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

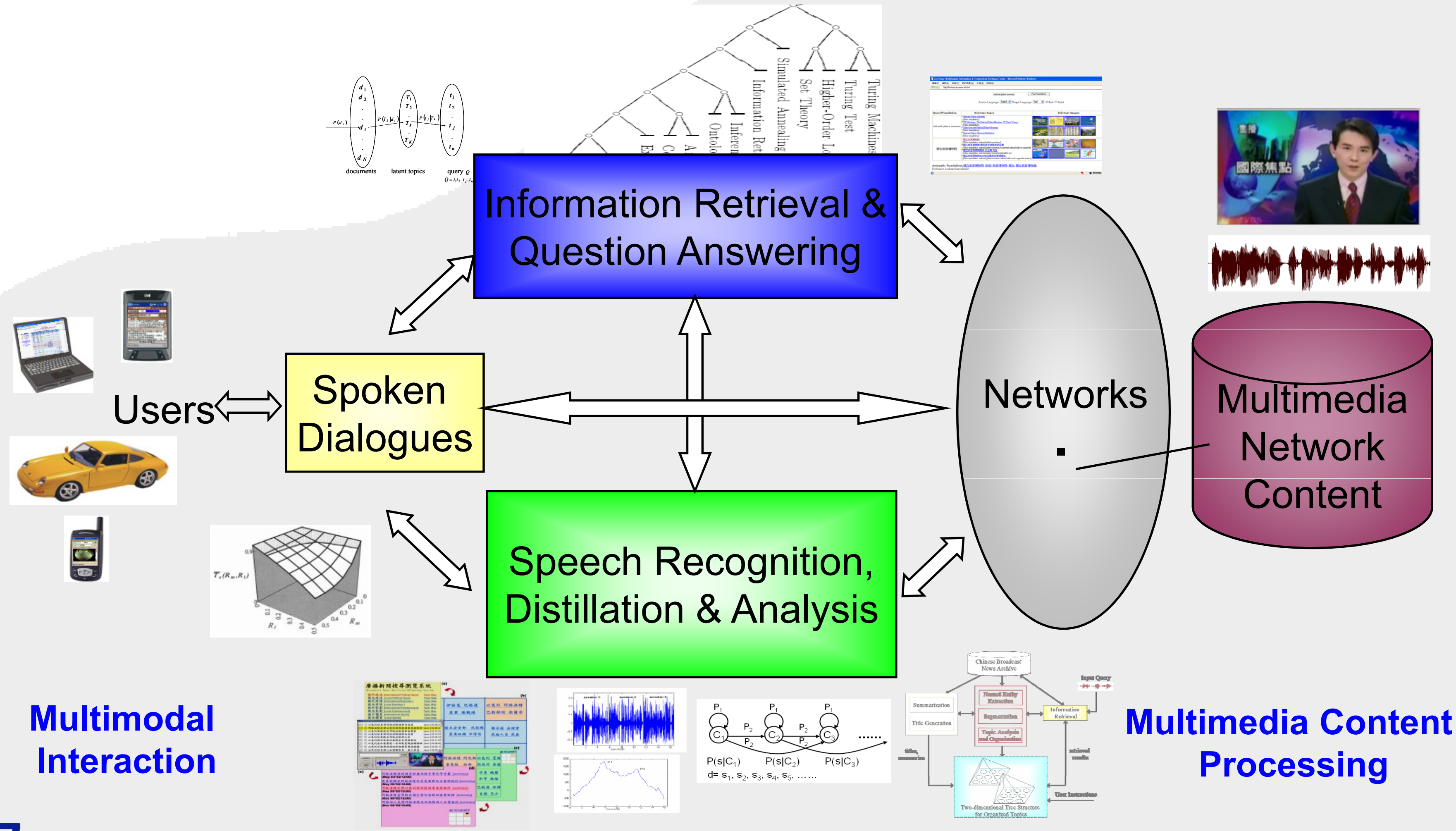
- On the other hand, speech is the primary and the most convenient means of communication between people
  – Speech provides a better (or natural) user interface in wireless environments
  – Especially helpful when using smaller hand-held devices with small screen sizes and limited keyboard entry capabilities

- 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

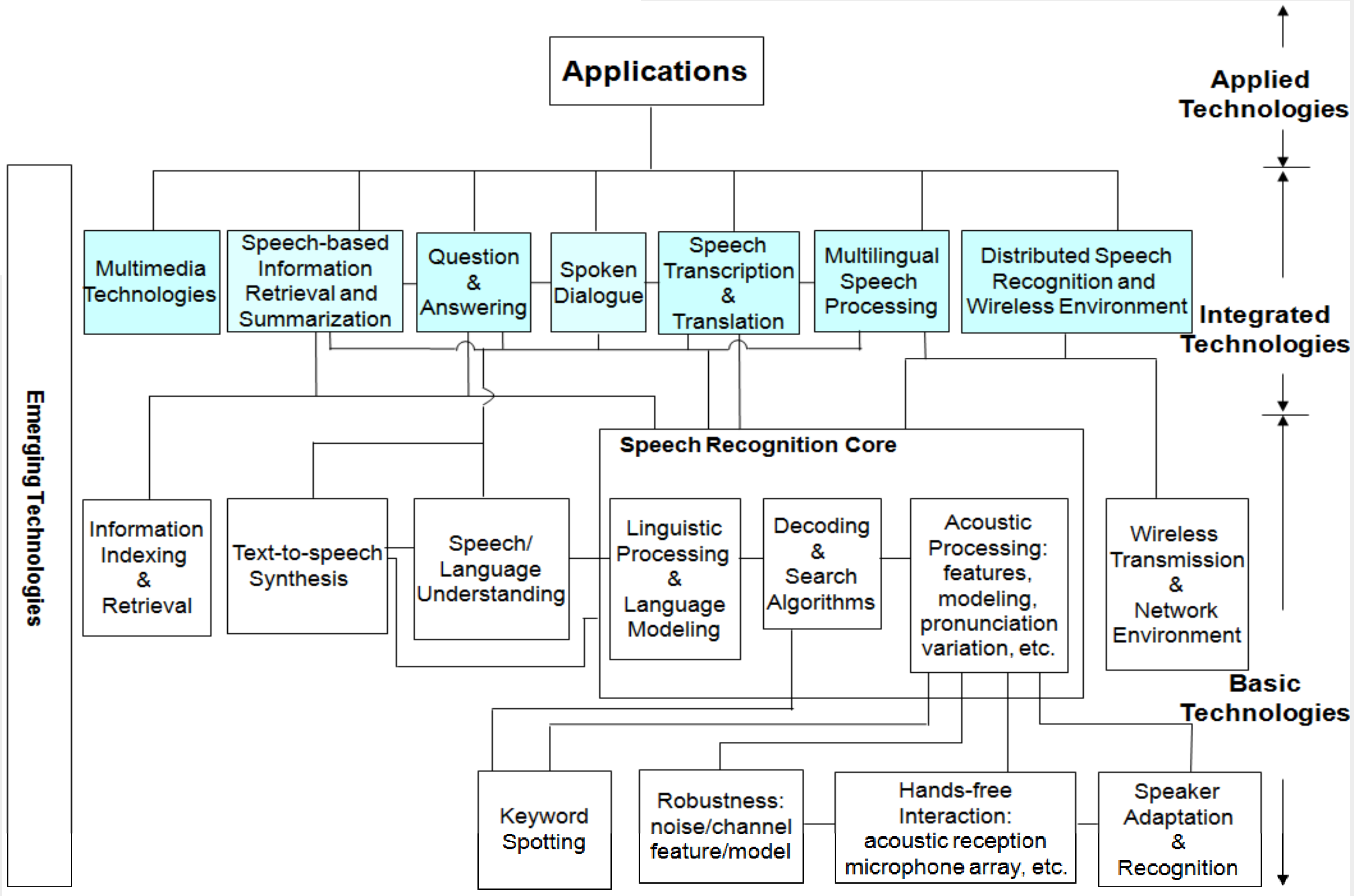
- There also are several research projects conducting on related spoken document processing tasks, e.g.,
  - Rich Transcription Project[1] in the United States (2002-)
    - Creation of recognition technologies that will produce transcriptions which are more readable by humans and more useful for machines
  - TC-STAR Project[2] (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

[1] http://www.nist.gov/speech/tests/rt/ [2] http://www.tc-star.org

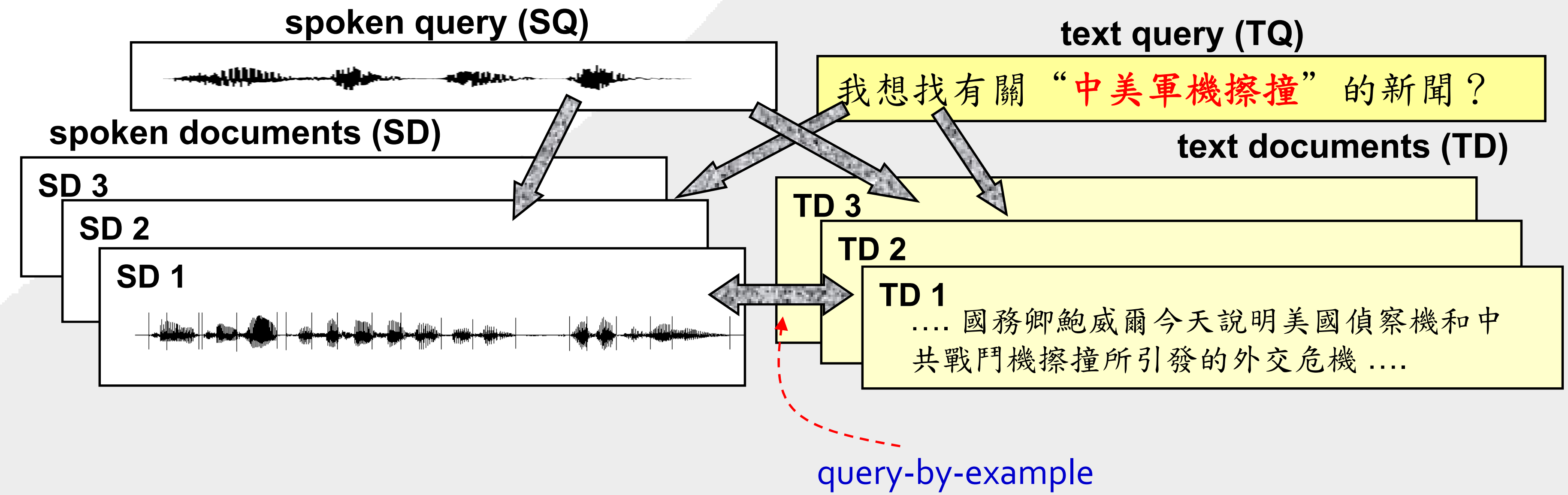
Evaluations of the Rich Transcription Project

