

# Spoken Document Recognition, Organization and Retrieval

語音文件辨識、組織與檢索



Berlin Chen (陳柏琳)

Associate Professor

Department of Computer Science & Information Engineering  
National Taiwan Normal University



2008/12/14

# Outline

---

- Introduction
- Related Research Work and Applications
- Key Techniques
  - Automatic Speech Recognition (ASR)
  - Information Retrieval (IR)
  - Spoken Document Summarization
  - Spoken Document Organization
- Prototype Systems Developed at NTNU
- Conclusions and Future Work

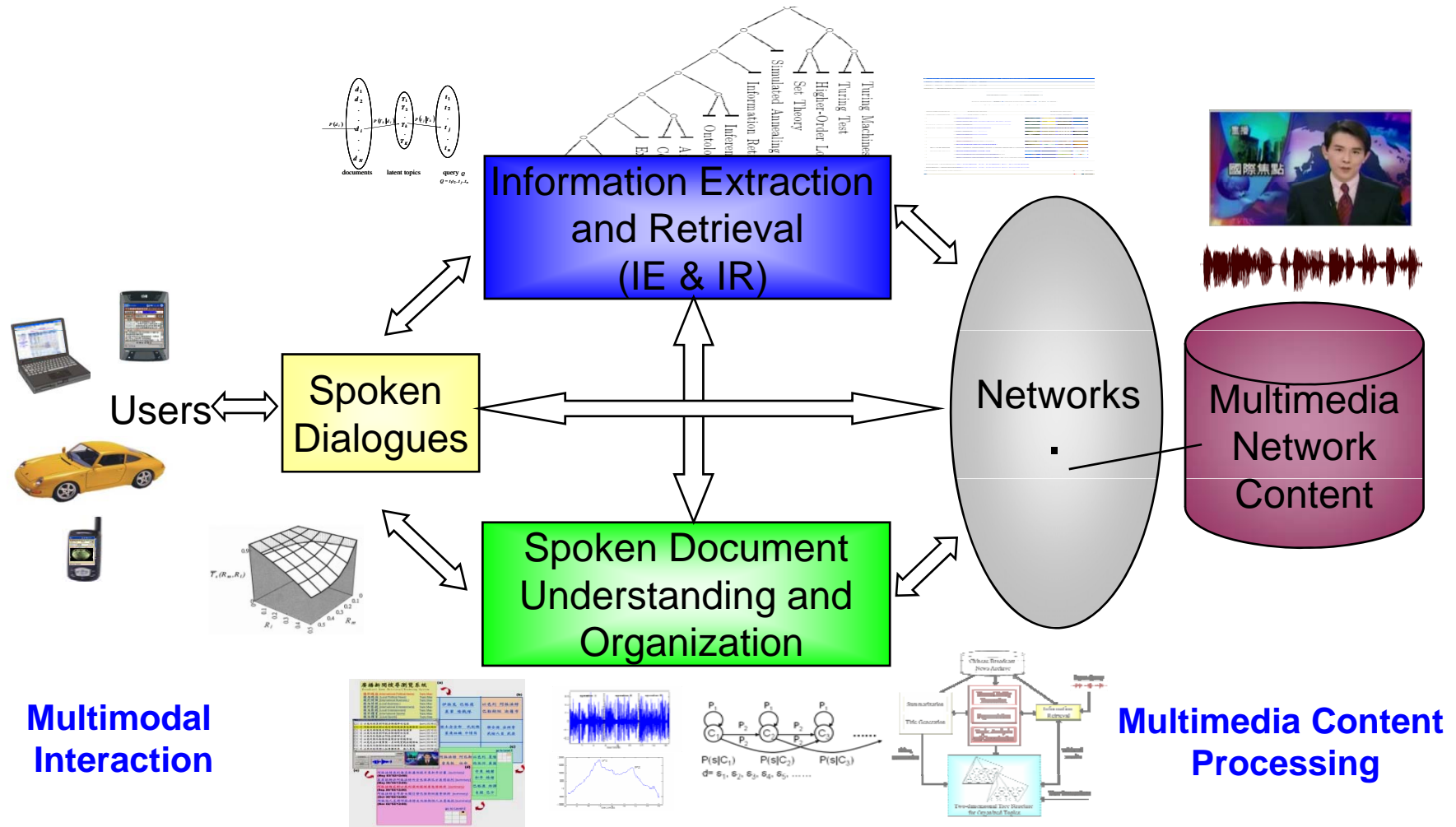
## Introduction (1/3)

---

- Multimedia associated with speech is continuously growing and filling our computers and lives
  - Such as broadcast news, lectures, historical archives, voice mails, (contact-center) conversations, etc.
  - Speech is the most semantic (or information)-bearing
- 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 and on smaller hand-held devices
- Speech will be the key for Multimedia information access in the near future

# Introduction (2/3)

- Scenario for Multimedia information access using speech



## Introduction (3/3)

---

- 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
    - 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 more attractive than text
    - Better approaches for efficient organization and retrieval of multimedia (spoken) documents are needed

# Related Research Work and Applications

---

- Substantial efforts have been paid to (multimedia) spoken document recognition, organization and retrieval in the recent past [R3, R4]
  - [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 Spoken Document Retrieval for Call-Center Conversations, Natural Language Call-Routing, Voicemail Retrieval*
  - [NTT Speech Communication Technology for Contact Centers](#)
  - [Google Voice Local Search](#)



- 
- Automatic Speech Recognition
    - 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
  - Audio Indexing and Information Retrieval
    - Robust representation of the spoken documents
    - Matching between (spoken) queries and spoken documents
  - Named Entity Extraction from Spoken Documents
    - Personal names, organization names, location names, event names
    - Very often out-of-vocabulary (OOV) words, difficult for recognition

## Key Techniques (2/2)

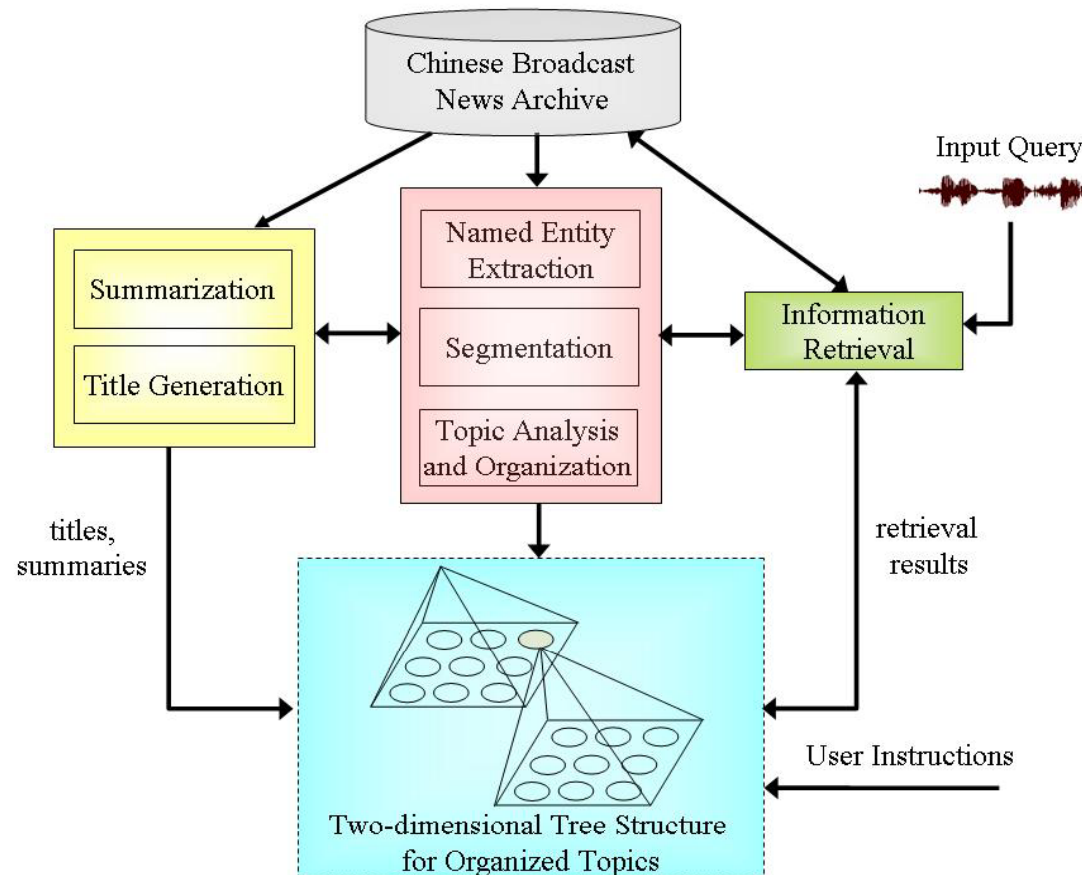
---

- Information Extraction for Spoken Documents
  - Extraction of key information such as who, when, where, what and how for the information described by spoken documents
- Summarization for Spoken Documents
  - Automatically generate a summary (in text or speech form) for each spoken document
- 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 the documents
  - Organize the subject topics of the documents into graphic structures for efficient browsing



# An Example System for Chinese Broadcast News (1/2)

- For example, a prototype system developed at NTU for efficient spoken document retrieval and browsing [R4]



# An Example System for Chinese Broadcast News (2/2)

- Users can browse spoken documents in top-down and bottom-up manners

**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.] 02.09.20  
 [ 2 ] 阿拉法特反對以色列保所提結束包圍條件 [sum.] 02.09.20  
 [ 3 ] 以色列部隊進攻阿拉法特總部後撤軍 [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區 [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.] 02.06.06  
 [ 8 ] 以色列巴勒斯坦就伯利恆撤軍達成協議 [sum.] 02.02.12  
 [ 9 ] 以色列坦克闖入加薩難民營 兩人喪生 [sum.] 02.04.20

