

# Recent Developments in Chinese Spoken Document Search and Distillation



Berlin Chen  
Department of Computer Science & Information Engineering  
National Taiwan Normal University



This talk was given at Google Taipei  
2009/01/21

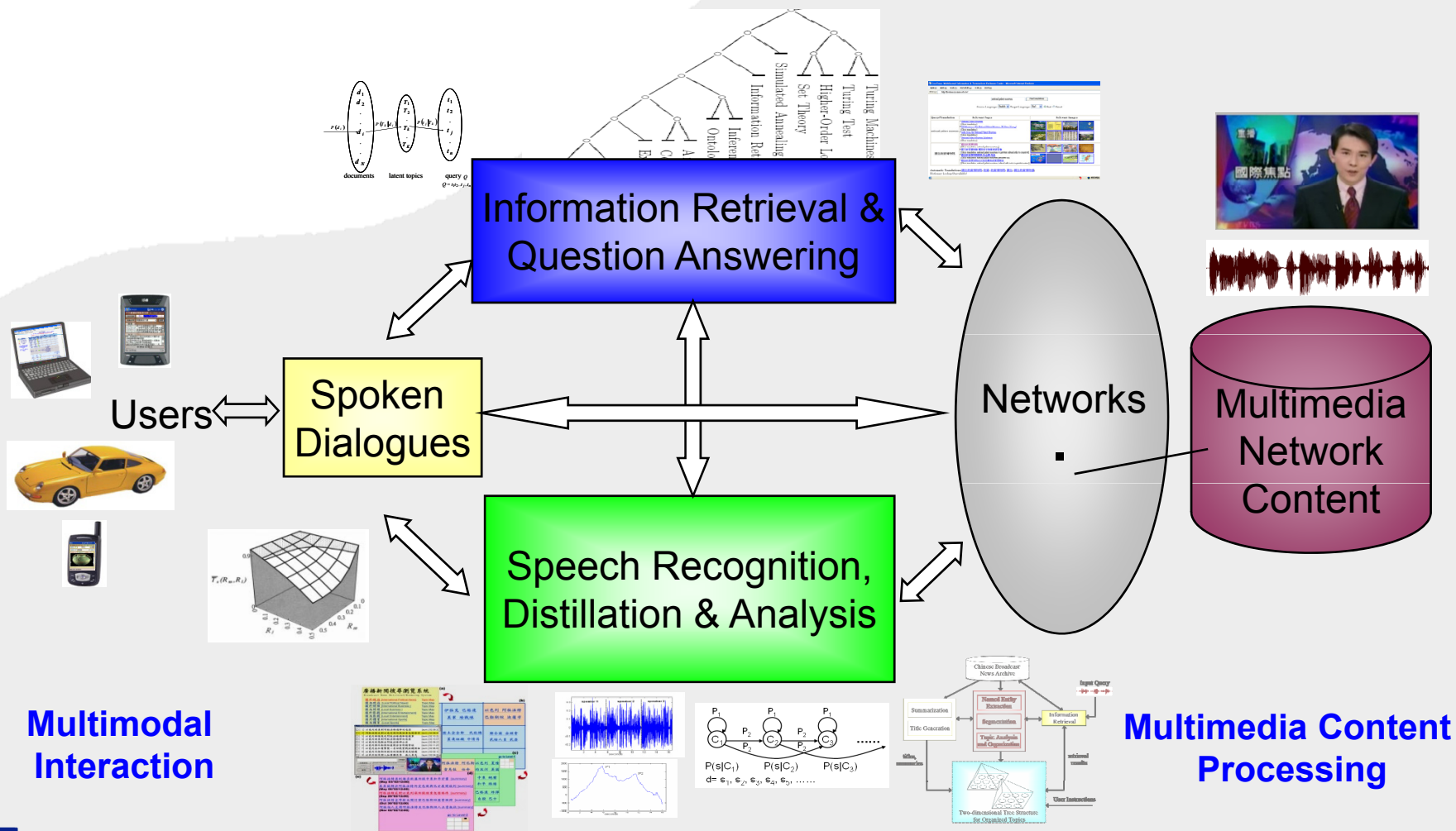
## Outline (1/2)

- Audio-visual contents associated with speech is continuously growing and filling our computers, networks and daily lives
  - Such as broadcast news, shows, podcasts, lecture videos, voice mails, (contact-center or meeting) conversations, etc.
  - Speech is one of the most semantic (or information)-bearing sources
- On the other hand, speech is the primary and the most convenient means of communication between people
  - Speech provides a better (or natural) user interface in wireless environments
  - Especially helpful when using smaller hand-held devices with small screen sizes and limited keyboard entry capabilities
- Speech will be the key for multimedia information access in the near future

## Outline (2/2)

- Organization and retrieval and of multimedia (or spoken) are much more difficult
  - Written text documents are better structured and easier to browse through
    - Provided with titles and other structure information (e.g., hyperlinks)
    - Easily shown on the screen to glance through (with visual perception)
  - Multimedia (Spoken) documents are just video (audio) signals
    - Users cannot efficiently go through each one from the beginning to the end during browsing, even if the they are automatically transcribed by automatic speech recognition
    - However, abounding **speaker**, **emotion** and **scene** information make them much more attractive than text
    - Better approaches for efficient organization and retrieval of multimedia (spoken) documents are highly demanded

# Multimodal Access to Multimedia in the Future



## Related Research Work and Applications

- Continuous and substantial efforts have been paid to (multimedia) speech recognition, distillation and retrieval in the recent past
  - [Informedia System at Carnegie Mellon Univ.](#)
  - [AT&T SCAN System](#)
  - [Rough'n'Ready System at BBN Technologies](#)
  - [SpeechBot Audio/Video Search System at HP Labs](#)
  - [IBM Speech Search for Call-Center Conversations & Call-Routing, Voicemails, Monitoring Global Video and Web News Sources \(TALES\)](#)
  - [Google Voice Search \(GOOG-411, Audio Indexing, Translation\)](#)
  - [Microsoft Research Audio-Video Indexing System \(MAVIS\)](#)
  - [MIT Lecture Browser](#)
  - [NTT Speech Communication Technology for Contact Centers](#)
  - [Some Prototype Systems Developed in Taiwan](#)

# World-wide Speech Research Projects

- There also are several research projects conducting on related spoken document processing tasks, e.g.,
  - **Rich Transcription Project**<sup>1</sup> in the United States (2002-)
    - Creation of recognition technologies that will produce transcriptions which are more readable by humans and more useful for machines
  - **TC-STAR Project**<sup>2</sup> (Technology and Corpora for Speech to Speech Translation) in Europe (2004-2007)
    - Translation of speeches recorded at European Parliament, between Spanish and English, and of broadcast news by Voice of America, from Mandarin to English
  - **“Spontaneous Speech: Corpus and Processing Technology” Project** in Japan (1999-2004)
    - 700 hours of lectures, presentations, and news commentaries
    - Automatic transcription, analysis (tagging), retrieval and summarization of spoken documents



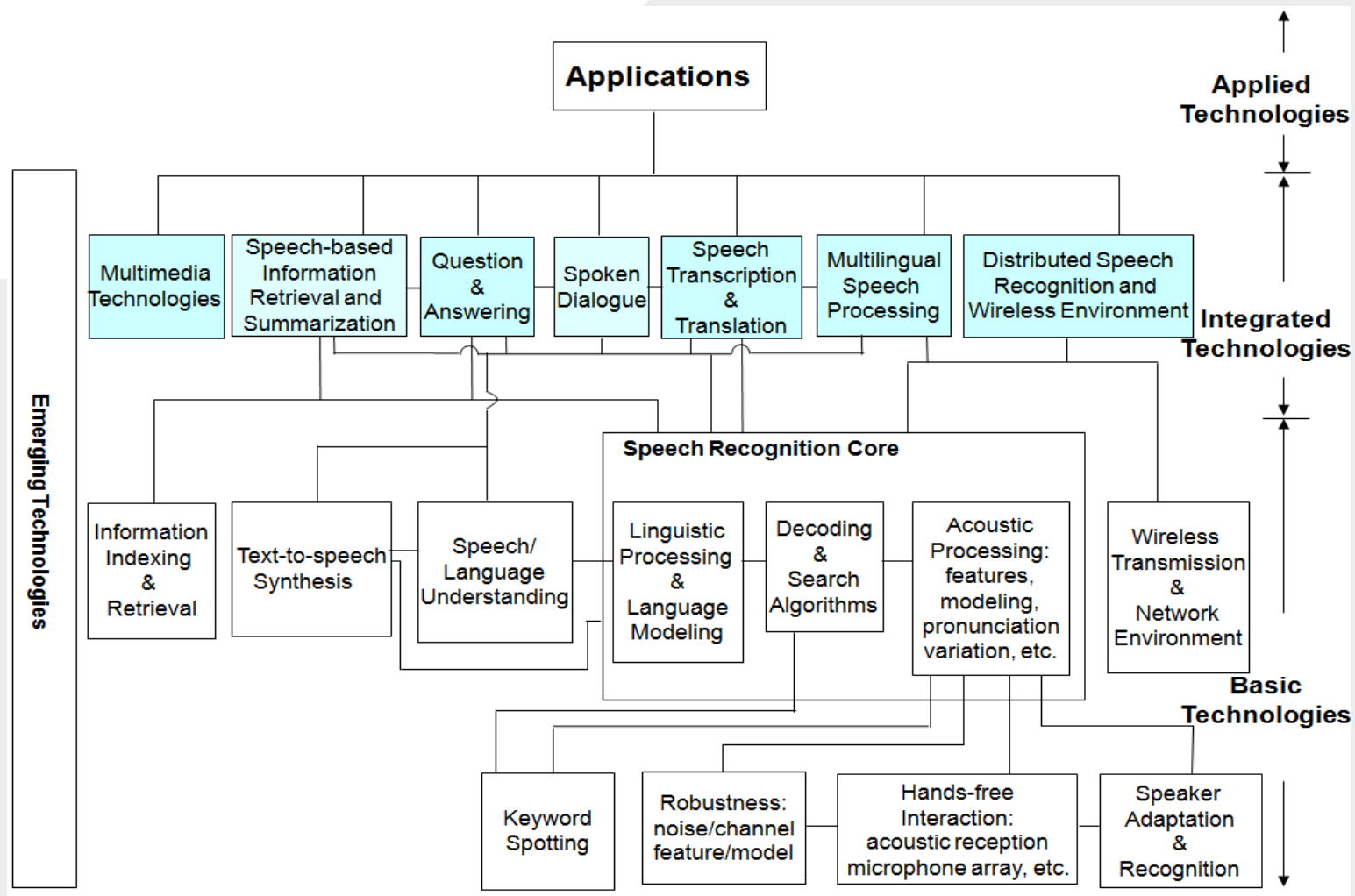
<sup>1</sup> <http://www.nist.gov/speech/tests/rt/>

<sup>2</sup> <http://www.tc-star.org>

# Evaluations of the Rich Transcription Project

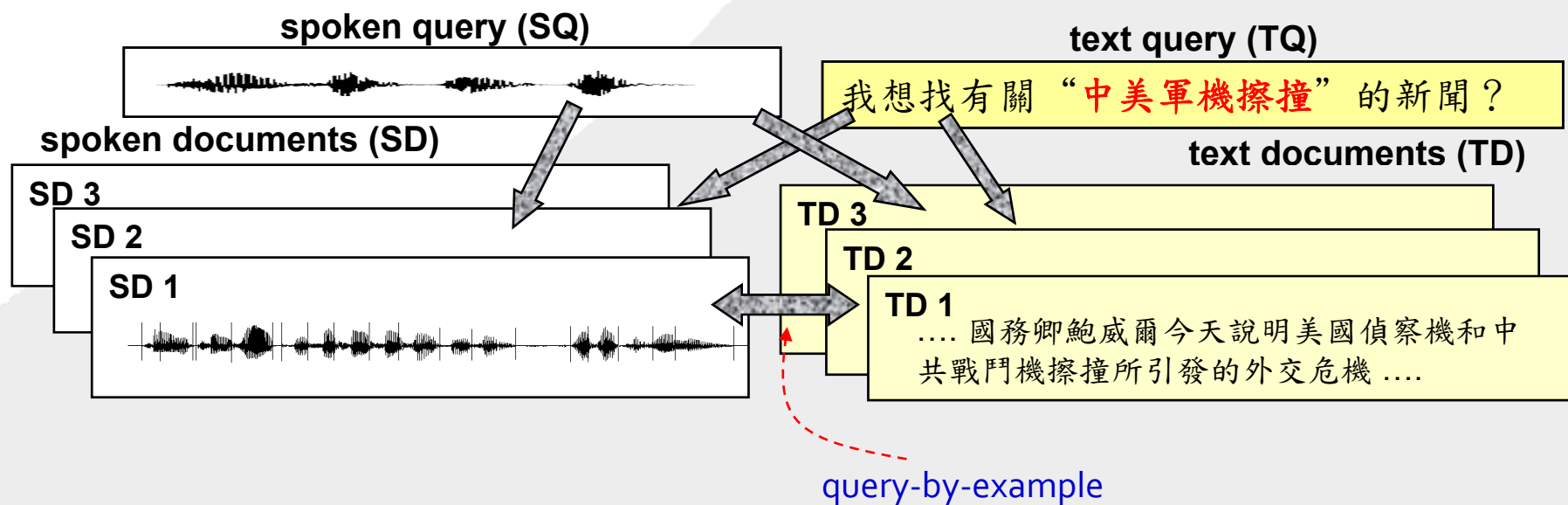
- GALE (Global Autonomous Language Exploitation) Translation: 2006 – present
  - Translates language data from a input source language (either Arabic or Chinese, in audio or text) into a target one (English in text).
- Spoken Term Detection: 2006 – present
  - Facilitate research and development of technology for finding short word sequences rapidly and accurately in large heterogeneous audio archives (three languages: Arabic, English, and Mandarin)
- TRECVID Event Detection: 2008 –
- Language Recognition Evaluation: 1996 –
- ...