- 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 (≧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 then using the keyword (Web) search scenario
- Findings
  - Retrieval performance is quite flat with ASR WER variations in the range of 10~35% (roughly ≦5% degradation in performance in comparison with the "approximately" manual transcriptions)
  - SDR of broadcast news speech has been thought of as "a successful story"

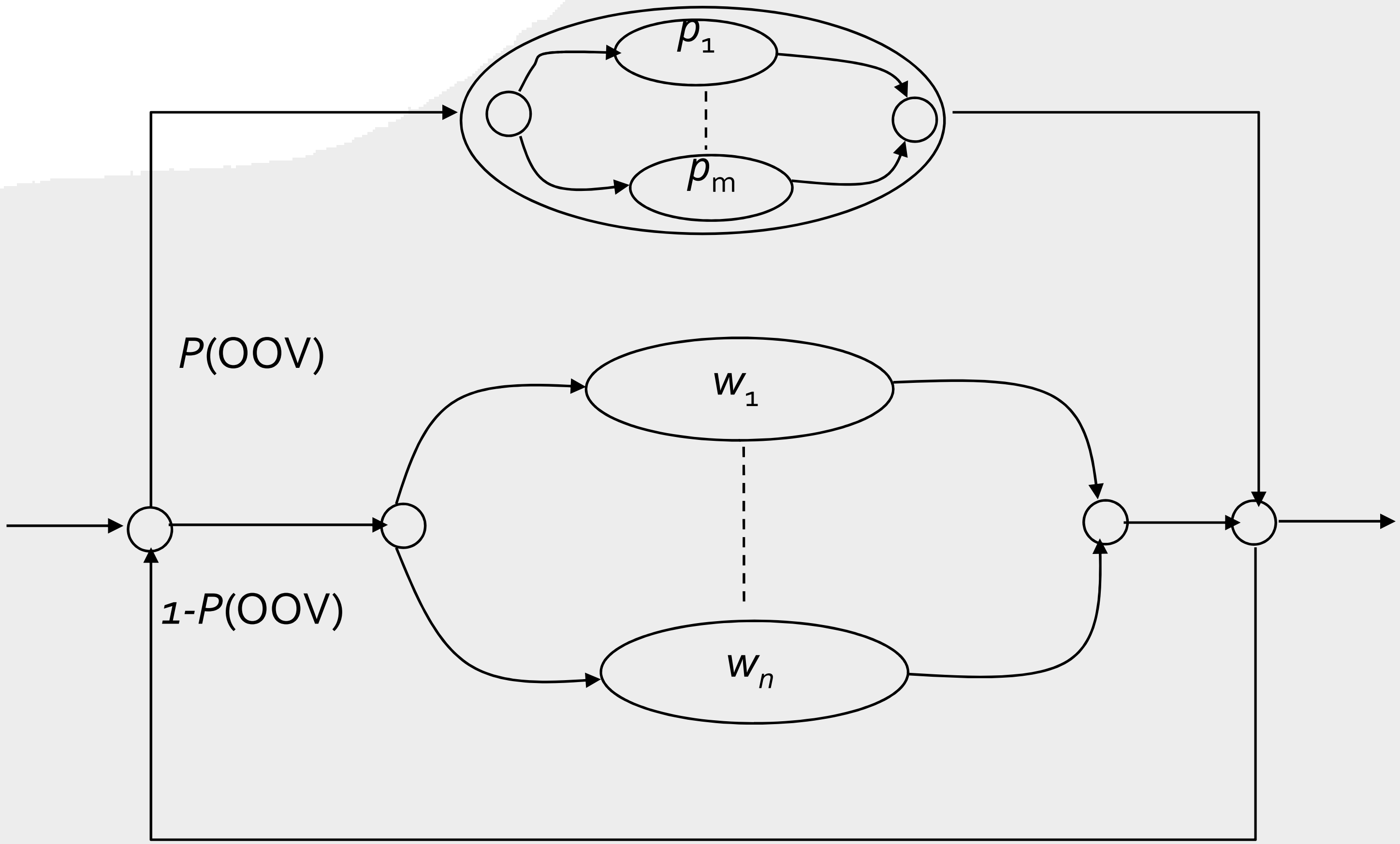
J.S. Garofolo, et al., "The TREC spoken document retrieval track: A success story," http://trec.nist.gov/pubs/trec8/./papers/trec8-sdr-overview.ps

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 expanse 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

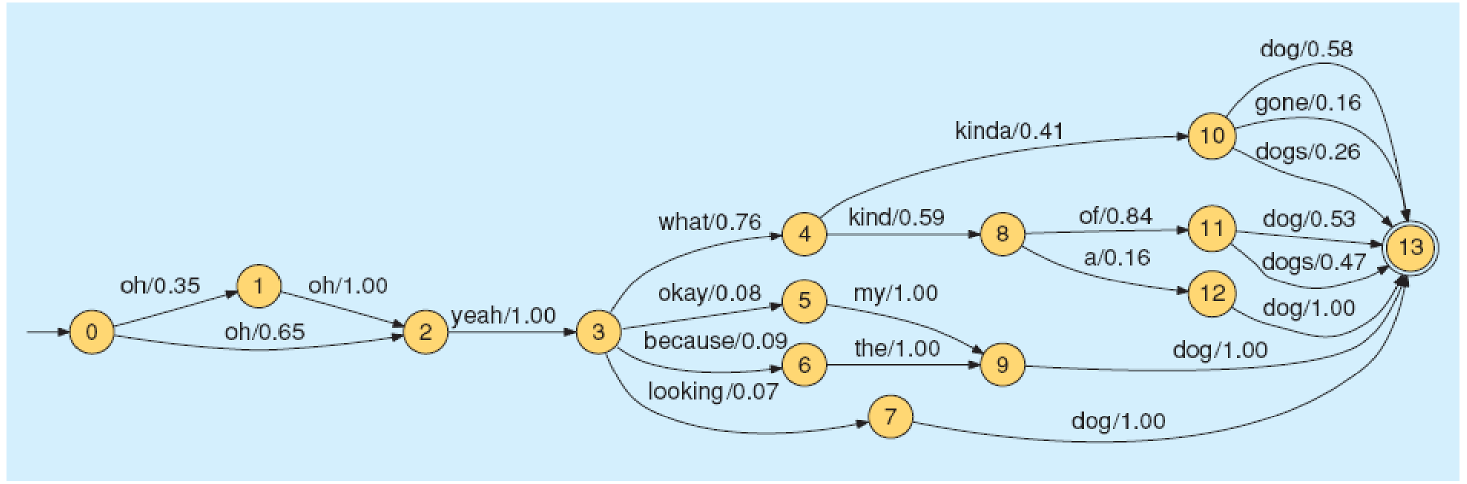


Evaluation Metrics

- 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)

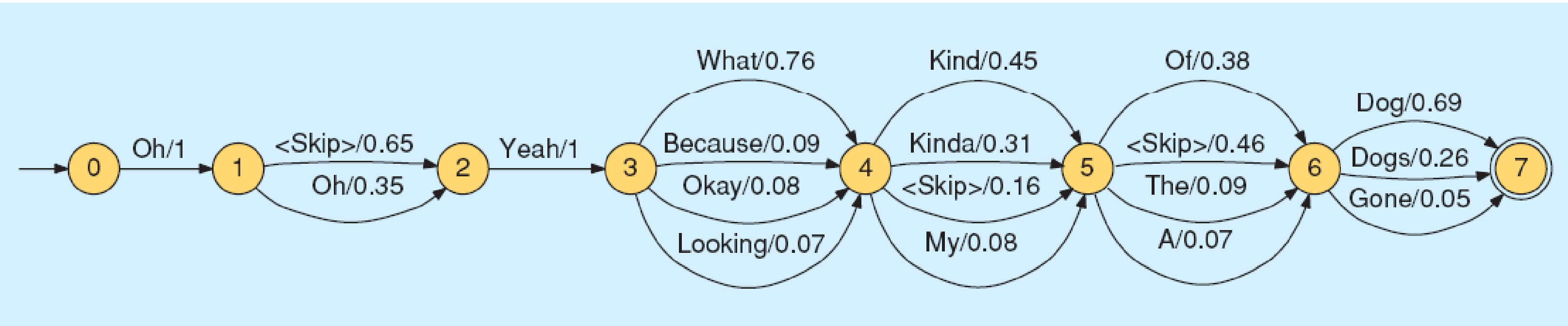
- Use of 1-best ASR output as the transcription to be indexed is suboptimal due to the high WER, which is likely to lead to low recall
- ASR lattices do provide much better WER, but the position information is not readily available (uncertainty of word occurrences) ?
- An example ASR Lattice



Adapted from: C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine* 25 (3), May 2008

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

Adapted from: C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine* 25 (3), May 2008

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 $P(w,l \mid LAT)$ of spoken queries and documents

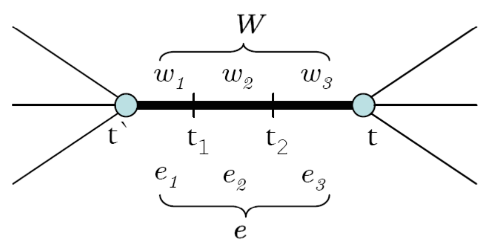
0		1		2		3		4		5		6		7	
Oh	1.0	**Yeah**	.65	**What**	.46	**Kind**	.27	Dog	.26	EOS	.34	**EOS**	.44	EOS	.16
—		Oh	.35	Yeah	.35	What	.27	**Of**	.23	**Dog**	.29	Dog	.09	—	
		—		Because	.06	Kinda	.19	Kind	.16	Dogs	.13	Dogs	.06		
				Okay	.05	The	.06	Kinda	.11	Of	.13	—			
				Looking	.05	My	.05	Dogs	.05	A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