伊拉克 巴格達 美軍 陸戰隊	以色列 阿拉法特 巴勒斯坦 迦薩市
國土安全部 民航機 蓋達組織 中情局	聯合國 安理會 武檢人員 武器

阿拉法特 阿巴斯  
雷馬拉 任命

以色列 夏隆  
約旦河 美國

中東 鮑爾  
和平 路線

巴格達 炸彈  
自殺 巴士

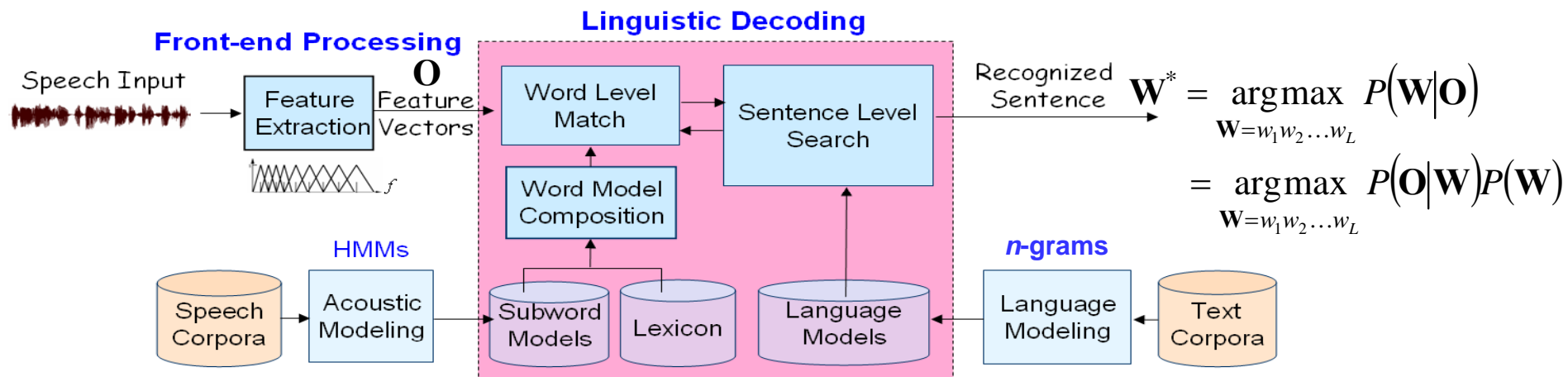
阿拉法特原則接受歐盟所提中東和平計畫 [summary] (May 03/02/12:00)  
 英美就解決阿拉法特所受包圍與巴方展開談判 [summary] (May 06/02/12:00)  
 阿拉法特反對以色列保所提結束包圍條件 [summary] (Sep 20/02/12:00)  
 阿拉法特宣布新內閣引發巴勒斯坦國會激辯 [summary] (Oct 30/02/12:00)  
 阿拉伯人支持阿拉法特及巴勒斯坦人正當抵抗 [summary] (Nov 02/02/12:00)



<http://sovideo.iis.sinica.edu.tw/NeGSST/Index.htm>

# Automatic Speech Recognition (1/3)

- Large Vocabulary Continuous Speech Recognition (LVCSR)





- The speech signal is converted into a sequence of feature vectors
- The pronunciation lexicon is structured as a tree
- Due to the constraints of  $n$ -gram language modeling, a word's occurrence is dependent on its previous  $n-1$  words

$$P(w_i | w_1 w_2 \dots w_{i-1}) \approx P(w_i | w_{i-n+1} w_{i-n+2} \dots w_{i-1})$$

- Search through all possible lexical tree copies from the start time to the end time of the utterance to find the best sequence among the word hypotheses

# Automatic Speech Recognition (2/3)

---

- Discriminative and Robust Speech Feature Extraction
  - Linear Discriminant Analysis (LDA), Heteroscedastic Linear Discriminant Analysis (HLDA) and Maximum Likelihood Linear Transformation (MLLT) for discriminative speech feature extraction *Interspeech 2005, 2008; ISCSLP2008*
  - Polynomial-fit Histogram Equalization (PHEQ) Approaches for robust speech feature extraction *Interspeech 2006, 2007, 2008; ICME 2007; ASRU 2007* 
- Acoustic Modeling
  - Lightly-Supervised Training of Acoustic Models *ICASSP 2004*
  - Data Selection for Discriminative Training of Acoustic Models (HMMs) *ICME 2007, ASRU 2007*
- Dynamic Language Model Adaptation
  - Minimum Word Error (MWE) Training *Interspeech 2005*
  - Word Topical Mixture Models (WTMM) *ICME2005; ICASSP 2007*
  - Position Information for Language Modeling *ISCSLP2008*
- Linguistic Decoding 
  - Syllable-level acoustic model look-ahead *ICASSP 2004*

# Automatic Speech Recognition (3/3)

## • Transcription of PTS (Taiwan) Broadcast News



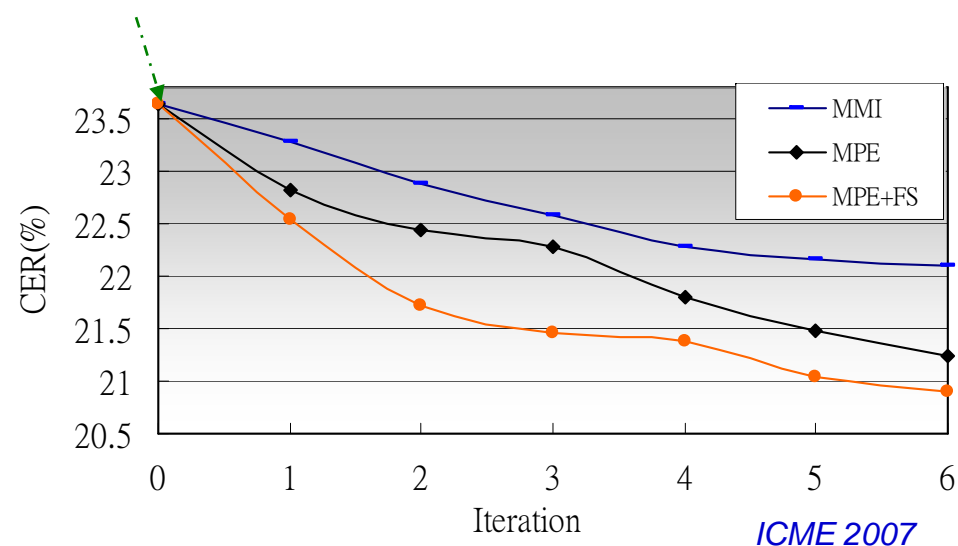
根據最新但雨量統計  
一整天下來  
費雪以試辦兩個水庫的雨量  
分別是五十三公里和二十九公厘  
對水位上升幫助不大  
不過就業機會期間也多在夜間  
氣象局也針對中部以北及東北部地區發佈豪雨特報  
因此還是有機會增加積水區的降雨量  
此外氣象局也預測  
華航又有另一道鋒面通過  
水利署估計如果這波鋒面能帶來跟著會差不多的雨水  
那個北台灣的第二階段限水時間  
渴望見到五月以後  
公視新聞當時匯率採訪報導

Automatic

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

Manual

10 Iterations of  
ML training



Relative Character Error Rate Reduction

- MMI: 6.5%
- MPE: 10.1%
- MPE+FS: 11.6%

# Information Retrieval Models

- Information retrieval (IR) models can be characterized by two different matching strategies
  - Literal term matching
    - Match queries and documents in an index term space
  - Concept matching
    - Match queries and documents in a latent semantic space



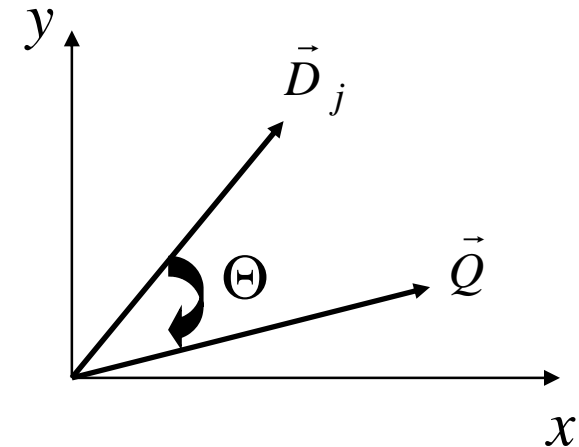
香港星島日報篇報導引述軍事觀察家的話表示，到二零零五年台灣將完全喪失空中優勢，原因是中國大陸戰機不論是數量或是性能上都將超越台灣，報導指出中國在大量引進俄羅斯先進武器的同時也得加快研發自製武器系統，目前西安飛機製造廠任職的改進型飛豹戰機即將部署尚未與蘇愷三十通道地對地攻擊住宅飛機，以督促遇到挫折的監控其戰機目前也已經取得了重大階段性的認知成果。根據日本媒體報導在台海戰爭隨時可能爆發情況之下北京方面的基本方針，使用高科技答應局部戰爭。因此，解放軍打算在二零零四年前又有包括蘇愷三十二期在內的兩百架蘇霍伊戰鬥機。

# IR Models: Literal Term Matching (1/2)

---

- Vector Space Model (VSM)
  - Vector representations are used for queries and documents
  - Each dimension is associated with a index term (TF-IDF weighting)
  - Cosine measure for query-document relevance

