

Latent Topic Modeling of Word Co-Occurrence Information for Spoken Document Retrieval

Berlin Chen



Department of Computer Science & Information Engineering
National Taiwan Normal University, Taiwan



2009/05/04

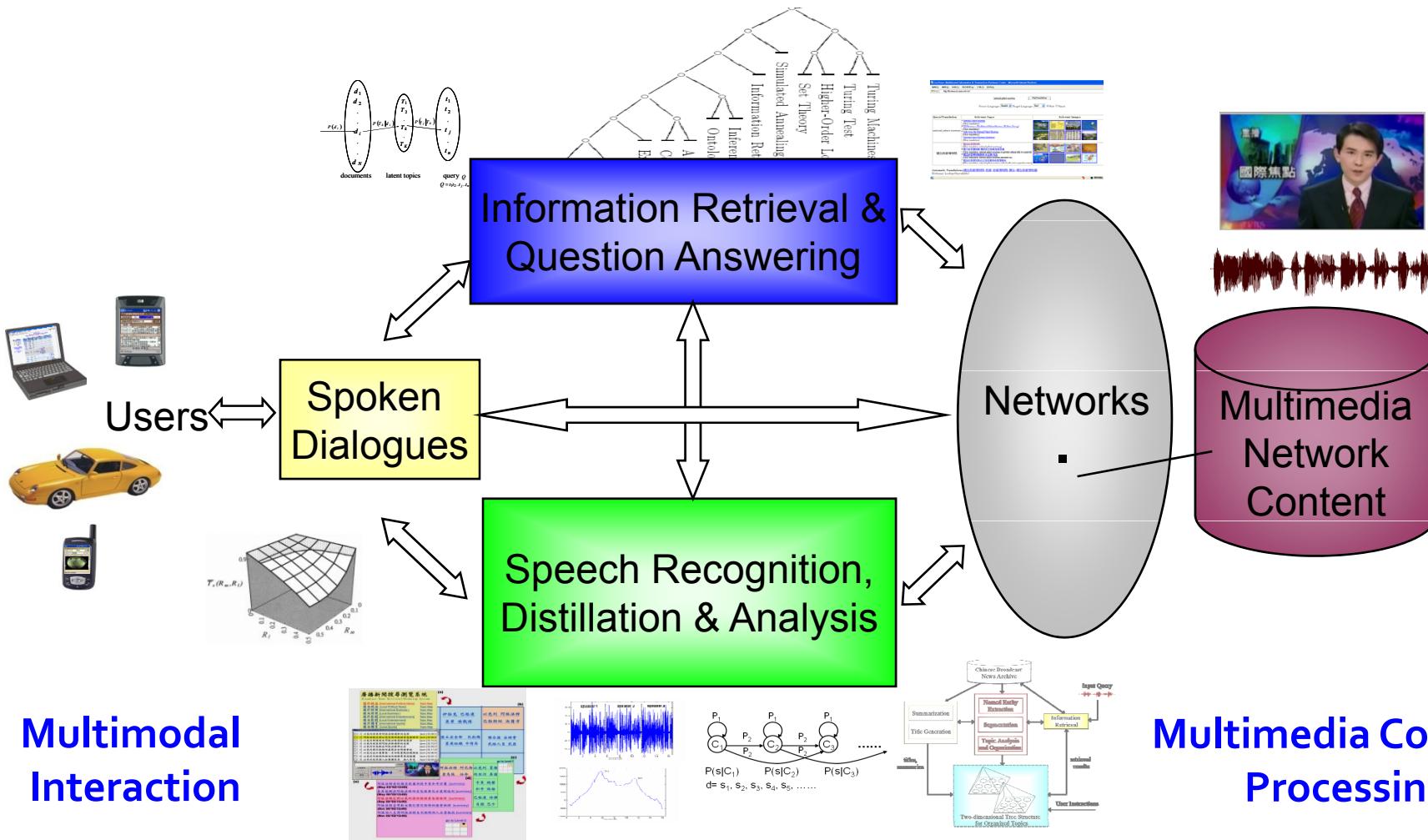
Outline

- Introduction
- Document Topic Models (DTM)
- Word Topic Model (WTM)
- Comparisons and Experiments on SDR
- Applications of WTM to Other Related Tasks
- Conclusions

Introduction

- Large volumes of multimedia associated with speech are now made available on the Internet
 - Speech is one of the most semantic (or information)-bearing sources
- Speech also is the most convenient means of communication between humans, or potentially between humans and machines
 - 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

Multimodal Access to Multimedia in the Future



Multimedia vs. Text

- 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 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

Related Research Work and Applications

- Continuous and substantial efforts have been paid to (multimedia) speech recognition, summarization and retrieval in the recent past
 - Informedia System at Carnegie Mellon Univ. ➔
 - AT&T SCAN System ➔
 - The *Rough'n'Ready* System at BBN Technologies ➔
 - The *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 ➔

Key Technologies (1/2)

- Automatic Speech Recognition (ASR)
 - Automatically convert speech signals into sequences of words or other suitable units for further processing
- Spoken Document Segmentation
 - Automatically segment speech signals (or automatically transcribed word sequences) into a set of documents (or short paragraphs) each of which has a central topic
- Named Entity Extraction from Spoken Documents
 - Personal names, organization names, location names, event names
 - Very often out-of-vocabulary (OOV) words, difficult for recognition
 - E.g., “蔡煌郎”, “九二共識”, etc.
- Audio Indexing and Retrieval (Voice Search)
 - Robust representation of the spoken documents
 - Matching between (spoken) queries and spoken documents

Key Technologies (2/2)

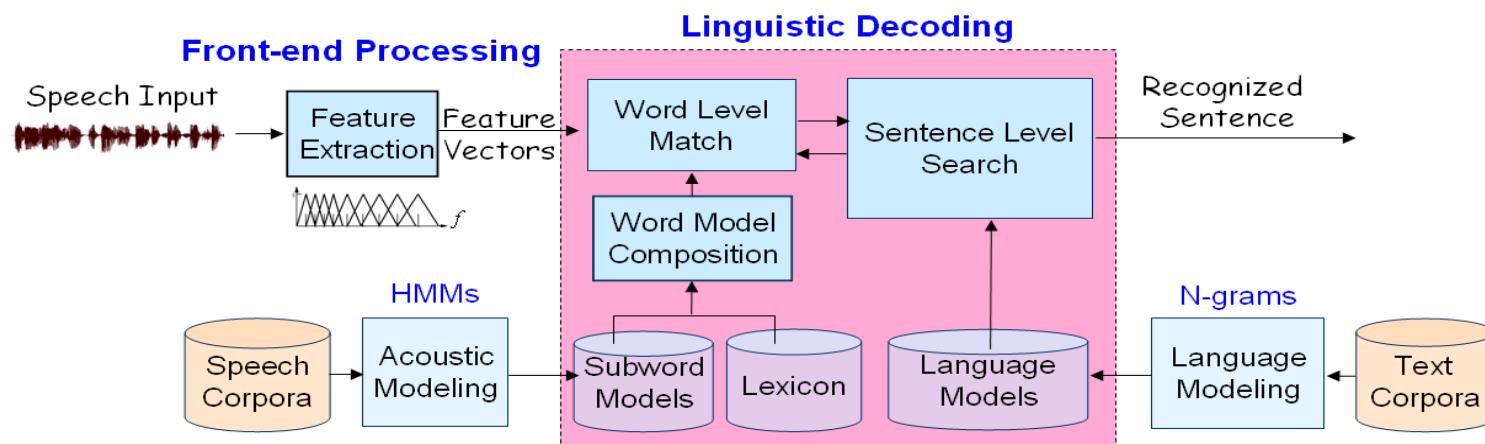
- Summarization for Spoken Documents
 - Automatically generate a summary (in text or speech form) for each spoken document or a set of topic-coherent documents
- Title Generation for Multi-media/Spoken Documents
 - Automatically generate a title (in text/speech form) for each short document; i.e., a very concise summary indicating the themes of the documents
- Topic Analysis and Organization for Spoken Documents
 - Analyze the subject topics for (retrieved) documents
 - Organize the subject topics of documents into graphic structures for efficient browsing
- Information Extraction for Spoken Documents
 - Extraction of key information such as who, when, where, what and how for the information described by spoken documents

World-wide Speech Research Projects

- There also are several research projects conducting on related spoken document processing tasks, e.g.,
 - Rich Transcription Project¹ 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 Project² (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

Automatic Speech Recognition (ASR)