# Related Research Areas of Speech Processing





# Scenario for Speech Search



- SQ/SD is the most difficult
- TQ/SD is studied most of the time

# Categorization of Speech Search Tasks

- **Spoken Document Retrieval (SDR)**
  - Find spoken documents that are (topically) “**relevant**” to a given query
  - Queries usually are very long topic descriptions (**query-by-example**)
  - Exploit LVCSR and text IR technologies
  - SDR is already regarded as a “**solved**” problem, especially for broadcast news (even with WER of more than 30%, retrieval using 1-best automatic transcripts are comparable to that using reference transcripts)
- **Spoken Term Detection (STD)**
  - Much like Web-style search
  - Queries are usually short (1-3 words), and find the “**matched**” documents where all query terms should be present
  - Then, relevance ranking are performed on the “**matched**” documents
  - Have drawn much attention recently in the speech processing community
    - Exploit word lattices or confusion networks consisting of multiple hypotheses to compensate for speech recognition errors

# TREC SDR Evaluation Plan

- A series of SDR tracks conducted during 1996-2000 (TREC-6 ~ TREC-9)
  - Focus on using broadcast news from various sources: Voice of America, CNN, ABC, PRI, etc., comprising more than 5 hundred hours of speech ( $\geq 20,000$  manually segmented documents, 250 words per document on average)
  - The queries are long and stated in plain English (e.g., a text news story) rather than using the keyword (Web) search scenario
- Findings
  - Retrieval performance is quite flat with ASR WER variations in the range of 10~35% (roughly  $\leq 5\%$  degradation in performance in comparison with the “approximately” manual transcriptions)
  - SDR of broadcast news speech has been thought of as “a successful story”

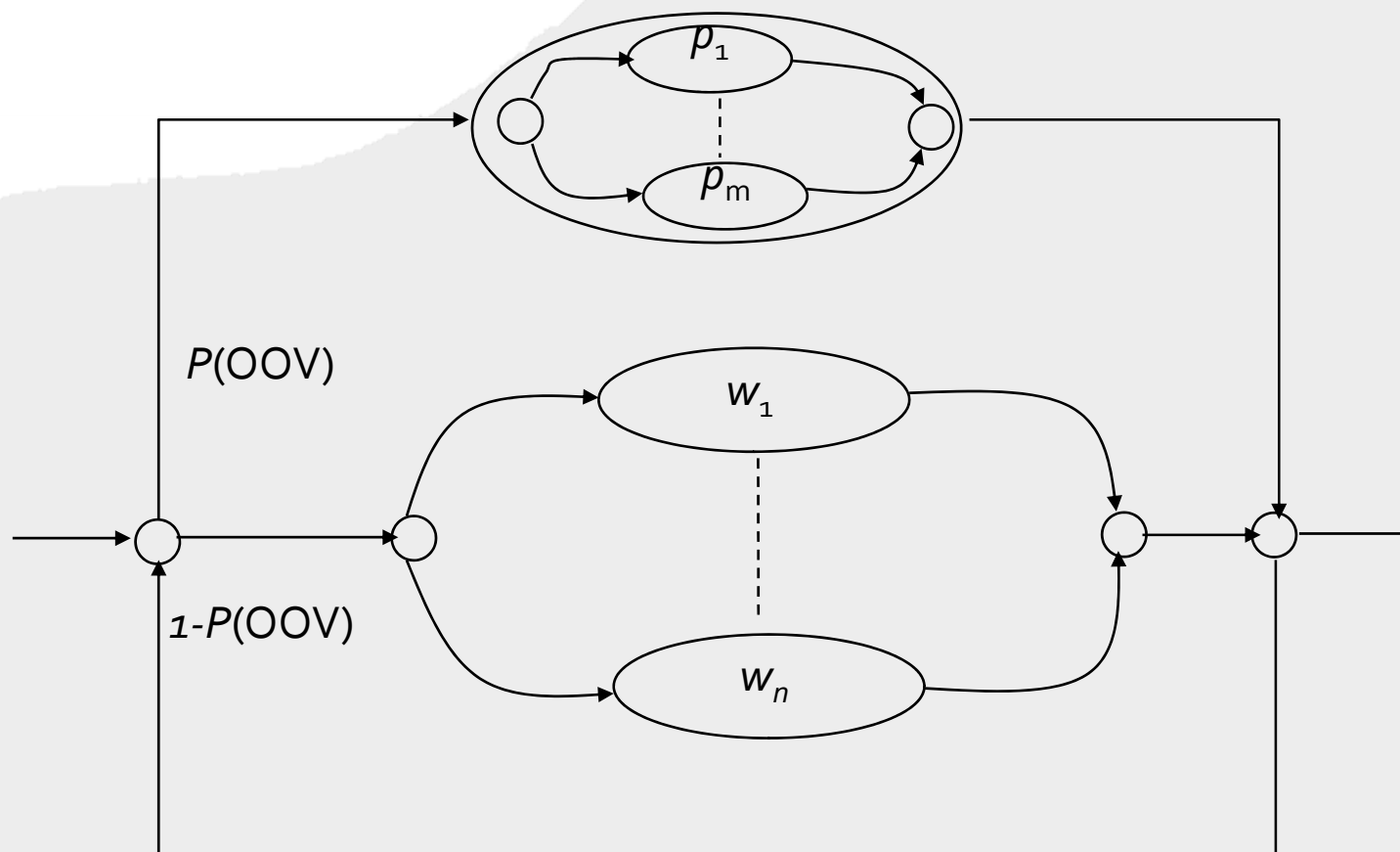


## Types of ASR Transcription (1/2)

- Word Sequences Produced by LVCSR
  - More accurate for audio indexing
  - Faced with the “OOV-word” problems (query terms are often less-frequent topic-specific words)
  - Tend to have lower recall
- Phonetic-Unit (or subword) Sequences Produced by Phone Recognizer
  - Bypass the “OOV-word” problems by locating spoken documents containing the phonetic sequences that match the pronunciations of the query words
  - Complicate the post-processing of the spoken documents for other IR-related applications
  - Tend to have higher recall at the expense of lower precision
- Hybrid Approach Blending Word and Phonetic Information

## Types of ASR Transcription (2/2)

- Represent the OOV region by a network of phonetic units

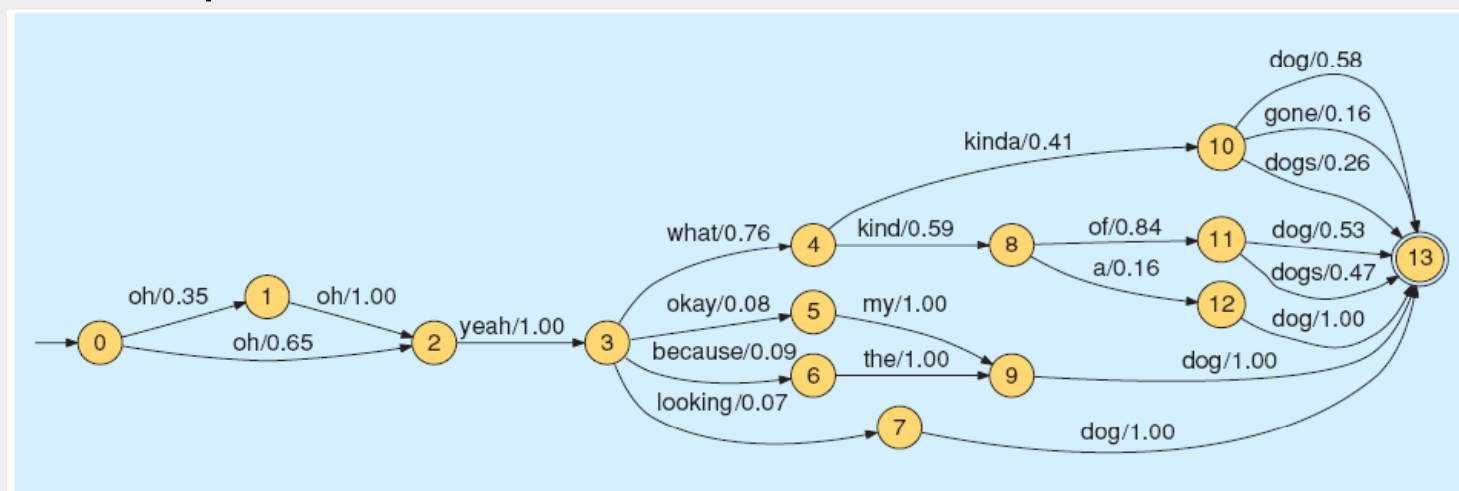


# Evaluation Metrics

- SDR and STD
  - Recall
  - Precision
  - F-measure (a harmonic mean of recall and precision)
  - *R*-precision
  - Precision at *N* document cutoff level
  - [Mean Average Precision \(MAP\)](#)
  - [Actual Term-Weighted Value \(ATWV\)](#)
  - ...
- ASR
  - WER
  - Lattice WER
  - OOV Rate
  - Query OOV Rate
  - ...

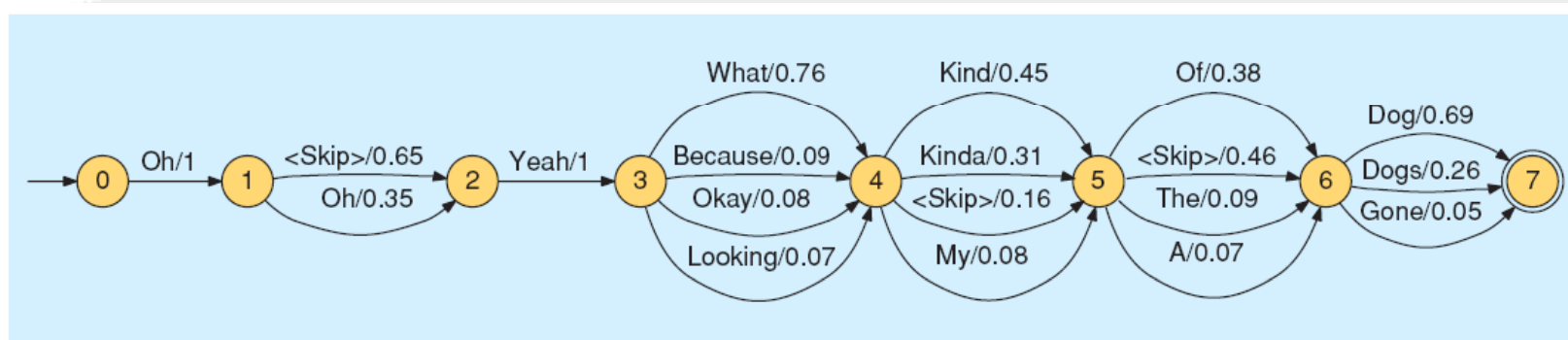
## STD: 1-bset Sequences vs. Lattices (1/5)

- Use of 1-best ASR output as the transcription to be indexed is suboptimal due to the high WER, which is likely to lead to low recall
- ASR lattices do provide much better WER, but the position information is not readily available (uncertainty of word occurrences) ?
- An example ASR Lattice



## STD: 1-bset Sequences vs. Lattices (2/5)

- Confusion/Consensus Networks (CN, also called “Sausages”)  
derived from the Lattice
  - Group the word arcs in the lattice into several strictly linear lists (clusters) of word alternatives



- L. Mangu, E. Brill, A. Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech & Language* 14(4), 2000



## STD: 1-bset Sequences vs. Lattices (3/5)

- Position-Specific Posterior Probability Lattices (PSPL)
  - Position information is crucial for being able to evaluate proximity when assigning a relevance score to a given document
  - Estimate the posterior probability of a word  $w$  at a specific position  $l$  in the lattices  $P(w, l | LAT)$  of spoken queries and documents

	0	1	2	3	4	5	6	7
Oh	1.0	Yeah .65	What .46	Kind .27	Dog .26	EOS .34	EOS .44	EOS .16
—		Oh .35	Yeah .35	What .27	Of .23	<b>Dog</b> .29	Dog .09	—
		—	Because .06	Kinda .19	Kind .16	Dogs .13	Dogs .06	
			Okay .05	The .06	Kinda .11	Of .13	—	
			Looking .05	My .05	Dogs .05	A .03		
			—	Dog .05	EOS .05	Gone .02		
				.....	...	.....	...	
						—		