  - 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

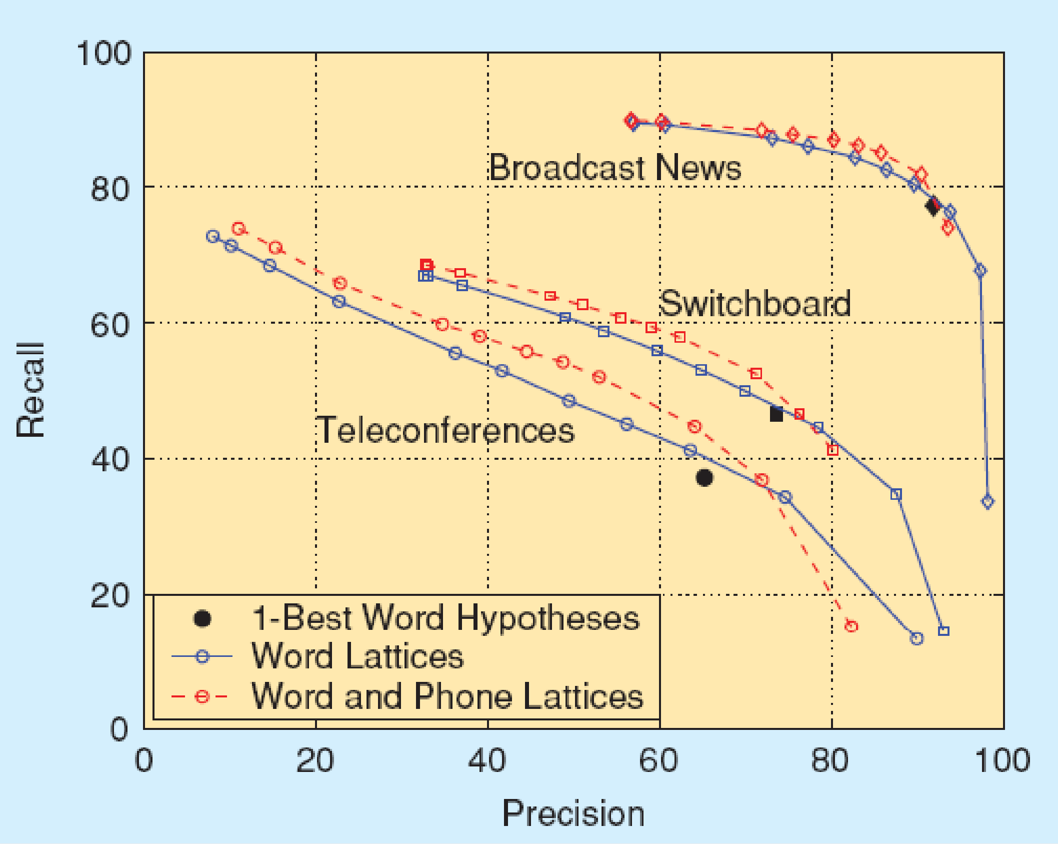


- P. Yu, F. Seide et al. at MSRA proposed alternative approaches that are analogous to PSPL
  - Time-based Merging for Indexing (TMI) for size reduction and Time-Anchored Lattice Expansion (TALE) for word-position mapping

  "Word-lattice based spoken-document indexing with stand text indexers," SSCS 2008 (in conjunction with SIGIR 2008) & SLT2008

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

- Comparison between indexing with 1-bset sequences and lattices



Adapted from: C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine* 25 (3), May 2008

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

Syllable pairs separated by M syllables

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

- 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)	0.9778 (0.7680, 0.9604)	0.9027 (0.8804, 0.9003)	0.9797
SQ/TD	0.8982 (0.4137, 0.8898)	0.8811 (0.6671, 0.8676)	0.7755 (0.7489, 0.7683)	0.9022
TQ/SD	0.7148 (0.3456, 0.7009)	0.6988 (0.5577, 0.6872)	0.6160 (0.5988, 0.6138)	0.7267
SQ/SD	0.6739 (0.3120, 0.6583)	0.6515 (0.5136, 0.6429)	0.5549 (0.5386, 0.5534)	0.6814

$S(N)$, N=1~3, $P(M)$, M=1~3

$S(N)$, N=1

$S(N)$, N=1~2

Chen et al., "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in Mandarin Chinese", *IEEE Transactions on Speech and Audio Processing* 10 (5), 2002

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
  1. Started with a set of indexing features consisting of single base syllables as the initial lexical segments (LSs)
  2. In each iteration, any two adjacent lexical segments with scores higher than the threshold become a new LS
     - Criterion: Forward-Backward Bigram (FB)

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

  3. All instances of these pairs are replaced by the new LSs
  4. 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 $(\Theta_{qry}, \Theta_{doc})$	Mean average precision	
	Queries	Documents			For devel. queries	For test queries
$\text{Ref} \xrightarrow{\text{smart}} \text{Ref}$	Exemplar reference	Reference	smart	–	0.8363	0.7781
$\text{1-best} \xrightarrow{\text{smart}} \text{1-best}$	Exemplar 1-best	1-best	smart	–	0.8271	0.7406
$\text{1-best} \xrightarrow{\text{smart}} \text{Lat}$	Exemplar 1-best	Lattices	smart	$(-, 140)$	0.8321	0.7499
$\text{Lat} \xrightarrow{\text{smart}} \text{1-best}$	Exemplar lattices	1-best	smart	$(240, -)$	0.8355	0.7487
$\text{Lat} \xrightarrow{\text{smart}} \text{Lat}$	Exemplar lattices	Lattices	smart	$(240, 160)$	0.8421	0.7569

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

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

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

$$P_{\text{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right) = \sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right)$$

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

$$P_{\text{WTM}}\left(Q \middle| \mathrm{M}_D\right) = \prod_{w_i \in Q} \left[\sum_{w_j \in D} P_{\text{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P_{\text{MLE}}\left(w_j \middle| D\right) \right]^{c(w_i, Q)}$$

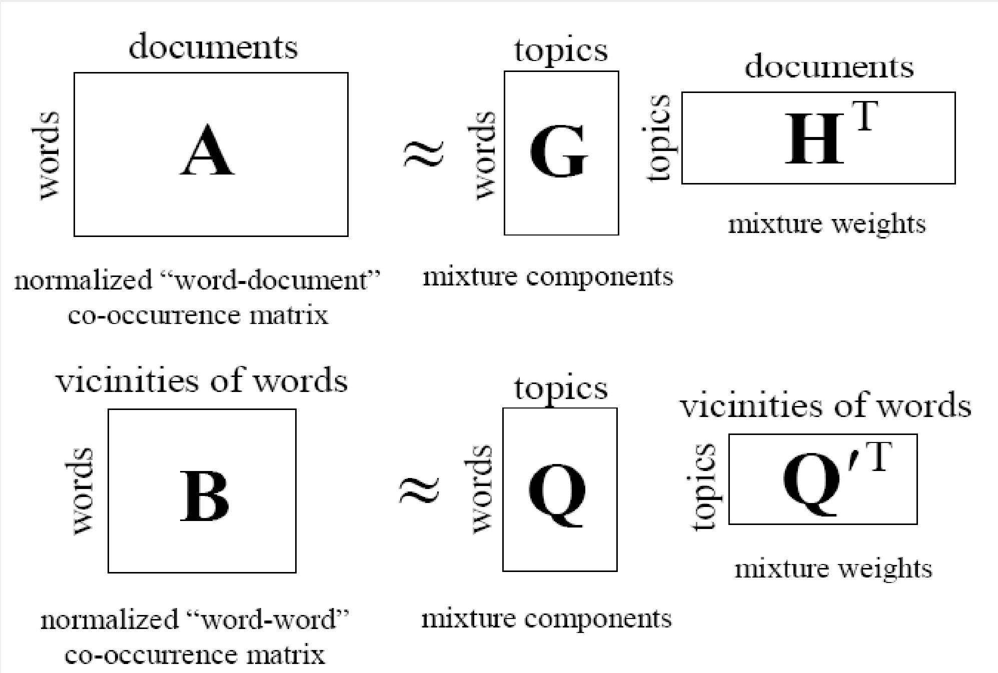
- B. Chen, "Word topic models for spoken document retrieval and transcription," *ACM TALIP*, March 2009
- B. Chen, "Latent topic modeling of word co-occurrence information for spoken document retrieval," *IEEE ICASSP 2009*

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

WTM



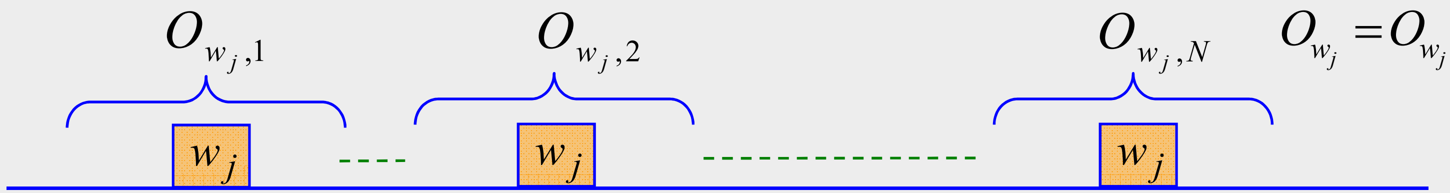
$$P_{\mathrm{PLSA}}\left(w_i \mid \mathrm{M}_D\right)=\sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_D\right)$$

$$P_{\mathrm{WTM}}\left(w_i \mid \mathrm{M}_{w_j}\right)=\sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_{w_j}\right)$$

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

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



$$O_{w_j} = O_{w_j,1}, O_{w_j,2}, \cdots, O_{w_j,N}$$

- Supervised training (WTM-S)
  - Maximize the log-likelihood of a set of training query exemplars generated by their relevant documents

$$\log L_{\mathbf{Q}_{TrainSet}} = \sum_{Q \in \mathbf{Q}_{TrainSet}} \sum_{D \in \mathbf{D}_{R \text{ to } Q}} \log P_{\text{WTM}}\left(Q \middle| \text{M}_D\right)$$