- Finding the most likely word sequence W in response to an input speech signal O



$$\hat{W} = \arg \max_W P(W | O)$$

Applying Bayes' Rule

$$= \arg \max_W \frac{p(O | W)P(W)}{P(O)}$$

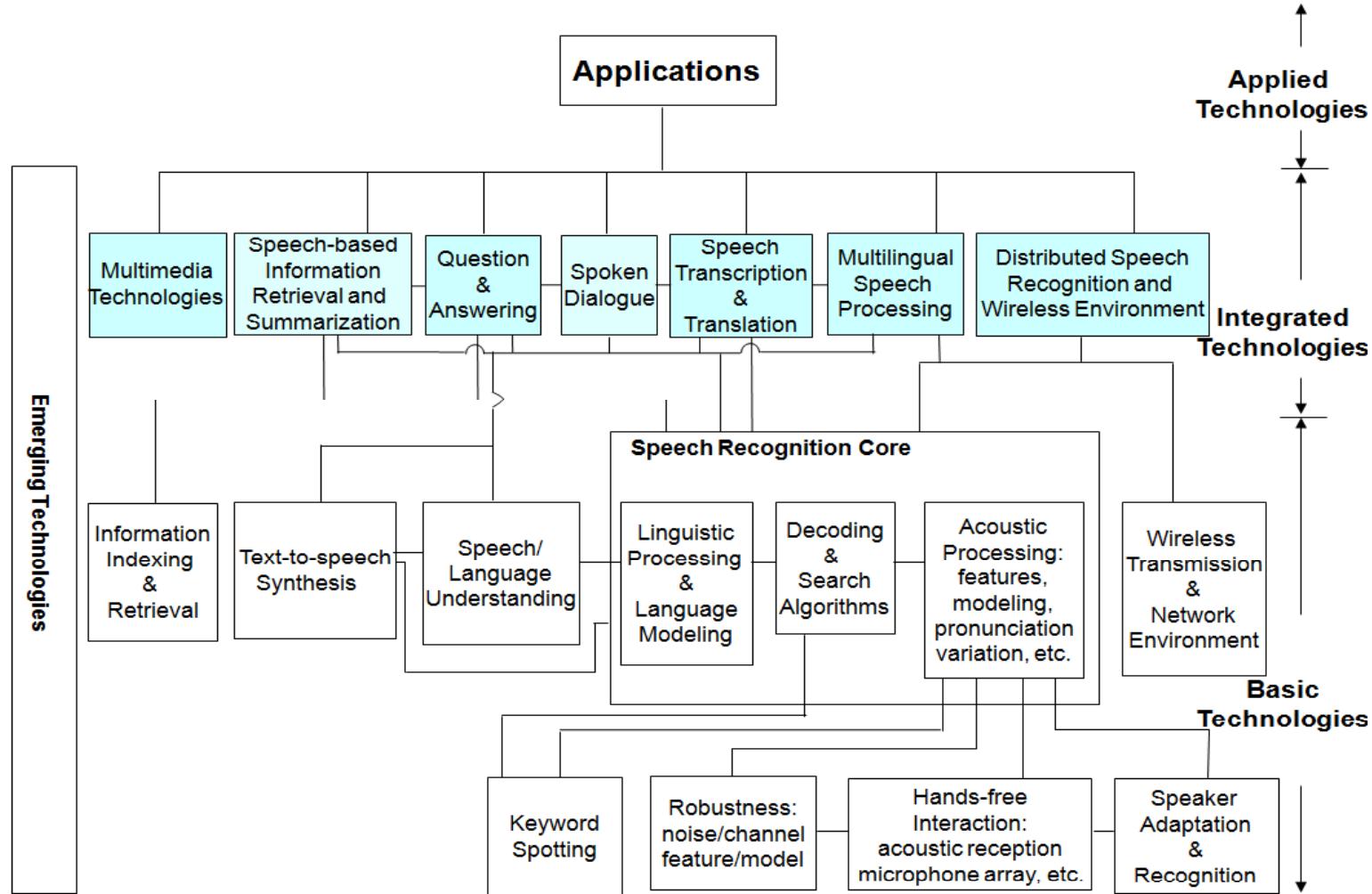
Linguistic Decoding (Search)

$$= \boxed{\arg \max_W p(O | W)P(W)}$$

Feature Extraction & Acoustic Modeling

Language Modeling

Related Research Areas of ASR

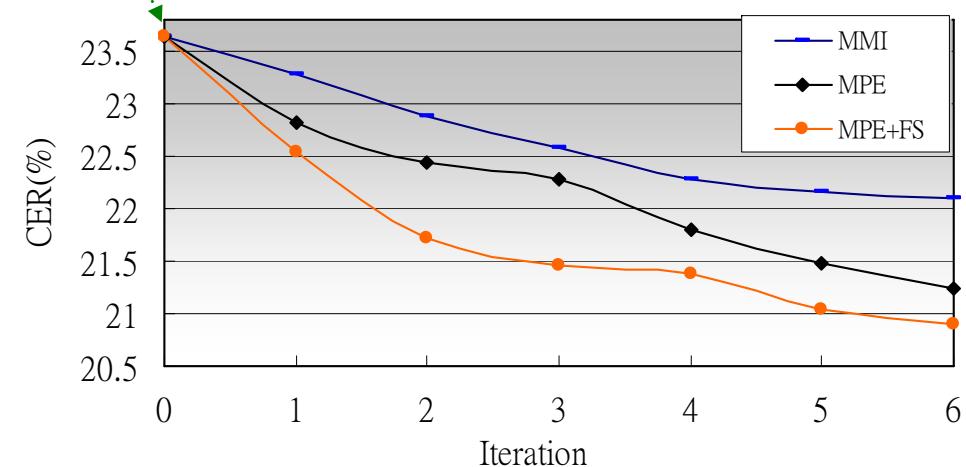


Broadcast Speech Transcription

Automatic
<p>根據最新但雨量統計 一整天下來 費雪以試辦兩個水庫的雨量 分別是五十三公里和二十九公里 對水位上升幫助不大 不過就業機會期間也多在夜間 氣象局也針對中部以北及東北部地區發佈豪雨特報 因此還是有機會增加積水區的降雨量 此外氣象局也預測 華航又有另一道鋒面通過 水利署估計如果這波鋒面能帶來跟著會差不多的雨水 那個北台灣的第二階段限水時間 渴望見到五月以後 公視新聞當時匯率採訪報導</p>

Manual
<p>根據最新的雨量統計 一整天下來 翡翠石門兩個水庫的雨量 分別是五十三公厘和二十九公厘 對水位上升幫助不大 不過由於集水區降雨多在夜間 氣象局也針對中部以北及東北部地區發布了豪雨特報 因此還是有機會增加集水區的降雨量 此外氣象局也預測 八號又有另一道鋒面通過 水利署估計如果這波鋒面能帶來跟這回差不多的雨水 那麼北台灣的第二階段限水時間 可望延到五月以後 公視新聞張玉菁陳柏諭採訪報導</p>

10 Iterations of
 ML training $F_{\text{ML}}(\Lambda) = \sum_r P_\Lambda(O_r | W_r)$



Discriminative acoustic model training can further improve the ASR performance:
 (Relative Character Error Rate Reduction)
 - MMI: 6.5% , MPE: 10.1%

$$F_{\text{MPE}}(\Lambda, \Gamma) = \sum_r \sum_{W_i \in \mathbf{W}^r} \frac{p_\Lambda(O_r | W_i) P_\Gamma(W_i) A(W_i, W_r)}{\sum_{W_k \in \mathbf{W}^r} p_\Lambda(O_r | W_k) P_\Gamma(W_k)}$$

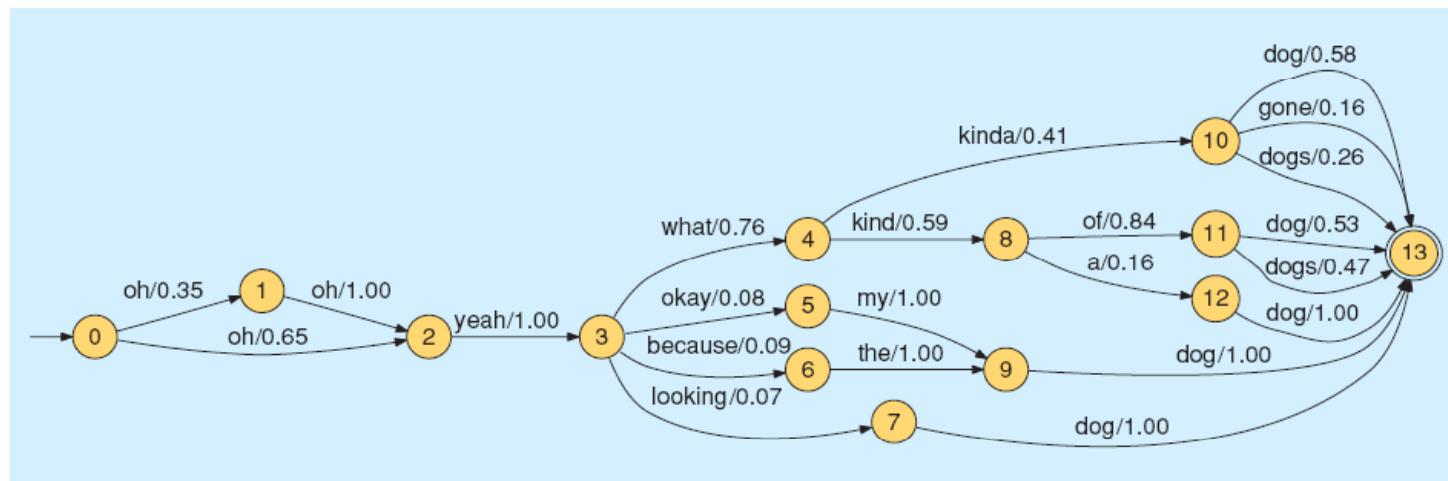
Task Definition for Voice Search

- Robustly Index spoken documents with speech recognition techniques
- Retrieve relevant spoken documents in response to a user query
 - Spoken Term Detection (STD)
 - Find “literally matched” spoken documents where all/most query terms should be present (much like Web search)
 - Spoken Document Retrieval (SDR)
 - Find spoken documents that are “topically related” to a given query

This talk focuses mainly on the document ranking models for SDR !