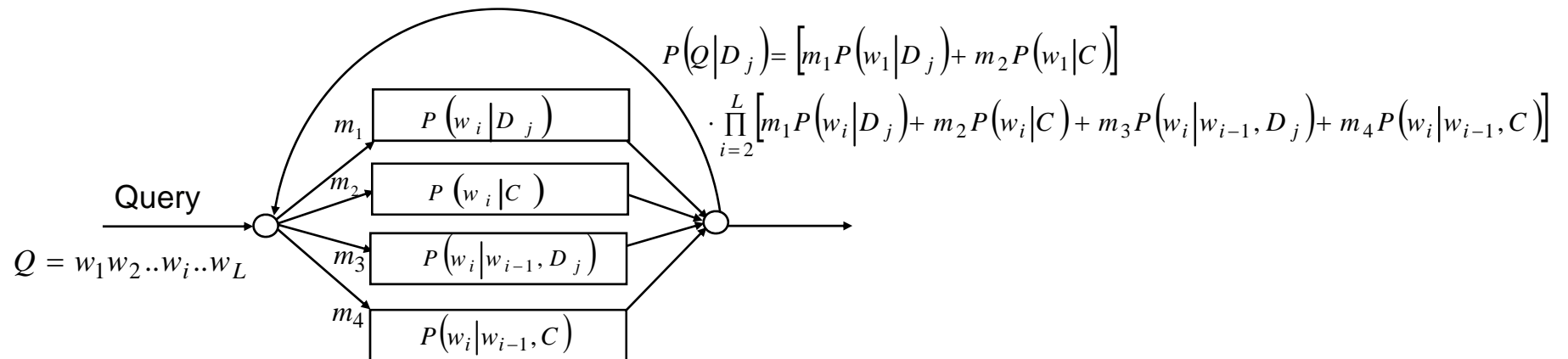
$$\begin{aligned} sim(D_j, Q) \\ &= \text{cosine}(\Theta) = \frac{\vec{D}_j \cdot \vec{Q}}{|\vec{D}_j| \times |\vec{Q}|} \\ &= \frac{\sum_{i=1}^n w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^n w_{i,j}^2} \times \sqrt{\sum_{i=1}^n w_{i,q}^2}} \end{aligned}$$



- VSM can be implemented with an inverted file structure for efficient document search

## IR Models: Literal Term Matching (2/2)

- Hidden Markov Models (HMM) [R1]
  - Also thought of as Language Modeling (LM) approaches
  - Each document is a probabilistic generative model consisting of a set of  $N$ -gram distributions for predicting the query

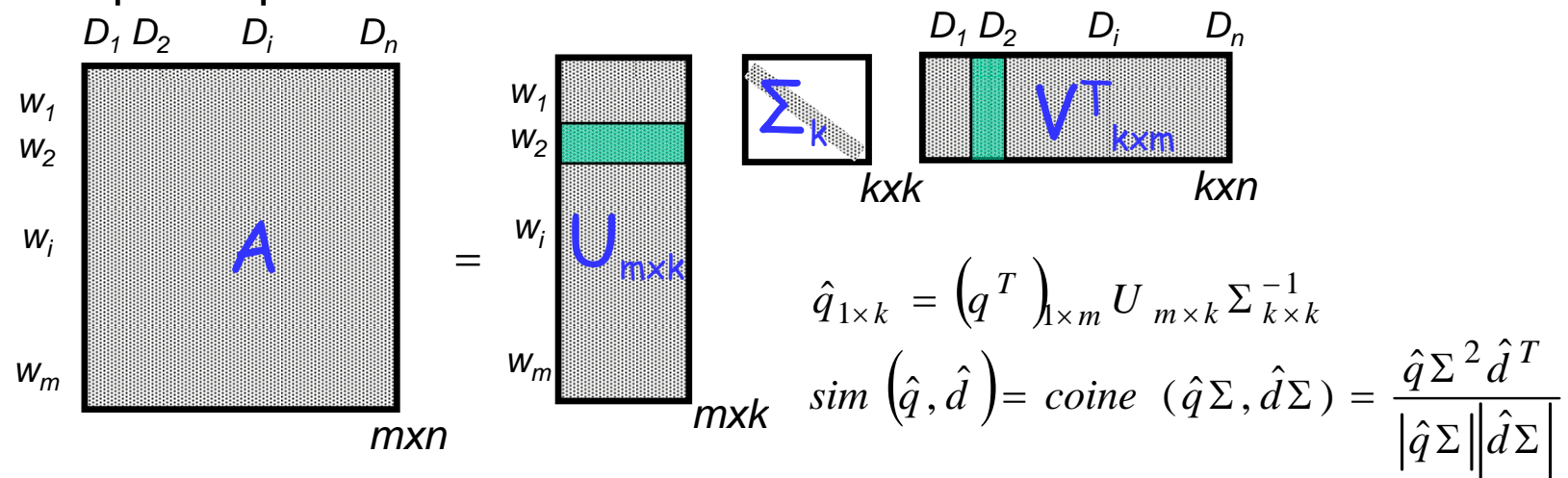


- Models can be optimized by the expectation-maximization (EM) or minimum classification error (MCE) training algorithms
- Such approaches do provide a potentially effective and theoretically attractive probabilistic framework for studying IR problems



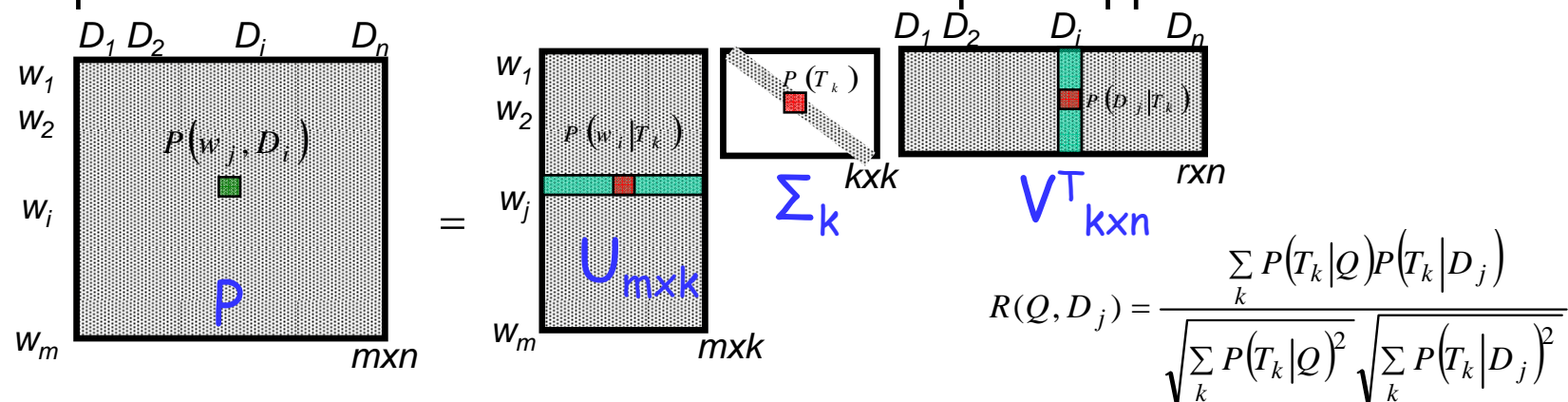
# IR Models: Concept Matching (1/3)

- Latent Semantic Analysis (LSA) [R2]
  - Start with a matrix describing the intra- and Inter-document statistics between all terms and all documents
  - Singular value decomposition (SVD) is then performed on the matrix to project all term and document vectors onto a reduced latent topical space
  - Matching between queries and documents can be carried out in this topical space



# IR Models: Concept Matching (2/3)

- Probabilistic Latent Semantic Analysis (PLSA) [R5, R6]
  - An probabilistic framework for the above topical approach

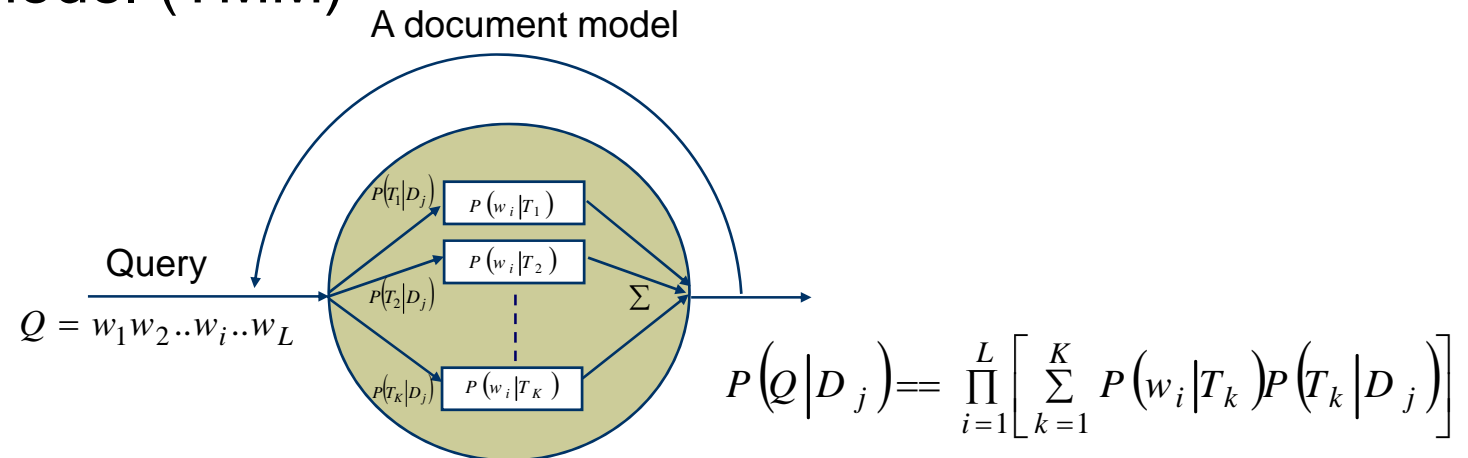


$$P(w_i, D_j) \approx \sum_k P(w_i | T_k) P(T_k) P(D_j | T_k)$$

- Relevance measure is not obtained directly from the frequency of a respective query term occurring in a document, but has to do with the frequency of the term and document in the latent topics
- A query and a document thus may have a high relevance score even if they do not share any terms in common