- [Technical Details of PSPL](#)

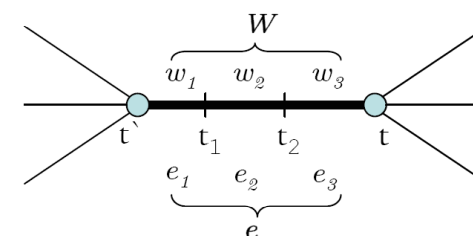


## STD: 1-bset Sequences vs. Lattices (4/5)

- Y.C. Pan and L.S. Lee at NTU extend PSPL to indexing subword-level (character & syllable) information for retrieval of Chinese broadcast news (using text queries)

“Analytical comparison between position specific posterior lattices and confusion network based on word and subword units for spoken document indexing,” ASRU 2007

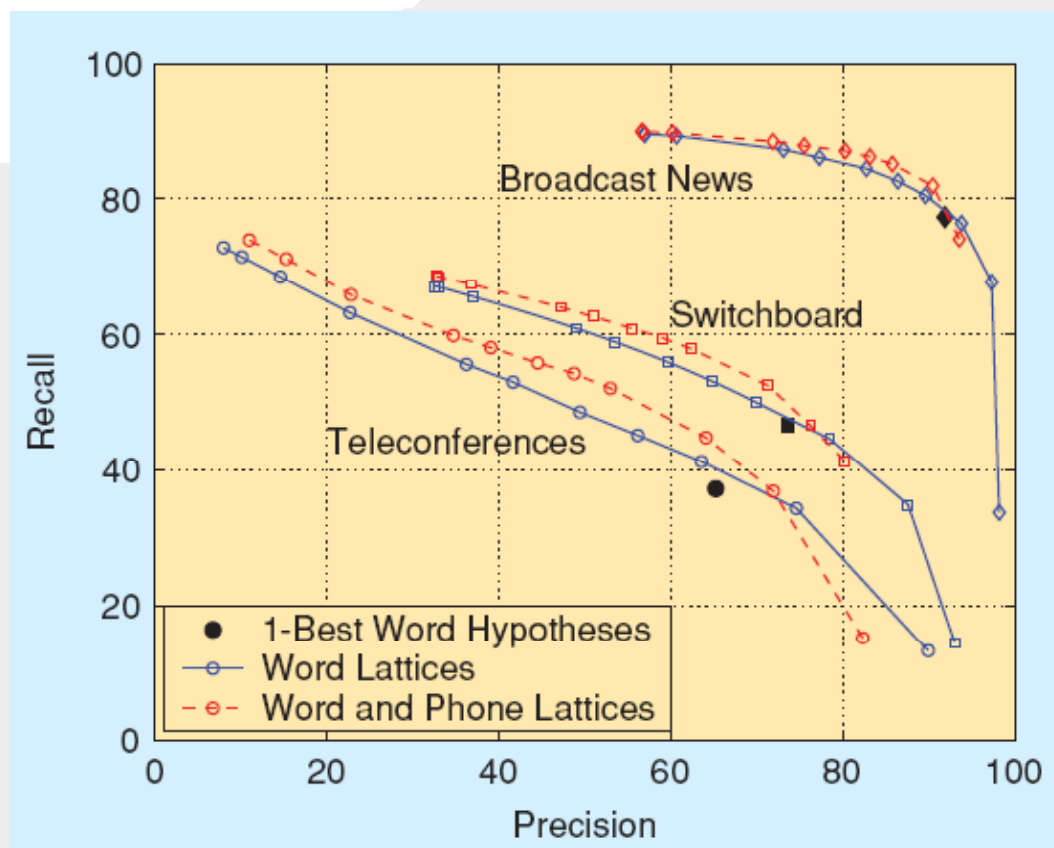
Subword Posterior Probability



- P. Yu, F. Seide et al. at MSRA proposed alternative approaches that are analogous to PSPL
    - Time-based Merging for Indexing (TMI) for size reduction and Time-Anchored Lattice Expansion (TALE) for word-position mapping
- “Word-lattice based spoken-document indexing with stand text indexers,” SSSC 2008 (in conjunction with SIGIR 2008) & SLT2008

## STD: 1-bset Sequences vs. Lattices (5/5)

- Comparison between indexing with 1-bset sequences and lattices



# STD: My Indexing Approaches, Probably Out-of-Date (1/2)

- Define several set of sub-word (character/phone) level features

Overlapping syllable segments with length  $N$  (free word boundaries)

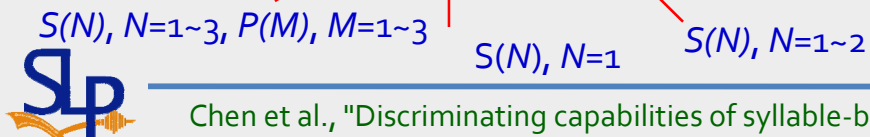
Syllable Segments	Examples
$S(N), N=1$	$(s_1) (s_2) \dots (s_{10})$
$S(N), N=2$	$(s_1 s_2) (s_2 s_3) \dots (s_9 s_{10})$
$S(N), N=3$	$(s_1 s_2 s_3) (s_2 s_3 s_4) \dots (s_8 s_9 s_{10})$
$S(N), N=4$	$(s_1 s_2 s_3 s_4) (s_2 s_3 s_4 s_5) \dots (s_7 s_8 s_9 s_{10})$
$S(N), N=5$	$(s_1 s_2 s_3 s_4 s_5) (s_2 s_3 s_4 s_5 s_6) \dots (s_6 s_7 s_8 s_9 s_{10})$

Syllable pairs separated by  $M$  syllables

Syllable Pair Separated by $M$ syllables	Examples
$P(M), M=1$	$(s_1 s_3) (s_2 s_4) \dots (s_8 s_{10})$
$P(M), M=2$	$(s_1 s_4) (s_2 s_5) \dots (s_7 s_{10})$
$P(M), M=3$	$(s_1 s_5) (s_2 s_6) \dots (s_6 s_{10})$
$P(M), M=4$	$(s_1 s_6) (s_2 s_7) \dots (s_5 s_{10})$

- Results on 10-hour broadcast news retrieval with short queries

Average Precision	Syllable-based (S)	Character-based (C)	Word-based (W)	S+C+W
TQ/TD	0.9740 (0.4743, 0.9656)	<b>0.9778</b> (0.7680, 0.9604)	0.9027 (0.8804, 0.9003)	0.9797
SQ/TD	<b>0.8982</b> (0.4137, 0.8898)	0.8811 (0.6671, 0.8676)	0.7755 (0.7489, 0.7683)	0.9022
TQ/SD	<b>0.7148</b> (0.3456, 0.7009)	0.6988 (0.5577, 0.6872)	0.6160 (0.5988, 0.6138)	0.7267
SQ/SD	<b>0.6739</b> (0.3120, 0.6583)	0.6515 (0.5136, 0.6429)	0.5549 (0.5386, 0.5534)	0.6814



## STD: My Indexing Approaches, Probably Out-of-Date (2/2)

- Let the spoken document collection “tells” us Which Are Important “Lexical Segments” for Indexing, Which Are Not
  - Started with a set of indexing features consisting of single base syllables as the initial lexical segments (LSs)
  - In each iteration, any two adjacent lexical segments with scores higher than the threshold become a new LS
    - Criterion: Forward-Backward Bigram (FB)

$$FB(u, v) = \sqrt{P_f(v | u)P_b(u | v)}$$

- All instances of these pairs are replaced by the new LSs

- Repeat to 2

Syllable Segments	Possible Words
a la fa te	阿拉法特 (Arafat)
ye lu sa leng	耶路撒冷 (Jerusalem)
a ken se zhou	阿肯色州 (Arkansas)
mai dang lao	麥當勞 (McDonald's)
jian pu zhai	柬埔寨 (Cambodia)

# SDR: Exploiting Lattices and Language Models

- T.K. Chia, H. Li, H.T. Ng et al., extended Chen et al.'s work on query-by-example (ACM TALIP 2004) to spoken queries, and also extended Lafferty and Zhai's Kullback-Leibler divergence based LMs for document modeling (SIGIR 2001)

“A lattice-based approach to query-by-example spoken document retrieval,”  
SIGIR 2008

System	Retrieval source		Stop word list	Pruning parameters ( $\Theta_{\text{qry}}, \Theta_{\text{doc}}$ )	Mean average precision	
	Queries	Documents			For devel. queries	For test queries
Ref $\xrightarrow{\text{smart}}$ Ref	Exemplar reference	Reference	smart	–	0.8363	0.7781
1-best $\xrightarrow{\text{smart}}$ 1-best	Exemplar 1-best	1-best	smart	–	0.8271	0.7406
1-best $\xrightarrow{\text{smart}}$ Lat	Exemplar 1-best	Lattices	smart	(–, 140)	0.8321	0.7499
Lat $\xrightarrow{\text{smart}}$ 1-best	Exemplar lattices	1-best	smart	(240, –)	0.8355	0.7487
Lat $\xrightarrow{\text{smart}}$ Lat	Exemplar lattices	Lattices	smart	(240, 160)	0.8421	0.7569

$$\begin{aligned} \text{Rel}_{\text{lat}}(\mathbf{d}, \mathbf{q}) &= \sum_{w \in \mathcal{V}} \Pr(w | \mathbf{q}) \log \Pr(w | \mathbf{d}) \\ &= \frac{1}{\mathbb{E}[\|\mathbf{q}\|]} \sum_{\substack{w \in \mathcal{V}, \\ \mathbb{E}[c(w; \mathbf{q})] > 0}} \mathbb{E}[c(w; \mathbf{q})] \log \Pr(w | \mathbf{d}) \end{aligned}$$

$$\Pr(w | \mathbf{d}) = (1 - \lambda) \frac{\mathbb{E}[c(w; \mathbf{d})] + \mu \Pr(w | \mathcal{C})}{\mathbb{E}[\|\mathbf{d}\|] + \mu} + \lambda \Pr(w | \mathcal{U})$$

## SDR: Word Topic Models (1/4)

- Each word of a language is treated as a word topic model (WTM) for predicting the occurrences of other words

$$P_{\text{WTM}}(w_i | M_{w_j}) = \sum_{k=1}^K P(w_i | T_k) P(T_k | M_{w_j})$$

- The relevance measure between a query and a document can be expressed by (a special kind of translation model)

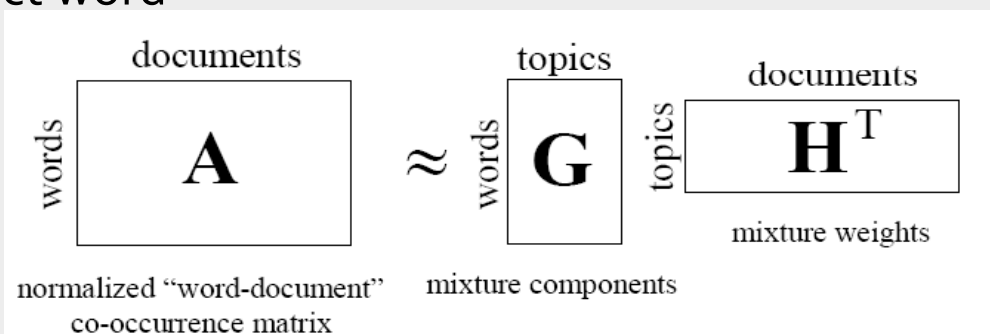
$$P_{\text{WTM}}(Q | M_D) = \prod_{w_i \in Q} \left[ \sum_{w_j \in D} P_{\text{WTM}}(w_i | M_{w_j}) P_{\text{MLE}}(w_j | D) \right]^{c(w_i, Q)}$$

- B. Chen, "Word topic models for spoken document retrieval and transcription," *ACM TALIP*, March 2009
- B. Chen, "Latent topic modeling of word co-occurrence information for spoken document retrieval," *IEEE ICASSP 2009*

## SDR: Word Topic Models (2/4)

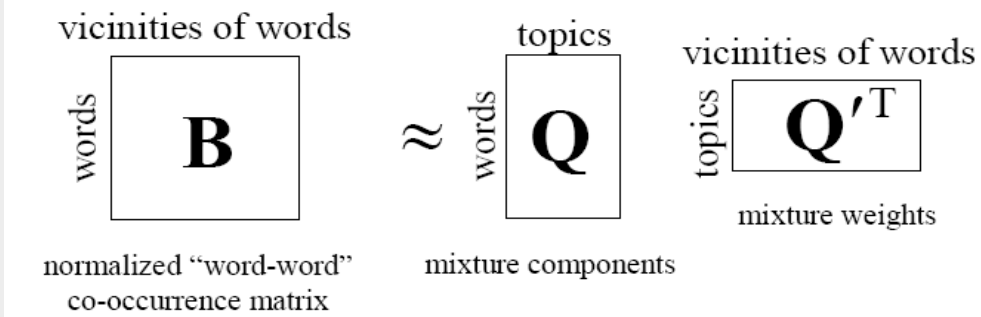
- WTM also can be viewed as a nonnegative factorization of a “word-word” matrix consisting probability entries (for unsupervised model training)
  - Each column encodes the vicinity information of all occurrences of a distinct word