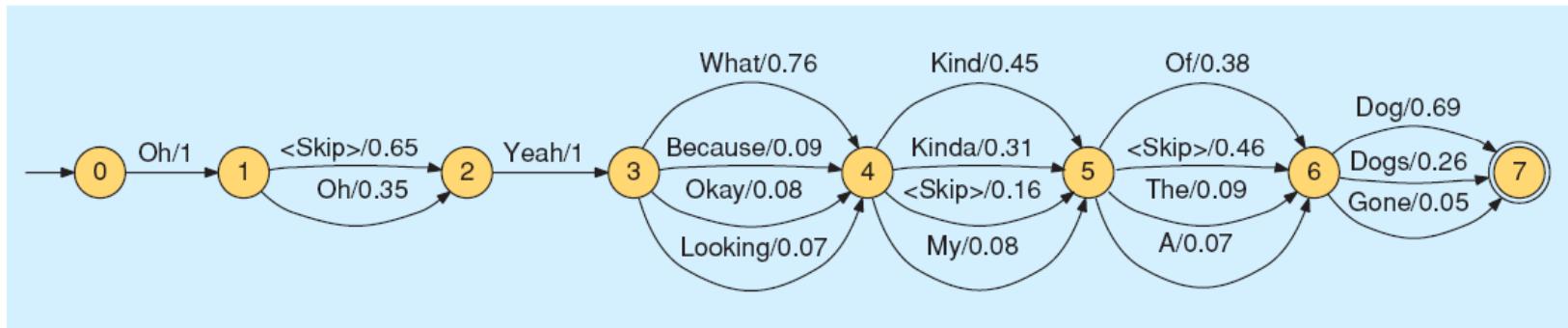
Robust Indexing: 1-bset Sequences vs. Lattices (1/3)

- 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



Robust Indexing: 1-bset Sequences vs. Lattices (2/3)

- 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



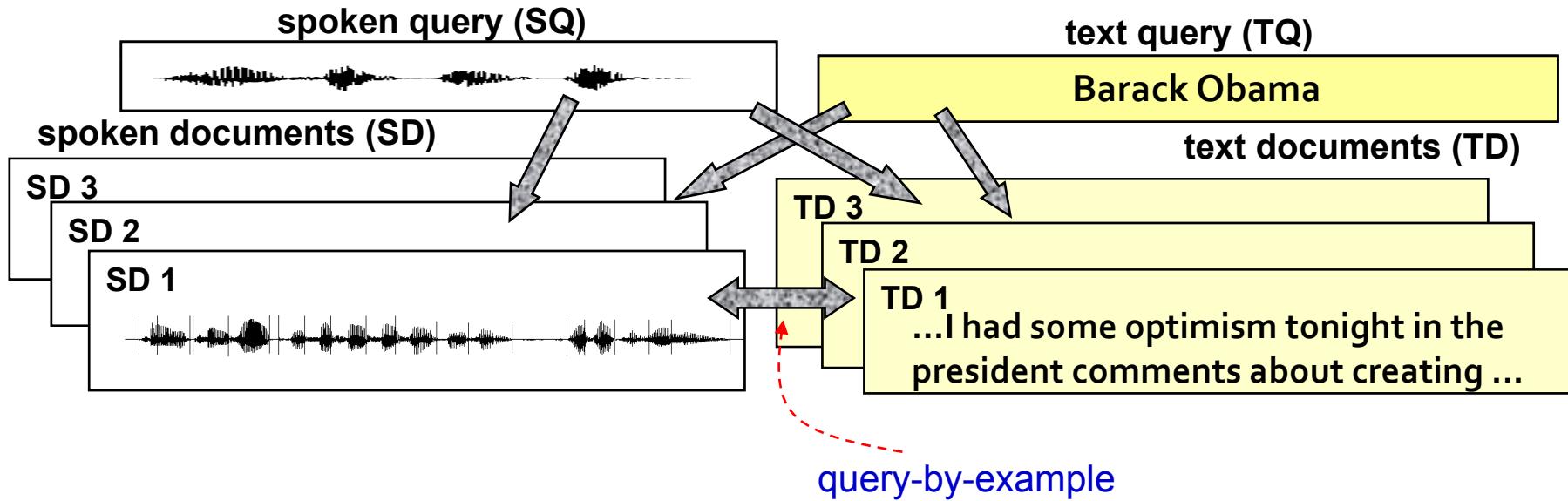
- L. Mangu, E. Brill, A. Stolcke, “Finding consensus in speech recognition: word error minimization and other applications of confusion networks,” *Computer Speech & Language* 14(4), 2000

Robust Indexing: 1-bset Sequences vs. Lattices (3/3)

- 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	Dog .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		
			—		

Scenarios for Spoken Document Retrieval (SDR)



- SQ/SD is the most difficult
- TQ/SD is studied most of the time
 - This research investigates using (Xinhua) text news to retrieve relevant (Voice of America) broadcast news
 - “query-by-example”
 - Useful for news monitoring and tracking

Language Modeling (LM) Approaches

- LM approaches have been introduced to IR (and SDR), and demonstrated with good success

$$P_{\text{LM}}(D|Q) = \frac{P(Q|M_D)P(D)}{P(Q)} \propto P(Q|M_D)$$

- A probabilistic framework for ranking documents given a query
- Each document is viewed as a language model for generating the query
- Those documents with higher **query-likelihoods** are more relevant to the query

LM for SDR: Two Matching Strategies

- **Literal Term Matching:** Each document offers a n -gram (usually unigram) distribution for observing a query word

$$P_{\text{Unigram}}(Q|M_D) = \prod_{i=1}^L [\lambda \cdot P(w_i|M_D) + (1-\lambda) \cdot P(w_i|M_C)]$$

- **Concept Matching:** Each document as a whole consists of a set of shared latent topics with different weights -- A document topic model (DTM)
 - Each topic offers a unigram (multinomial) distribution for observing a query word
- $$P_{\text{PLSA/LDA}}(Q|M_D) = \prod_{i=1}^L \left[\sum_{k=1}^K P(w_i | T_k) P(T_k | M_D) \right]$$
- PLSA ([Probabilistic Latent Semantic Analysis](#)) and LDA ([Latent Dirichlet Allocation](#)) are the two good examples
 - Mainly differ in inference of model parameters (fixed & unknown vs. Dirichlet distributed)

Word Topic Models (WTM)

- Each word of 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

$$P_{\text{WTM}}(Q|D) = \prod_{i=1}^L \left[\sum_{w_j \in D} P_{\text{WTM}}(w_i | M_{w_j}) P(w_j | D) \right]$$

- A spoken document can be viewed as a composite WTM
- WTM is a kind of LM for translating words in the document to words in the query
- $P(w_j | D)$ is estimated according to the frequency of w_j in D

Unsupervised Training of WTM

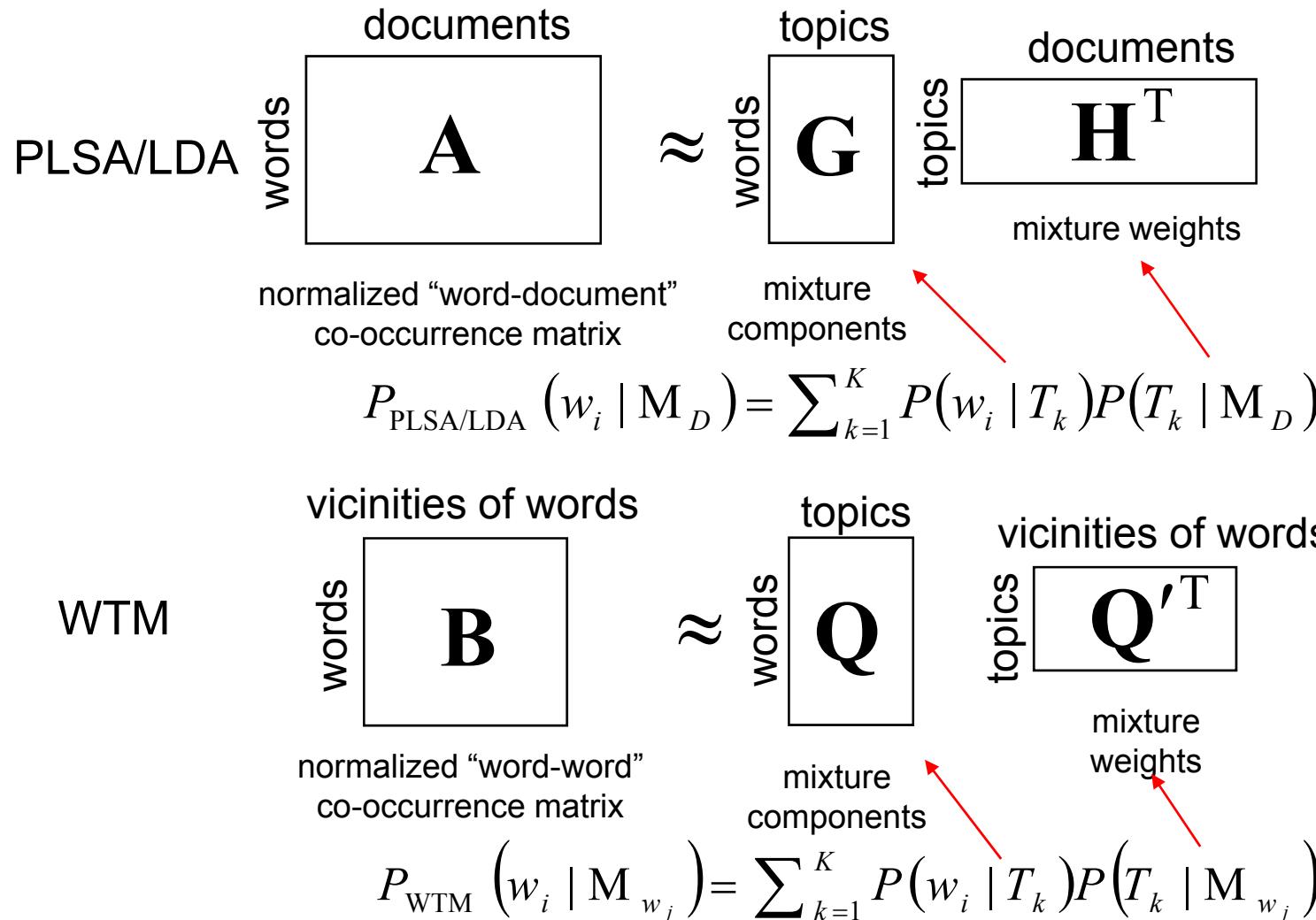
- The WTM $P_{\text{WTM}}(w_i | M_{w_j})$ of each word can be trained with maximum likelihood estimation (MLE)
 - By concatenating those words occurring within a context window around each occurrence of the word, which are assumed to be relevant to the word, to form the training observation



