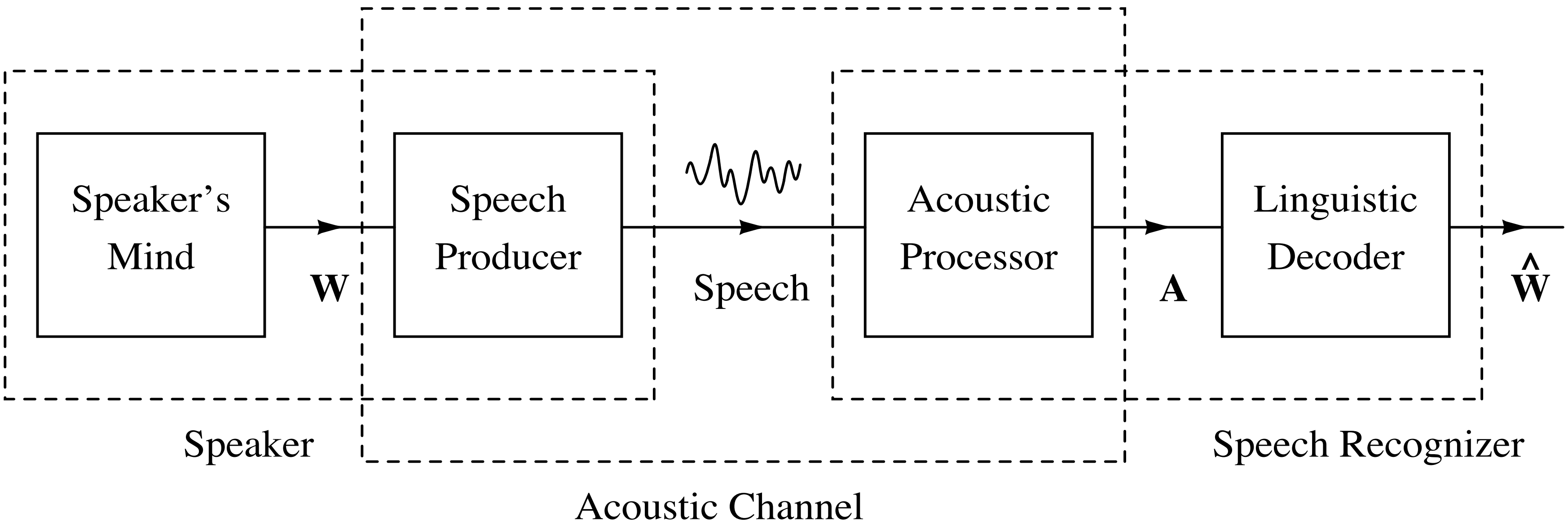
8 Rules of Matrix Multiplication:
The method for multiplying two matrices A and B to get $C = AB$ can be summarized as follows:
1) **Rule 8.1** To obtain the element in the r^{th} row and c^{th} column of C, multiply each element in the r^{th} row of A by the corresponding…

"I want to learn how to multiply matrices"

Speech Recognition for Spoken Documents

- Overview of Basic Speech Recognition Framework
- Language Modeling & Adaptation
- Acoustic Modeling & Adaptation

Speech Recognition: Probabilistic Framework



- **Find the most likely string of words, W, given the acoustic observations, A**

$$\max_W P(W|A)$$

Speech Recognition Evaluation

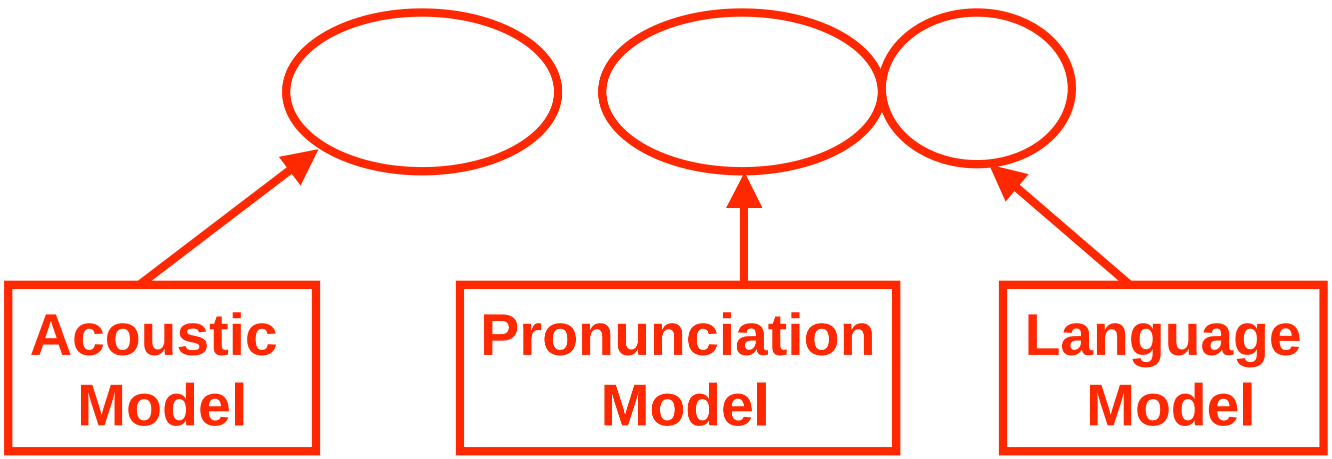
- **Word Error Rate (WER): counts substitutions/deletions/insertions in best string alignment**

TRN:	**UP**	**UPSTATE**	**NEW**	**YORK**	**SOMEWHERE**	**UH**		**OVER**	**OVER**	**HUGE**	**AREAS**
HYP:		**UPSTATE**	**NEW**	**YORK**	**SOMEWHERE**	**UH**	**ALL**	**ALL**	**THE**	**HUGE**	**AREAS**
	D	**0**	**0**	**0**	**0**	**0**	**I**	**S**	**S**	**0**	**0**
	1	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**

- **4 errors per 10 words in transcription; WER = 40%**
- **Evaluating WER reduction is computationally expensive; need to run recognizer**

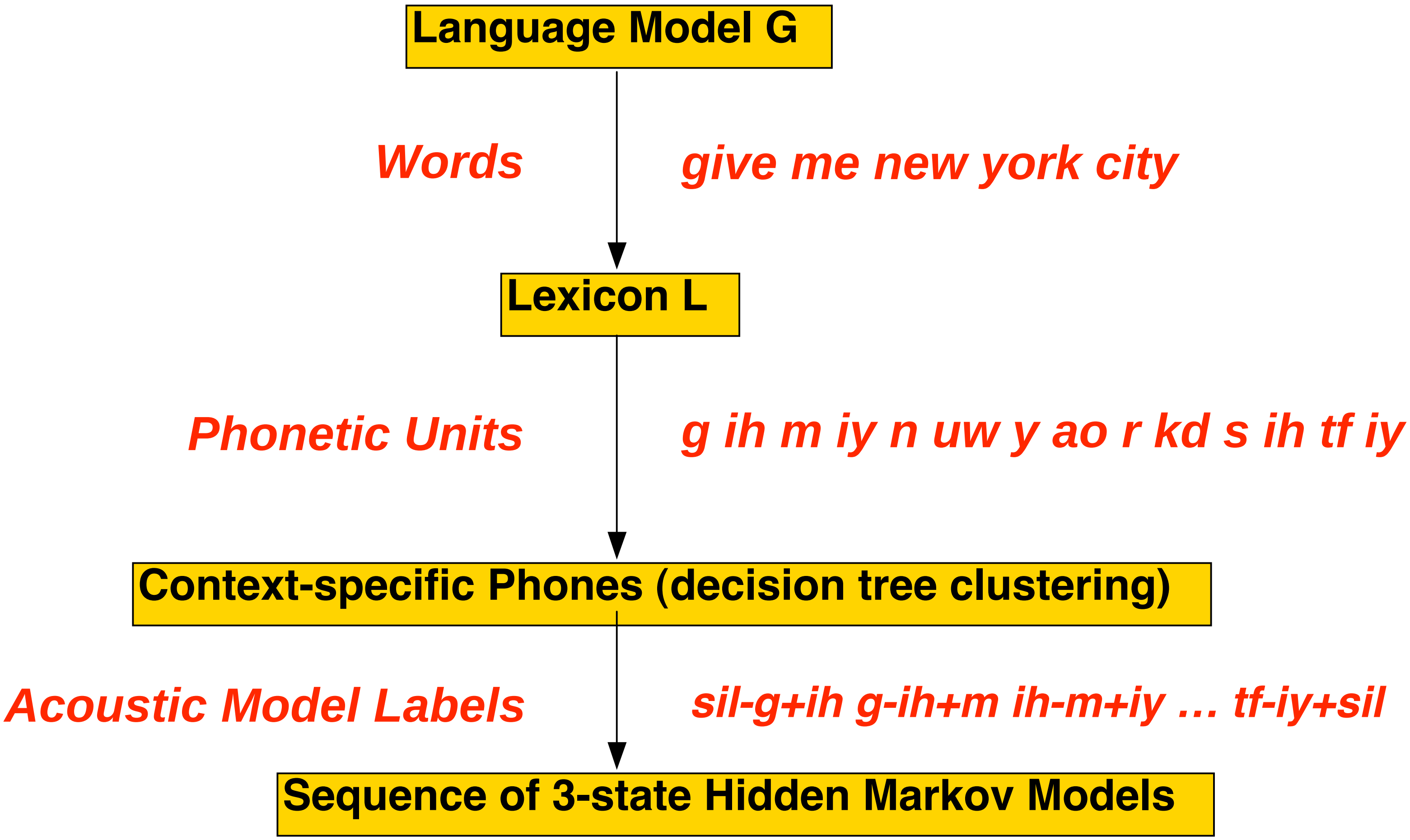
Speech Recognition: Probabilistic Framework

- Words are represented as sequence of phonetic units.
- Using phonetic units, U, expression expands to:



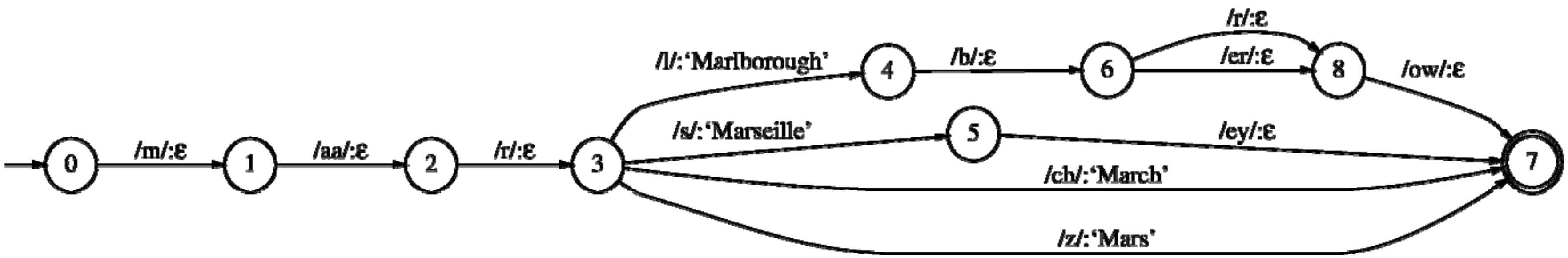
- Search must efficiently find most likely U and W
- Pronunciation, context specific phones (e.g. tri-phones), and language models are typically encoded using weighted finite state transducers/acceptors (Mohri et al., 2002)