PLSA/LDA



$$P_{\text{PLSA}}(w_i | M_D) = \sum_{k=1}^K P(w_i | T_k) P(T_k | M_D)$$

WTM



$$P_{\text{WTM}}(w_i | M_{w_j}) = \sum_{k=1}^K P(w_i | T_k) P(T_k | M_{w_j})$$

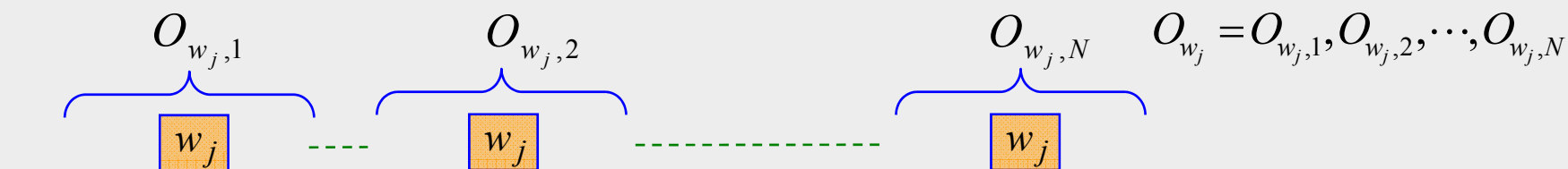


## SDR: Word Topic Models (3/4)

- Unsupervised training (WTM-U)

- The WTM of each word can be trained by concatenating those words occurring within a context window of size around each occurrence of the word, which are postulated to be relevant to the word

$$\log L_{\mathbf{w}} = \sum_{w_j \in \mathbf{w}} \log P_{\text{WTM}} \left( O_{w_j} \mid M_{w_j} \right) = \sum_{w_j \in \mathbf{w}} \sum_{w_i \in Q_{w_j}} c(w_i, O_{w_j}) \log P_{\text{WTM}} \left( w_i \mid M_{w_j} \right)$$



- Supervised training (WTM-S)

- Maximize the log-likelihood of a set of training query exemplars generated by their relevant documents

$$\log L_{\mathbf{Q}_{TrainSet}} = \sum_{Q \in \mathbf{Q}_{TrainSet}} \sum_{D \in \mathbf{D}_{R \text{ to } Q}} \log P_{\text{WTM}} (Q \mid M_D)$$

## SDR: Word Topic Models (4/4)

- Tested on TDT-2 & TDT-3 Collections (“query-by-example” tasks)
  - Results on TDT-2

Retrieval Model	VSM	LSA	SVM	HMM/ Unigram	HMM/ Bigram	PLSA-U	PLSA-S	WTM-U	WTM-S
TD	0.5548	0.5510	0.5797	0.6327	0.5427	0.6277	0.7243	0.6395	0.7672
SD	0.5122	0.5310	0.5317	0.5658	0.4803	0.5681	0.6652	0.5739	0.7558

- WTM also has been applied with good success to speech recognition and speech summarization
  - “Word topical mixture models for dynamic language model adaptation,” *ICASSP 2007*
  - “Word Topical Mixture Models for Extractive Spoken Document Summarization,” *ICME 2007*

# History of Text Summarization Research

- Research into automatic summarization of text documents dates back to the early 1950s
  - However, research work has suffered from a lack of funding for nearly four decades
- Fortunately, the development of the World Wide Web led to a renaissance of the field
  - Summarization was subsequently extended to cover a wider range of tasks, including multi-document, multi-lingual, and multi-media summarization

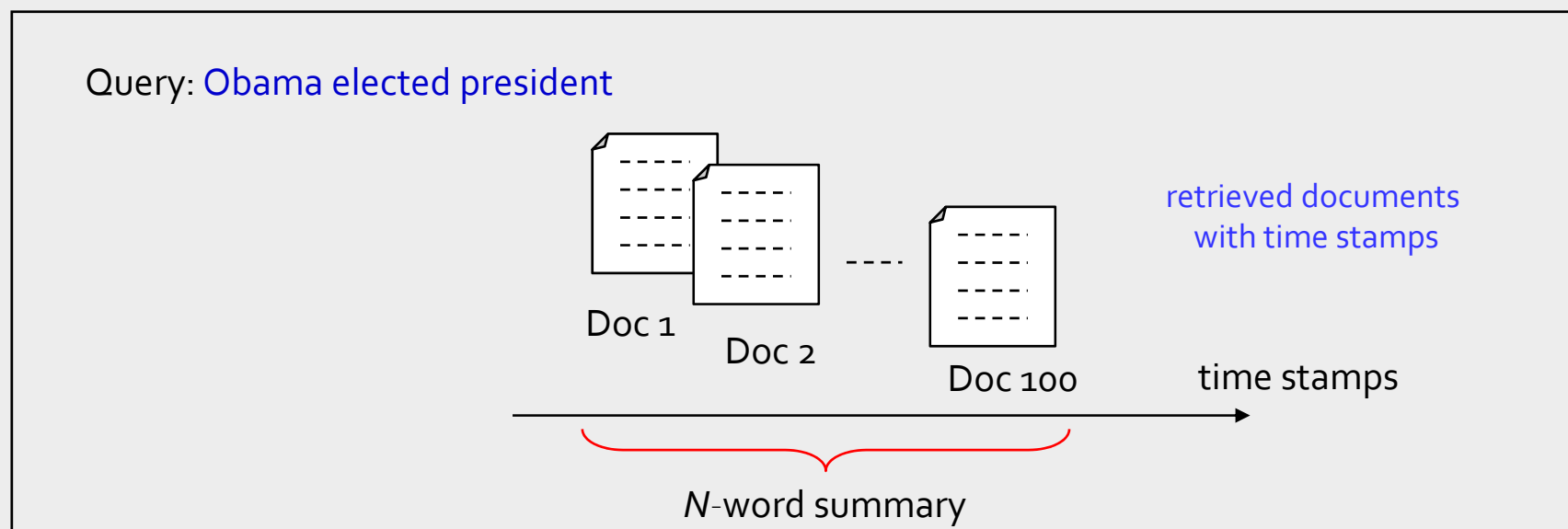
## Spectrum of Text/Speech Summarization Research (1/2)

- Extractive and Abstractive Summarization
  - **Extractive summarization** produces a summary by selecting indicative sentences, passages, or paragraphs from an original document according to a predefined target summarization ratio
    - This requires sentence ranking and compacting
  - **Abstractive summarization** provides a fluent and concise abstract of a certain length that reflects the key concepts of the document
    - This requires highly sophisticated techniques, including semantic representation and inference, as well as natural language generation

In recent years, researchers have tended to focus on extractive summarization.

## Spectrum of Text/Speech Summarization Research (2/2)

- Generic and Query-oriented Summarization
  - A **generic summary** highlights the most salient information in a document
  - A **query-oriented summary** presents the information in a document that is most relevant to the user's query



# A Probabilistic Generative Framework for Speech Summarization (1/2)

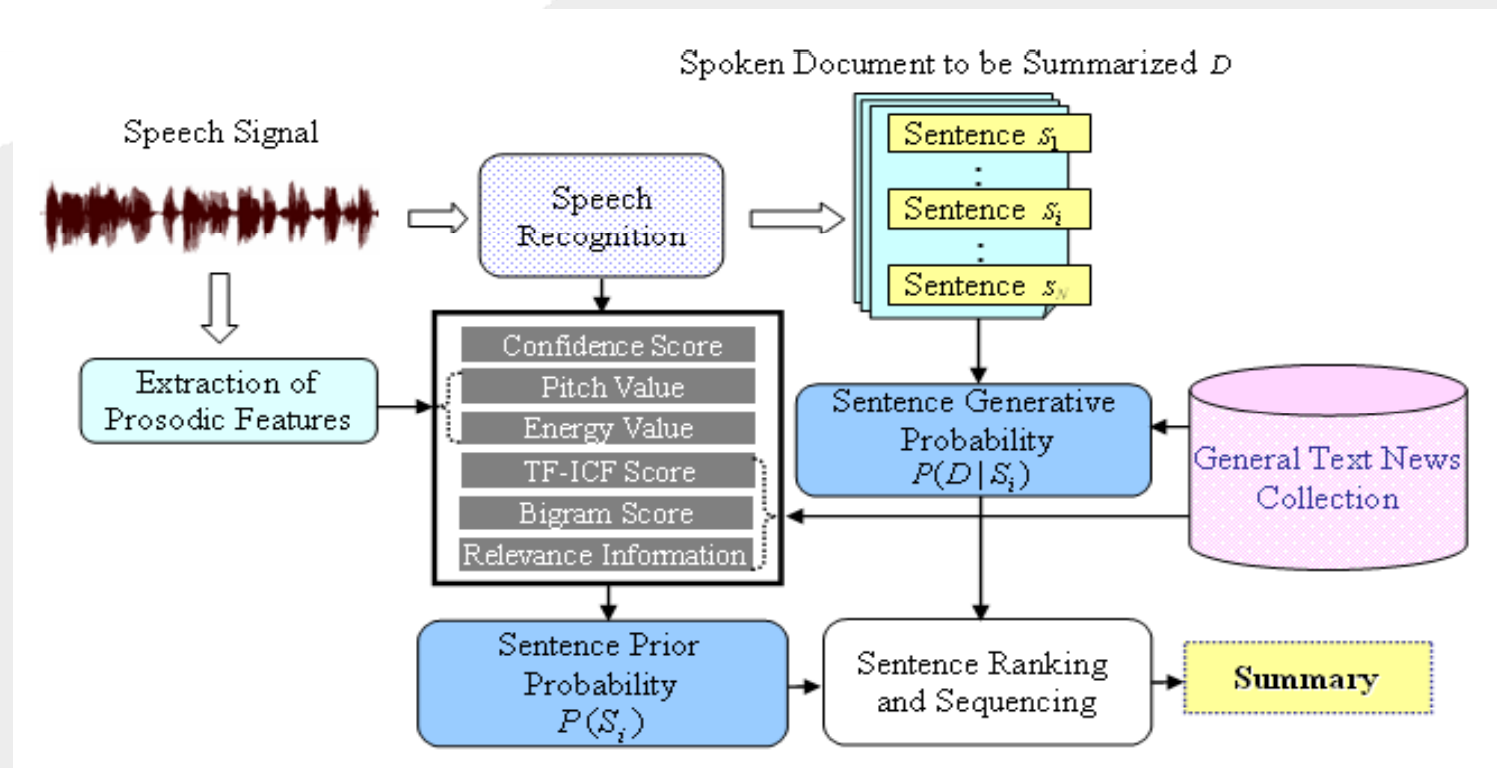
- Criterion: Ranking sentences by their posteriori probabilities

$$P(S_i|D) = \frac{P(D|S_i)P(S_i)}{P(D)} \stackrel{\text{rank}}{=} P(D|S_i)P(S_i)$$

- Sentence Generative Model,  $P(D|S_i)$ 
  - Each sentence of the document as a probabilistic generative model
  - Language Model (LM), Sentence Topic Model (STM) and Word Topic Model (WTM) are initially investigated
- Sentence Prior Distribution,  $P(S_i)$ 
  - The sentence prior distribution may have to do with sentence duration/position, correctness of sentence boundary, confidence score, prosodic information, etc.