$$\log L_w = \sum_{w_j \in \mathbf{W}} \log P_{\text{WTM}}(Q_{w_j} | M_{w_j}) = \sum_{w_j \in \mathbf{W}} \sum_{w_i \in Q_{w_j}} c(w_i, Q_{w_j}) \log P_{\text{WTM}}(w_i | M_{w_j})$$

- \mathbf{W} : the set of words in the language
 - WTM was trained to optimize its prediction power over the observation

Comparison Between WTM and DTM -- Probabilistic Matrix Decompositions



Comparison Between WTM and DTM

-- Spoken Document Retrieval

- Experiments were conducted on the TDT-2 spoken document collection (~50h broadcast news stories, 16 test queries)
 - Results were measured by Mean Average Precision (*mAP*)

PLSA		LDA		WTM		WTM-L	
TD	SD	TD	SD	TD	SD	TD	SD
0.627	0.568	0.641	0.570	0.636	0.573	0.644	0.574

- PLSA, LDA and WTM (8 topics) are all trained without supervision (without using additional query-document relevance information)
 - PLSA or LDA maximizes the collection likelihood
 - WTM maximizes the likelihood of words in each word's vicinity
- WTM-L: Further assume the parameters of WTM follow Dirichlet distributions

$$\hat{P}_{\text{DTM}/\text{WTM}}(w_i | M_D) = \rho_1 \cdot P_{\text{DTM}/\text{WTM}}(w_i | M_D) + \rho_2 \cdot P(w_i | M_D) + (1 - \rho_1 - \rho_2) \cdot P(w_i | M_C)$$

 -

Supervised Training of WTM

- Maximum Likelihood Estimation (MLE)
 - Maximize the log-likelihood of an outside training set of (~800) query exemplars generated by their relevant documents

$$\log L_{Q_{TrainSet}} = \sum_{Q \in Q_{TrainSet}} \sum_{D_r \in \mathbf{D}_R \text{ to } Q} \log P_{\text{WTM}}(Q | M_{D_r})$$

- Minimum Classification Error Training (MCE)
 - Given a training query exemplar, we can instead minimize the following error function

$$E(Q, D_r, D_{irr}) = \frac{1}{|Q|} \left[-\log P_{\text{WTM}}(Q | M_{D_r}) + \max_{D_{irr}} \log P_{\text{WTM}}(Q | M_{D_{irr}}) \right]$$

relevant document irrelevant document
Other irrelevant documents for the training query
can be into consideration

- Further converted to a loss function with a Sigmoid operator
- Corresponding parameters of WTM then are updated with a generalized probabilistic descent (GPD) procedure

Results of Supervised Training

	WTM				PLSA				Unigram	
	MIX-8		MIX-32		MIX-8		MIX-32			
	TD	SD	TD	SD	TD	SD	TD	SD	TD	SD
MLE	0.689	0.617	0.735	0.686	0.675	0.592	0.683	0.626	0.633	0.566
MCE	0.700	0.631	0.760	0.710	0.679	0.608	0.685	0.628	0.646	0.581

- For WTM, if training query-relevant document pairs were available, significantly better results could be achieved by either MLE or MCE
- PLSA and Unigram LM (i.e., the simple literal term matching model) can also be trained with supervision
- Notice also that, MCE seems to provide additional performance gains over MLE

Results of Various Vector Space Approaches

- Here we also list the results of retrieval using three popular vector space approaches

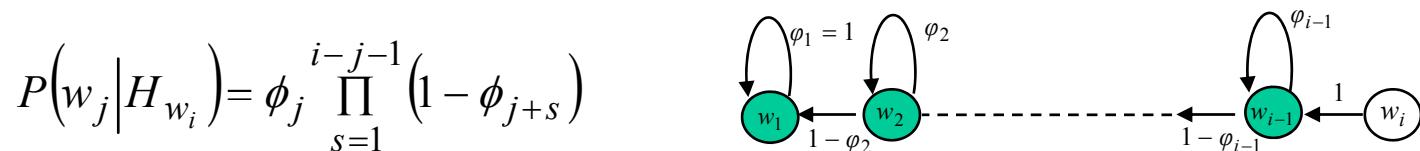
VSM		LSA		SVM	
TD	SD	TD	SD	TD	SD
0.555	0.512	0.551	0.531	0.580	0.532

- SVM (Support Vector Machine) treats IR as a classification problem
 - A set of 11 heterogeneous features is used to represent each spoken document given an input query
 - SVM was trained by leveraging the relevance information of the outside training query exemplars
- All LM-based retrieval approaches are significantly better than these vector space approaches

WTM Applied to Other Related Tasks

- Language Modeling in Speech Recognition

$$\begin{aligned} P(w_i | H_{w_i}) &= \sum_{j=1}^{i-1} P_{\text{WTM}}(w_i | M_{w_j}) P(w_j | H_{w_i}) \\ &= \sum_{j=1}^{i-1} P(w_j | H_{w_i}) \sum_{k=1}^K P(w_i | T_k) P(T_k | M_{w_j}) \end{aligned}$$



- Extractive Spoken Document Summarization

$$\begin{aligned} P(D|S) &= \prod_{i=1}^L \left[\sum_{w_j \in S} P_{\text{WTM}}(w_i | M_{w_j}) P(w_j | S) \right] \\ &= \prod_{i=1}^L \left[\sum_{w_j \in S} P(w_j | S) \sum_{k=1}^K P(w_i | T_k) P(T_k | M_{w_j}) \right] \end{aligned}$$

- For both tasks, WTM has preliminarily demonstrated good results as compared to existing approaches

Conclusions

- Multimedia information access (over the Web) using speech will be very promising in the near future
 - Speech is the key for multimedia understanding and organization
 - Several task domains still remain challenging
- This research presented a word topic modeling (WTM) approach for spoken document retrieval
 - Simple and easy to implement
- Various model inference techniques were studied for WTM and other document topic models (DTMs)
 - Given an outside training set of query exemplars with relevance labels, the LM-based retrieval models can be steadily improved

Thank You !

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
- **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 TALES
project