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

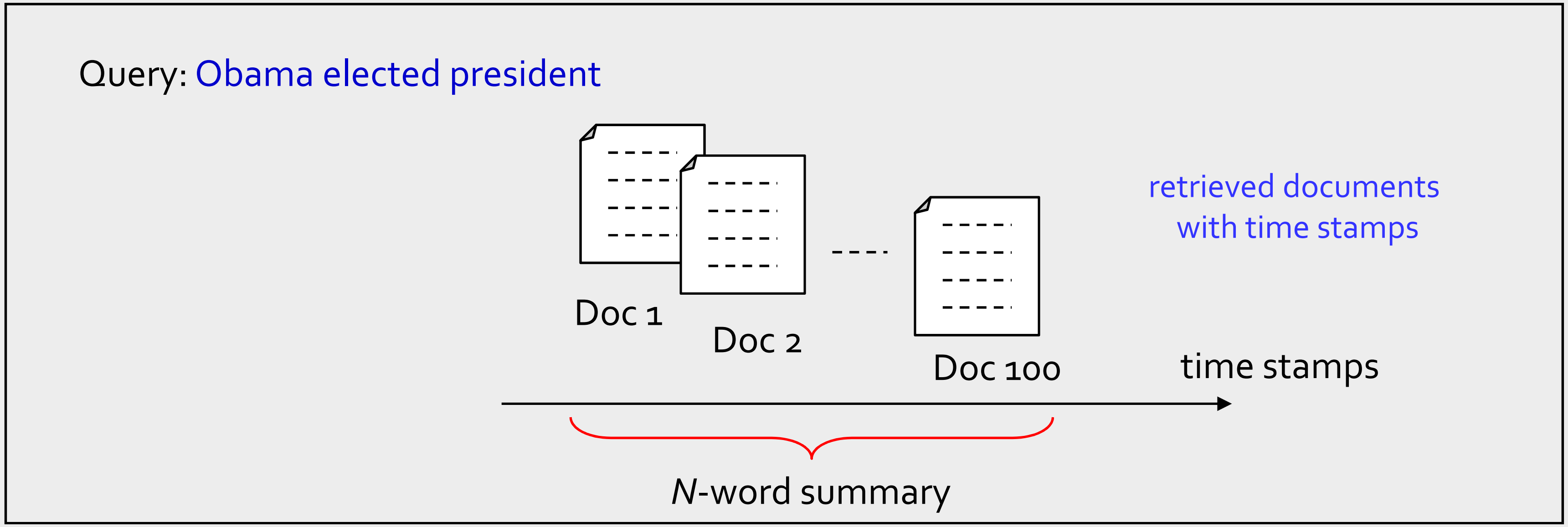
Spectrum of Text/Speech Summarization Research (1/2)

- 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

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

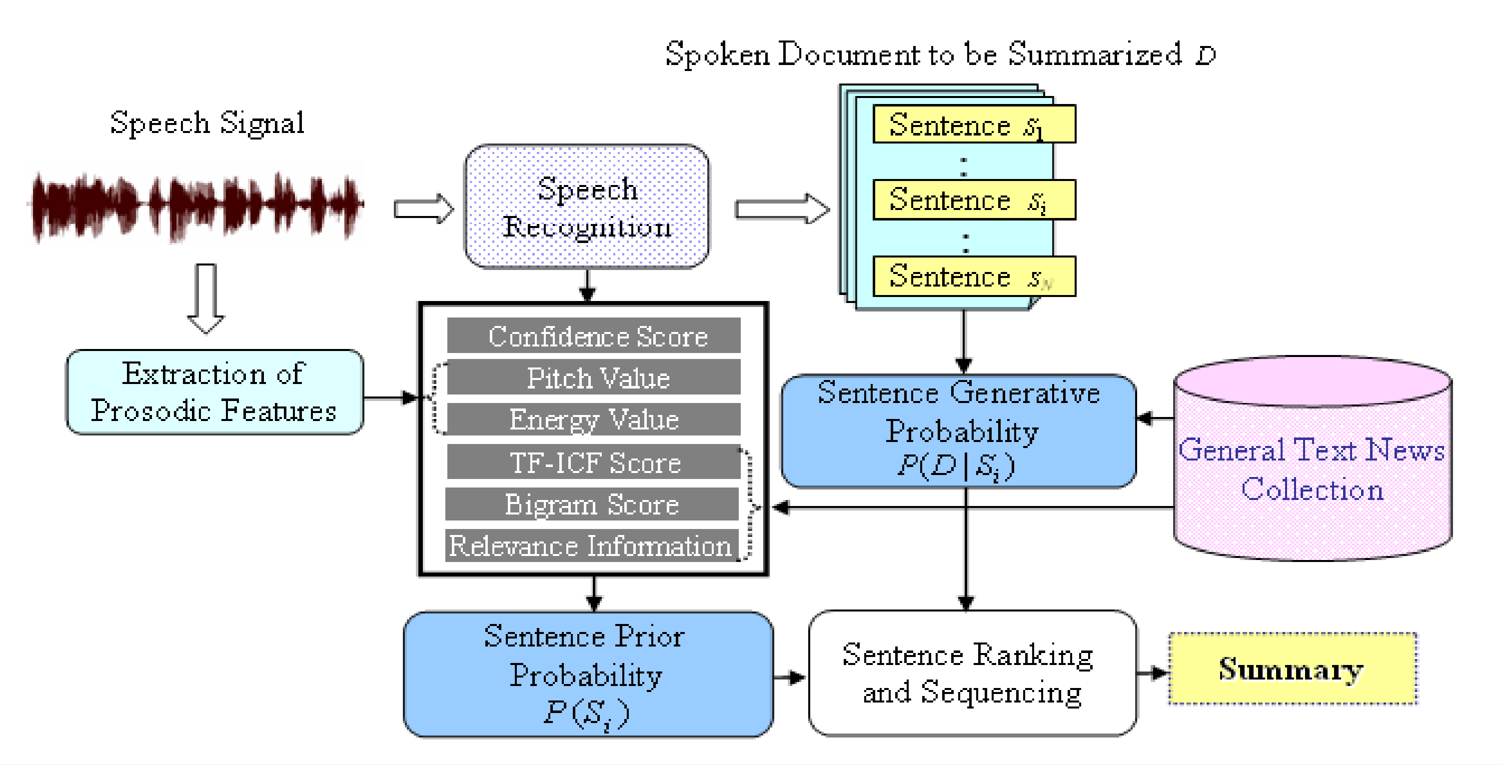
- Sentence Generative Model, $P(D|S_i)$
  - Each sentence of the document as a probabilistic generative model
  - Language Model (LM), Sentence Topic Model (STM) and Word Topic Model (WTM) are initially investigated

- 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.

Chen et al., "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Transactions on Audio, Speech and Language Processing* 17(1), 2009

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

- Speech summarization can be performed in a purely unsupervised manner



Chen et al., "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Transactions on Audio, Speech and Language Processing* 17(1), 2009

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

- Lin et al., "A comparative study of probabilistic ranking models for Chinese spoken document summarization," *ACM Transactions on Asian Language Information Processing 8(1)*, 2009

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 關鍵 在 哪裡
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
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
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
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 昆 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 娟娟 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
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 還 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

IBM TALES project

- **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 MASTOR project

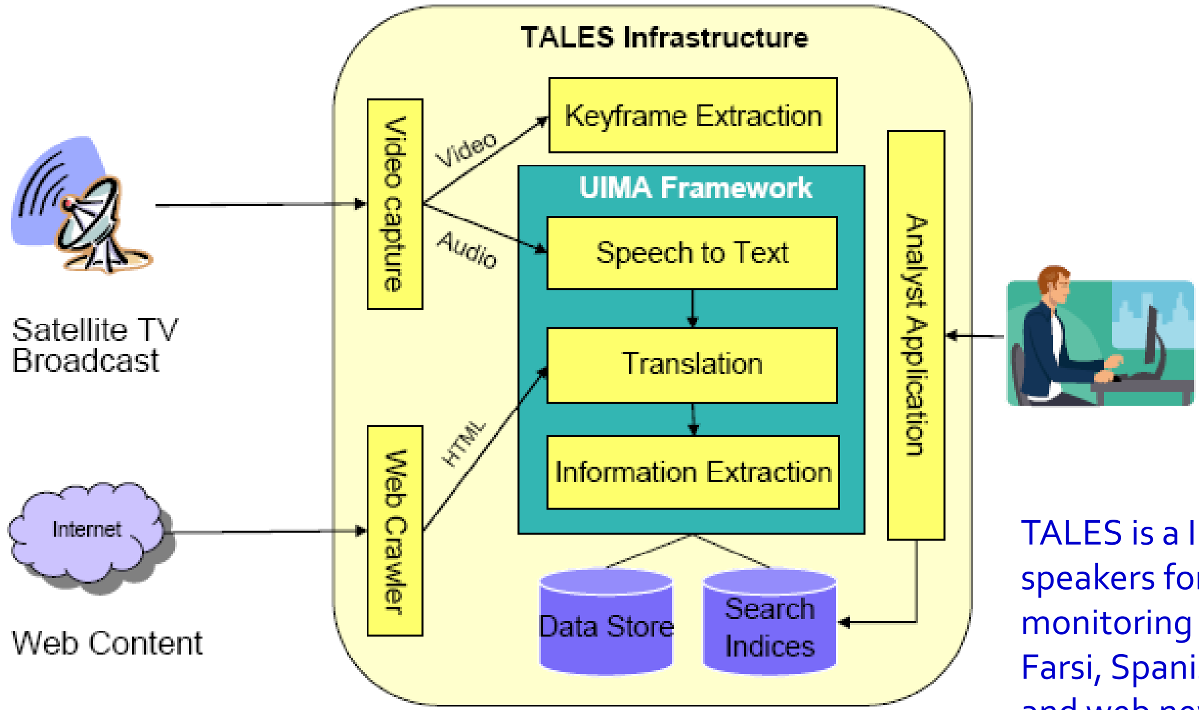
- **Speech-to-speech: for cross-lingual communication**

  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 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