# A Probabilistic Generative Framework for Speech Summarization (2/2)

- Speech summarization can be performed in a purely unsupervised manner



# Features Used for Speech Summarization

- We have also investigated using various supervised machine-learning models, such as SVM and CRF, to make use of these features

Structural Features (St)	<i>POSITION</i> : Sentence position <i>DURATION</i> : Duration of the preceding/current/following sentence
Lexical Features (Le)	<i>BIGRAM_SCORE</i> : Normalized bigram language model scores <i>SIMILARITY</i> : Similarity scores between a sentence and its preceding/following neighbor sentence <i>NUM NAME ENTITIES</i> : Number of named entities (NEs) in a sentence
Acoustic Features (Ac)	<i>PITCH</i> : Min/max/mean/difference pitch values of a spoken sentence <i>ENERGY</i> : Min/max/mean/difference value of energy features of a spoken sentence <i>CONFIDENCE</i> : Posterior probabilities
Relevance Features (Re)	<i>R-VSM</i> : Relevance score obtained by using the VSM summarizer <i>R-LSA</i> : Relevance score obtained by using the LSA summarizer

- Extra information cues
  - Inter-document similarity (IDS): similarity of documents in the relevance class of a given sentence
  - Inter-sentence similarity (ISS): similarity of sentences in a document





# An Example for Speech Summarization (1/2)

## reference transcript with correct sentence boundaries

行政院在今天對立法院三讀的案子提出覆議  
翻開憲政史其實並不多見  
事實上這一次的財劃法是行政院第七次行使覆議權  
在目前朝野立院席次相當接近的情況之下這次的覆議案會不會成功  
繼續是我們的報導  
覆議權可以說是憲法賦予行政部門反制立法權的重要手段  
以這回財劃法為例  
行政院認為立法院在一月二十五號函送的修正內容窒礙難行無法取得  
行政院依法必須在十天內對立法院提出覆議  
也就是說透過立委的表決將財劃法還原到沒有修法之前  
而今天二月六號就是行政院針對財劃法可以提出覆議案的法定期限  
依照憲法規定覆議案經過總統核可送到立法院  
立法院必須在十五天內召開院會來處理  
表決的時候必須要有半數以上的立委投下反對票  
如果過不了這個門檻那麼行政院就算是覆議成功  
換句話說這回在野黨必須動員一百一十三席來反對覆議案  
以目前朝野席次相差不多的情況來看國民黨想要捍衛黨版的財劃法  
不少的立委利用春節到國外去了國民黨光是動員就已經是一大難題  
再加上親民黨的態度游離國親兩黨還不見得會在這個案子上再度  
真的訴諸表決國民黨也難保優勢地位  
立法院現在已經確定將在二月十九號開議  
當天就聽取行政院長游錫堃對覆議案的報告並進行討論  
然後在第二天也就是二月二十號進行表決  
可以想見在過年期間朝野之間的角力戰將在臺面下悄悄展開  
公視新聞陳娟媽馬台興採訪報導

WER: 23.94%



## Automatic transcript with probably incorrect sentence boundaries

今天在今天被立法院三讀的案子提出覆議  
翻開線建設其實並不多見  
這是想再一次的財劃法是行政院第七次  
新竹市府民權在目前朝野議員席次相當接近的情況之下  
這次的覆議案會不會成功  
晚間在那裡在學術報導  
布希宣佈可以說是憲法賦予行政部門但距離大選的重要手段  
業為財劃法覆議的行政院認為立法院在一月二十五號函送的修正內容窒礙難行無法取得  
行政院依法必須在十天內對立法院提出覆議  
也就是說超過立委的表決將財劃法官員到沒有修法之前  
而今天二月六化  
九十七美元針對財劃法可以提出覆議案的法定期限  
依照憲法規定覆議案經過總統核可送到立法院  
立法院必須在十五天內召開年會來處理  
表決的時候必須要有半數以上的立委投下反對票  
中共公佈了這個門檻由行政院救災時覆議成功  
換句話說這位在野黨必須動員一百一十三席來反對覆議案  
以目前朝野席次相差不多的情況來看  
國民黨想要捍衛黨版的財劃法修正案  
南部地區相當高不少的立委利用春節到國外去了  
國民黨黃石公園就已經是一大難題  
在加上親民黨的態度有利  
波及兩黨還不見得會對這個案子已再度合作  
這個付諸表決國民黨也難保優勢地位  
立法院現在已經確定將在二月十九淘汰一  
當天就聽取行政院長游錫堃對覆議案的報導並進行討論  
然後在第二天依舊十二月二十號進行表決  
可以想見在過年期間朝野之間的角力戰將災害影響悄悄展開  
公視新聞成年人拉抬警方報案

# An Example for Speech Summarization (2/2)

## handcrafted summary

行政院在今天對立法院三讀的案子提出覆議 翻開憲政史其實並不多見 事實上這一次的財劃法是行政院第七次行使覆議權	Ratio:10%
行政院認為立法院在一月二十五號函送的修正內容窒礙難行無法配合 行政院依法必須在十天內對立法院提出覆議	Ratio:20%
換句話說這回在野黨必須動員一百一十三席來反對覆議案 以目前朝野席次相差不多的情況來看國民黨想要捍衛黨版的財劃法修正案難度的確相當高 可以想見在過年期間朝野之間的角力戰將在臺面下悄悄展開	Ratio:30%

## automatic summary

這是想再一次的財劃法是行政院第七次 新竹市府民權在目前朝野議員席次相當接近的情況之下 今天在今天被立法院三讀的案子提出覆議	Ratio:10%
這次的覆議案會不會成功 布希宣佈可以說是憲法賦予行政部門但距離大選的重要手段 九十七美元針對財劃法可以提出覆議案的法定期限	Ratio:20%
也就是說超過立委的表決將財劃法官員到沒有修法之前 業者為財劃法覆議的行政院認為立法院在一月二十五號函送的修正內容窒礙難行無法取得 翻開線建設其實並不多見	Ratio:30%

## Conclusions

- The use of word lattices (PSPL ,CN, et al.) has been an active area of research for robust audio indexing and retrieval
- Most of the research efforts devoted to spoken document retrieval focus on “text” queries but not “spoken” queries
- Given a query stating the user’s information need
  - Try to find “matched” spoken terms in documents or retrieve “relevant” documents ?
- Word topic models (WTM) have shown with good potential for spoken document recognition, search and summarization

# IBM's Research Activities in Speech Translation and Speech-based Multimedia Content Access

## Speech Translation

- **Speech-to-text: driven by foreign broadcast monitoring and information retrieval**

E.g., DARPA GALE program

1. ASR, MT
2. Broad domain coverage, formal languages
3. Large amount of training corpus
  - Hundreds of hours speech, hundreds of millions of words in training data
4. Rich computation resources
  - Servers, supercomputers
5. Allow response delay: minutes
6. Not dialog systems
7. Typical applications: intelligence, media companies

IBM TALES project

- **Speech-to-speech: for cross-lingual communication**

E.g. DARPA TransTac program

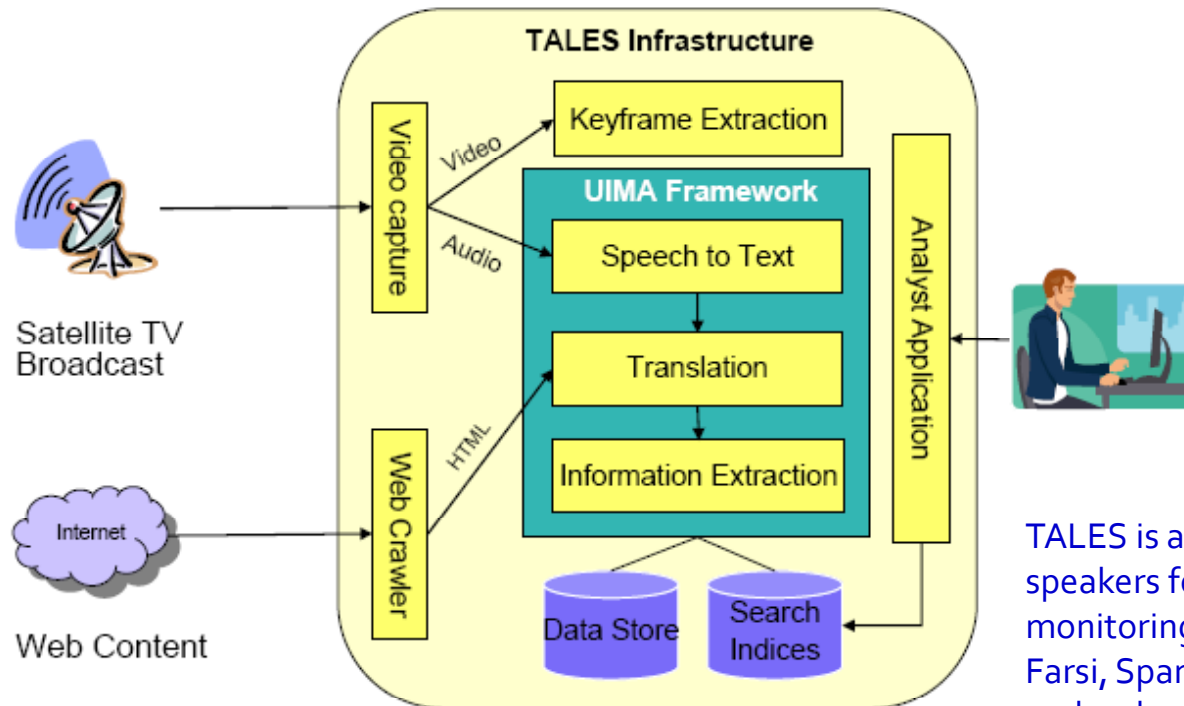
1. ASR, MT, TTS
2. Relative narrow domain coverage, conversational colloquial languages
3. Often have to deal with low resource languages and rapid development for such new languages
  - Much less data available
4. Very limited computation resources
  - Laptops, PDAs
5. Need real-time
6. Interactive dialog systems: allow repeats, confirmation
7. military, law enforcement, hospitals, business travelers, service industry

IBM MASTOR project



# IBM TALES (Translingual Automatic Language Exploitation System) Project (1/2)

## TALES Base Capabilities



TALES is a IBM solution for English speakers for global news monitoring for Arabic, Chinese, Farsi, Spanish, and English video and web news sources.

### Key Technologies

- Speech recognition
- Statistical machine translation
- Information extraction (Named entity and relationship detection - optional)
- UIMA framework
- OmniFind search engine

[http://domino.research.ibm.com/comm/research\\_projects.nsf/pages/tales.index.html](http://domino.research.ibm.com/comm/research_projects.nsf/pages/tales.index.html)



# IBM TALES (Translingual Automatic Language Exploitation System) Project (2/2)

## TALES Demo

### Foreign Broadcast Video Monitoring and Search



UIMA-based multi-lingual search technology:

- Speech-to-Text
- Machine Translation (English, Arabic, Chinese, Spanish)
- Advanced Text Analysis (language identification and translation, named entity extraction and translation)
- Cross-lingual Information Retrieval

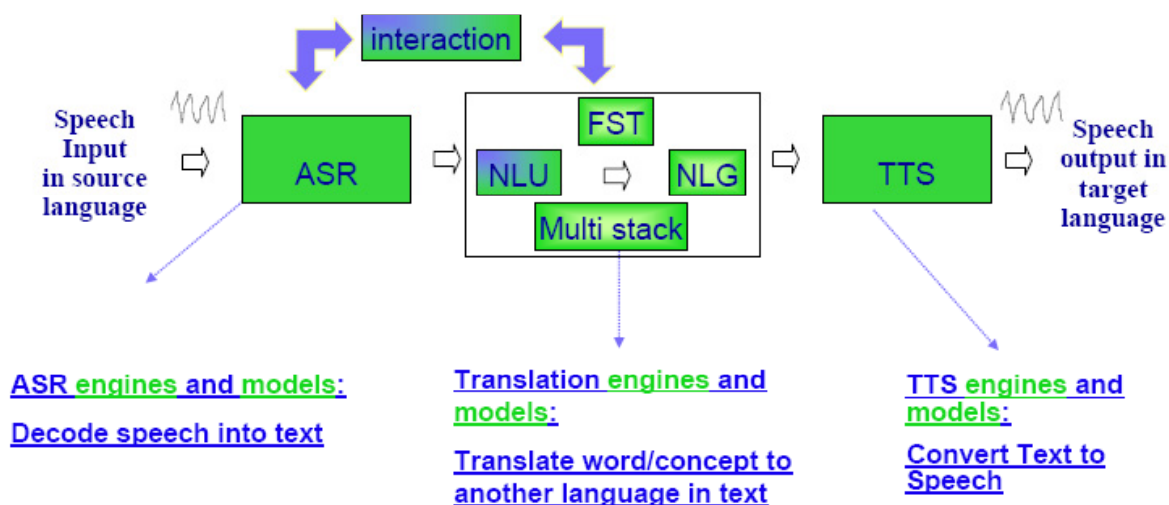
### Foreign Web Site Translation and Search



# IBM Mastor (Speech-to-Speech Translation) Project (1/2)

- MASTOR is a two-way, free form speech translator that assists human communication using natural spoken language for people who do not share a common language

## IBM Advanced Speech-to-Speech Translation Techniques

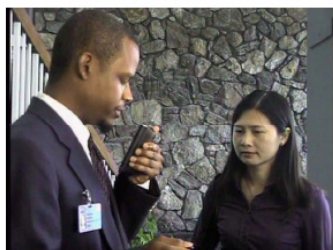
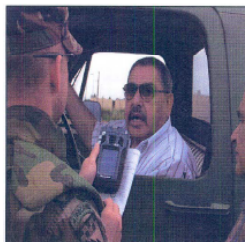