IBM MASTOR
project

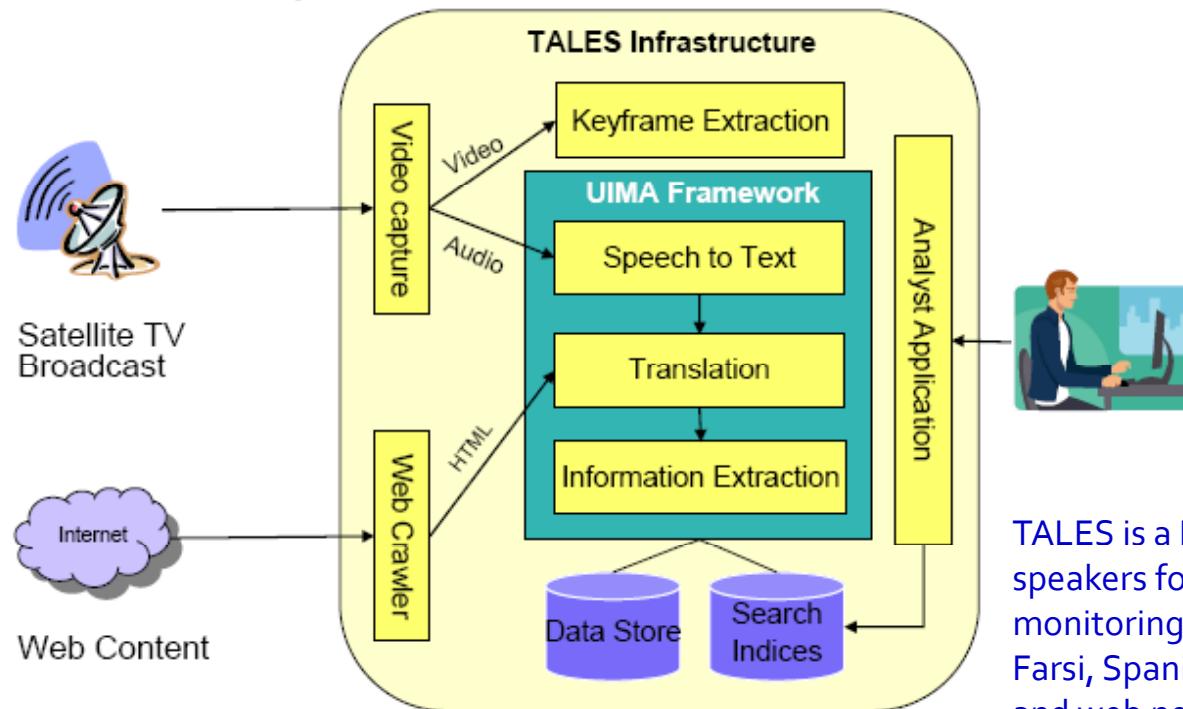


Adapted from the presentation slides of Dr. Yuqing Gao's at ISCSLP2008



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



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

TALES Demo

Foreign Broadcast Video Monitoring and Search

The screenshot shows a Microsoft Internet Explorer window with the URL <http://linc001.watson.ibm.com:9080/tales/>. The page title is "TALES Main Result Page - Microsoft Internet Explorer". On the left, there's a sidebar with "Headlines", "Usage Help", "Search History", "Bookmarks", "Media Type" (checkboxes for All, Video, Audio, WebPage), "Language" (checkboxes for All, English, Arabic, Chinese), "Sort By" (radio buttons for Date, Media, Language), "Media Player" (checkboxes for Reverse, Embedded), and "External". The main content area has a search bar with "evacuation hurricane". Below it, there are four video thumbnails with the following details:

- English Query:** Provider: Al-Jazeera, Duration: 01:00, Summary: ... at a time to play where the citizens and tourists low from areas which southern border in the state of Texas and is believed ... turned into a g...
- Arabic Text:** Provider: CCTV4, Duration: 00:57, Summary: germany both countries should hurricane northeas...
- Translated Speech:** Provider: CTV4, Duration: 00:56, Summary: ... results how game both countries should continue strong hurricane continue to ravage the gulf tranquility of the hurricane is originally arrived in mexico northeast and the united states of texas in southern 18,000 were local residents and many are stranded tourists but plenty of housing were damaged mickey hurricane caused economic losses amounting to tens us 1 million ...
- English Translation:** Provider: Al-Arabya, Duration: 01:01, Summary: ...

■ 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

The screenshot shows a Microsoft Internet Explorer window with the URL <http://news.bbc.co.uk/hi/arabic/news>. The page title is "Arabic2English Translation". The main content area shows a news article about a bombing in northern Iraq. The page includes a sidebar with links to "BBC in Arabic", "Aljazeera", "Linux links", "Aliyadh Newspaper", and "Al Ahram Newspaper". At the bottom, there are sections for "New Iraqi government begin tasks" and "The electoral battle in Brifida Enter the Last Day". A small logo in the bottom right corner says "SLP".

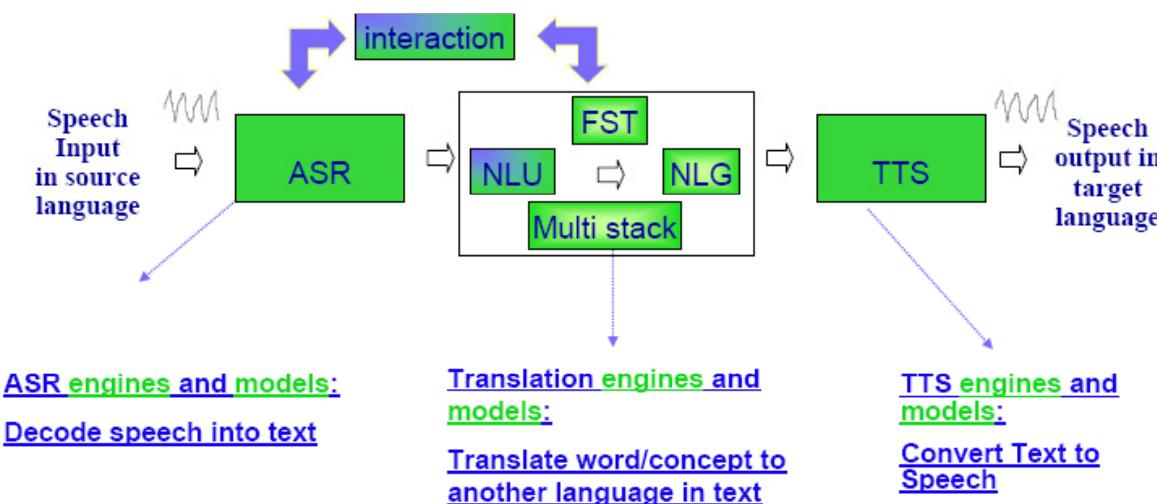


Adapted from the presentation slides of Dr. Yuqing Gao's at ISCSLP2008

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

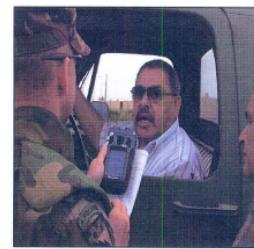
- MASTOR is a two-way, free form speech translator that assists human communication using natural spoken language for people who do not share a common language

IBM Advanced Speech-to-Speech Translation Techniques



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

Handheld System



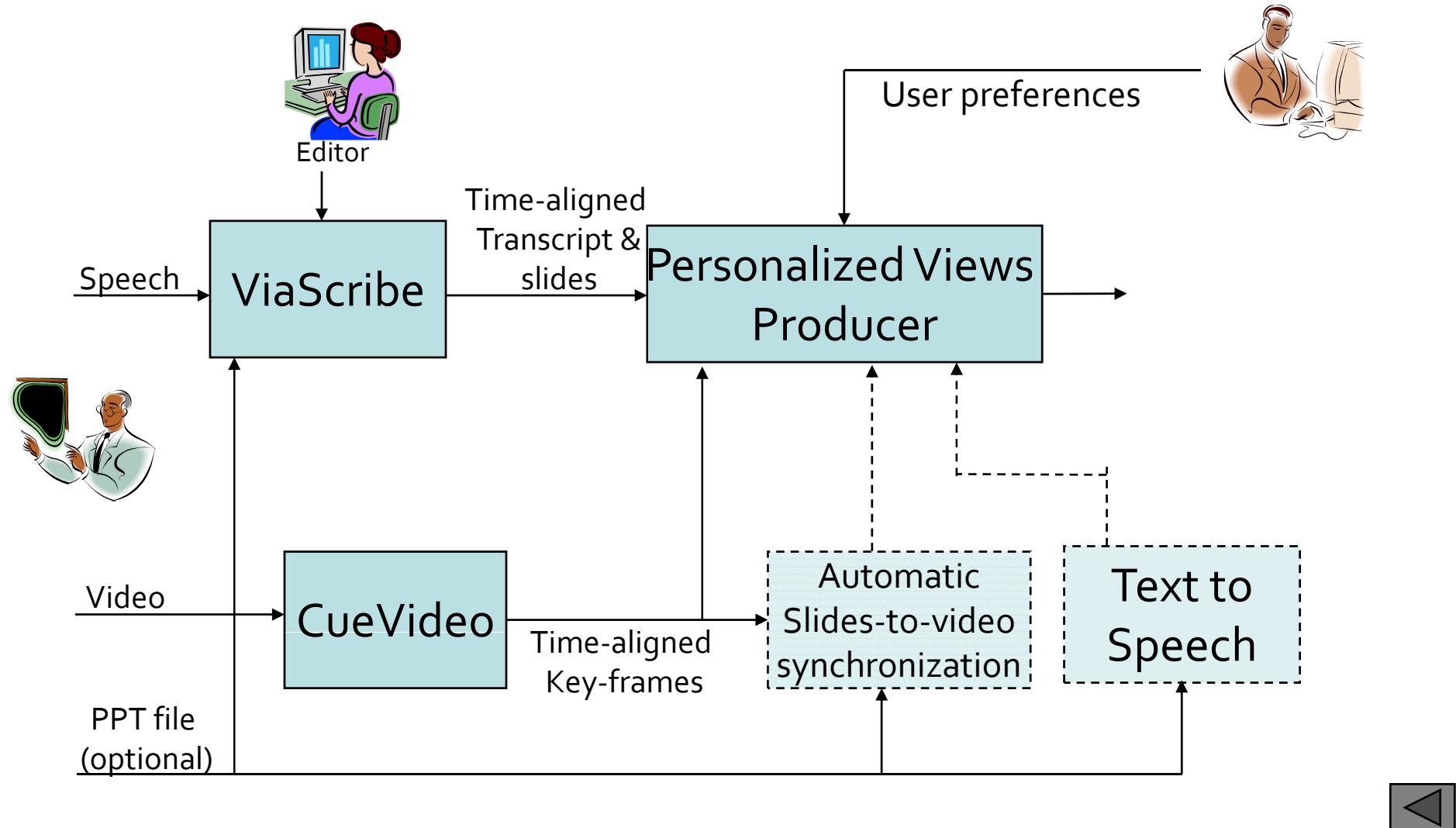
Laptop systems
- hands-free, eyes-free function