A Cascaded FST Recognizer



Finite State Transducer Example: Lexicon

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

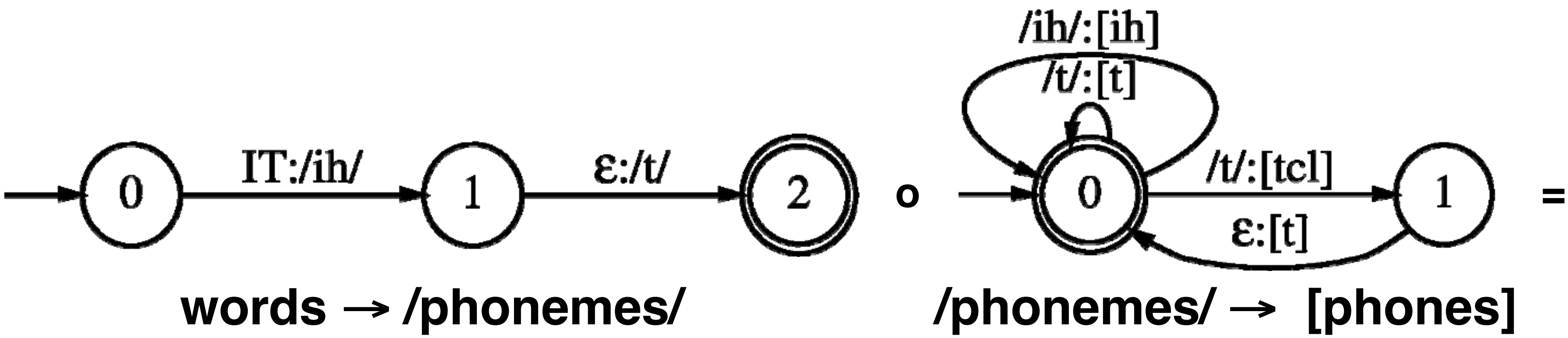


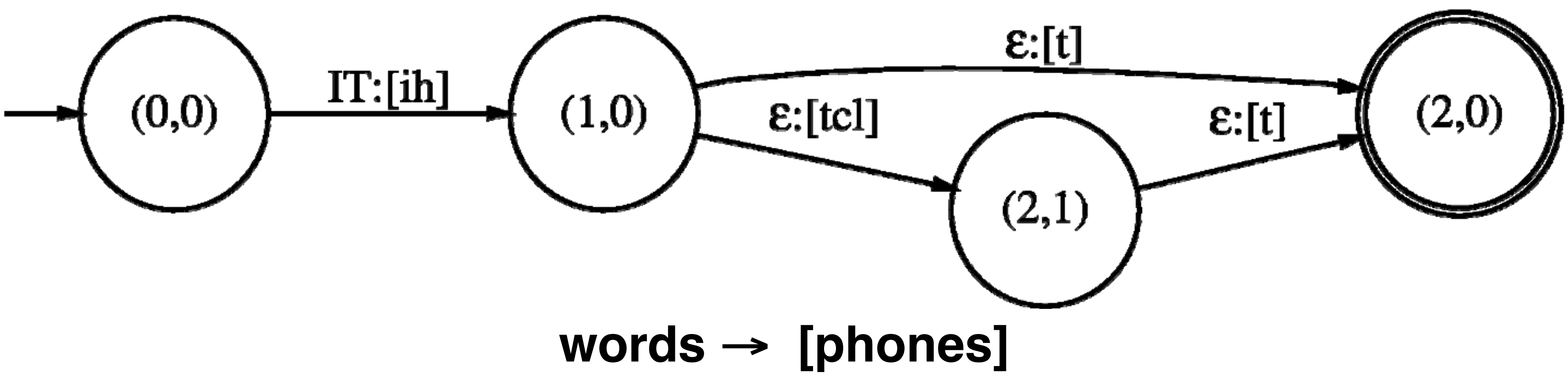
- Finite state transducers (FSTs) map input strings to new output strings
- Lexicon maps /phonemes/ to 'words'
- FSTs allow words to share parts of pronunciations
- Sharing at beginning beneficial to recognition speed because search can prune many words at once

FST Composition

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

- Composition (o) combines two FSTs to produce a single FST that performs both mappings in single step





Defining a Vocabulary

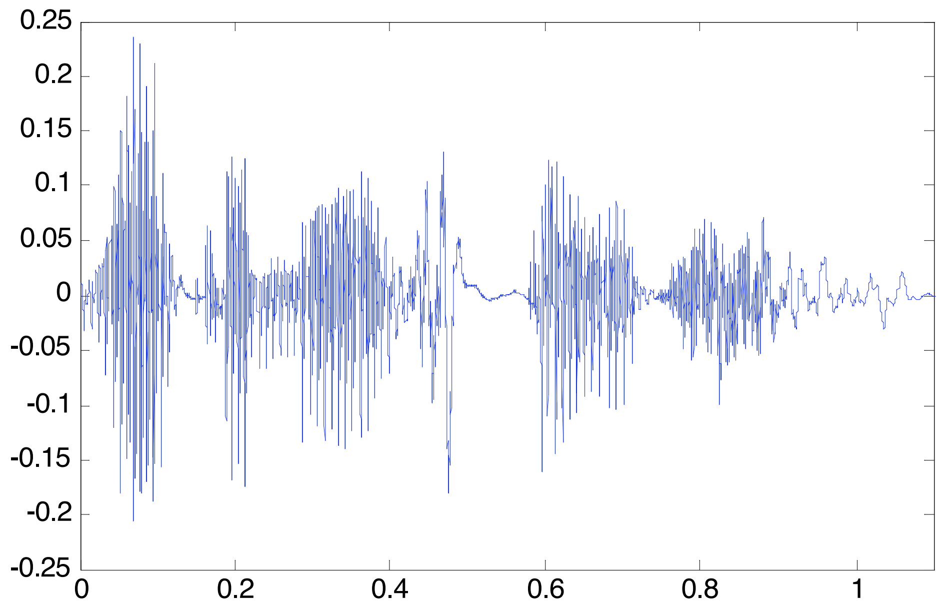
- Words not in a system's vocabulary can not be recognized
- State-of-the-art recognizers attack the out-of-vocabulary (OOV) problem using (very) large vocabularies
  - LVCSR: Large vocabulary continuous speech recognition
  - Typical systems use lexicons of 30K to 100K words
  - Diminishing returns from larger vocabularies when using WER as evaluation metric
- For spoken document search, it is the **query-side out-of-vocabulary rate (Q-OOV)** what matters
  - typically much higher than the OOV rate on the document side

Lexicon

- Typically start with manually created pronunciations for words in vocabulary
- Also needed: an algorithm for automatically generating pronunciations for out-of-vocabulary words
- FST encoding not necessarily deterministic in either direction:
  - READ (inf.): r ih d
  - READ (past tense): r ae d
  - RED: r ae d

Why a Language Model?

(Thanks to Asela Gunawardana, Microsoft Research)

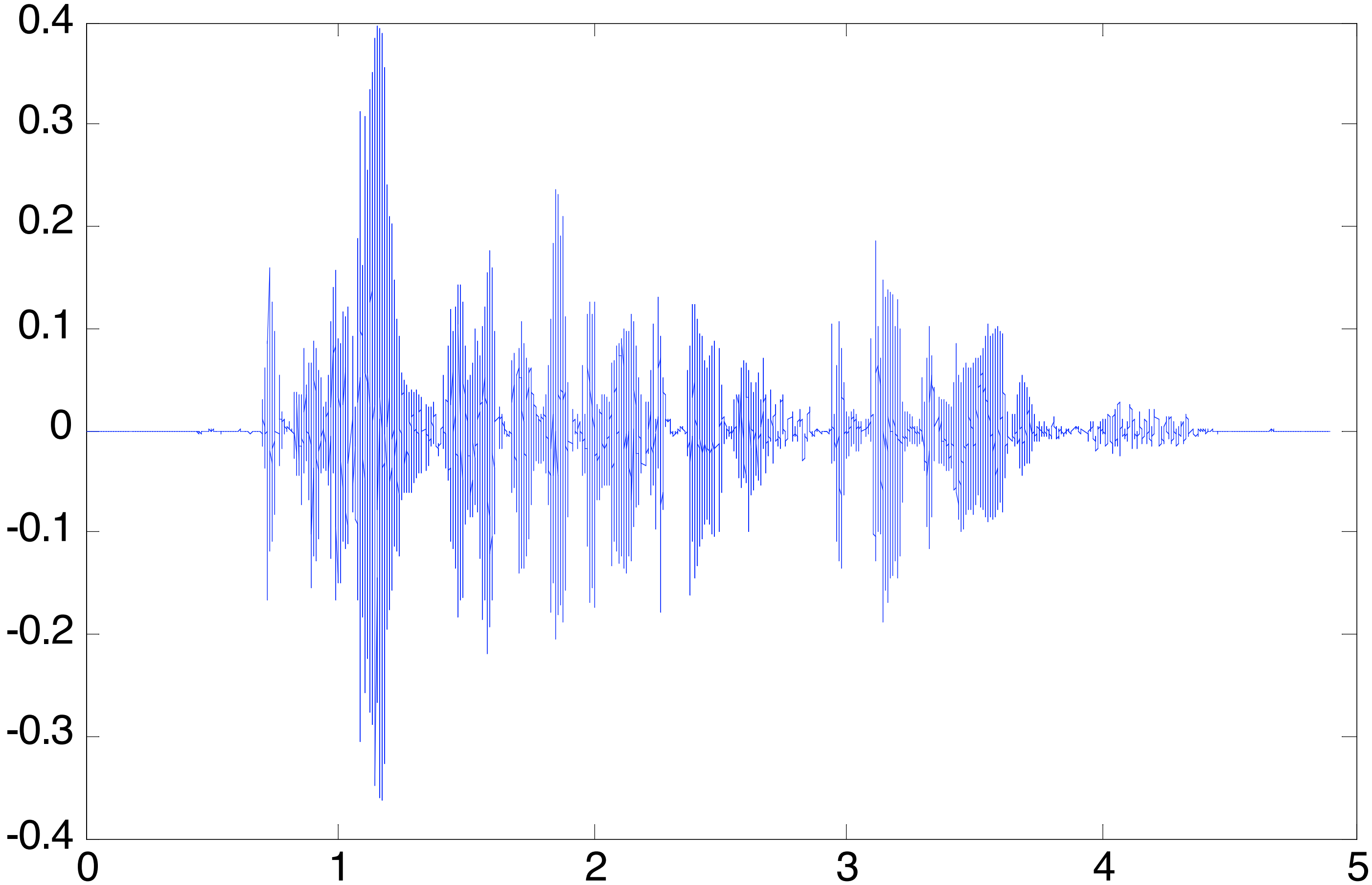


“wreck a nice beach” or

“recognize speech”?

Why a Language Model?

(Thanks to Asela Gunawardana, Microsoft Research)



N-gram Language Modeling

- An *n*-gram model is a statistical language model
- Predicts current word based on previous *n-1* words
- Trigram model expression:

$$P(w_n | w_{n-2}, w_{n-1})$$

- Examples

P(*beach* | *a nice*)

P(*speech* | *to recognize*)

- An *n*-gram model allows any sequence of words…
- …but prefers sequences common in training data.