# IBM Mastor (Speech-to-Speech Translation) Project (2/2)

Laptop systems  
- hands-free, eyes-free function



Handheld System

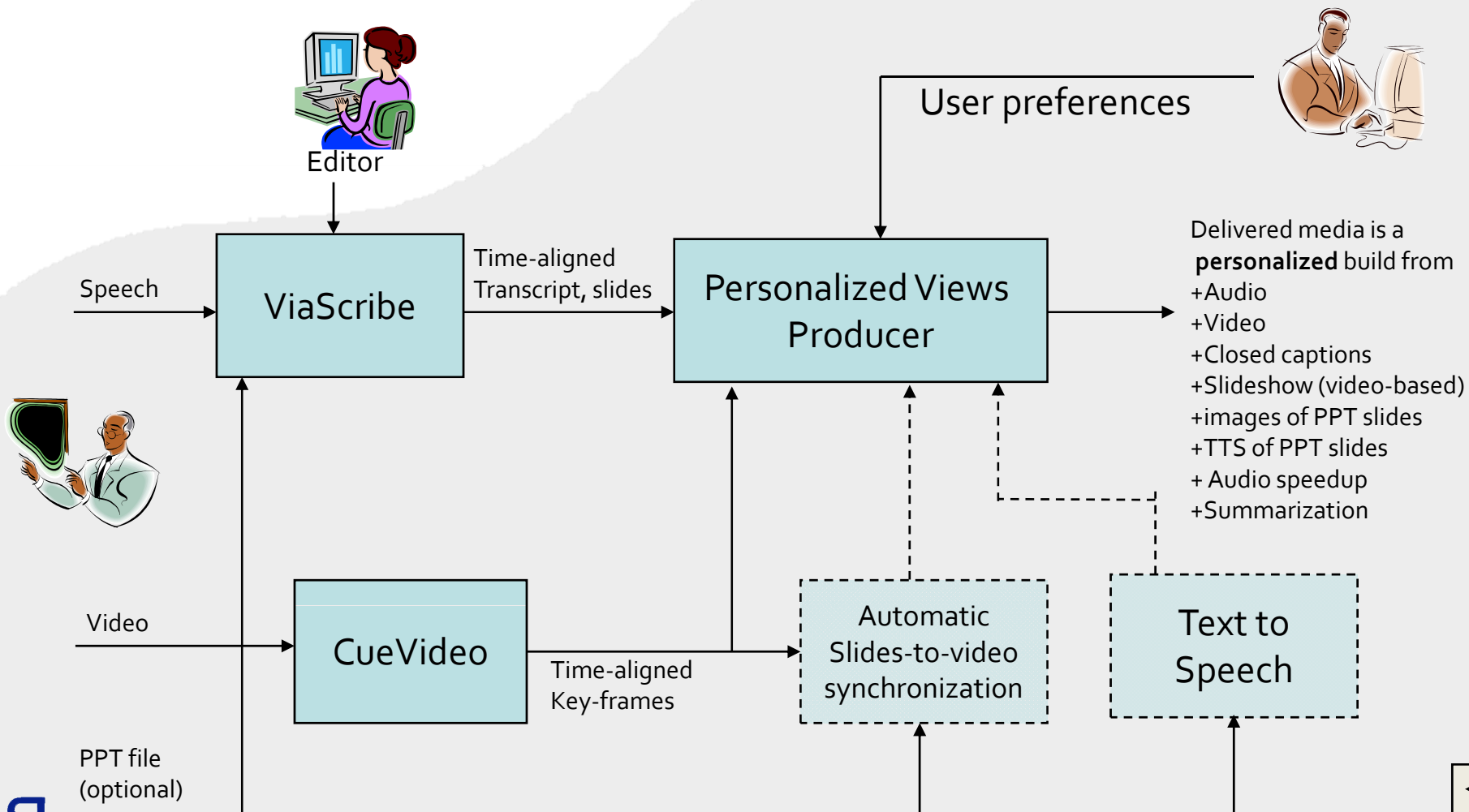


MASTOR Demo



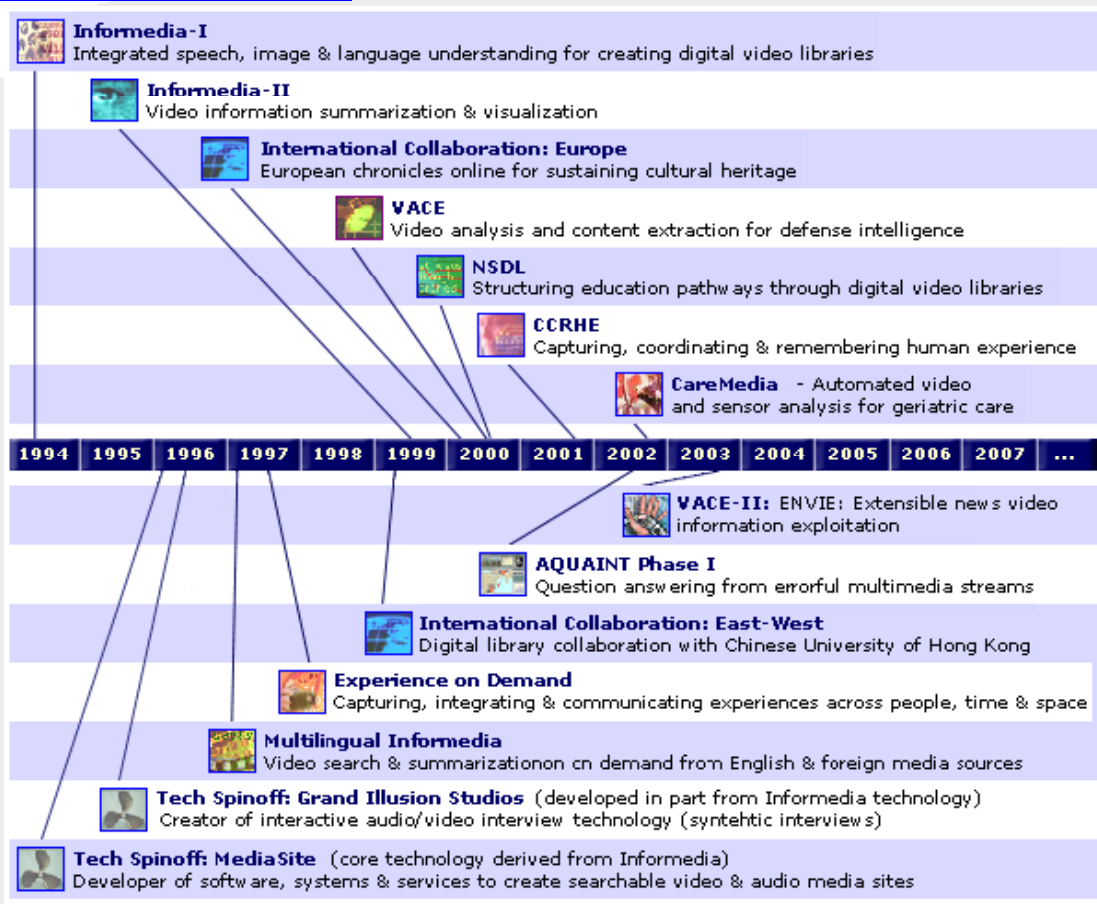
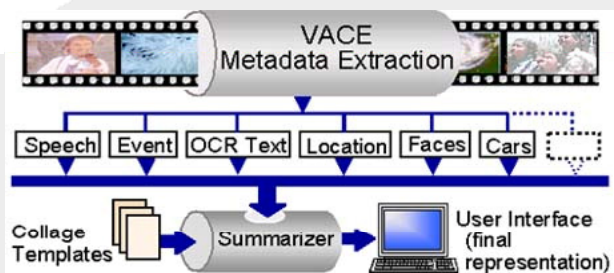


# IBM's Audio-Visual Search Solutions



# The Informedia System at CMU

- Video Analysis and Content Extraction (VACE)
  - <http://www.informedia.cs.cmu.edu/>



# AT&T SCAN System

SCAN - Speech Content Based Audio Navigator

File Search Scan

QUERY:  SEARCH CLEAR

RESULTS - "What is the status of the trade deficit with Japan"

RANK	PROGRAM	DATE	STORY	SCORE	LENGTH	HITS
1	NPR All Things Considered	05/31	3	15.63	27.65	6
2	NPR All Things Considered	05/10	15	13.89	512.42	16
3	NPR/PRI Marketplace	06/14	4	13.82	166.40	14
4	ABC World News Now	06/13	6	13.44	30.00	3
5	NPR All Things Considered	05/21	4	11.14	13.62	3
6	NPR All Things Considered	05/31	3	10.92	17.02	3
7	NPR/PRI Marketplace	06/14	3	10.87	30.00	4
8	CNN Headline News	06/07	18	9.83	183.55	6
9	NPR/PRI Marketplace	06/11	23	9.82	203.21	11
10	NPR/PRI Marketplace	06/14	6	9.41	90.33	4

Prev Doc Next Doc

OVERVIEW - NPR All Things Considered 05/10

deficit  
status  
japan  
trade

ASR TRANSCRIPTS - NPR All Things Considered 05/10

"expanding defense cooperation span is a part of our pacific democracy defense program will strengthen are lines and serve on mutual interest that while president clinton is earth credit for renewing inspecting those ties on his recent trip the administration's amateurs and in a factory posturing on trade disputes"

"buster and those ties and assess state of the president's recent attempt of damage control in nineteen ninety four that lead administration for both a trade war and lost and then declared victory even though present but received nothing the clinton a station shows funk war dead and then contradictory tactics"

"did not work for the force camp and saving deregulation competition and economic reform the result has been an increase in both the bilateral trade deficit and japanese trade nationalism the merchandise trade that has no sacred is anthony here no but i do not agree with president clinton's decision"

"the normal eyes relations with vietnam until they could have and should receive more returned from vietnam the decision has been made the case is not closed there are many outstanding issues in our relationship with vietnam was shared economic and other enters can only be realized"

"after the outcome achieved fullest possible accounting for a missing servicemen and vietnam must understand that further progress on the field of the a. m. i. a. issue remain are biased bilateral priority now it is simply that i think we all saw to be very forthright flat out but i have fun"

"that out neo from about are commercial relations with china was incredible is right the nineteen ninety four when a funny decided extension of most favored nation status was the best way to promote are long term interest in china"

Selection Length:  seconds Stop Audio

AT&T Labs Research

Design and evaluate user interfaces to support retrieval from speech archives



# BBN Rough'n'Ready System

## Automatic Structural Summarization for Broadcast News

### Distinguished Architecture for Audio Indexing and Retrieval

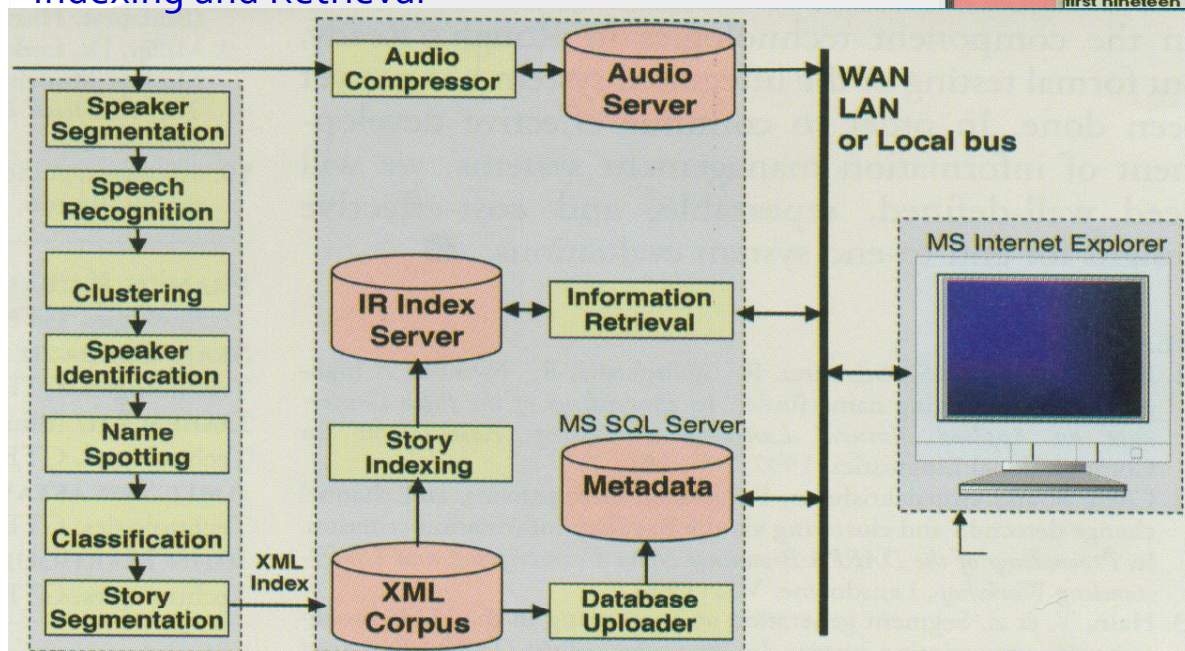
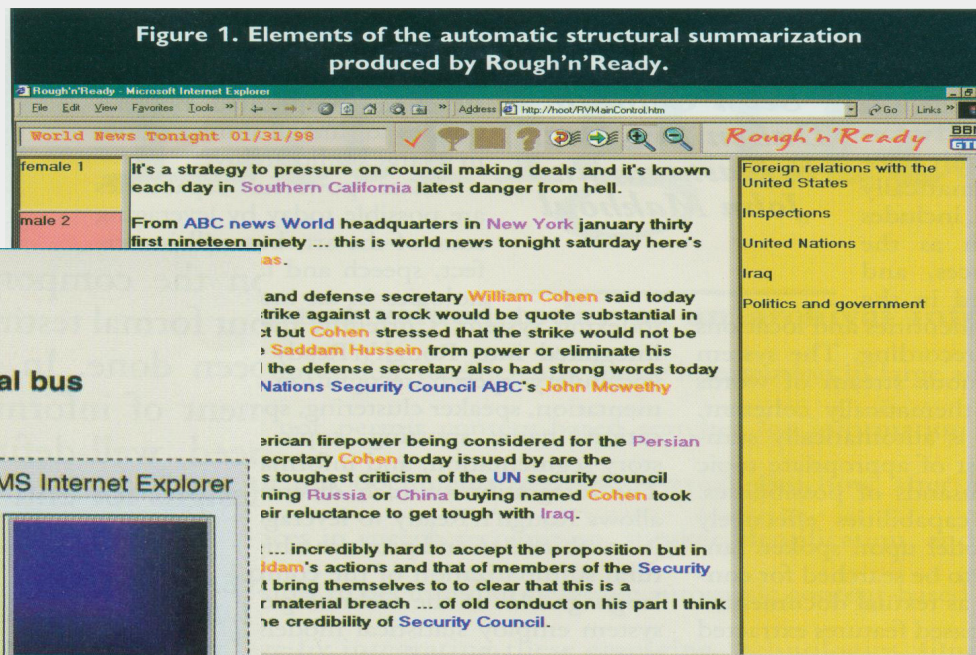


Figure 1. Elements of the automatic structural summarization produced by Rough'n'Ready.