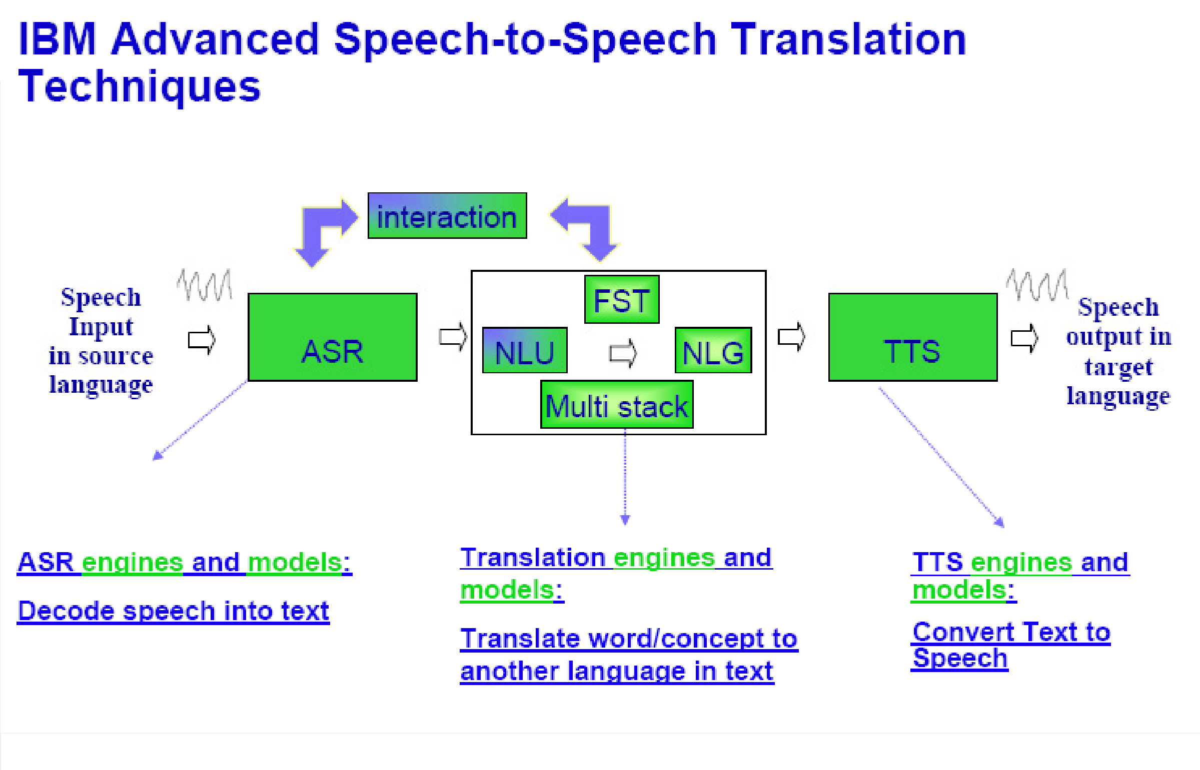
Foreign Web Site Translation and Search

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

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

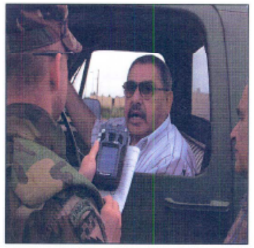


http://domino.research.ibm.com/comm/research_projects.nsf/pages/mastor.index.html

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

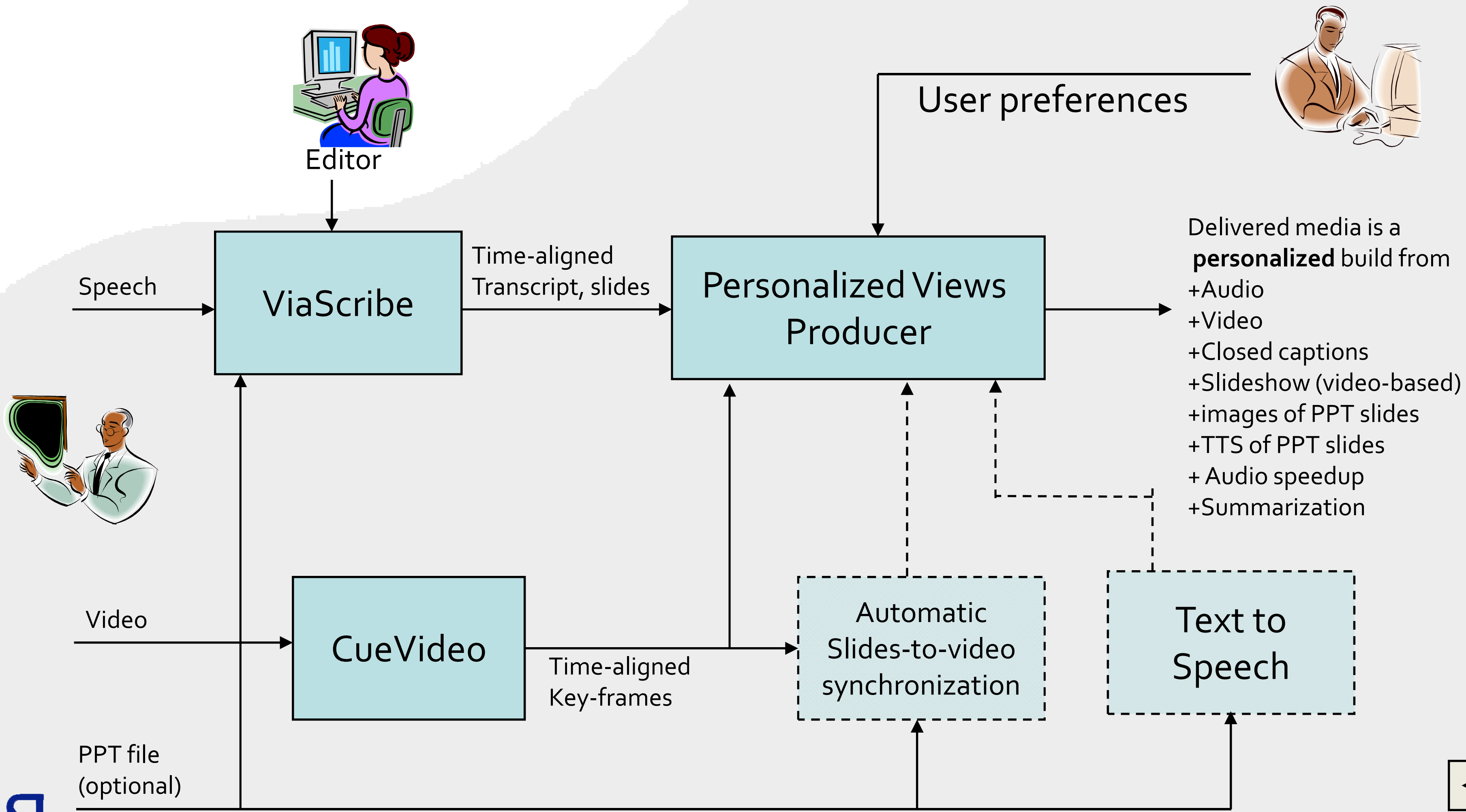
Laptop systems
- hands-free, eyes-free function

Handheld System



[MASTOR Demo]

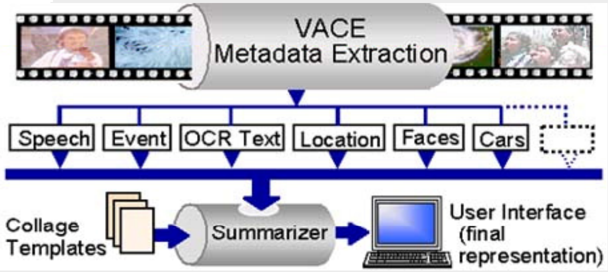
IBM's Audio-Visual Search Solutions

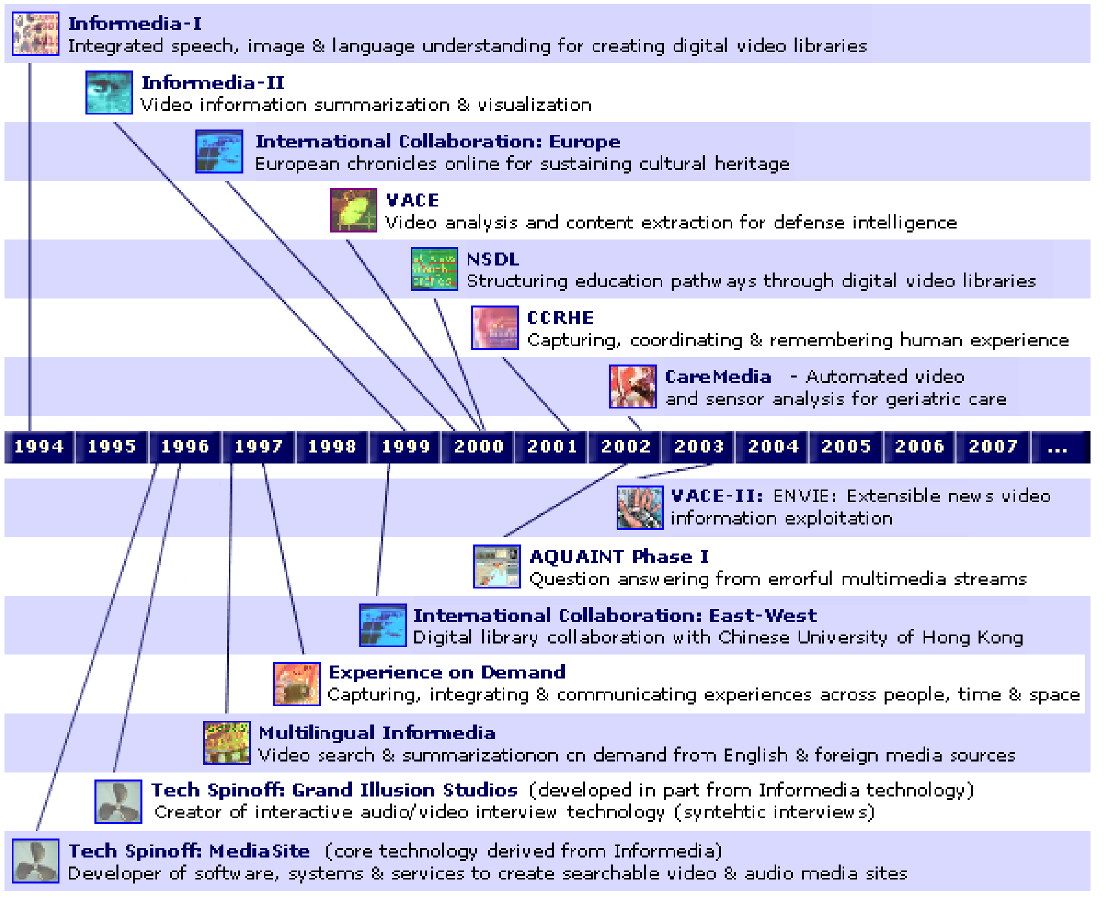


Adapted from the presentation slides of Dr. Basson et al.