# IR Models: Concept Matching (3/3)

- PLSA also can be viewed as an HMM model or a topical mixture model (TMM)



- Explicitly interpret the document as a mixture model used to predict the query, which can be easily related to the conventional HMM modeling approaches widely studied in speech processing community (topical distributions are tied among documents)
- Thus quite a few of theoretically attractive model training algorithms can be applied in supervised or unsupervised manners

# IR Evaluations

---

- Experiments were conducted on TDT2/TDT3 spoken document collections [R6]
  - TDT2 for parameter tuning/training, while TDT3 for evaluation
  - E.g., mean average precision (*mAP*) tested on TDT3

	VSM	LSA	TMM	HMM	PLSA
TD	0.6505	0.6440	0.7870	0.7174	0.6882
SD	0.6216	0.6390	0.7852	0.7156	0.6688

*TALIP2004; Interspeech2004, 2005*

- HMM/PLSA/TMM are trained in a supervised manner
- Language modeling approaches (TMM/PLSA/HMM) are evidenced with significantly better results than that of conventional statistical approaches (VSM/LSA) in the above spoken document retrieval (SDR) task

# Spoken Document Summarization

---

- Spoken document summarization (SDS), which aims to generate a summary automatically for the spoken documents, is the key for better speech understanding and organization
- **Extractive** vs. **Abstractive** Summarization
  - **Extractive summarization** is to select a number of indicative sentences or paragraphs from original document and sequence them to form a summary
  - **Abstractive summarization** is to rewrite a concise abstract that can reflect the key concepts of the document
  - Extractive summarization has gained much more attention in the recent past

# SDS: Proposed Framework (1/2)

---

- **A Probabilistic Generative Framework for Sentence Selection (Ranking)**

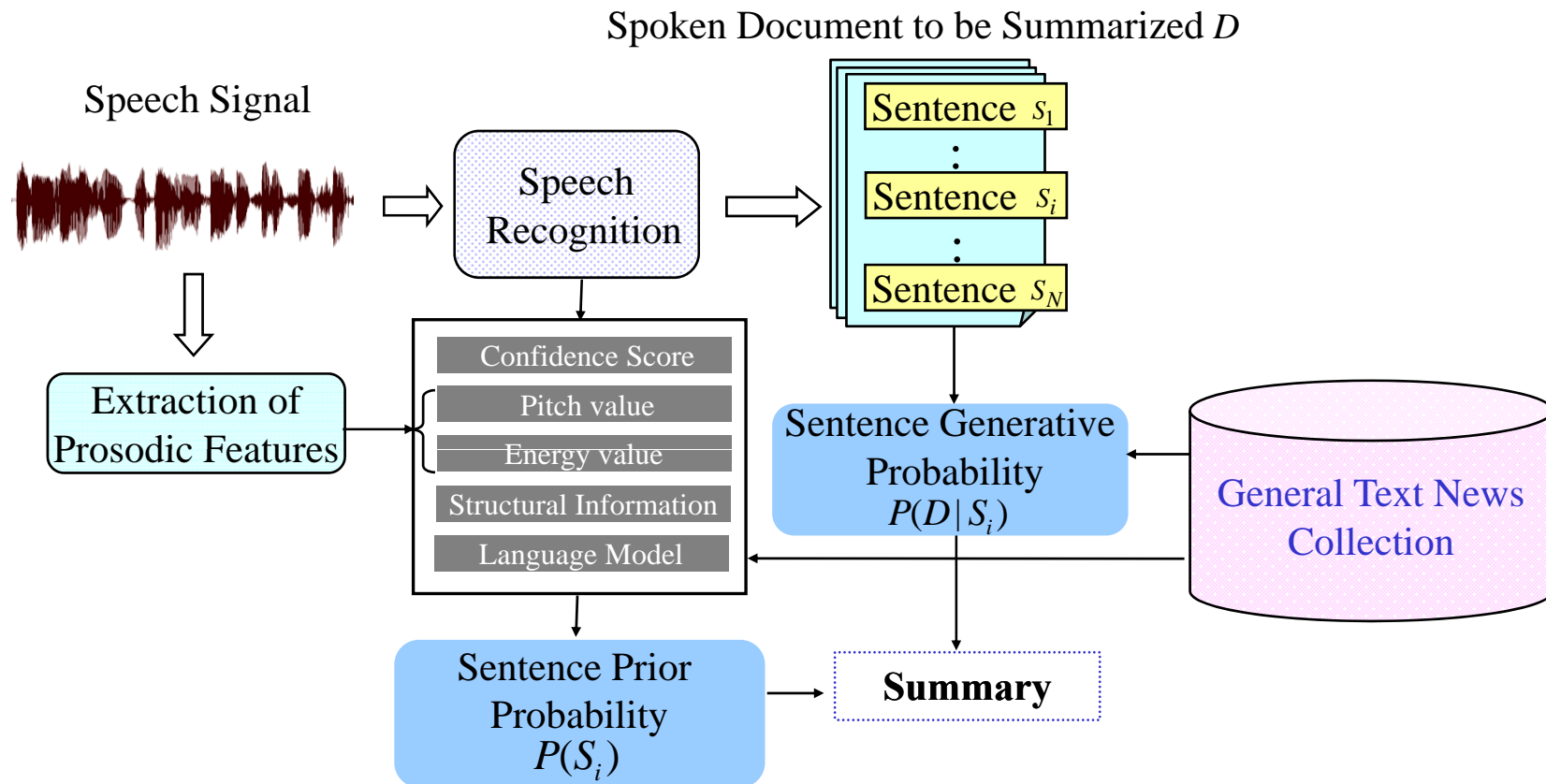
- Maximum a posteriori (MAP) criterion

$$\arg \max_s P(S|D_i) = \arg \max_s \frac{P(D_i|S)P(S)}{P(D_i)} = \arg \max_s P(D_i|S)P(S)$$

- **Sentence Generative Model,  $P(D_i|S)$**  ICASSP2006, ISCSLP2006, ICME 2007  
ICASSP2008
  - Each sentence of the document as a probabilistic generative model
  - Hidden Markov Model (HMM), Topical Mixture Model (TMM) and Word Topical Mixture Model (WTMM) are initially investigated
- **Sentence Prior Distribution,  $P(S)$**  Interspeech2007, ASRU 2007
  - The sentence prior distribution may have to do with sentence *duration/position, correctness of sentence boundary, confidence score, prosodic information*, etc. (information sources integrated by Whole-Sentence Maximum Entropy Model)

# SDS: Proposed Framework (2/2)

- A flowchart for our proposed framework



# SDS: Evaluation

- Preliminary tests on 200 radio broadcast news stories collected in Taiwan (automatic transcripts with 14.17% character error rate)
  - ROUGE-2 measure was used to evaluate the performance levels of different models

	VSM	MMR	LSA	SIG	HMM	TMM	WTMM	Random
10%	0.2845	0.2875	0.2755	0.2760	0.2989	0.3043	0.3193	0.1122
20%	0.3110	0.3218	0.2911	0.3190	0.3295	0.3345	0.3437	0.1263
30%	0.3435	0.3493	0.3081	0.3491	0.3670	0.3688	0.3716	0.1834
50%	0.4565	0.4668	0.4070	0.4804	0.4743	0.4753	0.4676	0.3096

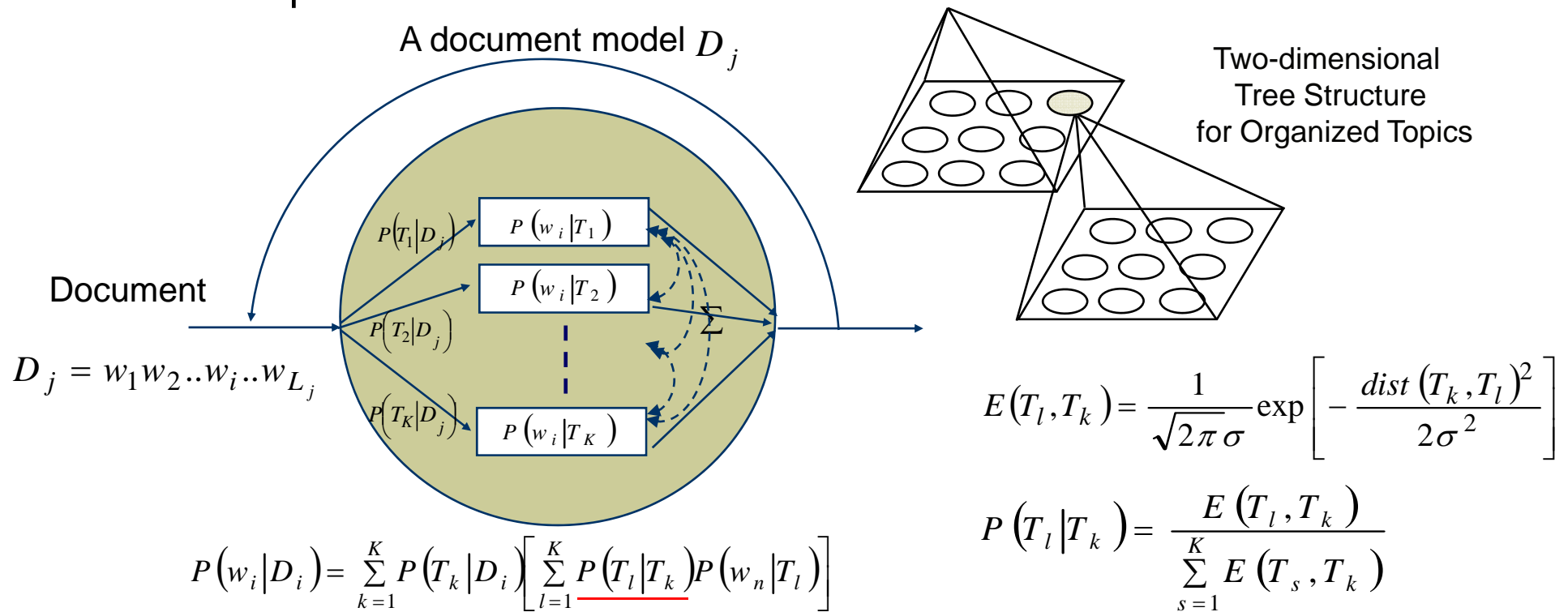


- Proposed models (HMM, TMM, WTMM) are consistently better than the other models at lower summarization ratios
  - HMM, TMM, WTMM are trained without any document-summary relevance information (labeled data)