N-gram Model Smoothing

- **For a bigram model, relative frequency estimate is often 0**
- **We want smooth models**

$$f(w_n|w_{n-1}) = 0$$

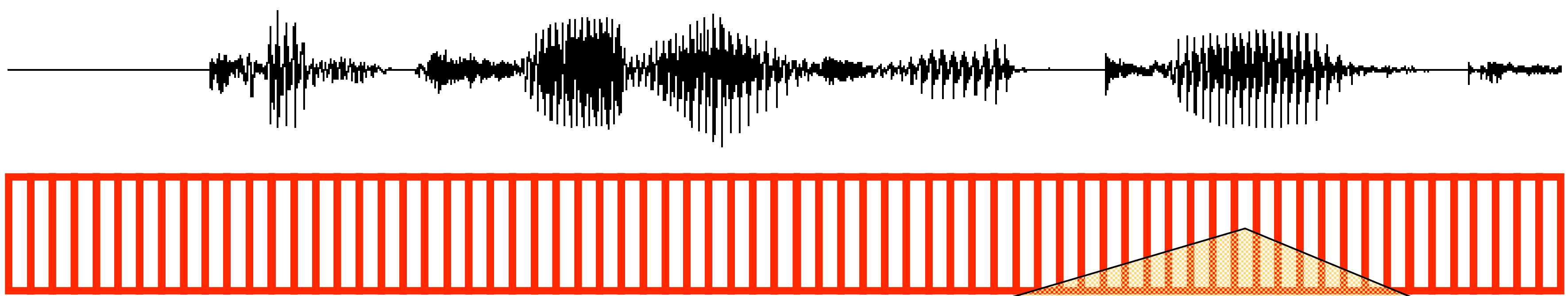
- **To avoid sparse training data problems, we can recursively make use of the lower order model:**

•Wide range of smoothing methods available (Katz, Kneser- Ney) determine the exact way of mixing various N-gram orders, (Goodman 2001)

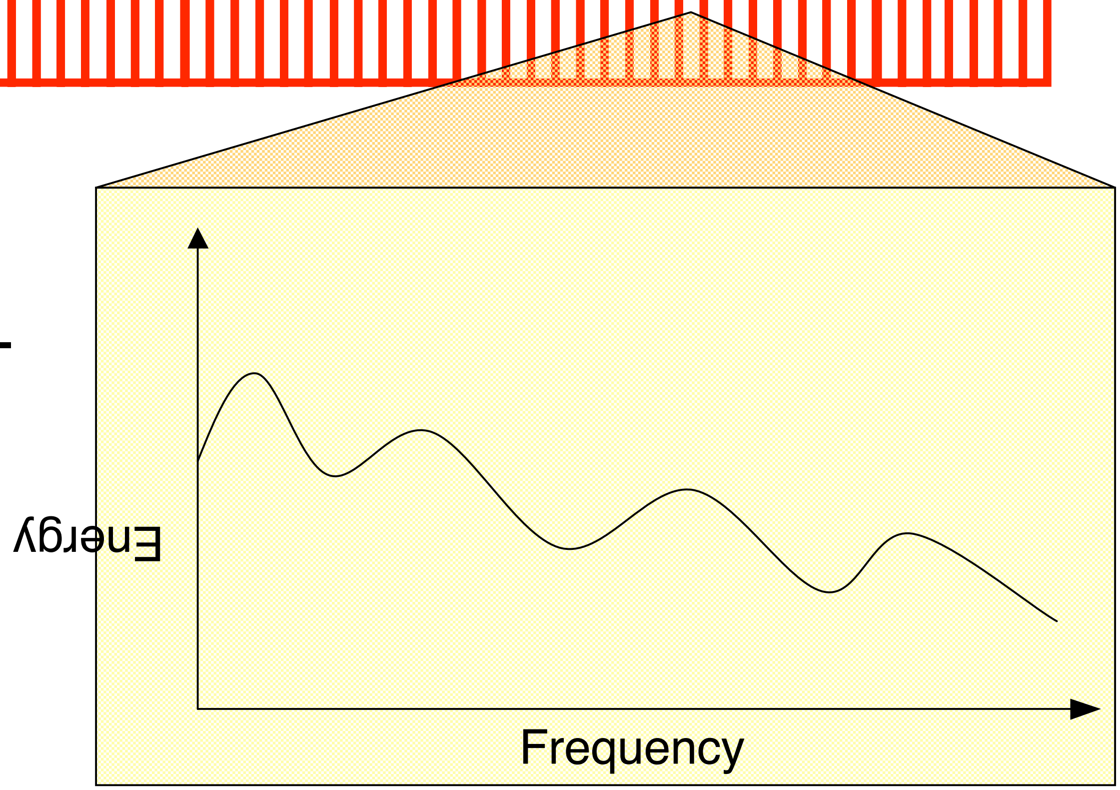
Acoustic Feature Extraction for Recognition

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

Waveform



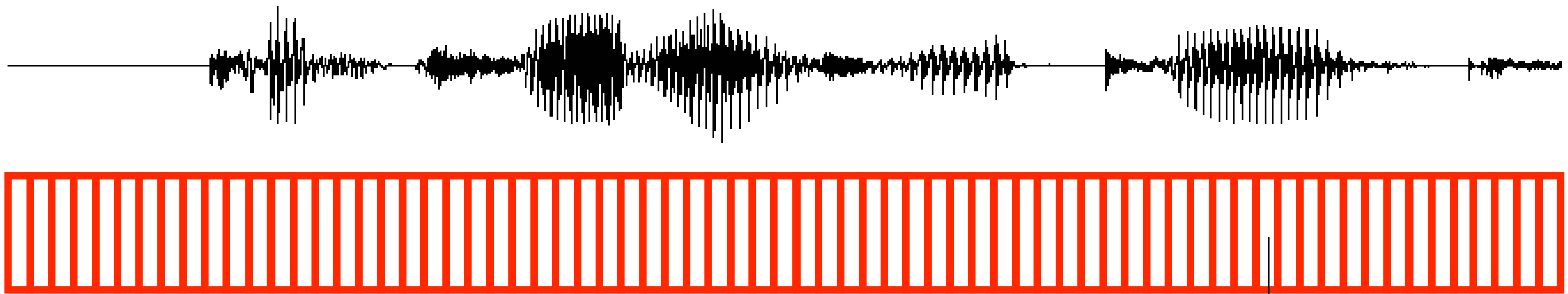
- **Frame-based spectral feature vectors (typically every 10 milliseconds)**
- **Efficiently represented with Mel-frequency scale cepstral (MFCCs)**
  - **Typically ~13 MFCCs used per frame + 1st and 2nd order differences: a total of 39 MFCC coeffs./frame**



Acoustic Feature Scoring for Recognition

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

Waveform

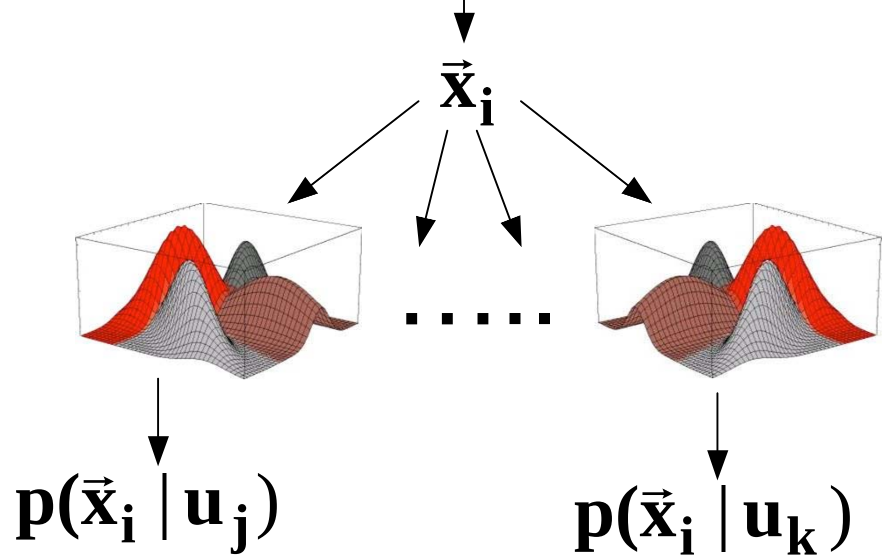


- **Feature vector scoring:**

$$P(X \mid U) = \prod_{i=0}^{N} P(\vec{x}_i \mid u_i)$$

- **Each phonetic unit modeled w/ a mixture of Gaussians:**

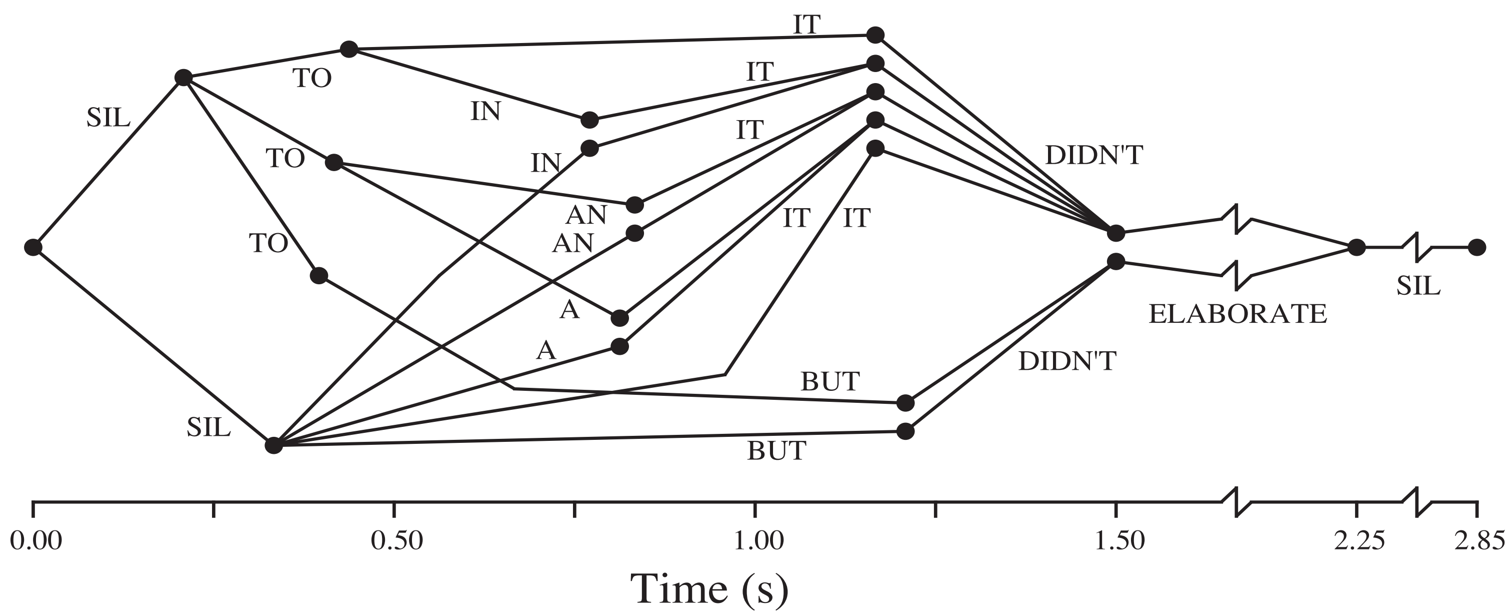
$$P(\vec{x} \mid u) = \sum_{j=0}^{M} w_j N(\vec{x} \mid \mu_j, \Sigma_j)$$