MASTOR Demo

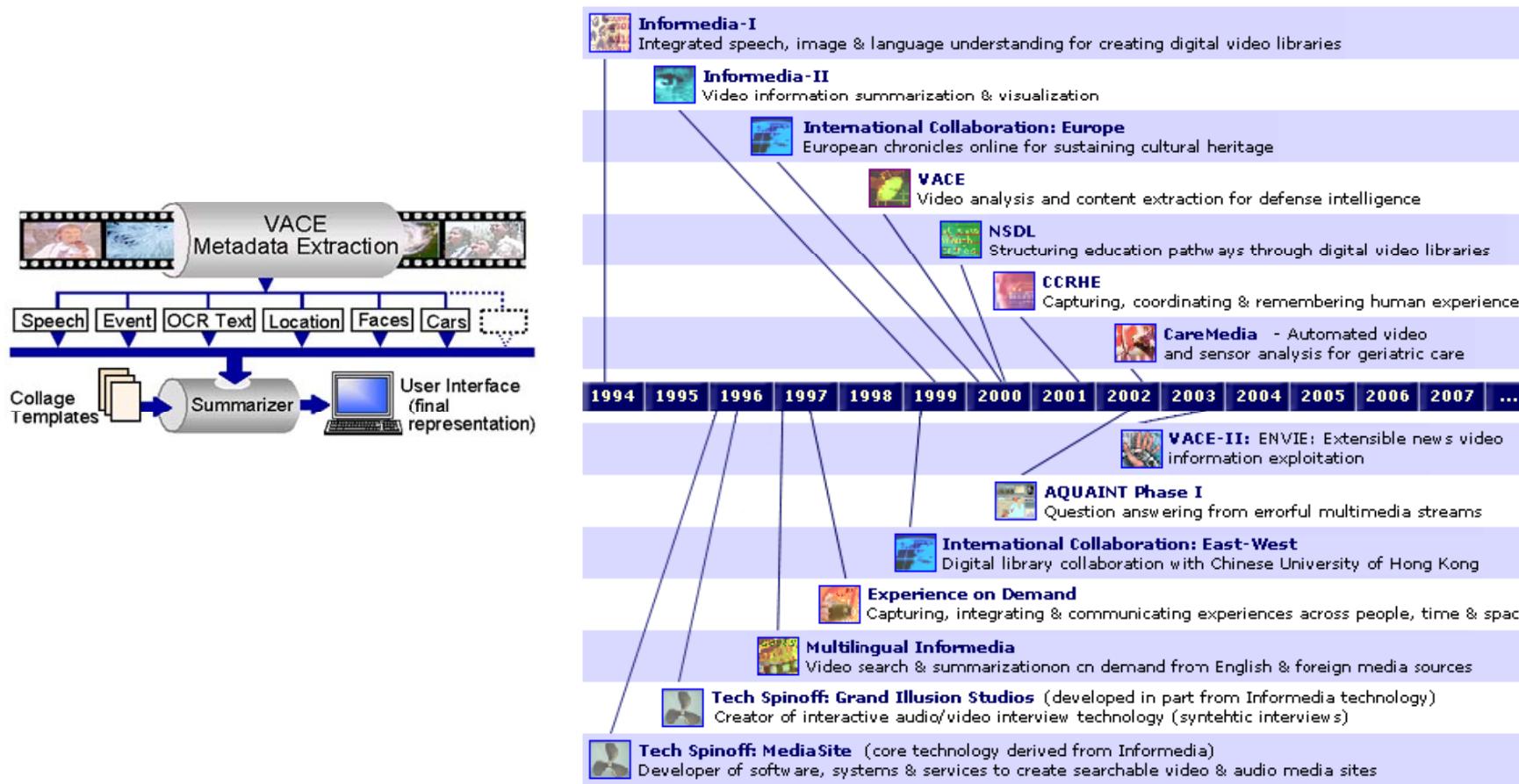


IBM's Audio-Visual Search Solutions



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: What is the status of the trade deficit with Japan?

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	1	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	50.33	4

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 amotcurs 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 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. l. 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: 19.1699 seconds Stop Audio

AT&T Labs Research

AT&T

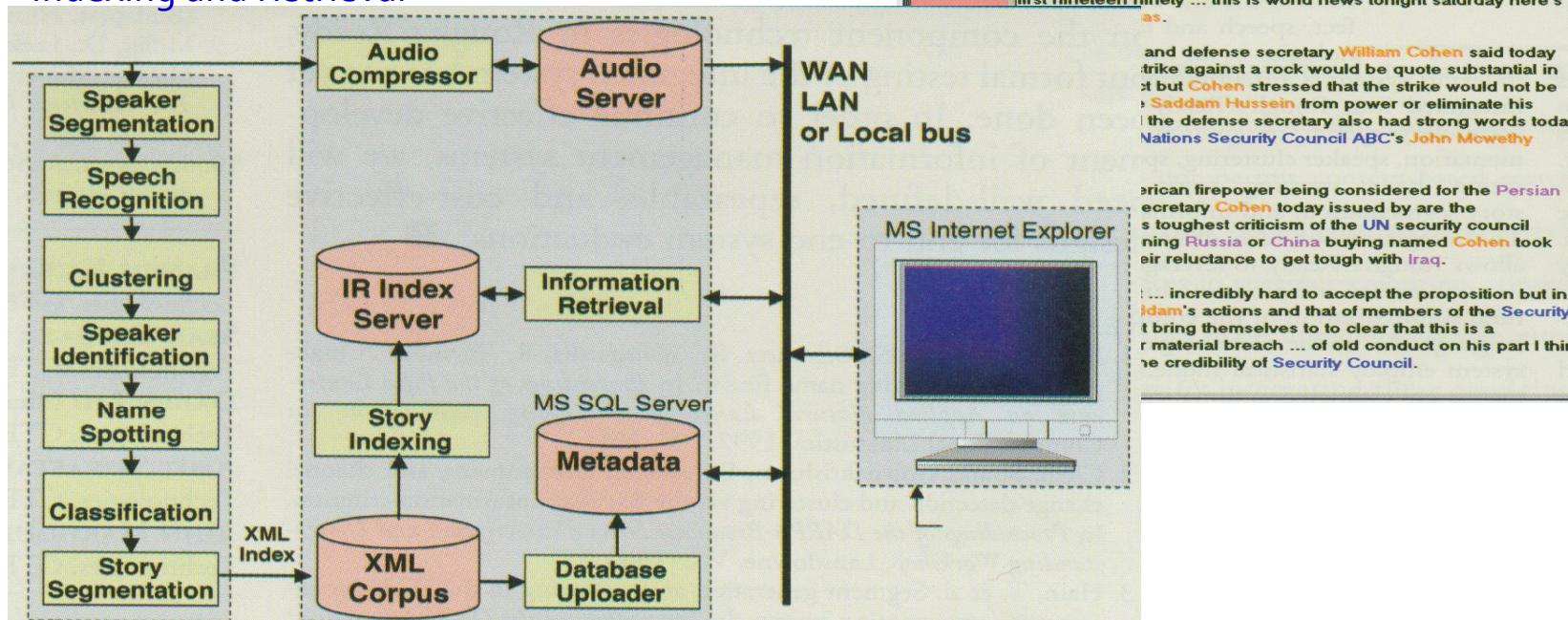
Design and evaluate user interfaces to support retrieval from speech archives



Julia Hirschberg, Fernando Pereira, Amit Singhal et al., "SCAN: designing and evaluating user interfaces to support retrieval from speech archives," SIGIR 1999

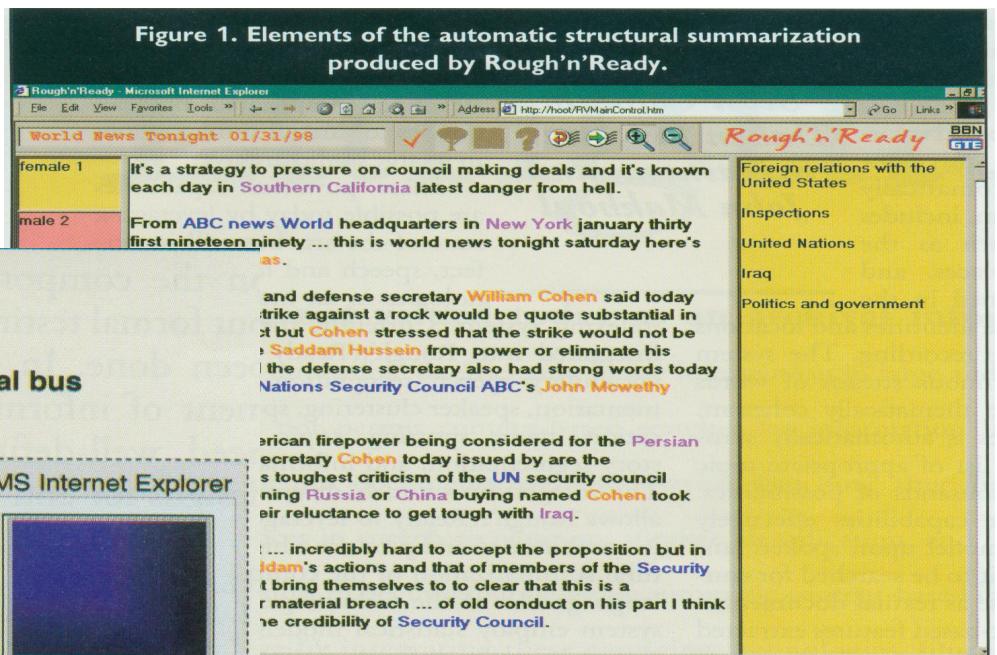
BBN Rough'n'Ready System

Distinguished Architecture for Audio Indexing and Retrieval



Automatic Structural Summarization for Broadcast News

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



Google Voice Search



Google-411:
Finding and connecting to local business

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](#).