# Spoken Document Organization (1/3)

- Each document is viewed as a TMM model to generate itself
  - Additional transitions between topical mixtures have to do with the topological relationships between topical classes on a 2-D map



## Spoken Document Organization (2/3)

---

- Document models can be trained in an unsupervised way by maximizing the total log-likelihood of the document collection

$$L_T = \sum_{j=1}^n \sum_{i=1}^V c(w_i, D_j) \log P(w_i | D_j)$$

- Initial evaluation results (conducted on the TDT2 collection)

Model	Iterations	dist <sub>Between</sub> /dist <sub>Within</sub>
TMM	10	1.9165
SOM	100	2.0604

- TMM-based approach is competitive to the conventional Self-Organization Map (SOM) approach

# Spoken Document Organization (3/3)

- Each topical class can be labeled by words selected using the following criterion

$$\text{Sig}(w_i, T_k) = \frac{\sum_{j=1}^n c(w_i, D_j) P(T_k | D_j)}{\sum_{i=1}^n c(w_i, D_j) [1 - P(T_k | D_j)]}$$

- An example map for international political news

<p>聯邦調查局 執法 劃歸 空對空飛彈            安全部 艾希克羅 蓋達組織 接種            等級 民航機 認出 輻射性            劫機 主謀 重警旗鼓 歐瑪            穆勒 國土 黃色 塞門            美國境內 中情局 天花 丙吉</p>	<p>僑界 僑務 台商 會長            僑胞 呼吸 雙十國慶 酒會            立委 舉辦 國慶 聯誼會            經文 履新 組長 衛生            餐會 春節 滬太華 後援            中華 僑團 華僑 鄉親</p>	<p>法輪 鈴木宗男 巫統 中國共產黨            李光耀 挪用 書記 交替            班子 馬哈地 一邊 李顯龍            吳作棟 新疆 論說 軍委            政治局 標題 馬來人 早報            格局 資政 接班 報章</p>
<p>檢查人員 檢查員 動武 最後通牒            安理會 布里克斯 決議 精密            武檢 聯合國 授權 沙丹·            銷毀 遠禁 解除 武檢人員            檢查 首席 武器 決議案            胡笙 禁航區 導引 毀滅性</p>	<p>西非 衛隊 巴格達機場 伊拉克部隊            伊拉克南部 賴比瑞亞 伊北 科威特            步兵 辛格 庫德族 斯拉            法新社 翁山蘇姬 庫克 蒙羅維亞            巴格達 陸戰隊 轟炸 激戰            卡達 克里 市中心 基爾</p>	<p>林東源 金大中 漢城 南北            多邊 正常化 長官 平壤            分界線 會談 鐵路 南韓統一部            韓美 燃料 南韓 懸案            金正日 盧武鉉 朝鮮半島 打撈            黃海 銜接 核子 北韓</p>
<p>普查 支領 王太 王室            登基 會計年度 小島內閣 瑪格麗特            問卷 靈樞 溫莎堡 英鎊            西敏寺大廳 白金漢宮 社會勞工黨 王太后            加班 女王 降至 百分點            享年 伊麗莎白 太后 太爾</p>	<p>自殺 加薩市 炸彈 巴勒斯坦            威鎮 約旦河 巴勒斯坦人 哈瑪斯            震生 耶路撒冷 阿拉法特 約旦河西岸            以色列 伯利恆 槍手 加薩走廊            夏隆 黎區 西岸 受傷            特立維夫 以色列部隊 包圍 巴士</p>	<p>中美洲 決選 薩爾瓦多 哥斯大黎加            中間 兼職 雷朋 宏都拉斯            羅育 馬達加斯加 史瓦濟蘭 翁岳生            王金平 動章 院長 金哥納            馬拉坎南官 游錫方 右派 雅羅            查維斯 喬斯班 孟代爾 方士</p>



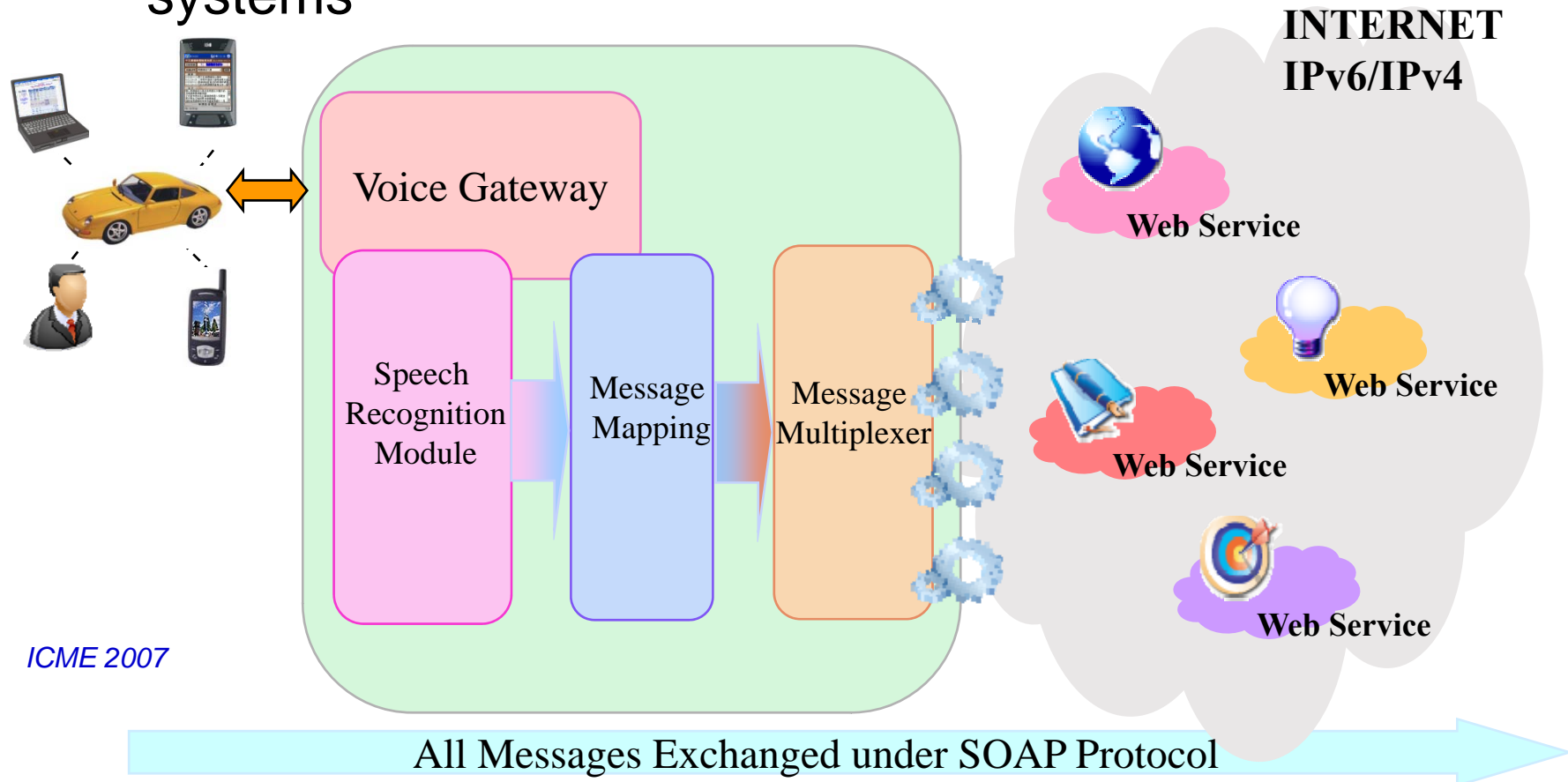
# Prototype Systems Developed at NTNU (2/3)

- Speech Retrieval and Browsing of Digital Archives



# Prototype Systems Developed at NTNU (3/3)

- Speech-based Driving/Trip Information Retrieval for ITS systems



ICME 2007

- Projects supported by Taiwan Network Information Center (TWNIC)

# Conclusions and Future Work

---

- Multimedia information access 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
- Spoken document retrieval (SDR) provides good assistance for companies in
  - Contact (Call)-center conversations: monitor agent conduct and customer satisfaction, increase service efficiency
  - Content-providing service such as MOD (Multimedia on Demand): provide a better way to retrieve and browse described program contents

*Thank You!*



# References

---

- [R1] B. Chen et al., "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," *ACM Transactions on Asian Language Information Processing*, Vol. 3, No. 2, June 2004
- [R2] J.R. Bellegarda, "Latent Semantic Mapping," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005
- [R3] K. Koumpis and S. Renals, "Content-based access to spoken audio," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005
- [R4] L.S. Lee and B. Chen, "Spoken Document Understanding and Organization," *IEEE Signal Processing Magazine*, Vol. 22, No. 5, Sept. 2005
- [R5] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, Vol. 42, 2001
- [R6] B. Chen, "Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval," *Pattern Recognition Letters*, Vol. 27, No. 1, Jan. 2006
- [R7] B. Chen, Y.T. Chen, "Extractive Spoken Document Summarization for Information Retrieval," *Pattern Recognition Letters*, Vol. 29, No. 4, March 2008.