ASR Lattices as a Decoding Side-product

- **Compact way to represent the probability distribution**

$$P(W, A)$$



- **Each link has a start time, end time, word label and associated acoustic and language model scores (probabilities)**
- **Keep only paths with high probability**

Issues in Language Modeling: Mismatch Train/Test

- The vocabulary, N-gram skeleton, N-gram probabilities are all estimated from large amounts of training data "expected to be similar to the test data"

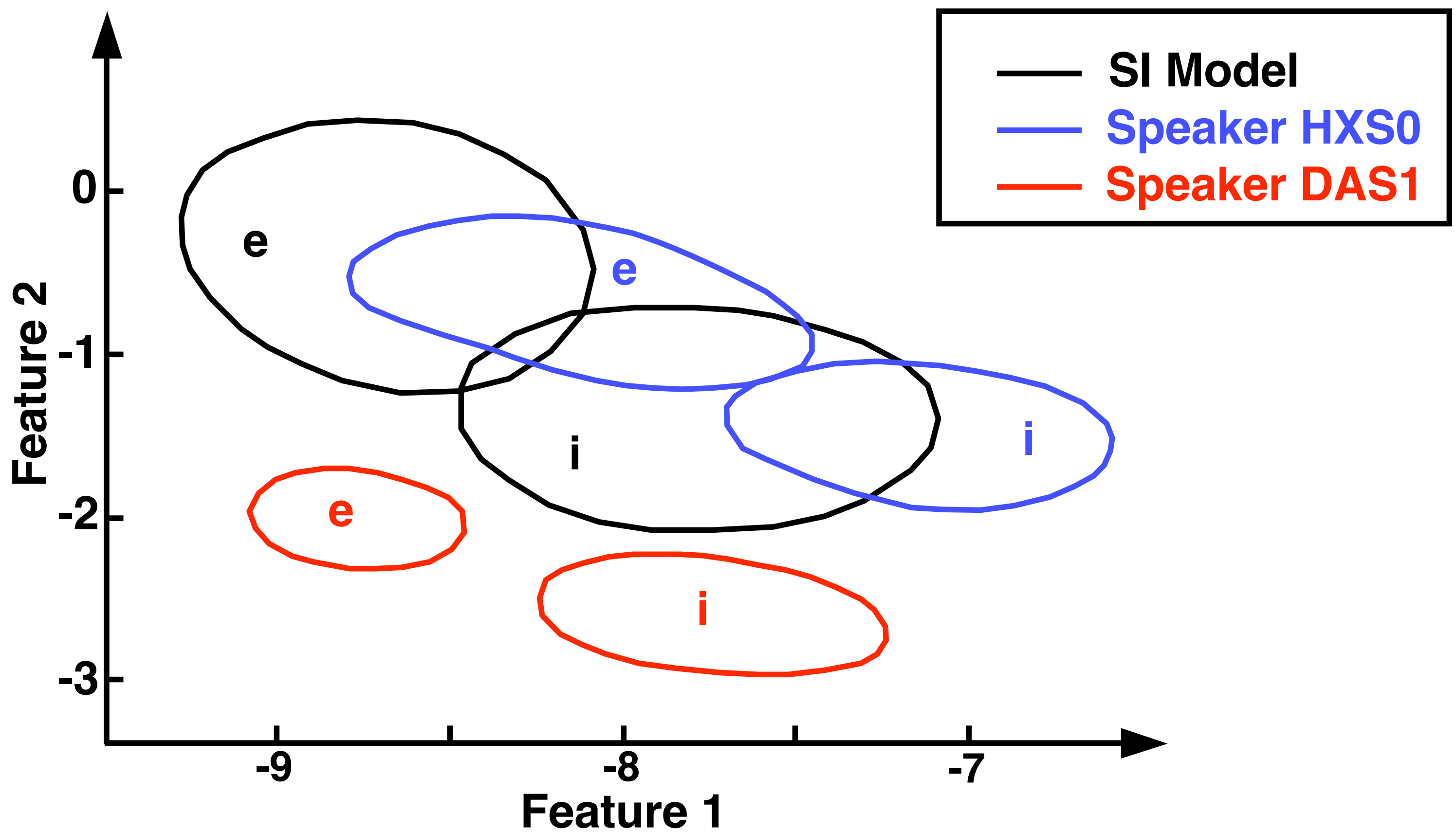
- Assuming a small amount of adaptation data is available, identifying such data is very hard, even if plenty (Tera words) available

- Research issue: Language Model Adaptation to mismatched test data:
  - What is a good vocabulary?
  - What new N-grams would be needed?
  - How should one adjust the N-gram probabilities such that it performs best on the test data?

Issues in Acoustic Modeling: Variability

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

- Plot of isometric likelihood contours for phones [i] and [e]
- One SI model and two speaker dependent (SD) models
- SD contours are tighter than SI and correlated w/ each other



MLLR Adaptation

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

- **Maximum Likelihood Linear Regression (MLLR) is a common transformational adaptation techniques (Leggetter & Woodland, 1995)**
- **Idea: Adjust models parameters using a transformation shared globally or across different units within a class**
- **Global mean vector translation:**

$$\forall p, \mu_p^{sa} = \mu_p^{si} + v$$

adapt mean vectors of all phonetic models

shared translation vector

- **Global mean vector scaling, rotation and translation:**

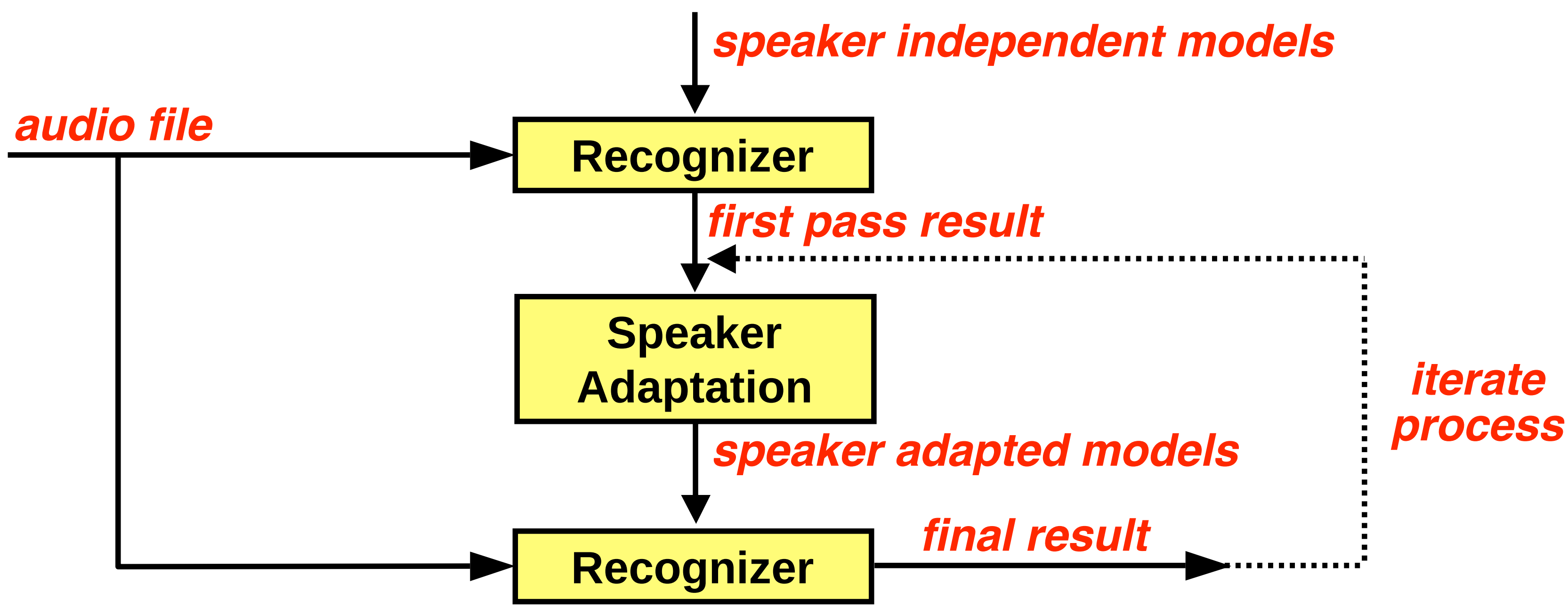
$$\forall p, \mu_p^{sa} = \mathcal{R}\mu_p^{si} + v$$

shared scaling and rotation matrix

Transform chosen to maximize likelihood of adaptation or test data

Unsupervised Adaptation Architecture

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)



Importance of Adaptation

(work by TJ Hazen, MIT, Spoken Lecture Processing Project)

- **Experiment: Examine performance of recognizer on one lecture from a non-native speaker**
- **Perform adaptation:**
  - Adapt language model by adding course textbook to LM training data
  - Adapt acoustic model by adding 38 previous lectures to AM training data
- **Acoustic model adaptation helps much more than language model adaptation in this case**

Adaptation	WER (%)
None	46.8
Language Model Only	45.2
Acoustic Model Only	20.5
AM and LM	19.5

Unsupervised AM Adaptation

(work by Asela Gunawardana, Interspeech 2003)

- Initial model WSJ-0, Sennheiser close talking microphone
- Test data is Aurora-II (TI-digits with a lot of of noise and telephone/cell phone channel)
- Idea is to adapt a generic AM to a task with no supervision
  - the adaptation is actually retraining the AMs completely on the test data, but with automatically derived transcriptions or with lattices

	Word Accuracy
WSJ-0 baseline	**59.97%**
4 its 1-best	**76.36%**
4 its lattice training	**83.72%**
Cheating (supervised)	**88.84%**
Aurora-II system	**92.28%**

Spoken Document Retrieval: Outline

- Brief overview of text retrieval algorithms
- Integration of IR and ASR using lattices
- Query Processing
- Relevance Scoring
- Evaluation
- User Interface

- Try to balance overview of work in the area with experimental results from our own work
- Active area of research:
  - Emphasize known approaches as well as interesting research directions
  - No established way of solving these problems as of yet

Text Retrieval

- **Collection of documents:** $\mathcal{D} = D_1, \ldots, D_N$
  - **“large” N: 10k-1M documents or more (videos, lectures)**
  - **“small” N: < 1-10k documents (voice-mails, VoIP chats)**
- **Query:** $\mathcal{Q} = q_1 \ldots q_Q$
  - **Ordered set of words in a large vocabulary** $\mathcal{V}$
  - **Restrict ourselves to keyword search; other query types are clearly possible:**
    * Speech/audio queries (match waveforms)
    * Collaborative filtering (people who watched X also watched…)
    * Ontology (hierarchical clustering of documents, supervised or unsupervised)

Text Retrieval: Vector Space Model

- Build a term-document co-occurrence (LARGE) matrix (Baeza-Yates, 99)
  - Rows indexed by word
  - Columns indexed by documents

$$(t_{ij})_{\substack{i=1\ldots V \\ j=1\ldots D}}$$

$$t_{ij} = \underbrace{f_{ij}}_{TF} \cdot log \underbrace{N/n_i}_{IDF}$$

- TF (term frequency): frequency of word in document
  - Could be normalized to maximum frequency in a given document
- IDF (inverse document frequency): if a word appears in all documents equally likely, it isn't very useful for ranking
  - (Bellegarda, 2000) uses normalized entropy

$$H(D|w_i)/log(N)$$

Text Retrieval: Vector Space Model (2)