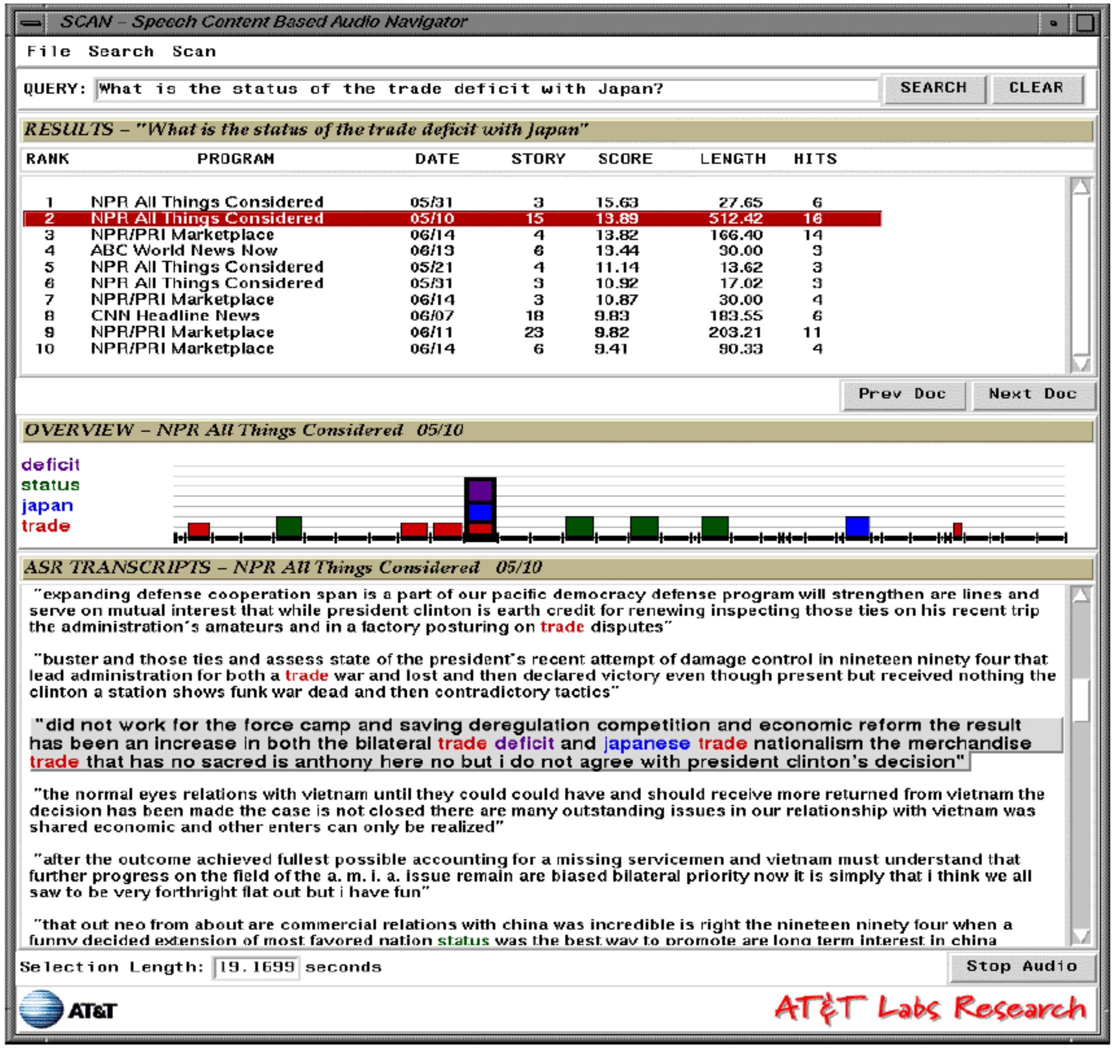
The Informedia System at CMU

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





AT&T *SCAN* System

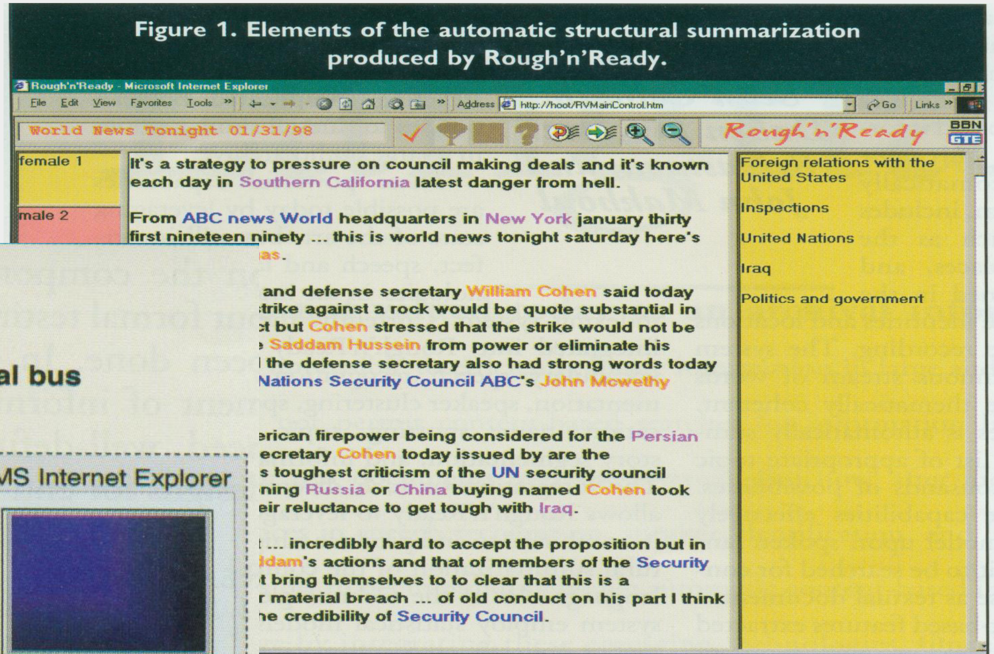


Design and evaluate user interfaces to support retrieval from speech archives

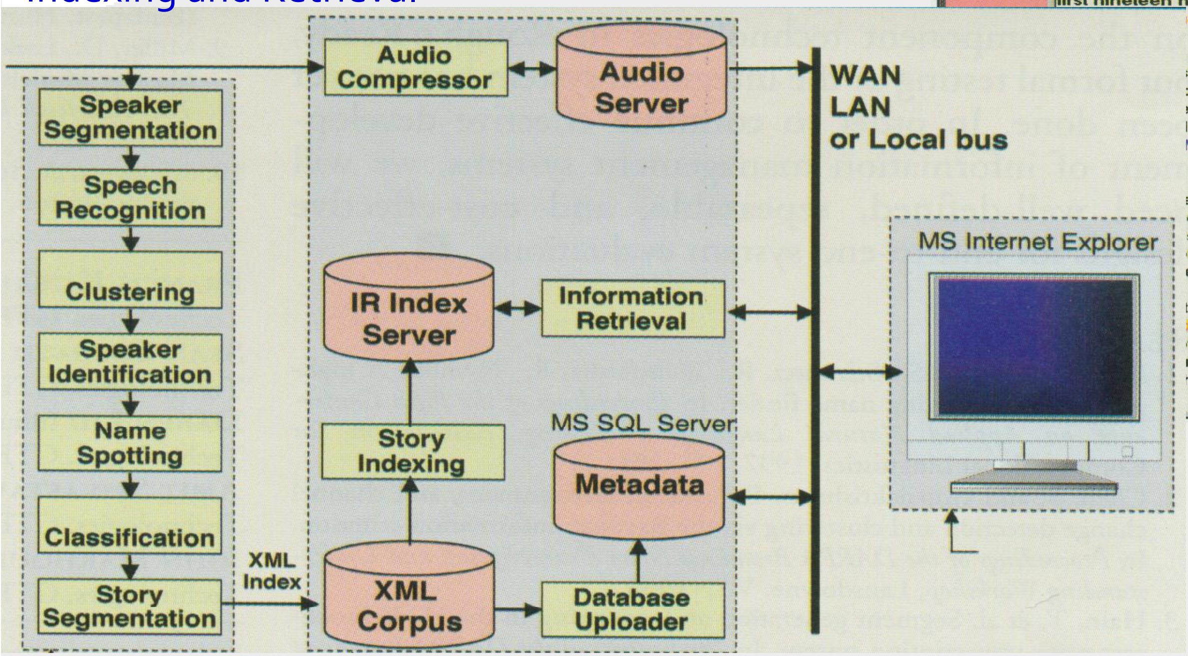
Julia Hirschberg, Fernando Pereira, Amit Singhal et al., "SCAN: designing and evaluating user interfaces to support retrieval from speech archives," SIGIR 1999