# Representative Publications (1/5)

---

- **Spoken Document Transcription (Language & Acoustic Modeling)**
  - B. Chen\*, “Word topic models for spoken document retrieval and transcription,” to appear in *ACM Transactions on Asian Language Information Processing*, March 2009
  - H.-S. Chiu, G.-Y. Chen, C.-J. Lee, B. Chen, "Position Information for Language Modeling in Speech Recognition," *ISCSLP 2008*
  - S.-H. Liu, F.H. Chu, S.H. Lin, H.S. Lee, B. Chen, “Training Data Selection for Improving Discriminative Training of Acoustic Models,” *ASRU2007*
  - H.S. Chiu, B. Chen, “Word Topical Mixture Models for Dynamic Language Model Adaptation,” *ICASSP2007*
  - S.H. Liu, F.H. Chu, S.H. Lin, B. Chen, "Investigating Data Selection for Minimum Phone Error Training of Acoustic Models," *ICME2007*
  - J.W. Kuo, S.H. Liu, H.M. Wang, B. Chen, "An Empirical Study of Word Error Minimization Approaches for Mandarin Large Vocabulary Speech Recognition," *IJCLCLP Sept. 2006*.
  - T.H. Chen, B. Chen, H.M. Wang, "On Using Entropy Information to Improve Posterior Probability-based Confidence Measures," *ISCSLP2006*
  - J. W. Kuo, B. Chen, “Minimum Word Error Based Discriminative Training of Language Models,” *Interspeech2005*
  - B. Chen, “Dynamic Language Model Adaptation using Latent Topical Information and Automatic Transcripts,” *ICME 2005*
  - B. Chen, J.W. Kuo, W.H. Tsai, "Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription, *ICASSP2004; IJCLCLP Mar. 2005*
  - B. Chen, W.H. Tsai, J.W. Kuo, “Statistical Language Model Adaptation for Mandarin Broadcast News Transcription,” *ISCSLP2004*

# Representative Publications (2/5)

---

- **Speech Feature Extraction and Robustness**

- S.-H. Lin, B. Chen\*, Y.-M. Yeh, “Exploring the Use of Speech Features and their Corresponding Distribution Characteristics for Robust Speech Recognition,” Accepted, to appear in *IEEE Transactions on Audio, Speech and Language Processing*
- H. S. Lee, B. Chen, “Linear Discriminant Feature Extraction Using Weighted Classification Confusion Information,” *ISCSLP2008*
- H. S. Lee, B. Chen, “Linear Discriminant Feature Extraction Using Weighted Classification Confusion Information,” *Interspeech2008*
- W.H. Chen, S.H. Lin, B. Chen, “Exploiting Spatial-Temporal Feature Distribution Characteristics for Robust Speech Recognition,” *Interspeech2008*
- S.H. Lin, Y.M. Yeh, B. Chen, “Investigating the Use of Speech Features and their Corresponding Distribution Characteristics for Robust Speech Recognition,” *ASRU2007*
- S.H. Lin, Y.M. Yeh, B. Chen, “Cluster-based Polynomial-Fit Histogram Equalization (CPHEQ) for Robust Speech Recognition,” *Interspeech2007*
- S.H. Lin, Y.M. Yeh, B. Chen, “A Comparative Study of Histogram Equalization (HEQ) for Robust Speech Recognition,” *IJCLCLP June 2007*
- S.H. Lin, Y.M. Yeh, B. Chen, “Improved Histogram Equalization (HEQ) for Robust Speech Recognition,” *ICME2007*
- S.H. Lin, Y.M. Yeh, B. Chen, “Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition,” *Interspeech2006*

# Representative Publications (3/5)

---

- **Spoken Document Navigation, Retrieval and Organization**
  - S.-H. Lin, B. Chen\*, H.-M. Wang, “A comparative study of probabilistic ranking models for Chinese spoken document summarization,” to appear in *ACM Transactions on Asian Language Information Processing*, March 2009
  - Y.-T. Chen, B. Chen\*, H.-M. Wang, “A Probabilistic Generative Framework for Extractive Broadcast News Speech Summarization,” Accepted, to appear in *IEEE Transactions on Audio, Speech and Language Processing*
  - S.H. Lin, Y.T. Chen, H.M. Wang, B. Chen, “A Comparative Study of Probabilistic Ranking Models for Spoken Document Summarization,” *ICASSP2008*
  - B. Chen, Y.T. Chen, “Extractive Spoken Document Summarization for Information Retrieval,” *Pattern Recognition Letters* Vol. 29, No. 4, March 2008
  - Y.T. Chen, S.H. Lin,, B. Chen, “Spoken Document Summarization Using Relevant Information,” *ASRU 2007*
  - Y.T. Chen, H.S. Chiu, H.M. Wang, B. Chen, “A Unified Probabilistic Generative Framework for Extractive Spoken Document Summarization,” *Interpseech2007*
  - B. Chen, Y.T. Chen, “Word Topical Mixture Models for Extractive Spoken Document Summarization,” *ICME2007*
  - B. Chen, H.M. Wang, L.S. Lee, “Spoken Document Retrieval and Summarization,” Chapter 13 of the book “*Advances in Chinese Spoken Language Processing*,” World Scientific Publisher, December 2006
  - Y.T. Chen, S. Yu, H.M. Wang, B. Chen, “Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models,” *ISCSLP2006*

# Representative Publications (4/5)

---

- **Spoken Document Navigation, Retrieval and Organization (cont.)**

- B. Chen, Y.M. Yeh, Y.M. Huang, Y.T. Chen, "Chinese Spoken Document Summarization Using Probabilistic Latent Topical Information," *ICASSP2006*
- B. Chen, "Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval," *Pattern Recognition Letters*, January 2006
- B. Chen, "Voice Retrieval of Mandarin Broadcast News Speech," *International Journal of Pattern Recognition and Artificial Intelligence*, February 2006.
- B. Chen, Y.T. Chen, C.H. Chang, H.B. Chen, "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," *Interspeech2005*
- T.H. Li, M.H. Lee, B. Chen, L.S. Lee, "Hierarchical Topic Organization and Visual Presentation of Spoken Documents Using Probabilistic Latent Semantic Analysis (PLSA) for Efficient Retrieval/Browsing Applications," *Interspeech2005*
- L.S. Lee and B. Chen, "Spoken Document Understanding and Organization," *IEEE Signal Processing Magazine*, September 2005
- B. Chen, J.W. Kuo, Y.M. Huang, H.M. Wang, "Statistical Chinese Spoken Document Retrieval Using Latent Topical Information," *Interspeech2004*
- B. Chen, H.M. Wang, L.S. Lee, "A Discriminative HMM/N-Gram-Based Retrieval Approach for Mandarin Spoken Documents," *ACM Transactions on Asian Language Information Processing*, June 2004
- H. Meng, B. Chen, S. Khudanpur, G.A. Levow, W.K. Lo, D. Oard, P. Schone, K. Tang, H.M. Wang, J. Wang, "Mandarin-English Information (MEI): Investigating Translingual Speech Retrieval," *Computer Speech and Language*, April 2004

# Representative Publications (5/5)

---

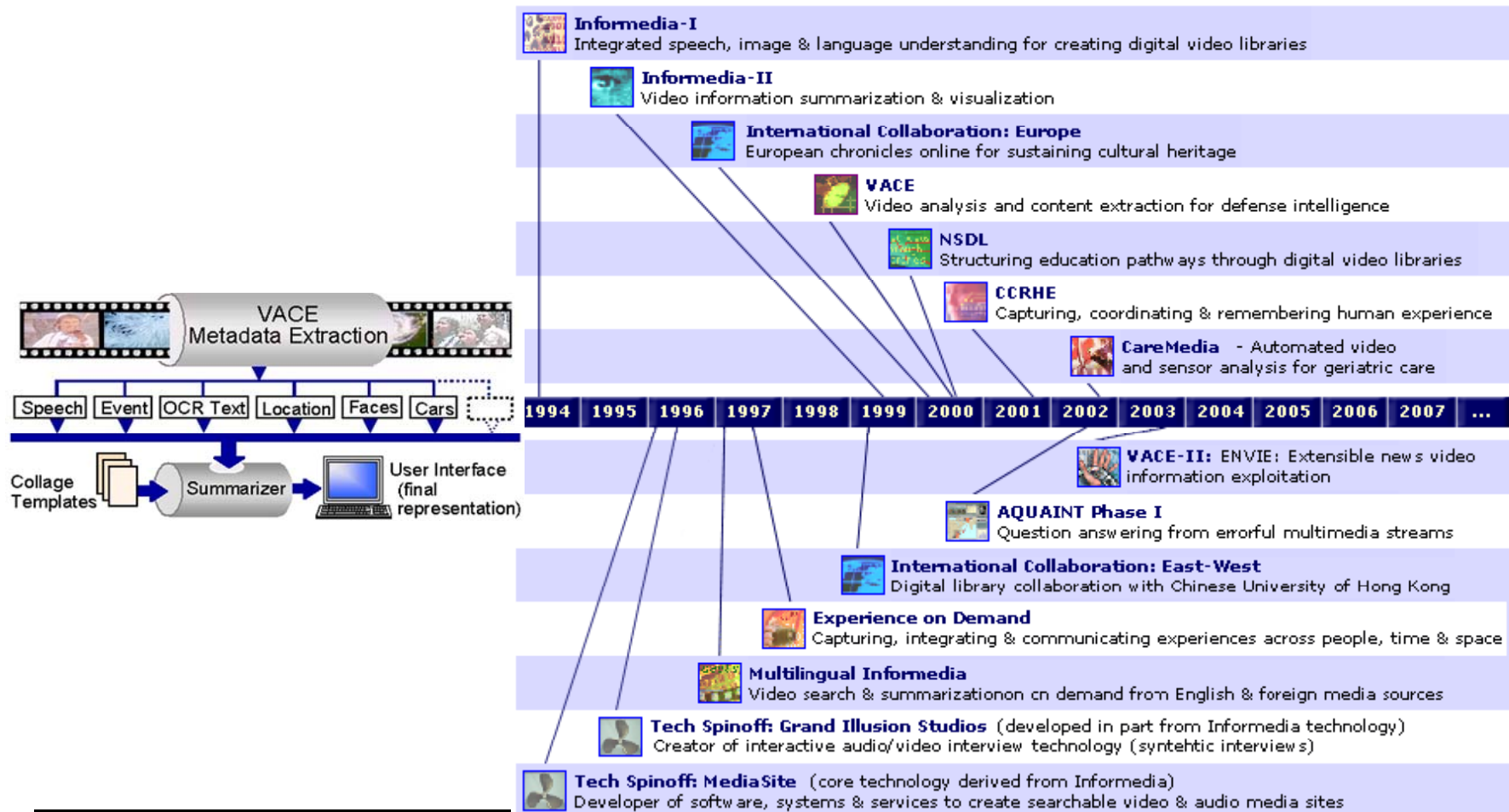
- **Spoken Document Navigation, Retrieval and Organization (cont.)**