- For retrieval/ranking one ranks the documents in decreasing order of the relevance score:

- The query weights have minimal impact since queries are very short, so one often uses a simplified relevance score:

$$S(D_j, Q) = \frac{\sum_{i=1}^{Q} w_{ij}}{norm(\underline{w}_j)}$$

Text Retrieval: TF-IDF Shortcomings

- **Hit-or-Miss:**
  - **Only documents containing the query words are returned**
  - **A query for Coca Cola will not return a document that reads:**
    - "… its **Coke** brand is the most treasured asset of the **soft drinks** maker …"
- **Cannot do phrase search: "Coca Cola"**
  - **Needs post processing to filter out documents not matching the phrase**
- **Ignores word order and proximity**
  - **A query for Object Oriented Programming:**
    - **" … the object oriented paradigm makes programming a joy … "**
    - **" … TV network programming transforms the viewer in an object and it is oriented towards…"**

Vector Space Model: Query/Document Expansion

- Correct the Hit-or-Miss problem by doing some form of expansion on the query and/or document side
  - add similar terms to the ones in the query/document to increase number of terms matched on both sides
  - corpus driven methods: TREC-7 (Singhal et al,. 99) and TREC-8 (Singhal et al,. 00)
- Query side expansion works well for long queries (10 words)
  - short queries are very ambiguous and expansion may not work well
- Expansion works well for boosting Recall:
  - very important when working on small to medium sized corpora
  - typically comes at a loss in Precision

Vector Space Model: Latent Semantic Indexing

- **Correct the Hit-or-Miss problem by doing some form of dimensionality reduction on the TF-IDF matrix**
  - **Singular Value Decomposition (SVD) (Furnas et al., 1988)**
  - **Probabilistic Latent Semantic Analysis (PLSA) (Hoffman, 1999)**
  - **Non-negative Matrix Factorization (NMF)**
- **Matching of query vector and document vector is performed in the lower dimensional space**
- **Good as long as the magic works**
- **Drawbacks:**
  - **still ignores WORD ORDER**
  - **users are no longer in full control over the search engine**

Humans are very good at crafting queries that'll get them the documents they want and expansion methods impair full use of their natural language faculty

Probabilistic Models (Robertson, 1976)

- Assume one has a probability model for generating queries and documents $P(D, Q)$
- We would like to rank documents according to the point-wise mutual information

- One can model $P(Q|D_j)$ using a language model built from each document (Ponte, 1998)
- Takes word order into account
  - models query N-grams but not more general proximity features
  - expensive to store

Ad-Hoc (Early Google) Model (Brin,1998)

- HIT = an occurrence of a query word in a document
- **Store context in which a certain HIT happens (including integer position in document)**
  - **Title hits are probably more relevant than content hits**
  - **Hits in the text-metadata accompanying a video may be more relevant than those occurring in the speech reco transcription**
- **Relevance score for every document uses proximity info**
  - **weighted linear combination of counts binned by type**
    - * proximity based types (binned by distance between hits) for multiple word queries
    - * context based types (title, anchor text, font)
- **Drawbacks:**
  - **ad-hoc, no principled way of tuning the weights for each type of hit**

Text Retrieval: Scaling Up

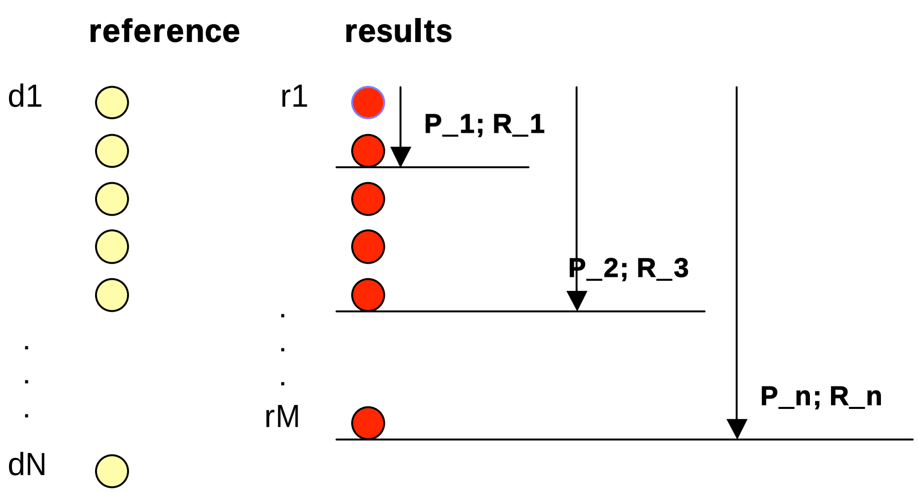
- Linear scan of document collection is not an option for compiling the ranked list of relevant documents
  - Compiling a short list of relevant documents *may* allow for relevance score calculation on the document side
- Inverted index is critical for scaling up to large collections of documents
  - think index at end of a book as opposed to leafing through it!

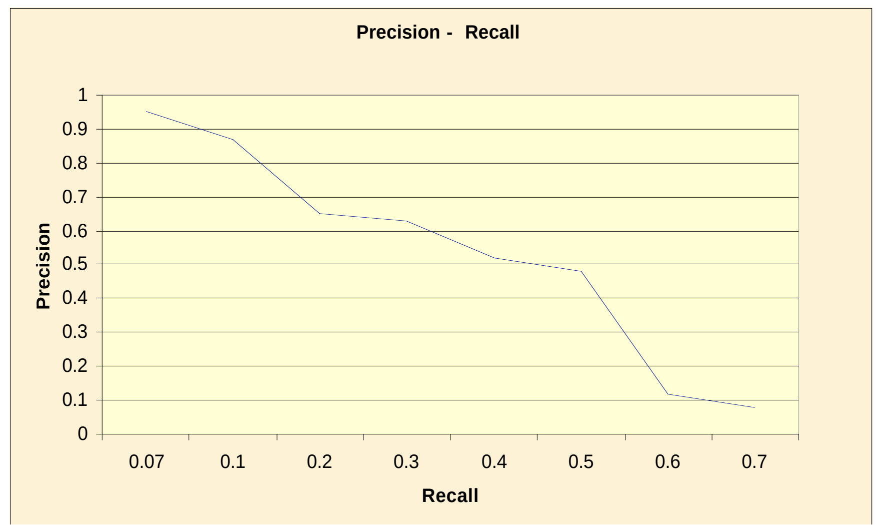
All methods are amenable to some form of indexing:

- TF-IDF/SVD: compact index, drawbacks mentioned
- LM-IR: storing all N-grams in each document is very expensive
  - significantly more storage than the original document collection
- Early Google: compact index that maintains word order information and hit context
  - relevance calculation, phrase based matching using only the index

Text Retrieval: Evaluation

- trec_eval (NIST) package requires reference annotations for documents with binary relevance judgments for each query
  - Standard Precision/Recall and Precision@N documents
  - Mean Average Precision (MAP)
  - R-precision (R=number of relevant documents for the query)





□Ranking on reference side is flat (ignored)

Search in Spoken Documents

- TREC-SDR approach:
  - treat both ASR and IR as black-boxes
  - run ASR and then index 1-best output for retrieval
  - evaluate MAP/R-precision against human relevance judgments for a given query set
- **Issues with this approach:**
  - **1-best WER is usually high when ASR system is not tuned to a given domain**
    - 0-15% WER is unrealistic
    - iCampus experiments (lecture material) using a general purpose dictation ASR system show 50% WER!
  - **OOV query words at a rate of 5-15% (frequent words are not good search words)**
    - average query length is 2 words
    - 1 in 5 queries contains an OOV word

Evaluation for Search in Spoken Documents

- In addition to the standard IR evaluation setup one could also use the output on transcription
- Reference list of relevant documents to be the one obtained by running a state-of-the-art text IR system
- How close are we matching the text-side search experience?
  - Assuming that we have transcriptions available
- Drawbacks of using trec_eval in this setup:
  - Precision/Recall, Precision@N, Mean Average Precisision (MAP) and R-precision: they all assume binary relevance ranking on the reference side
  - Inadequate for large collections of spoken documents where ranking is very important
- (Fagin et al., 2003) suggest metrics that take ranking into account using Kendall's tau and Spearman's footrule

Out-of-Vocabulary (OOV) Query Terms

- Map OOV query words to some sub-word representation, e.g. phonetic pronunciation
- Need to generate phone lattices as well as word lattices
  - Mixed word+phone lattices also possible - see (Bazzi, 2001)
- General issues with phone lattices:
  - not as accurate as word-level recognition; anecdotal evidence shows that a very good way to get phone lattices is to run word-level ASR and then map down to phones (Saraclar, 2004)
  - Do not match word boundaries well; critical for high quality retrieval
  - Inverted indexing is not very efficient unless one indexes N-phones (N > 3) but then index becomes very large
  - Combining word level and phone level information is hard – (Logan et al., 2002)

Domain Mismatch Hurts Retrieval Performance

SI BN system on BN data

Percent Total Error	**=**	**22.3%**	**(7319)**
Percent Substitution	=	15.2%	(5005)
Percent Deletions	=	5.1%	(1675)
Percent Insertions	=	1.9%	(639)

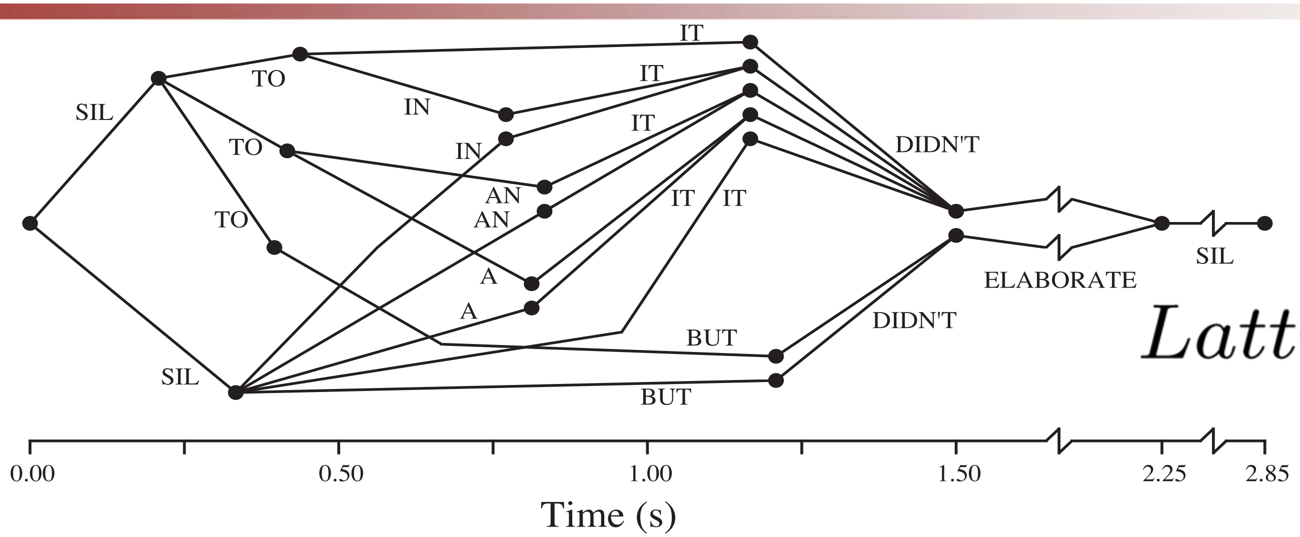
1: 61 -> a ==> the (1.2%)
2: 61 -> and ==> in
3: 35 -> (%hesitation) ==> of
4: 35 -> in ==> and
5: 34 -> (%hesitation) ==> that
6: 32 -> the ==> a
7: 24 -> (%hesitation) ==> the
8: 21 -> (%hesitation) ==> a
9: 17 -> as ==> is
10: 16 -> that ==> the
11: 16 -> the ==> that
12: 14 -> (%hesitation) ==> and
13: 12 -> a ==> of
14: 12 -> two ==> to
15: 10 -> it ==> that
16: 9 -> (%hesitation) ==> on
17: 9 -> an ==> and
18: 9 -> and ==> the
19: 9 -> that ==> it
20: 9 -> the ==> and