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

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

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

網址① http://labs.google.com/gaudi

Google [G] google labs 開始 搜尋 番號 PageRank ABC 拼字檢查 傳送到... google 設定 lenovo

Search what the politicians are saying
Obama grandmother Search videos Learn more

Audio Indexing

All Politicians | McCain | Obama | Debates

- Seniors for Obama 1 month ago - 11:25 - about 9 mentions
- Barack Obama on Veterans in Sioux Falls SD 5 months ago - 09:04 - about 3 mentions
- Barack Obama at American University 9 months ago - 13:38 - about 2 mentions
- Barack Obama on Equal Pay in Albuquerque NM 2 months ago - 09:19 - about 2 mentions
- Barack Obama in Louisville KY 5 months ago - 04:35 - about 4 mentions

Seniors for Obama

Obama grandmother Search inside this video

Goooooooole ►
1 2 3 4 5 6 Next

...I was born and where he and my grandmother help raise I think you...
min 0
...issues come and go Senator Obama called the saying absolutely correct...
min 1
...the problem the bigger question Barack Obama does not want to privatize...
min 3



<http://labs.google.com>



Hold the Phone: It's Google Voice and It's Free

Posted Mar 16, 2009

Google last week released Google Voice, an application offering a number of features for the telecommunications space.

Among its offerings, the service allows users to place domestic calls for free from a PC, place low-cost calls internationally, switch phones during a call, record conversations, share voicemails, and transcribe voicemails to email or SMS with automatic recognition technology.

Google Voice is built on GrandCentral, a service that Google bought in the summer of 2007 and allows a user to tie all of her phones to one phone number. According to an email from Sara Jew-Lim, a member of the global communications team for Google, the automatic recognition engine was developed in-house and is the same engine used in Goog411, Google's business directory service, and in its mobile voice search products.

The launch of Google Voice is just another sign, like the launch of the Google mobile operating system (OS), Android phone, or Goog411, that Google is serious about making a name for itself in the telecommunications space. The move also suggests that Google is looking to include new modalities, like voice, in its search capabilities.

.....



Source: <http://www.speechtek.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 shows a Microsoft SharePoint-like web interface for 'Audio Search'. A search bar at the top contains the query 'virtualization hypervisor'. Below it, a heading reads 'Click-to-play snippets navigate directly into video'. The main area displays 'Audio Search Results' with two items:

- The Next Server Wave [MS StudioCasts]**: A thumbnail image of a man speaking, with a summary of the video content below it. A blue bracket on the left labeled 'Metadata Hits' points to this item.
- DTTS-Feb 2007-Tech Edition-Longhorn-Technical Overview [Academy Mobile Podcasts]**: Another thumbnail image of a man speaking, with a summary of the video content below it. A blue bracket on the left labeled 'Audio Hits' points to this item.

At the bottom of the results page, it says 'Your search took 0.23 seconds.' and 'Fast search speeds!' with a blue arrow pointing to the text. The footer of the page reads 'Microsoft Research 2008'.



Microsoft Launches Mobile Voice Search Application

Microsoft this week launched [Microsoft Recite](#), a voice search application for Windows mobile devices.

The offering—which Microsoft calls “search technology for your voice”—allows users to capture, search, and retrieve spoken notes and reminders via voice commands.

“The idea behind Recite was to help make a busy mobile lifestyle easier by providing [users] an easy method for remembering, searching, and retrieving mental notes and reminders from their mobile phones using the sound of their voice—without navigating menus or tapping out text,” writes Stathis Papaefstathiou, product unit manager at Microsoft, in an email to Speech Technology. “Just speak your mental note to store it and later, when you need the information, just search and retrieve the remembrance using your voice.”

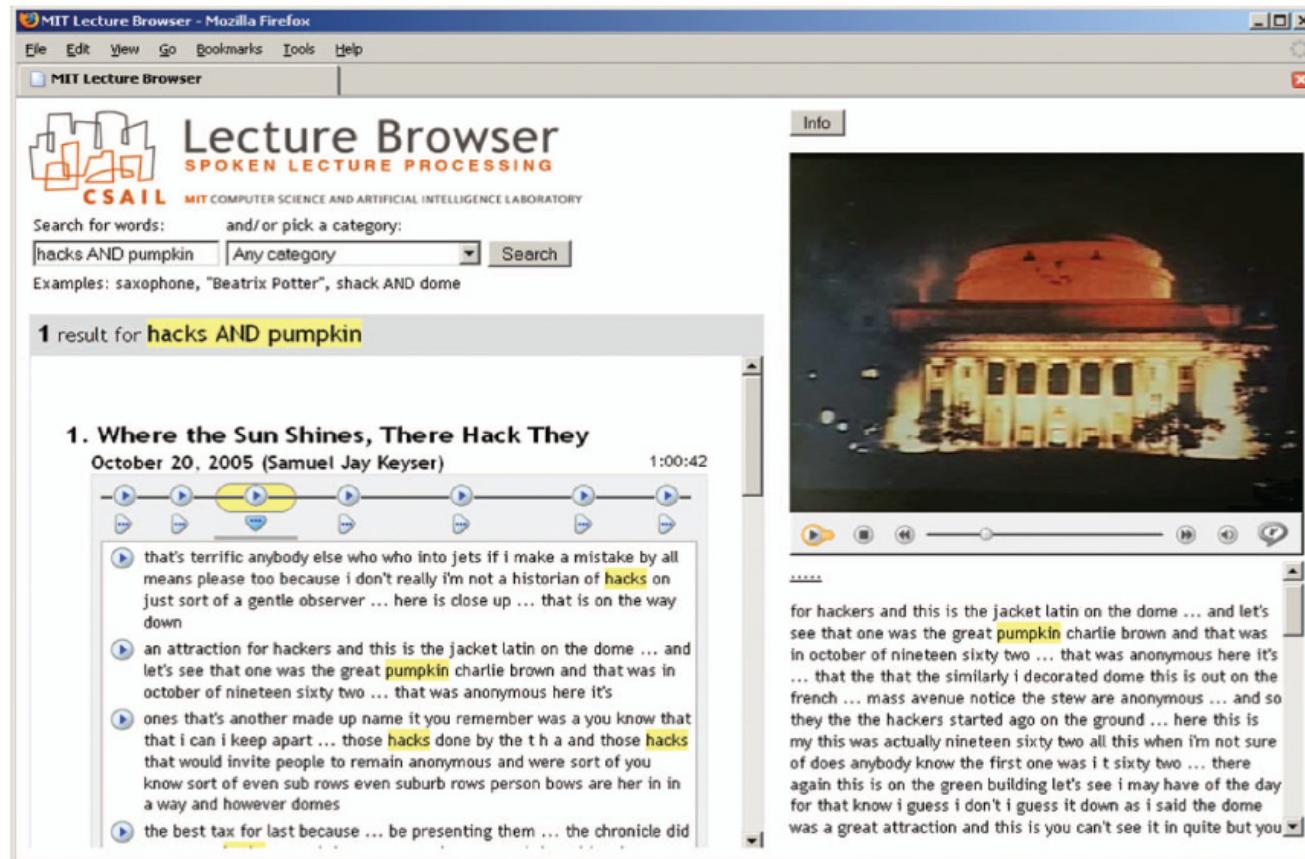
Microsoft Recite can store hundreds of spoken notes and uses voice pattern matching, which is based on speaker inflection, to analyze speech and search recordings.

.....