- B. Chen, H.M. Wang, L.S. Lee, "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*, July 2002
- B. Chen, Y.T. Chen, C.H. Chang, H.B. Chen, "Speech Retrieval of Mandarin Broadcast News", C.J. Wang, B. Chen, and L.S. Lee, "Improved Chinese Spoken Document Retrieval with Hybrid Modeling and Data-driven Indexing Features," *Interspeech2002*
- H. Meng, W.K. Lo, B. Chen, K. Tang, "Generating Phonetic Cognates to Handle Named Entities in English-Chinese Cross-Language Spoken Document Retrieval," *ASRU2001*
- B. Chen, H.M. Wang, L.S. Lee, " Improved Spoken Document Retrieval by Exploring Extra Acoustic and Linguistic Cues," *Interspeech2001*
- B. Chen, H.M. Wang, and L.S. Lee, "An HMM/N-gram-based Linguistic Processing Approach for Mandarin Spoken Document Retrieval," *Interspeech2001*
- B. Chen, H.M. Wang, L.S. Lee, " Retrieval Of Broadcast News Speech in Mandarin Chinese Collected In Taiwan Using Syllable-Level Statistical Characteristics," *ICASSP2000*
- B. Chen, H.M. Wang, L.S. Lee, " Retrieval of Mandarin Broadcast News Using Spoken Queries," *Interspeech2000*

# 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 (1999)

The screenshot shows the AT&T SCAN interface. At the top, there is a search bar with the query "What is the status of the trade deficit with Japan?". Below the search bar is a table of results. The second result is highlighted in red, indicating it is the selected document. Below the table is an overview section with a bar chart showing the distribution of terms like "deficit", "status", "japan", and "trade". Below the chart is an ASR transcript of the selected document, with terms highlighted in red and blue. At the bottom, there is a selection length of 19.1699 seconds and a "Stop Audio" button.

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

OVERVIEW - NPR All Things Considered 05/10

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 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: 19.1699 seconds

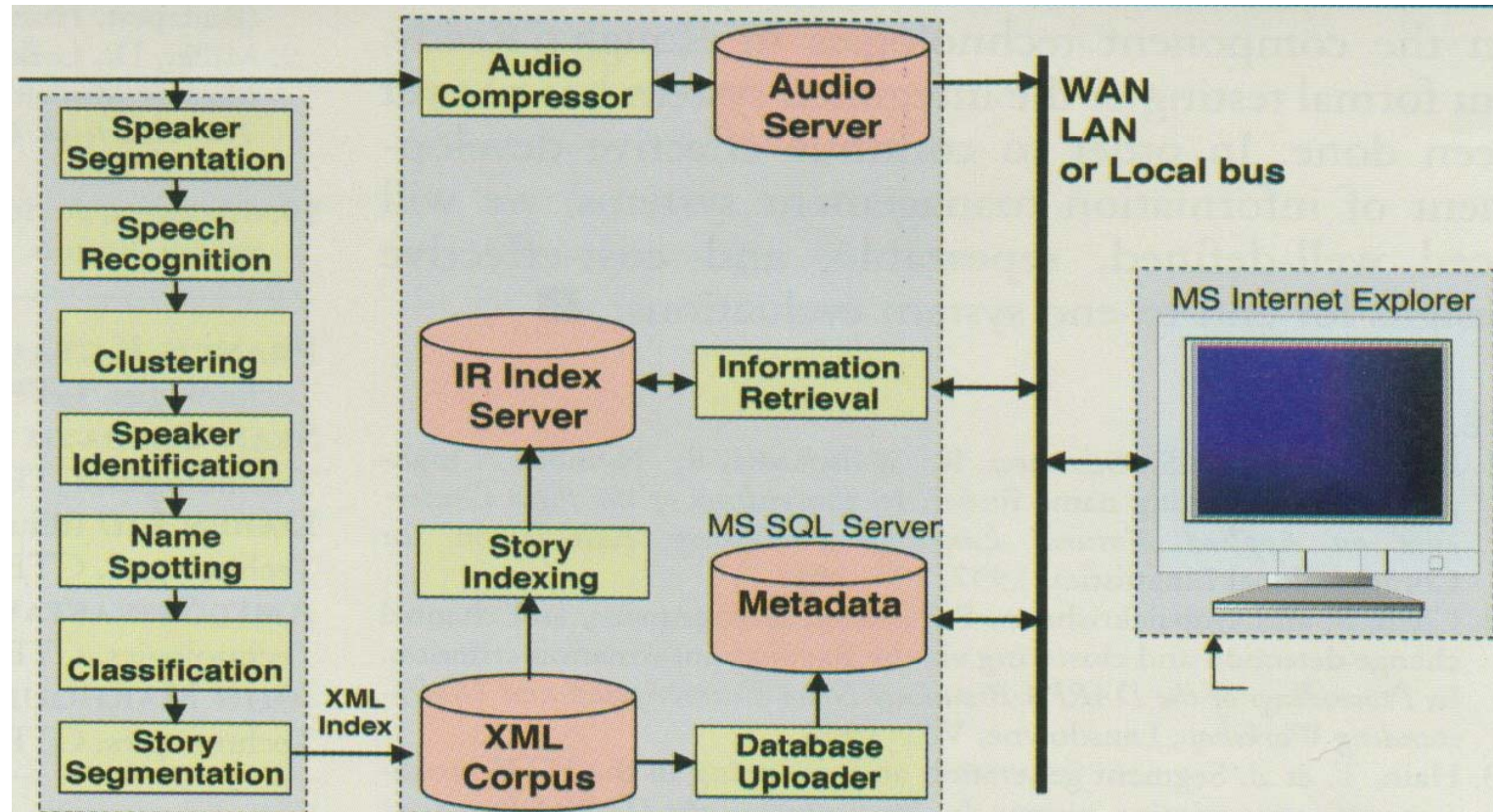
AT&T Labs Research

Design and evaluate user interfaces to support retrieval from speech archives



# BBN *Rough'n'Ready* System (1/2)

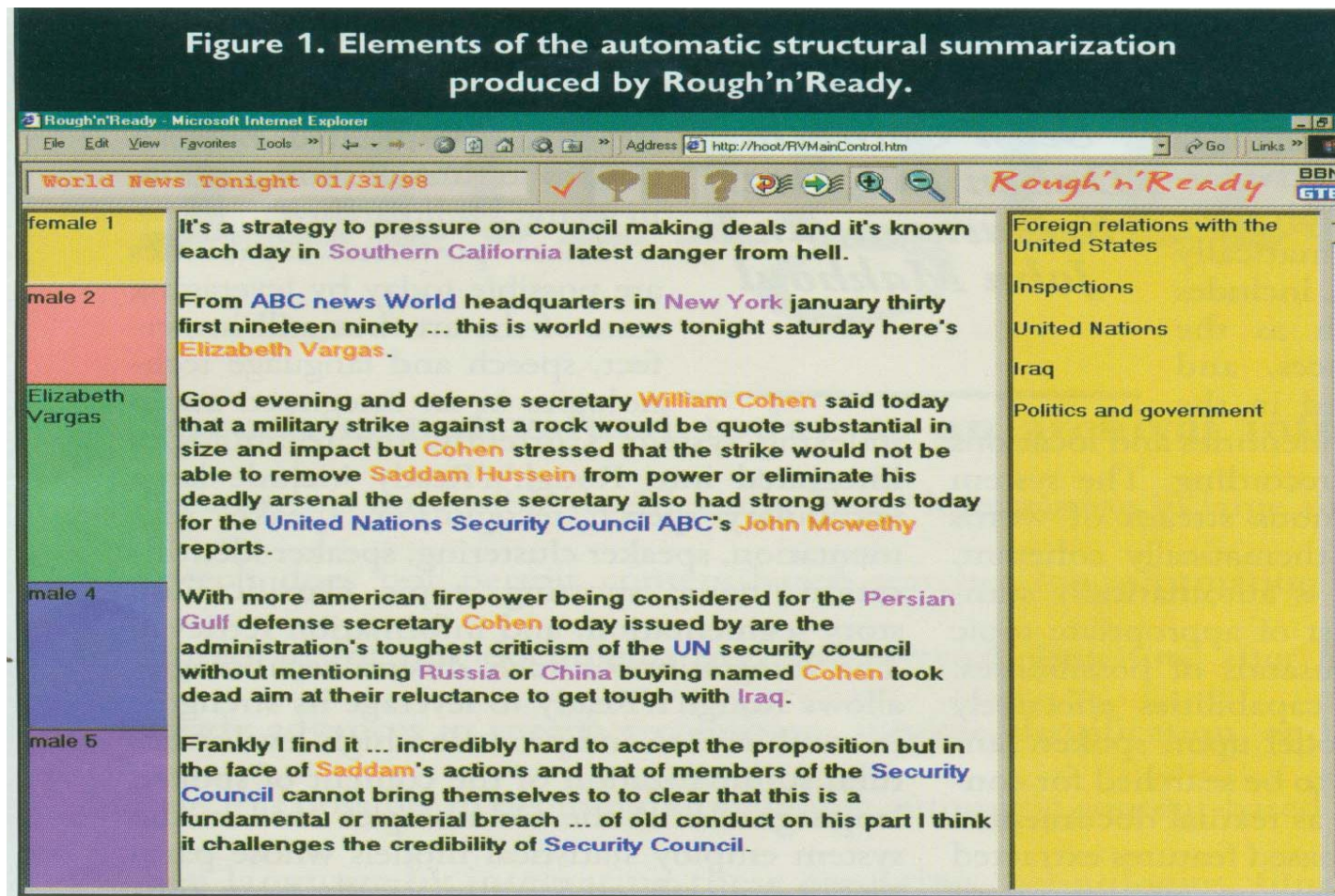
- Distinguished Architecture for Audio Indexing and Retrieval (2002)