SI BN system on MIT lecture
Introduction to Computer Science

Percent Total Error	**=**	**45.6%**	**(4633)**
Percent Substitution	=	27.8%	(2823)
Percent Deletions	=	13.4%	(1364)
Percent Insertions	=	4.4%	(446)

1: 19 -> lisp ==> list (0.6%)
2: 16 -> square ==> where
3: 14 -> the ==> a
4: 13 -> the ==> to
5: 12 -> ok ==> okay
6: 10 -> a ==> the
7: 10 -> root ==> spirit
8: 10 -> two ==> to
9: 9 -> square ==> this
10: 9 -> x ==> tax
11: 8 -> and ==> in
12: 8 -> guess ==> guest
13: 8 -> to ==> a
14: 7 -> about ==> that
15: 7 -> define ==> find
16: 7 -> is ==> to
17: 7 -> of ==> it
18: 7 -> root ==> is
19: 7 -> root ==> worried
20: 7 -> sum ==> some

ASR Lattices for Search in Spoken Documents



$$Lattice \rightarrow P(W|A)$$

Error tolerant design

Lattices contain paths with much lower WER than ASR 1-best:

- **-dictation ASR engine on iCampus (lecture material) 55% lattice vs. 30% 1-best**
- **-sequence of words is uncertain but may contain more information than the 1-best**

Cannot easily evaluate:

- **-counts of query terms or Ngrams**
- **-proximity of hits**

Vector Space Models Using ASR Lattices

- Straightforward extension once we can calculate the sufficient statistics "expected count in document" and "does word happen in document?"
  - Dynamic programming algorithms exist for both

- One can then easily calculate term-frequencies (TF) and inverse document frequencies (IDF)
- Easily extended to the latent semantic indexing family of algorithms
- (Saraclar, 2004) show improvements using ASR lattices instead of 1-best

Vector Space Models for SDR (Pros and Cons)

- Compact word level index
- Abundant literature in ASR community for calculating expected counts --- “confidence scoring” --- at both word and/or phone level and integrating in IR vector model: (James, 1995), (Jones et al., 1996), (Ng, 2000) to name a few

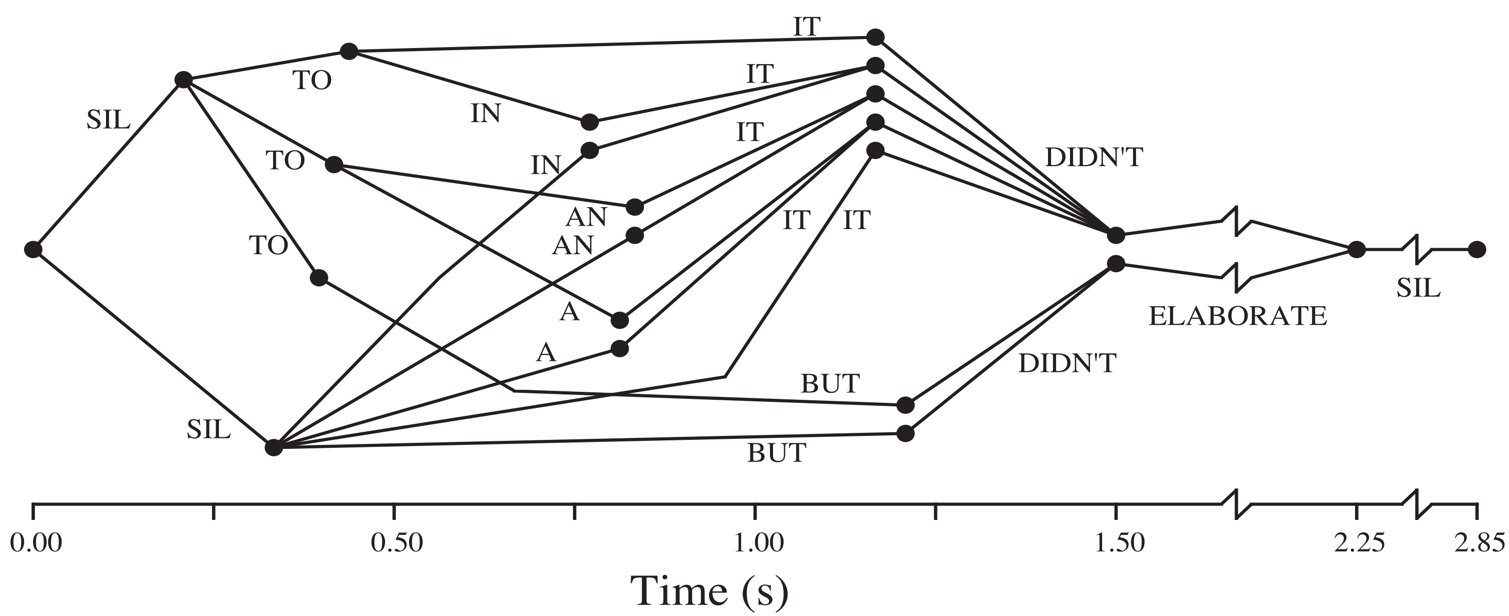
- Calculating word posteriors for OOV words needs the entire lattice: forced to do linear scan over documents/lattices
- Could speed up using some form on N-phone indexing; index size becomes an issue (Seide, 2004)
- Hard to combine word and sub-word information in a good way (Logan et al., 2002)
- Same drawbacks as those listed for TF-IDF on text documents

Probabilistic IR Models Using ASR Lattices

- Would need to estimate a language model from counts derived from $P(W|A)$ (lattice) rather than from text
- GRM library (Allauzen et al., 2003) allows this type of LM estimation
- Not yet applied to word-level IR; storing the LMs is likely to be a problem
- Phone-level IR: (Seide, 2004) uses such an approach to propose a candidate of phone lattices that are then going to be used for exact word posterior calculation

- Drawback: does not scale up for large collections of documents if one wants to use N-grams of order higher than 1 (equivalent to indexing 2-grams, 3-grams etc.)

SOFT-HITS for Ad-Hoc SDR



0	1		2		3		4	
SIL	TO	0.5	IT	0.2	DIDN'T	0.5	ELABORATE	0.7
	IN	0.3	IN	0.2	IT	0.4	DIDN'T	0.2
	AN	0.1	A	0.1	ELABORATE	0.1	SIL	0.1
	BUT	0.1	BUT	0.1				
			DIDN'T	0.1				

Soft-Indexing of ASR Lattices

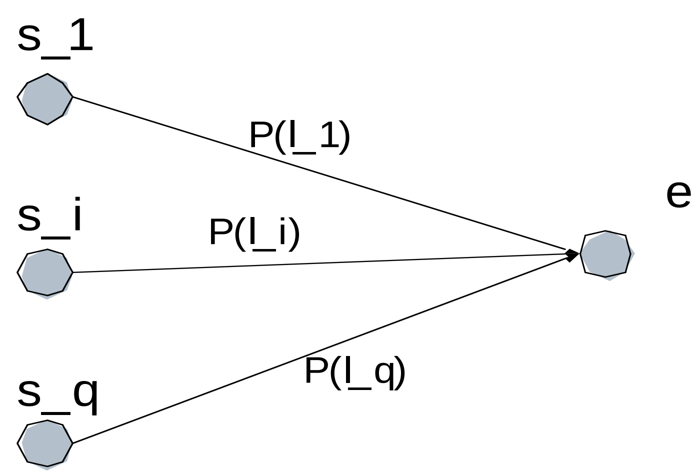
- Lossy encoding of ASR recognition lattices (Chelba, 2005)
- Preserve word order information without indexing N-grams
- SOFT-HIT: posterior probability that a word w happens at a position n in the spoken document A

$$P(w, n|LAT(A))$$

- Minor change to text inverted index: store probability along with regular hits
- Can easily evaluate proximity features ("is query word i within three words of query word j?") and phrase hits
- Drawbacks:
  - approximate representation of posterior probability $P(W|A)$
  - unclear how to integrate phone- and word-level hits

Position-Specific Word Posteriors

- **Split forward probability based on path length**
- **Link scores are flattened**



$$\alpha_n[l] = \sum_{\pi:end(\pi)=n,length(\pi)=l} P(\pi)$$

$$\alpha_e[l+1] = \sum_{i=1}^{q} \alpha_{s_i}[l+\delta(l_i,\epsilon)] \cdot P(l_i)$$

$$P(n,l|LAT) = \frac{\alpha_n[l] \cdot \beta_n}{norm(LAT)}$$

Experiments on iCampus Data

- Our own work (Chelba 2005) (Silva et al., 2006)
  - Carried out while at Microsoft Research
- Indexed 170 hrs of iCampus data
  - lapel mic
  - transcriptions available
- dictation AM (wideband), LM (110Kwds vocabulary, newswire text)
- dvd1/L01 - L20 lectures (Intro CS)
  - 1-best WER ~ 55%, Lattice WER ~ 30%, 2.4% OOV rate
  - *.wav files (uncompressed) 2,500MB
  - 3-gram word lattices 322MB
  - soft-hit index (unpruned) 60MB (20% lat, 3% *wav)
  - transcription index 2MB

Document Relevance using Soft Hits (Chelba, 2005)

- Query $Q = q_1 \ldots q_Q$
- N-gram hits, N = 1 … Q
- full document score is a weighted linear combination of N-gram scores
- Weights increase linearly with order N but other values are likely to be optimal
- Allows use of context (title, abstract, speech) specific weights

$$
\begin{aligned}
S(D, q_i \ldots q_{i+N-1}) &= \log\left[1 + \sum_{segment} \sum_{s\, position\ k} \prod_{l=0}^{N-1} P(w_{k+l}(s) = q_{i+l} | D)\right] \\
S_{N-gram}(D, Q) &= \sum_{i=1}^{Q-N+1} S(D, q_i \ldots q_{i+N-1}) \\
S(D, Q) &= \sum_{N} w_N \cdot S_{N-gram}(D, Q)
\end{aligned}
$$

Retrieval Results

ACL (Chelba, 2005)

How well do we bridge the gap between speech and text IR?