BBN *Rough'n'Ready* System

Automatic Structural Summarization for Broadcast News

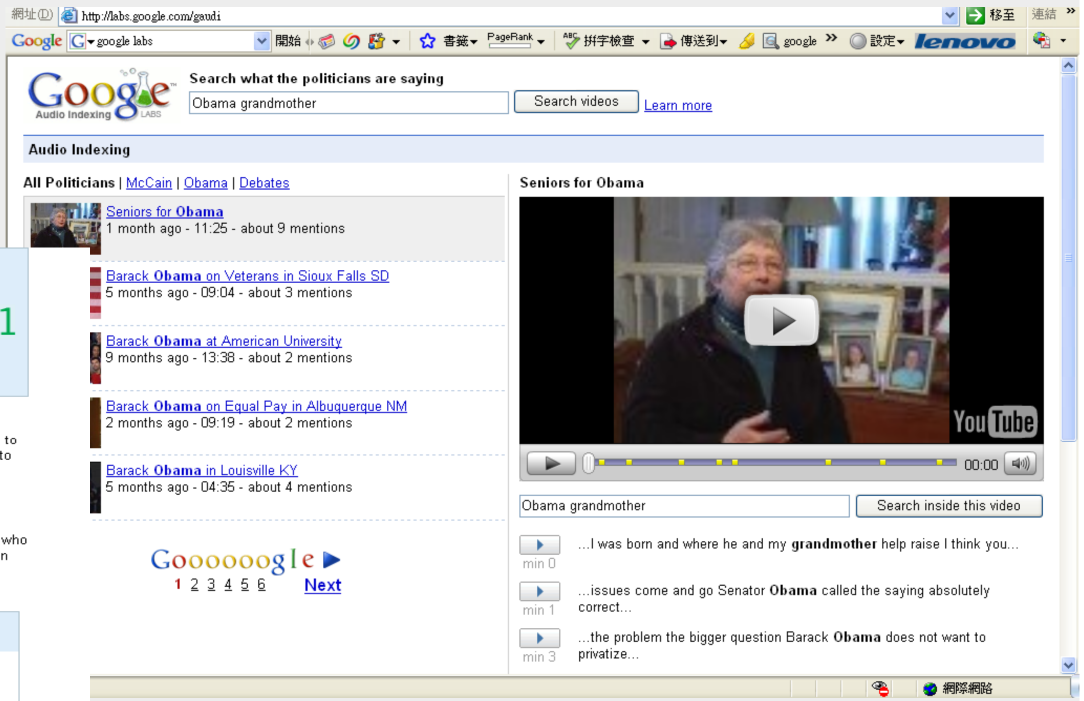


Distinguished Architecture for Audio Indexing and Retrieval



Google Voice Search

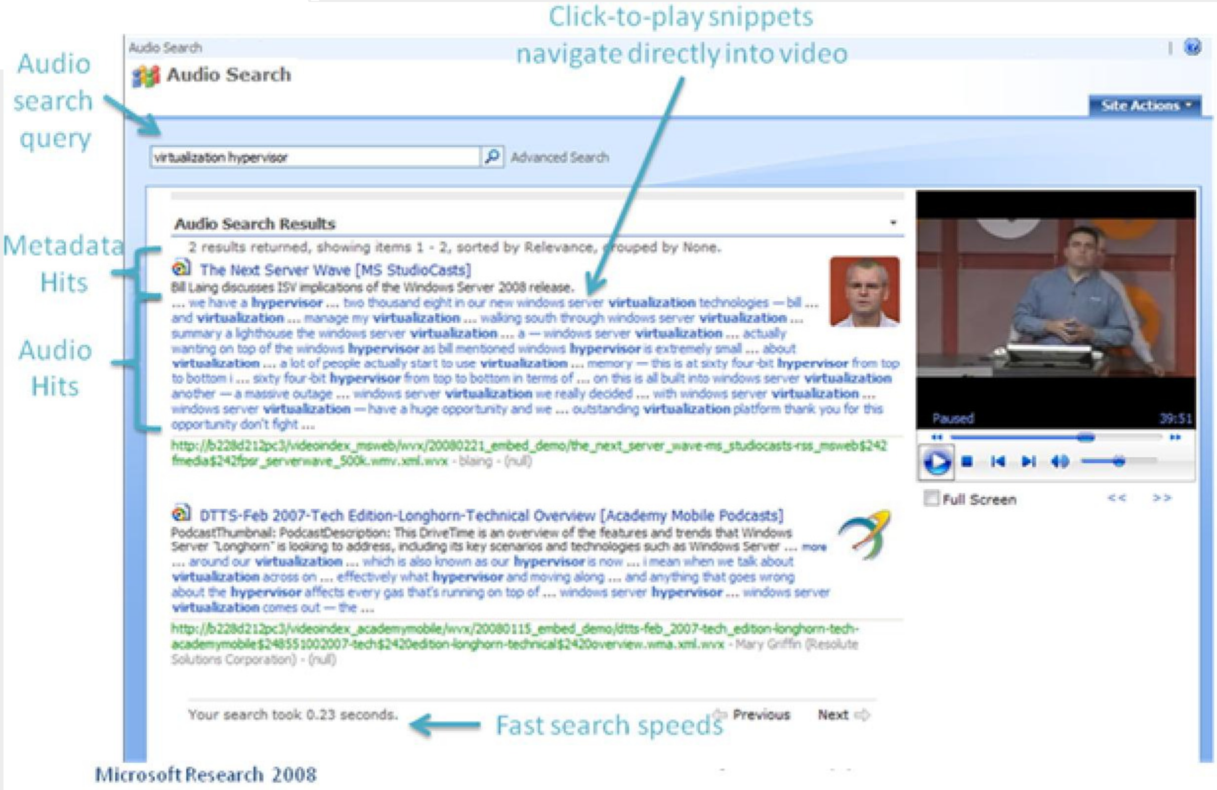
Google Audio Indexing:
Searching what people are saying inside YouTube videos
(currently only for what the politicians are saying)



Google Voice Local Search

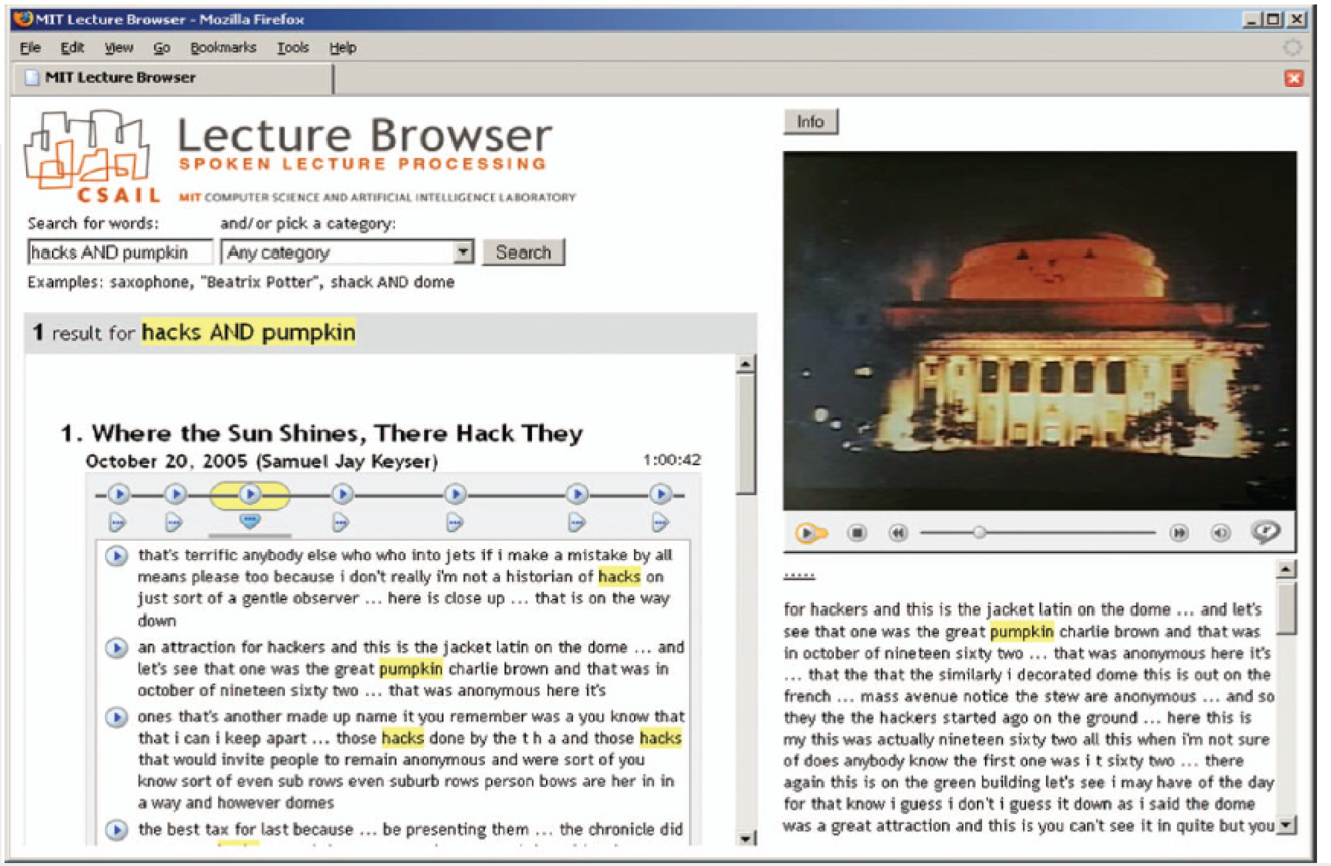
Microsoft Research Audio-Video Indexing System (MAVIS)