# Google Voice Search

## Google Audio Indexing: Searching what people are saying inside YouTube videos (currently only for what the politicians are saying)

## Google Voice Local Search



Dial from any phone  
**1-800-GOOG-411**  
(1-800-466-4411)

**About GOOG-411**  
Google's new 411 service is free, fast and easy to use. Give it a try now and see how simple it is to find and connect with local businesses for free.

[Learn more](#) - [FAQ](#)

**Liked the video?** Want to comment or guess who the voice of GOOG-411 is? Post your opinion on our [YouTube page](#).

<p>1 Dial 1-800-GOOG-411 from any phone</p>	<p>2 State the location and business type</p>	<p>3 Connect to the business for free</p>	<p>4 Done!</p>
---	---	---	----------------

©2007 Google - [Terms of Service](#) - [Privacy Policy](#) - [Google Home](#) - [Mobile Home](#)



<http://labs.google.com/>



# Microsoft Research Audio-Video Indexing System (MAVIS)

- MAVIS uses speech recognition technology to index spoken content of recorded conversations, like meetings, conference calls, voice mails, lectures, Internet videos

The screenshot displays the 'Audio Search' interface. At the top, a search bar contains the query 'virtualization hypervisor'. Below the search bar, the results are listed under 'Audio Search Results'. The first result is 'The Next Server Wave [MS StudioCasts]' by Bill Laing, discussing ISV implications of the Windows Server 2008 release. The second result is 'DTTS-Feb 2007-Tech Edition-Longhorn-Technical Overview [Academy Mobile Podcasts]'. A video player is visible on the right side of the interface, showing a man speaking. Annotations with arrows point to various parts of the interface: 'Audio search query' points to the search bar; 'Metadata Hits' points to the search results; 'Audio Hits' points to the text of the search results; 'Click-to-play snippets navigate directly into video' points to the video player; and 'Fast search speeds' points to the search time indicator at the bottom.



# MIT Lecture Browser

- Retrieval and browsing of academic lectures of various categories

MIT Lecture Browser - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

MIT Lecture Browser

**Lecture Browser**  
SPOKEN LECTURE PROCESSING  
CSAIL MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Search for words: and/or pick a category:  
hacks AND pumpkin Any category Search

Examples: saxophone, "Beatrix Potter", shack AND dome

1 result for hacks AND pumpkin

**1. Where the Sun Shines, There Hack They**  
October 20, 2005 (Samuel Jay Keyser) 1:00:42

▶ that's terrific anybody else who who into jets if i make a mistake by all means please too because i don't really i'm not a historian of hacks on just sort of a gentle observer ... here is close up ... that is on the way down

▶ an attraction for hackers and this is the jacket latin on the dome ... and let's see that one was the great pumpkin charlie brown and that was in october of nineteen sixty two ... that was anonymous here it's

▶ ones that's another made up name it you remember was a you know that that i can i keep apart ... those hacks done by the t h a and those hacks that would invite people to remain anonymous and were sort of you know sort of even sub rows even suburb rows person bows are her in in a way and however domes

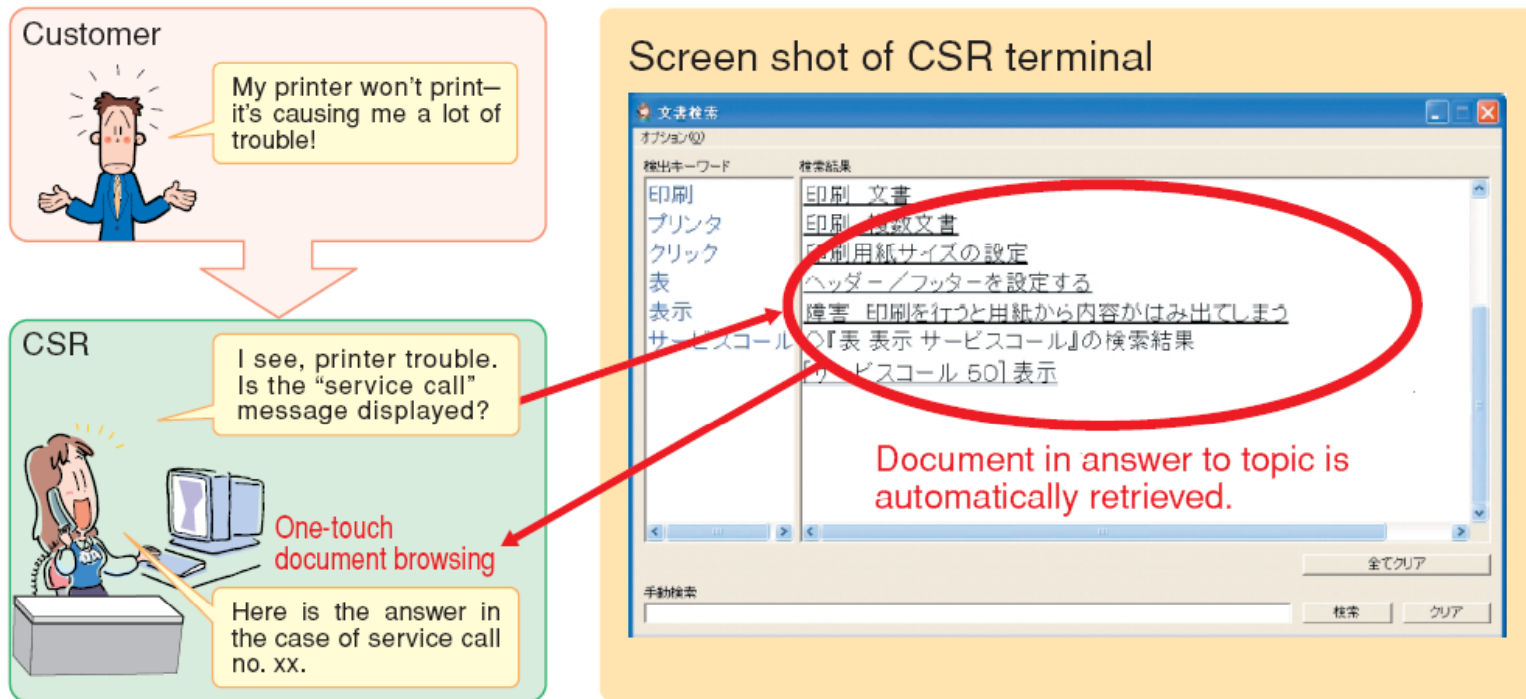
▶ the best tax for last because ... be presenting them ... the chronicle did

Info

for hackers and this is the jacket latin on the dome ... and let's see that one was the great pumpkin charlie brown and that was in october of nineteen sixty two ... that was anonymous here it's ... that the that the similarly i decorated dome this is out on the french ... mass avenue notice the stew are anonymous ... and so they the the hackers started ago on the ground ... here this is my this was actually nineteen sixty two all this when i'm not sure of does anybody know the first one was i t sixty two ... there again this is on the green building let's see i may have of the day for that know i guess i don't i guess it down as i said the dome was a great attraction and this is you can't see it in quite but you

# NTT Speech Communication Technology for Contact Centers

## Automatic document-retrieval by speech recognition

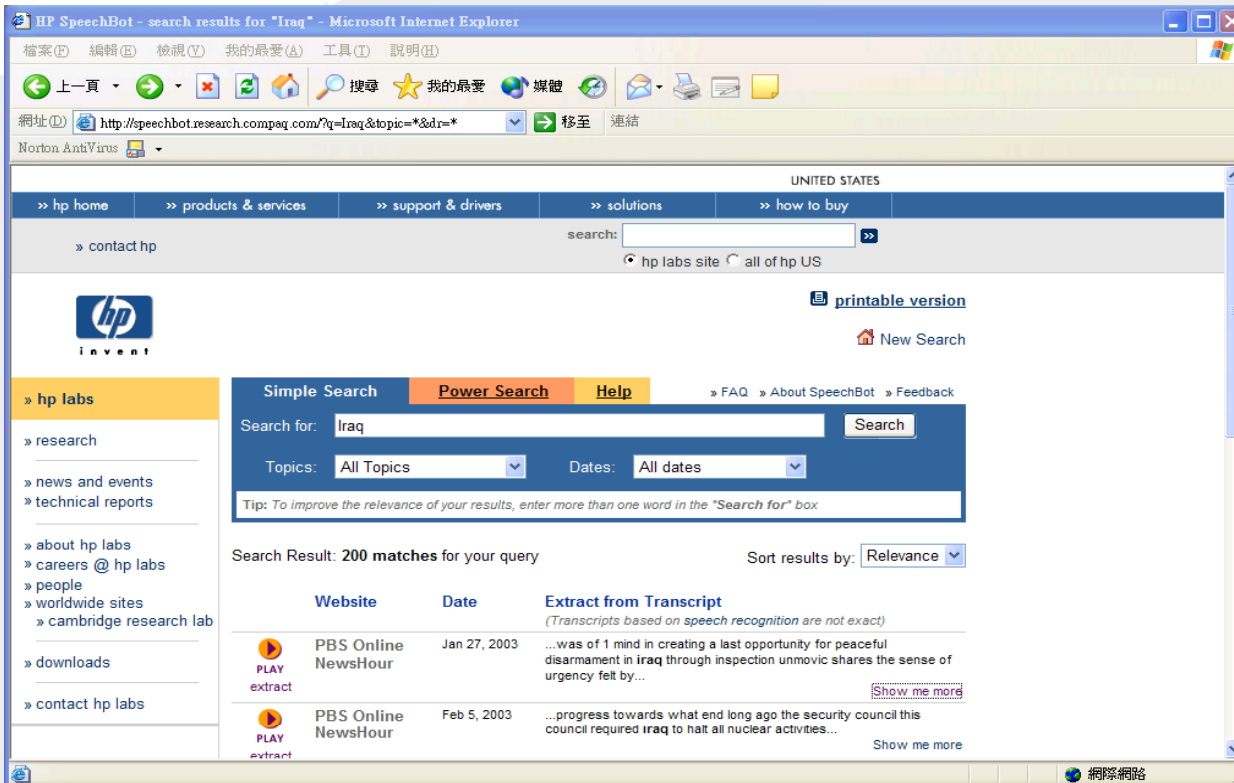


– CSR: Customer Service Representative





# SpeechBot Audio/Video Search System at HP Labs

- An experimental Web-based tool from HP Labs that used voice-recognition to create searchable keyword transcripts from thousands of hours of audio content



The screenshot shows a Microsoft Internet Explorer browser window displaying the HP SpeechBot search results for the keyword "Iraq". The browser's address bar shows the URL: [http://speechbot.research.compaq.com/?q=Iraq&topic=&id=\\*](http://speechbot.research.compaq.com/?q=Iraq&topic=&id=*). The page features a navigation menu with links to "hp home", "products & services", "support & drivers", "solutions", and "how to buy". A search bar is visible with the text "search:" and a search button. Below the search bar, there are options for "hp labs site" and "all of hp US". The main content area displays the HP logo and a search interface with tabs for "Simple Search", "Power Search", and "Help". The search results show 200 matches for the query "Iraq", sorted by relevance. The results are presented in a table with columns for "Website", "Date", and "Extract from Transcript".

Website	Date	Extract from Transcript
 PBS Online NewsHour	Jan 27, 2003	...was of 1 mind in creating a last opportunity for peaceful disarmament in Iraq through inspection unmovic shares the sense of urgency felt by...
 PBS Online NewsHour	Feb 5, 2003	...progress towards what end long ago the security council this council required Iraq to halt all nuclear activities...