Mean Average Precision

- **REFERENCE= Ranking output on transcript using TF-IDF IR engine**
- **116 queries: 5.2% OOV word rate, 1.97 words/query**
- **Removed queries w/ OOV words for now (10/116)**

Our ranker	**transcript**	**1-best**	**lattices**
MAP	**0.99**	**0.53**	**0.62** (17% over 1-best)

Retrieval Results: Phrase Search

How well do we bridge the gap between speech and text IR?

Mean Average Precision

- **REFERENCE= Ranking output on transcript using our own engine (to allow phrase search)**
- **Preserved only 41 quoted queries:**
  - "OBJECT ORIENTED" PROGRAMMING
  - "SPEECH RECOGNITION TECHNOLOGY"

Our ranker	**1-best**	**lattices**
MAP	**0.58**	**0.73** **(26% over 1-best)**

Why Would This Work?

[30]:
BALLISTIC = -8.2e-006
MISSILE = -11.7412
A = -15.0421
TREATY = -53.1494
ANTIBALLISTIC = -64.189
AND = -64.9143
COUNCIL = -68.6634
ON = -101.671
HIMSELF = -107.279
UNTIL = -108.239
HAS = -111.897
SELL = -129.48
FOR = -133.229
FOUR = -142.856
[…]

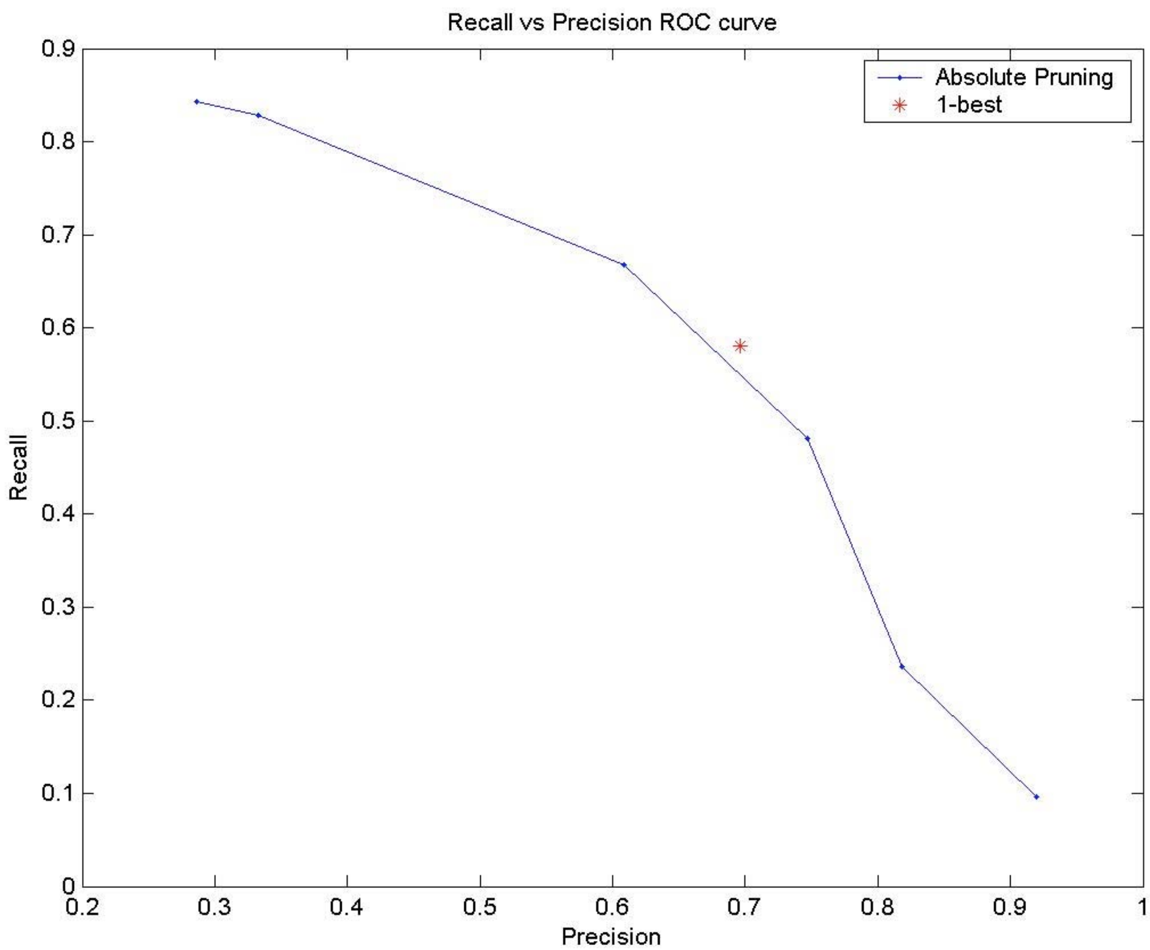
[31]:
MISSILE = -8.2e-006
TREATY = -11.7412
BALLISTIC = -15.0421
AND = -53.1726
COUNCIL = -56.9218
SELL = -64.9143
FOR = -68.6634
FOUR = -78.2904
SOFT = -84.1746
FELL = -87.2558
SELF = -88.9871
ON = -89.9298
SAW = -91.7152
[…]

[32]:
TREATY = -8.2e-006
AND = -11.7645
MISSILE = -15.0421
COUNCIL = -15.5136
ON = -48.5217
SELL = -53.1726
HIMSELF = -54.1291
UNTIL = -55.0891
FOR = -56.9218
HAS = -58.7475
FOUR = -64.7539
</s> = -68.6634
SOFT = -72.433
FELL = -75.5142
[…]

Search for "ANTIBALLISTIC MISSILE TREATY" fails on 1-best but succeeds on PSPL.

Precision/Recall Tuning (runtime)

(Joint Work with Jorge Silva Sanchez, UCLA)

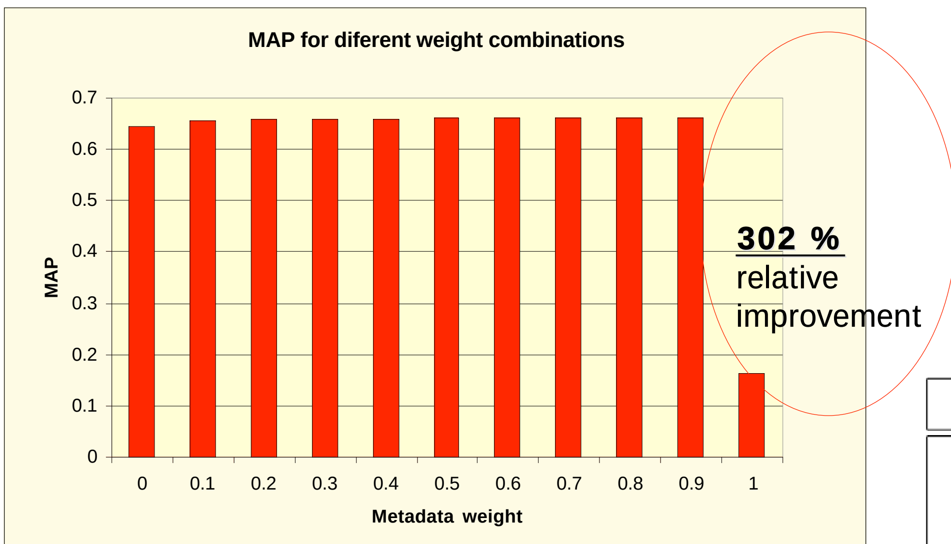


- **User can choose Precision vs. Recall trade-off at query run-time**

Speech Content or just Text-Meta Data?

(Joint Work with Jorge Silva Sanchez, UCLA)

- **Corpus:**
  - MIT iCampus: **79 Assorted MIT World seminars (89.9 hours)**
  - Metadata: **title, abstract, speaker bibliography (less than 1% of the transcription)**

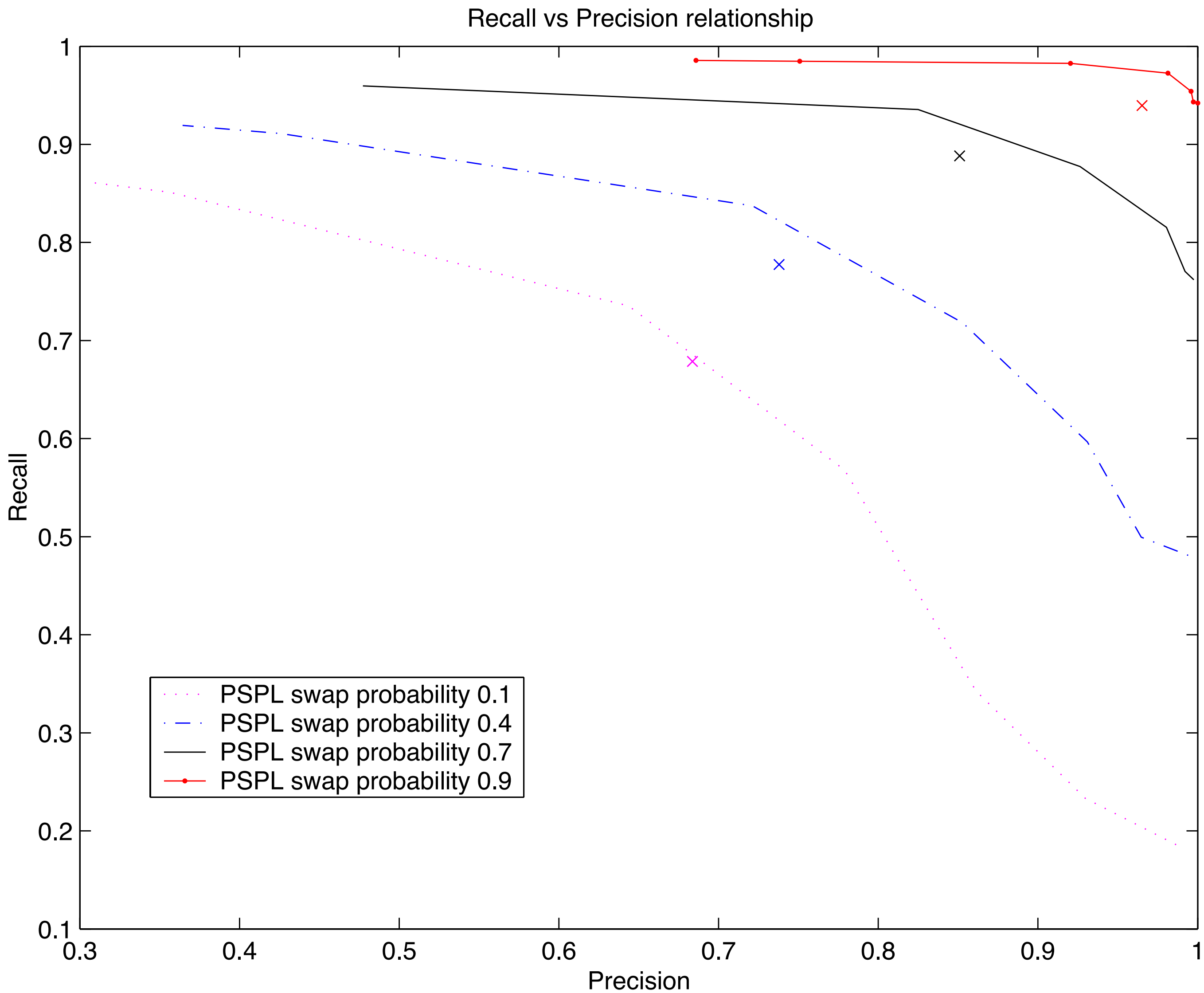


- **Multiple data streams**
  - **similar to (Oard et al., 2004):**
  - speech: **PSPL word lattices from ASR**
  - metadata: **title, abstract, speaker bibliography (text data)**
  - **linear interpolation of relevance scores**