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

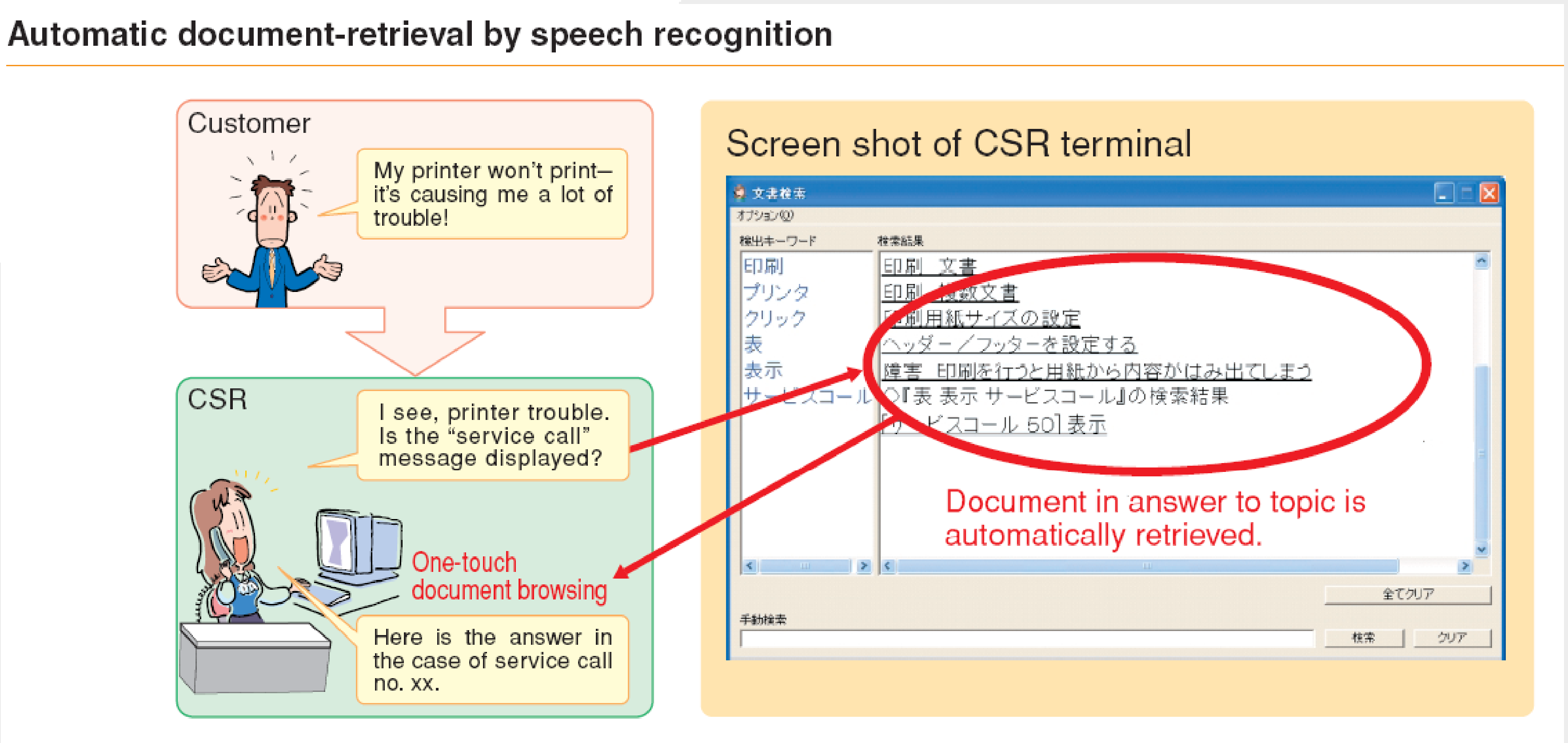


MIT Lecture Browser

- Retrieval and browsing of academic lectures of various categories



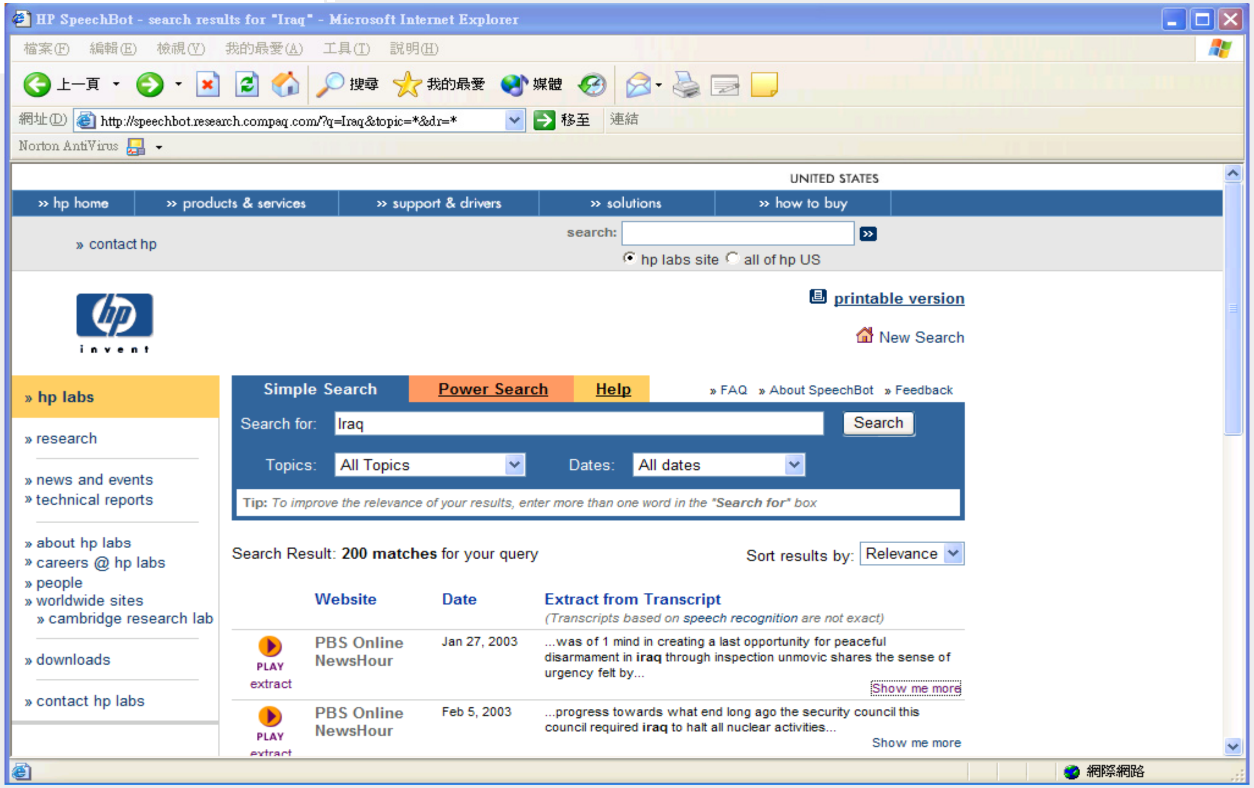
NTT Speech Communication Technology for Contact Centers



– CSR: Customer Service Representative

SpeechBot Audio/Video Search System at HP Labs

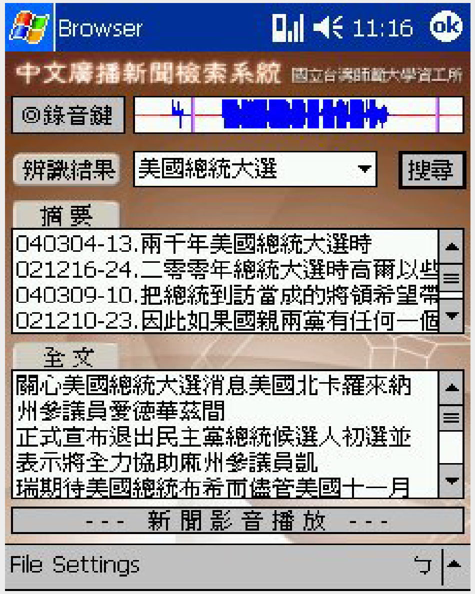
- An experimental Web-based tool from HP Labs that used voice-recognition to create searchable keyword transcripts from thousands of hours of audio content



Some Prototype Systems Developed in Taiwan

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

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



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

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

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

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

n_{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

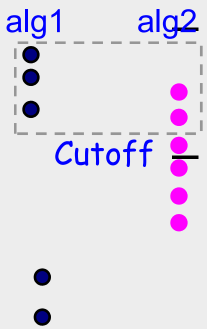
β : empirically set parameter (e.g., 1000)

Appendix A: Mean Average Precision (*m*AP) (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



- 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 (*m*AP) (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 (*m*AP)

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

- 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)

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

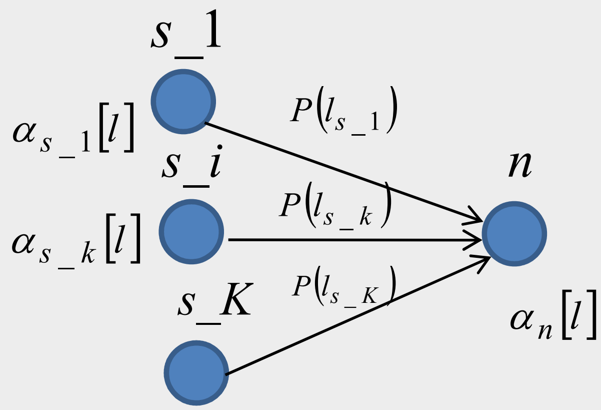
  - Defined by the sum of the insertion (*Ins*), deletion (*Del*), and substitution (*Sub*) errors between the recognized and reference word strings, divided by the total number of words in the reference string (*Ref*)

Position-Specific Posterior Probability Lattices (1/6)

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

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

position/length along the partial path traversed



- A modified forward procedure

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

insertion penalty

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

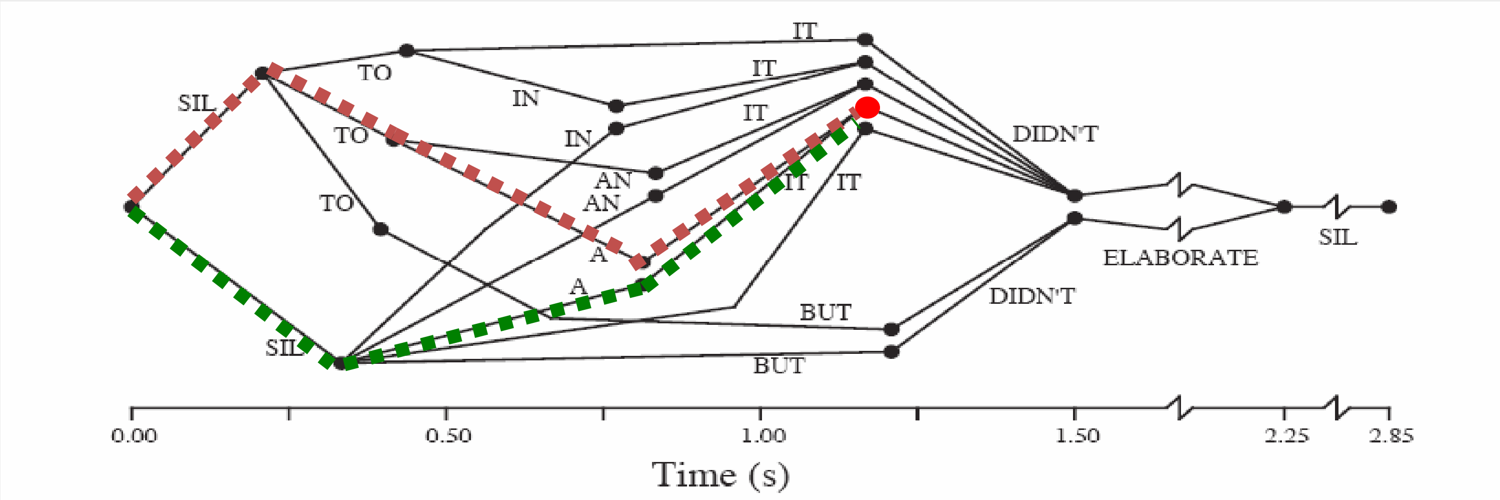
C. Chelba, J. Silva and A. Acero, "Soft indexing of speech content for search in spoken documents," *Computer Speech & Language* 21(3), 2007

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*)

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



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 $Q = q_1, q_2, \cdots, q_M$

unigram matching

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

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

$$P(w, l \mid LAT)$$

N-gram (or phrase) matching

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

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

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

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

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

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

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

    - 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

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

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

"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.

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

Uniform posterior prob. 1.0/#entries in each position

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