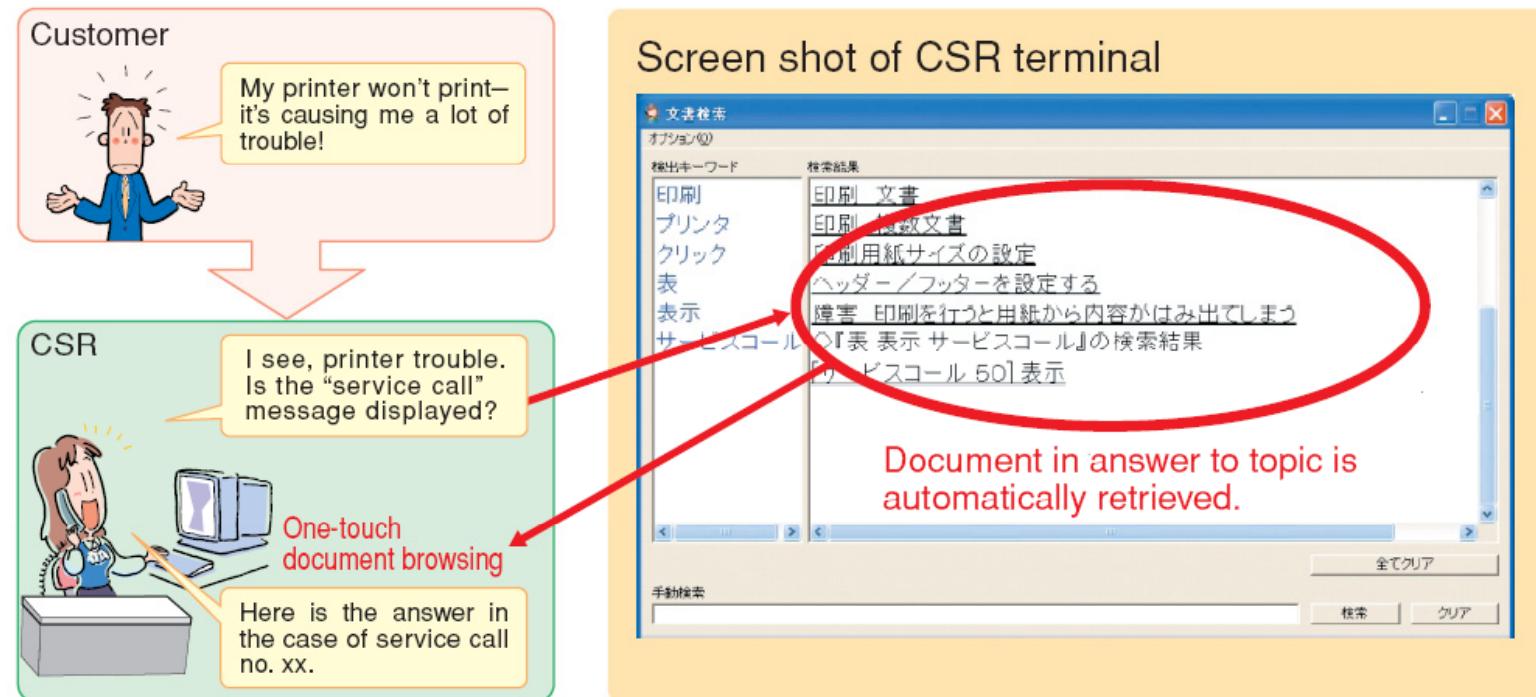
MIT Lecture Browser

- Retrieval and browsing of academic lectures of various categories



NTT Speech Communication Technology for Contact Centers

Automatic document-retrieval by speech recognition

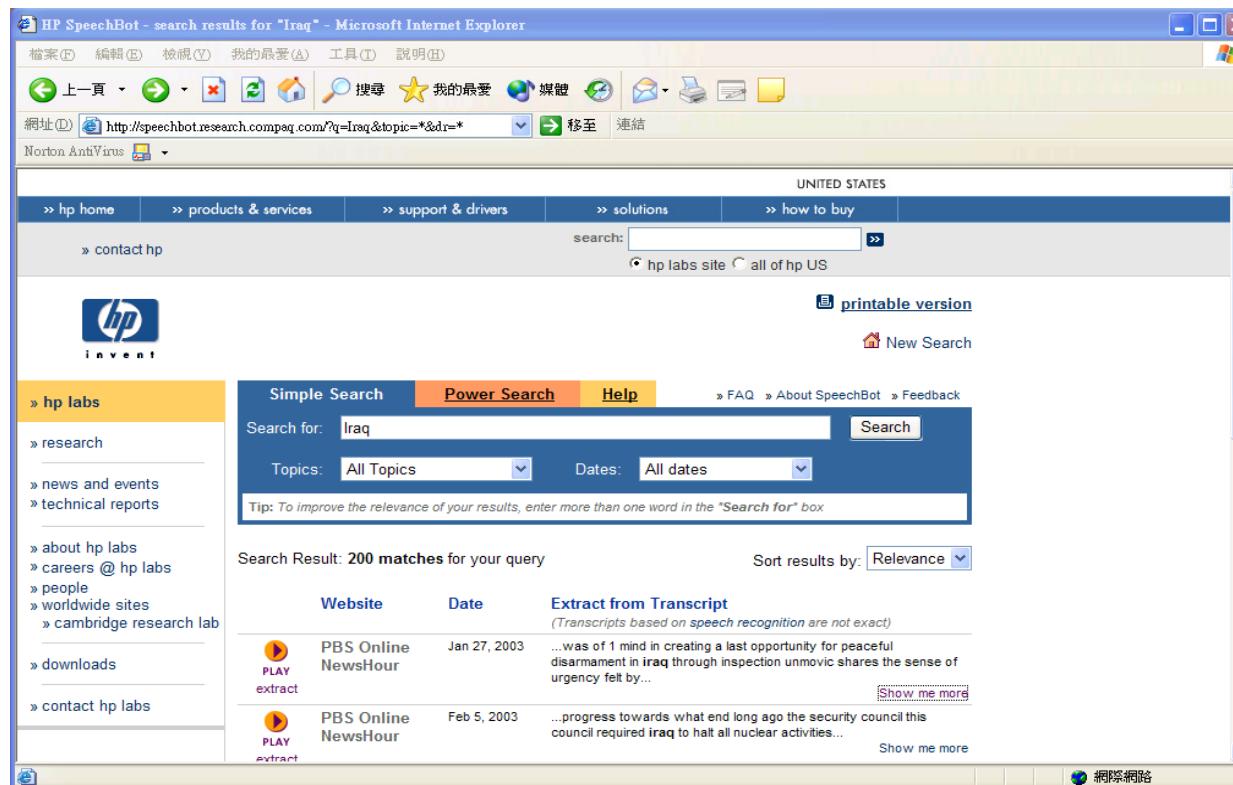


- CSR: Customer Service Representative



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

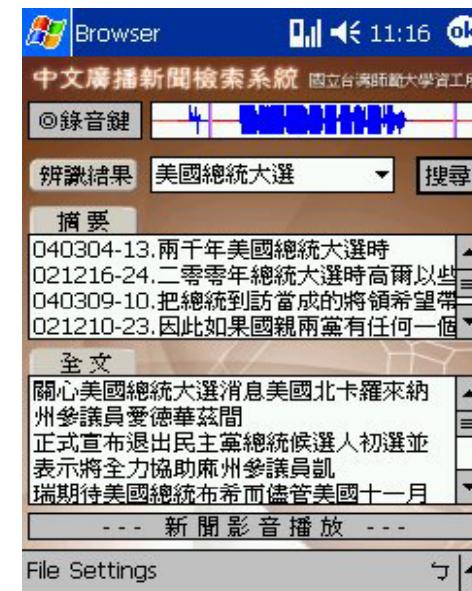


Some Prototype Systems Developed in Taiwan

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



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



Detailed Statistics of the TDT-2 Collection

No. of spoken documents	2,265 stories, 46 hrs of audio		
No. of distinct text test queries	16 Xinhua text stories (Topics 20001~20096)		
	Min.	Max.	Mean
Document length (characters)	23	4841	287
Query length (characters)	183	2623	533
No. of relevant documents per query	2	95	29

WTM vs. Other LMs

- If the context window for modeling the vicinity information of WTM is reduced to one word ($S=1$), WTM can be either degenerated to:
 - A **unigram model** as the latent topic number K is set to 1
 - Or, viewed as analogous to a **bigram model** (as $K=V$), or an **aggregate Markov model** (as $1 < K < V$)
- WTM is also close in spirit to the word class based model (WCBM) as well, by relating the latent topics of the former to the word classes of the latter
 - WTM differs from WCBM in that WTM **disregards word order** information and **leverages word co-occurrence statistics from longer text spans**
 - Whereas most of WCBM approaches are based purely on modeling word bigram sequences

WTM vs. IBM Translation Model (1/2)

- WTM

$$\begin{aligned} P_{\text{WTM}}(Q|D) &= \prod_{i=1}^L \left[\sum_{w_j \in D} P_{\text{WTM}}(w_i | M_{w_j}) P(w_j | D) \right] = \prod_{i=1}^L \left[\sum_{w_j \in D} \sum_{k=1}^K P(w_i | T_k) P(T_k | M_{w_j}) P(w_j | D) \right] \\ &= \prod_{i=1}^L \left[\sum_{k=1}^K P(w_i | T_k) \sum_{w_j \in D} P(T_k | M_{w_j}) P(w_j | D) \right] \\ &= \prod_{i=1}^L \left[\sum_{k=1}^K P(w_i | T_k) \hat{P}(T_k | M_D) \right] \quad \text{WTM has the same document ranking form as PLSA/LDA} \end{aligned}$$

- Both unsupervised and supervised training
- IBM Translation Model

$$P_{\text{IBM}}(Q|D) = \prod_{i=1}^L \left[\sum_{w_j \in D} P_{\text{TRANS}}(w_i | w_j) P(w_j | D) \right]$$

- Supervised training (e.g., using title-document corresponding relationships)

WTM vs. IBM Translation Model (2/2)

- Compare WTM with IBM Translation Model (IBM-1)
 - Supervised training: using outside training query exemplars

WTM (MIX-8)		WTM (MIX-32)		IBM-1	
TD	SD	TD	SD	TD	SD
0.689	0.617	0.735	0.686	0.700	0.645

WTM vs. PLSA: Language Model Adaptation

- Tested on broadcast news transcription task

		CER (%)	PP		
Baseline (Background Trigram Model)		15.22%	752.49		
		WTM	PLSA		
Adaptation Corpus	No. Latent Topic	CER (%)	PP	CER (%)	PP
Texts	16	14.77	566.10	14.83	588.51
	32	14.69	553.88	14.73	571.46
	64	14.60	540.62	14.58	552.80
	128	14.44	524.15	14.53	527.41
	256	14.38	508.29	14.47	510.20
Automatic Transcripts	16	14.87	574.60	14.99	591.21
	32	14.90	568.76	14.92	580.80
	64	14.85	564.56	14.82	569.93
	128	14.81	563.25	14.87	562.45
	256	14.96	567.53	14.92	565.85