Scenario	Precision	Recall
Metadata	1	0.056
Speech	0.319	0.815
Meta - Speech.	0.323	0.826

Enriching Meta-data

(Joint Work with Jorge Silva Sanchez, UCLA)



- **Artificially add text meta-data to each spoken document by sampling from the document manual transcripti on**

Indexing Lattices: Related Work

- (Siegler, 1999) shows improvements by using N-best lists
  - Does not take into account word posteriors
- (Saraclar et al., 2004) HLT-NAACL also shows improvements from using lattices
  - Build inverted index for full lattice (start/end node, score)
  - Adjacency information and posterior probability are fully preserved
  - Can easily evaluate N-gram posterior counts
  - Hard to evaluate proximity hits of type "are two hits within a window of 5 words from each other?"
  - PSPL is probably more compact although no formal comparison has been carried out

Spoken Document Retrieval: Conclusion

- Tight Integration between ASR and TF-IDF technology holds great promise for general SDR technology
  - Error tolerant approach with respect to ASR output
  - ASR Lattices
  - Better solution to OOV problem is needed
- Better evaluation metrics for the SDR scenario:
  - Take into account the ranking of documents on the reference side
  - Use state of the art retrieval technology to obtain reference ranking
- Integrate other streams of information
  - Links pointing to documents (www)
  - Slides, abstract and other text meta-data relevant to spoken document
  - Collaborative filtering

User Experience

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

- **Scanning information in spoken documents is difficult**
  - **Quickly scanning text is far easier**
  - **Spontaneously generated speech not as well organized as text or prepared broadcast news stories**
    - * Can't always listen to first few sentences to "catch the drift"
- **Want to enable users to browse documents for relevance without requiring them to listen to audio**
  - **Unformatted ASR transcriptions may be difficult to scan**
    - * High error rates
    - * Lack of capitalization, punctuation, sentence boundaries
  - **Topic detection and summarization may help**
- **Problem still has many open questions**
  - **Extensive user studies needed to find optimal approach**
  - **Best approach may be application and scenario specific**

Recognition: What's Good Enough for Browsing?

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

- **Text-based browsing is more efficient than audio browsing**
  - Accurate transcriptions help users identify relevant material
- **Some data points on what may be sufficient accuracy:**
  - For court stenographers to become *Certified Real-Time Reporters* they must transcribe with 95% accuracy
  - The Liberated Learning Consortium found transcription error rates of up to 15% are acceptable for comprehension of real-time speech recognition outputs in classrooms
  - Closed captioning WER was measured to be in the 10-15% WER (Garofolo, 2000)
- **User prefer ASR output that is formatted with capitalization, punctuation, etc. (Jones *et al*, 2003)**
  - But this formatting may not lead to improved comprehension

Spoken Document Summarization

(Thanks to TJ Hazen, MIT, Spoken Lecture Processing Project)

- Summarization from audio generally follows this approach
  - Generate automatic transcription with confidence levels
  - Extract "important" sentences w/ high recognition confidences
  - Compact text representation removing redundant information and unimportant words
- Importance of words/phrases/sentences is measured from a combination of features:
  - Term frequency - inverse document frequency (TF-IDF)
  - Part-of-speech, e.g., nouns are more important than adverbs
  - Prosodic prominence (Inoue et al, 2003)
- Example efforts:
  - Broadcast news (McKeown et al, 2005)
  - Conference presentations (Furui et al, 2004)
  - Voice-mail (Koumpis & Renals, 2003)

Summary

- Large amounts of audio-visual data is now online, but tools are needed to efficiently annotate, search & browse it
- Speech transcription key points:
  - Accurate speech transcription requires knowledge of topic
  - Content words often reliably recognized (if in vocabulary)
  - Adaptation contributes significant improvements
- Spoken document retrieval key points:
  - Tight integration between ASR and text retrieval technology holds great promise for general SDR technology
  - Better evaluation metrics for the SDR scenario
  - Integrate other streams of information
- User interface key points:
  - Generation of readable transcriptions
  - Topic segmentation and summarization

References

- J. Allan, "Robust techniques for organizing and retrieving spoken documents", *EURASIP Journal on Applied Signal Processing*, no. 2, pp. 103-114, 2003.
- C. Allauzen, M. Mohri, and B. Roark, "A general weighted grammar library", in *Proc. of International Conf. on the Implementation and Application of Automata*, Kingston, Canada, July 2004.
- M. Bacchiani, et al, "SCANMail: audio navigation in the voicemail domain," in *Proc. of the HLT Conf.*, pp. 1-3, San Diego, 2000.
- R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, chapter 2, pages 27-30. Addison Wesley, New York, 1999.
- I. Bazzi and J. Glass. "Modeling out-of-vocabulary words for robust speech recognition", in *Proc. of ICSLP*, Beijing, China, October, 2000.
- I. Bazzi and J. Glass. "Learning units for domain independent out-of-vocabulary word modeling", in *Proc. of Eurospeech*, Aalborg, Sep. 2001.
- S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems, Vol. 30, pp. 107-117, 1998.
- C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech", In Proc. of the Annual Meeting of the ACL (ACL'05), pp. 443-450, Ann Arbor, Michigan, June 2005.

References

- R. Fagin, R. Kumar, and D. Sivakumar. “Comparing top k lists”, In *SIAM Journal of Discrete Math*, vol. 17, no. 1, pp. 134-160, 2003.
- S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, “Speech-to-text and speech-to-speech summarization of spontaneous speech”, *IEEE Trans. on Speech and Audio Processing*, vol. 12, no. 4, pp. 401-408, July 2004.
- G. Furnas, et al. “Information retrieval using a singular value decomposition model of latent semantic structure”, in *Proc. of ACM SIGIR Conf.*, pp. 465-480 Grenoble, France, June 1988.
- J. Garofolo, C. Auzanne, and E. M. Voorhees, “The TREC spoken document retrieval track: A success story,” in *Proc. 8th Text REtrieval Conference (1999)*, vol. 500-246 of *NIST Special Publication*, pp. 107–130, NIST, Gaithersburg, MD, USA, 2000.
- J. Gauvain and C. Lee, “Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains”, *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 2, pp. 291-298, April 1994.
- J. Glass, T. Hazen, L. Hetherington and C. Wang, “Analysis and processing of lecture audio data: Preliminary investigations”, in *Proc. of the HLT-NAACL 2004 Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval*, pp. 9-12, Boston, May 2004.

References

- T. Hofmann, “Probabilistic latent semantic analysis”, in *Proc. of Uncertainty in Artificial Intelligence (UAI'99)*, Stockholm, 1999.
- A. Inoue, T. Mikami and Y. Yamashita, “Prediction of sentence importance for speech summarization using prosodic features”, in *Proc. Eurospeech*, 2003.
- R. Iyer and M. Ostendorf, “Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models”, in *Proc. ICSLP*, Philadelphia, 1996.
- D. James, *The Application of Classical Information Retrieval Techniques to Spoken Documents,* PhD thesis, University of Cambridge, 1995.
- D. Jones, et al, “Measuring the readability of automatic speech-to-text transcripts”, in *Proc. Eurospeech*, Geneva, Switzerland, September 2003.
- G. Jones, J. Foote, K. Spärck Jones, and S. Young, “Retrieving spoken documents by combining multiple index sources”, In *Proc. of ACM SIGIR Conf.*, pp. 30-38, Zurich, Switzerland, 1996.
- K. Koumpis and S. Renals, “Transcription and summarization of voicemail speech”, in *Proc. ICSLP*, Beijing, October 2000.
- C. Leggetter and P. Woodland, “Maximum likelihood linear regression for speaker adaptation on continuous density hidden Markov Models”, *Computer Speech and Language*, vol. 9, no. 2, pp. 171-185, April 1995.

References

- B. Logan, P. Moreno, and O. Deshmukh, “Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio”, in *Proc. of HLT*, San Diego, March 2002.
- I. Malioutov and R. Barzilay, “Minimum cut model for spoken lecture segmentation”, in *Proc. of COLING-ACL*, 2006.
- S. Matsoukas, et al, “BBN CTS English System,” available at http:www.nist.gov/speech/tests/rt/rt2003/spring/presentations.
- Kenney Ng, *Subword-Based Approaches for Spoken Document Retrieval,* PhD thesis, Massachusetts Institute of Technology, 2000.
- NIST. The TREC evaluation package available at: http://www-nlpir.nist.gov/projects/trecvid/trecvid.tools/trec_eval
- Douglas W. Oard, et al, “Building an information retrieval test collection for spontaneous conversational speech”, In *Proc. ACM SIGIR Conf.*, pp. 41--48, New York, 2004.
- J. Ponte and W. Croft, “A language modeling approach to information retrieval”, *Proc. ACM SIGIR)*, pp. 275--281, Melbourne, Australia, August1998.

References

- J. Silva Sanchez, C. Chelba, and A. Acero, “Pruning analysis of the position specific posterior lattices for spoken document search”, in *Proc. of ICASSP*, Toulouse, France, May 2006.
- M. Saraclar and R. Sproat, “Lattice-based search for spoken utterance retrieval”, In *Proc. of HLT-NAACL 2004*, pp. 129-136, Boston, May 2004.
- F. Seide and P. Yu, “Vocabulary-independent search in spontaneous speech”, in *Proc. of ICASSP*, Montreal, Canada, 2004.
- F. Seide and P. Yu, “A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech”, in *Proc. of ICSLP*, Jeju, Korea, 2004.
- M. Siegler, *Integration of Continuous Speech Recognition and Information Retrieval for Mutually Optimal Performance,* PhD thesis, Carnegie Mellon University, 1999.
- A. Singhal, J. Choi, D.Hindle, D. Lewis and F. Pereira, “AT&T at TREC-7”, in *Text REtrieval Conference*, pages 239-252, 1999.
- A. Singhal, S. Abney, M. Bacchiani, M. Collins, D. Hindle and F. Pereira, “AT&T at TREC-8”. In *Text REtrieval Conference*, pp. 317-330, 2000.

References

- J. M. Van Thong, et al, “SpeechBot: An experimental speech-based search engine for multimedia content on the web”, *IEEE Trans. on Multimedia*, Vol. 4, No. 1, March 2002.
- A. Waibel, et al, “Advances in automatic meeting record creation and access,” in Proc. of ICASSP, Salt Lake City, May 2001.
- P. Woodland, S. Johnson, P. Jourlin, and K. Spärck Jones, “Effects of out of vocabulary words in spoken document retrieval”, In *Proc. of SIGIR*, pp. 372-374, Athens, Greece, 2000.
- J. Goodman. A Bit of Progress in Language Modeling, Extended Version Microsoft Research Technical Report MSR-TR-2001-72.
- Mehryar Mohri, Fernando C. N. Pereira, and Michael Riley, "Weighted Finite-State Transducers in Speech Recognition" Computer Speech and Language, 16(1):69-88, 2002.
- L. R. Bahl, F. Jelinek and R. L. Mercer. A maximum likelihood approach to continuous speech recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 179-190, 1983.