# Some Prototype Systems Developed in Taiwan

NTU Broadcast News Retrieval and Browsing System  
(Prof. Lin-shan Lee), 2004~

NTNU PDA Broadcast News Retrieval System  
(Dr. Berlin Chen), 2003~2004

**廣播新聞搜尋瀏覽系統**  
Broadcast News Retrieval/Browsing System

國外政治 [International Political News] Topic Map  
國內政治 [Local Political News] Topic Map  
國外財經 [International Business] Topic Map  
國內財經 [Local Business] Topic Map  
國外影劇 [International Entertainment] Topic Map  
國內影劇 [Local Entertainment] Topic Map  
國外體育 [International Sports] Topic Map  
國內體育 [Local Sports] Topic Map

1 以色列結束對阿拉法特總部的包圍 [sum.] 02.09.21  
2 阿拉法特反對以色列保所提結束包圍條件 [sum.] 02.09.21  
3 以色列部隊進攻阿拉法特總部復撤軍 [sum.] 02.10.22  
4 以色列結束對阿拉法特總部的包圍 [sum.] 02.10.01  
5 以色列坦克撤出阿拉法特辦公室 [sum.] 02.09.21  
6 以色列與巴勒斯坦展開安全問題會議 [sum.] 02.11.23  
7 以色列在加薩擊斃一名回教聖戰組織領袖 [sum.] 02.06.05  
8 以色列巴勒斯坦領袖伯拉撤軍達成協議 [sum.] 02.02.12  
9 以色列坦克闖入加薩難民營 兩人喪生 [sum.] 02.04.21

阿拉法特 阿巴斯 以色列 夏隆  
雷馬拉 任命 約旦河 美國  
中東 鮑爾 和平 路線  
巴格達 炸彈 自殺 巴士

阿拉法特原則接受歐盟所提中東和平計畫 [summary] (May 03/02/12:00)  
英美就解決阿拉法特所受包圍與巴方展開談判 [summary] (May 06/02/12:00)  
阿拉法特反對以色列保所提結束包圍條件 [summary] (Sep 20/02/12:00)  
阿拉法特宣布新內閣引發巴勒斯坦國會激辯 [summary] (Oct 30/02/12:00)  
阿拉伯人支持阿拉法特及巴勒斯坦人正當抵抗 [summary] (Nov 02/02/12:00)

Browser 11:16 ok

中文廣播新聞檢索系統 國立台灣師範大學資工所

錄音鍵

辨識結果 美國總統大選 搜尋

摘要

040304-13.兩千年美國總統大選時  
021216-24.二零零年總統大選時高爾以些  
040309-10.把總統到訪當成將領希望帶  
021210-23.因此如果國親兩黨有任何一個

全文

關心美國總統大選消息美國北卡羅來納州參議員愛德華茲間正式宣布退出民主黨總統候選人初選並表示將全力協助麻州參議員凱瑞期待美國總統布希而儘管美國十一月

新聞影音播放



## Appendix A: Actual Term Weighted Value (2/2)

- Actual Term Weighted Value (ATWV) is a metric defined in the NIST Spoken Detection (STD) 2006 evaluation plan

$$\text{ATWV} = 1 - \frac{1}{Q} \sum_{q=1}^Q \{P_{\text{miss}}(q) + \beta P_{\text{FA}}(q)\}$$

$$P_{\text{miss}}(q) = 1 - \frac{C(q)}{R(q)} \quad P_{\text{FA}}(q) = \frac{A(q) - C(q)}{n_{\text{tps}} \times T_{\text{speech}} - C(q)}$$

$T_{\text{speech}}$  = duration of speech (in sec.)

$n_{\text{tps}}$  = number of trials per sec. of speech

$R(q)$  = total number of times examples of a specific term (phrase)  $q$  actually appears

$C(q)$  = total number of times examples of a specific term (phrase)  $q$  detected  
by the system that are actually correct

$A(q)$  = total number of times examples of a specific term (phrase)  $q$  detected  
by the system

$\beta$  : empirically set parameter (e.g., 1000)

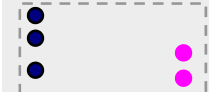
## Appendix A: Mean Average Precision ( $mAP$ ) (1/2)

- Average Precision at Seen Relevant Documents
  - A single value summary of the ranking by averaging the precision figures obtained after each new relevant doc is observed

1. $d_{123}$ • ( $P=1.0$ )	6. $d_9$ • ( $P=0.5$ )	11. $d_{38}$
2. $d_{84}$	7. $d_{511}$	12. $d_{48}$
3. $d_{56}$ • ( $P=0.66$ )	8. $d_{129}$	13. $d_{250}$
4. $d_6$	9. $d_{187}$	14. $d_{113}$
5. $d_8$	10. $d_{25}$ • ( $P=0.4$ )	15. $d_3$ • ( $P=0.3$ )

$(1.0+0.66+0.5+0.4+0.3)/5=0.57$

alg1    alg2



Cutoff

It favors systems which retrieve relevant docs quickly (early in the ranking)

But when doc cutoff levels were used

- An algorithm might present a good average precision at seen relevant docs but have a poor performance in terms of overall recall



## Appendix A: Mean Average Precision (*mAP*) (2/2)

- Averaged at relevant docs and across queries
  - E.g. relevant docs ranked at 1, 5, 10, precisions are  $1/1, 2/5, 3/10,$ 
    - **non-interpolated average precision** (or called Average Precision at Seen Relevant Documents in textbook)  $= (1/1 + 2/5 + 3/10) / 3$
  - Mean average Precision (*mAP*)

$$\frac{1}{|Q|} \sum_{q=1}^{|Q|} (\text{non - interpolated average precision})_q$$

- **Widely used in IR performance evaluation**

## Appendix A: Word Error Rate (WER) (2/2)

- The speech recognition experiments are usually evaluated in terms of word error rate (WER)

$$\text{WER} = \frac{\textit{Ins} + \textit{Sub} + \textit{Del}}{\textit{Ref}}$$

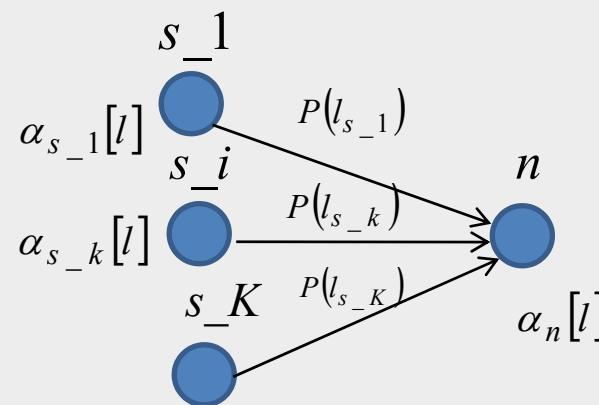
- Defined by the sum of the insertion (*Ins*), deletion (*Del*), and substitution (*Sub*) errors between the recognized and reference word strings, divided by the total number of words in the reference string (*Ref*)

# Position-Specific Posterior Probability Lattices (1/6)

- **Soft-hit:** indexing of the occurrence of each word  $n$  in the lattice

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

position/length along the partial path traversed



- A modified forward procedure

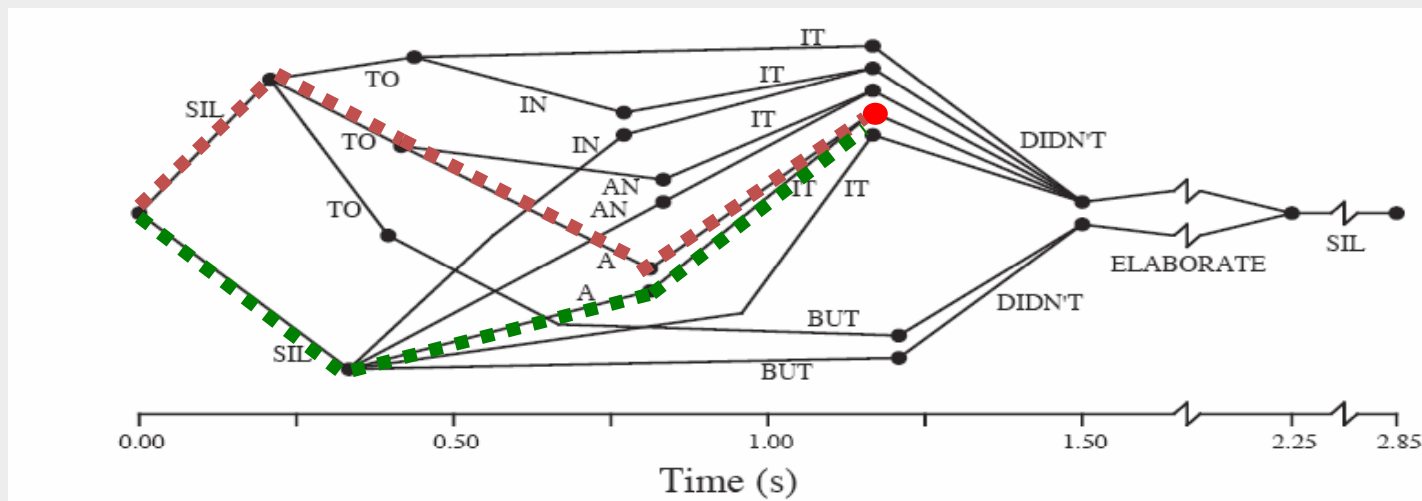
$$\alpha_n[l+1] = \sum_{i=1}^q \alpha_{s_i}[l + \delta(l_{s_i}, \varepsilon)] P(l_{s_i})$$

$$\log P(l_{s_i}) = \eta \cdot \left( \frac{1}{\kappa} \log P_{AM}(l_{s_i}) + \log P_{LM}(l_{s_i}) - \frac{1}{\kappa} \log P_{IP} \right)$$

## Position-Specific Posterior Probability Lattices (2/6)

- The backward procedure follows the original definition
- The posterior probability for a word  $w$  at position  $l$  is expressed as (i.e., expected counts of  $w$  at position  $l$ )

$$P(w, l | LAT) = \sum_{n \text{ s.t. } \alpha_n[l] \cdot \beta_n > 0} \frac{\alpha_n[l] \cdot \beta_n}{\beta_{start}} \delta(w, word(n))$$





## Position-Specific Posterior Probability Lattices (3/6)

- A document  $D$  can be first divided into several segments
- Then, calculate the expected count of a given query term according to the PSPL probability distribution for each segment  $s$  of document  $D$

Query  $Q = q_1, q_2, \dots, q_M$

unigram matching

$$S(D, q_i) = \log \left[ 1 + \sum_s \sum_l P(w_l(s) = q_i | D) \right]$$

$$S_{1\text{-gram}}(D, Q) = \sum_{i=1}^M S(D, q_i)$$

$P(w, l | LAT)$

$N$ -gram (or phrase) matching

$$S(D, q_i \dots q_{i+N-1}) = \log \left[ 1 + \sum_s \sum_l \prod_{r=0}^{N-1} P(w_{l+r}(s) = q_{i+r} | D) \right]$$

$$S_{N\text{-gram}}(D, Q) = \sum_{i=1}^{M-N+1} S(D, q_i \dots q_{i+N-1})$$

$$S(D, Q) = \sum_{N=1}^M \lambda_N \cdot S_{N\text{-gram}}(D, Q)$$

## Position-Specific Posterior Probability Lattices (4/6)

- “Relative Pruning” of PSPL lattices
  - For a given position bin  $l$ , the relative pruning first finds the most likely word entry given by

$$w_l^* = \arg \max_{w \in V} p(w_l(s) = w | D)$$

- Word entries have test values lower than or equal to the threshold are retained in the position bin of the PSPL lattice

$$W_l = \{w \in V : \log \frac{P(w_l(s) = w_l^* | D)}{P(w_l(s) = w | D)} \leq \tau_r\} \quad \tau_r \in [0, \infty)$$

- As the threshold decreased to zero, the pruned PSPL is reduced “approximately” to the 1-best output
- Then, the posterior probability of words (bin entries)  $W_l$  in each bin are renormalized

## Position-Specific Posterior Probability Lattices (5/6)

- “Absolute Pruning” of PSPL lattices
  - Retrain the word entries in each bin  $l$  that have log posterior probability higher than an absolute threshold

$$\bar{P}(w_l(s) = q|D) = P(w_l(s) = q|D) \cdot 1_{\{\log P(w_l(s) = q|D) \geq \tau_{abs}\}}$$

$$\tau_{abs} \in (-\infty, 0]$$

“Absolute Pruning” can be performed at query run-time

# Position-Specific Posterior Probability Lattices (6/6)

- Corpus: MIT *i*Campus Corpus (169 h, recorded using lapel microphone)
  - 116 test text queries (Q-OOV rate: 5.2%; avg. query length: 1.97 words)

Table 3

Retrieval performance on indexes built from PSPL lattices under various PSPL probability assignments

	lat	raw	noP	unif	1-best
# docs retrieved	4971	4971	4971	4971	3206
# relevant docs	1416	1416	1416	1416	1416
# rel retrieved	1301	1301	1301	1301	1088
MAP	0.62	0.60	0.47	0.57	0.53
R-precision	0.58	0.56	0.42	0.52	0.53

without flattening of word prob.

$$\log P(l_{s\_i}) = \eta \cdot \left( \frac{1}{\kappa} \log P_{AM}(l_{s\_i}) + \log P_{LM}(l_{s\_i}) - \frac{1}{\kappa} \log P_{IP} \right)$$

without using posterior prob.  
(hard-index, more than one word occurs at the same position)

Uniform posterior prob.  
1.0/#entries in each position