# BBN *Rough'n'Ready* System (2/2)

- Automatic Structural Summarization for Broadcast News



# SpeechBot Audio/Video Search System at HP Labs

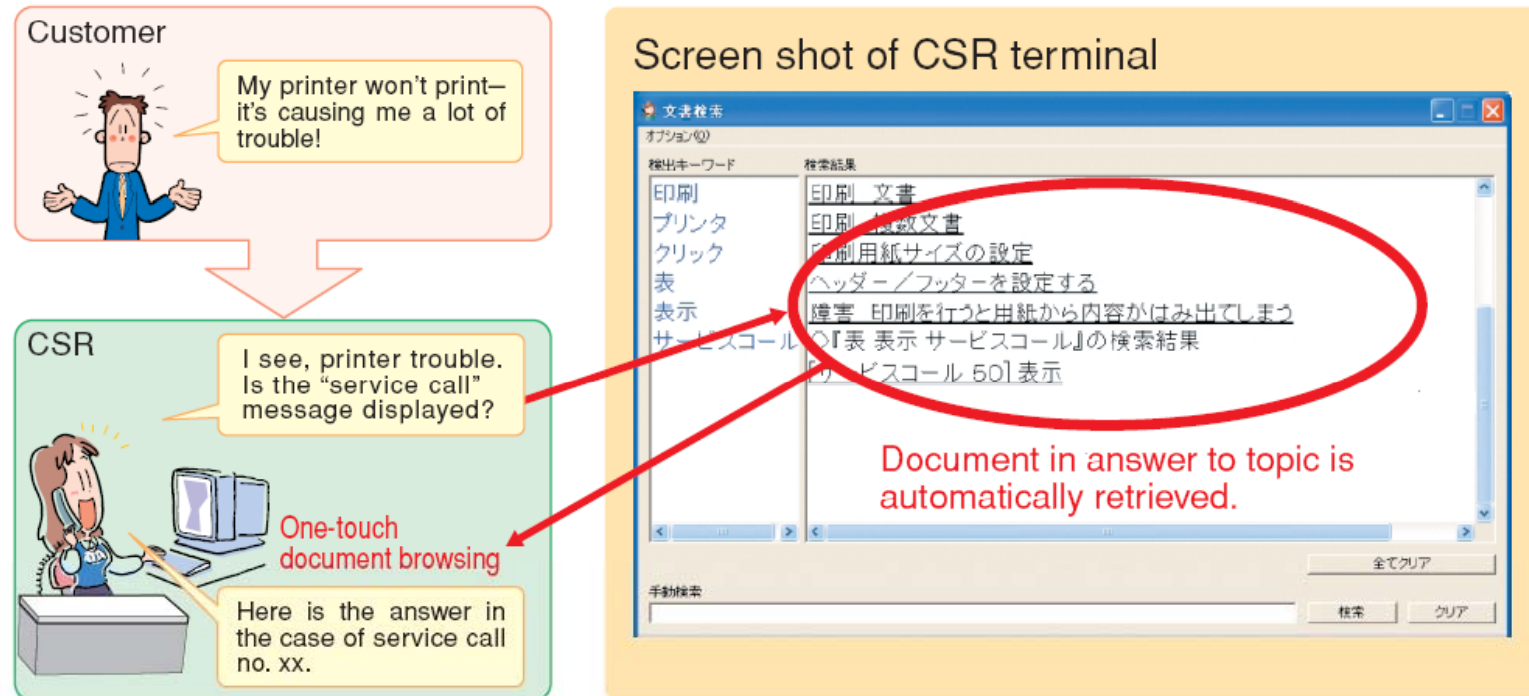
- An experimental web-based tool from HP Labs that used voice-recognition to create searchable keyword transcripts from thousands of hours of audio content

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=\\*&dr=\\*](http://speechbot.research.compaq.com/?q=Iraq&topic=*&dr=*). The page features a navigation menu with links to "hp home", "products & services", "support & drivers", "solutions", and "how to buy". A search bar is present 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 HP logo is visible on the left side of the page. The search results section shows "Search Result: 200 matches for your query" and "Sort results by: Relevance". The results are displayed 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 <b>iraq</b> 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 <b>iraq</b> to halt all nuclear activities...

# NTT Speech Communication Technology for Contact Centers

## Automatic document-retrieval by speech recognition



- CSR: Customer Service Representative

# Google Voice Local Search



Google Voice Local Search

<http://labs1.google.com/gvs.html>

Home

[About the service](#)

[FAQ](#)

[Cheat Sheet](#)

[Terms of Service](#)

[Privacy Policy](#)

[User Group](#)

[Send Feedback](#)

## Welcome to Google Voice Local Search

Google Voice Local Search is Google's experimental service to make local-business search accessible over the phone.

**To try this service, just dial 1-800-GOOG-411 (1-800-466-4411) from any phone.**

Using this service, you can:

- search for a local business by name or category.  
You can say "Giovanni's Pizzeria" or just "pizza".
- get connected to the business, free of charge.
- get the details by SMS if you're using a mobile phone.  
Just say "text message".

And it's free. Google doesn't charge you a thing for the call or for connecting you to the business. Regular phone charges may apply, based on your telephone service provider.

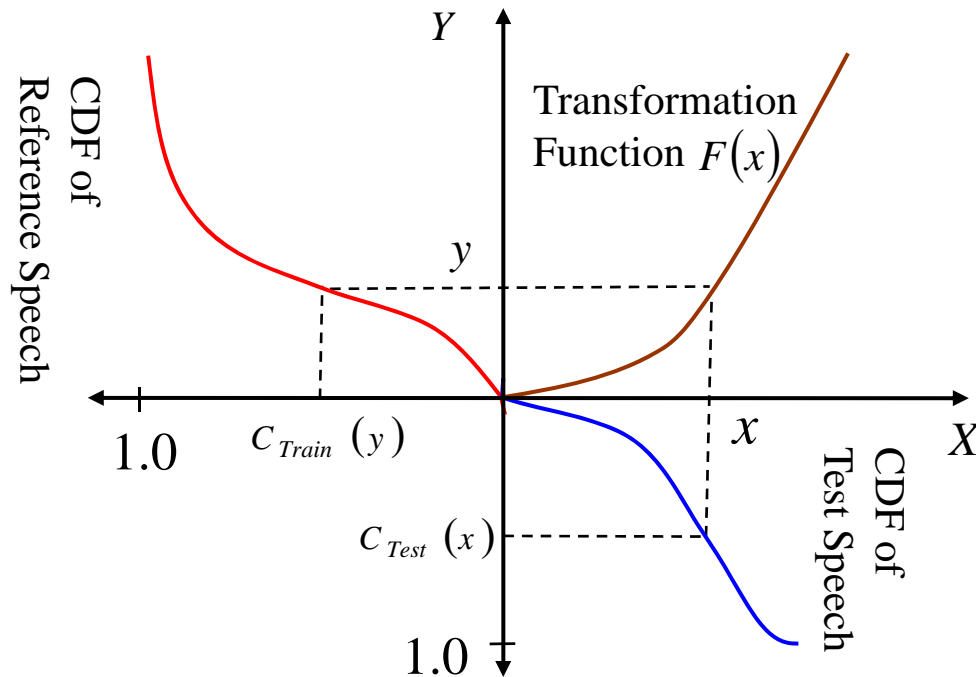
Note: Google Voice Local Search is still in its experimental stage. It may not be available at all times and may not work for all users. We're fine-tuning the service to get better at recognizing your requests. It's currently only available in English, in the US, for US business listings.

©2007 Google - [Home](#) - [About Google](#)



# Theoretical Roots of Histogram Equalization (HEQ)

- HEQ is a general non-parametric method to make the cumulative distribution function (CDF) of some given data match a reference one
  - E.g., Equalizing the CDF of test speech to that of training (reference) speech



$$\begin{aligned}
 C_{Test}(x) &= \int_{-\infty}^x p_{Test}(x') dx' \\
 &= \int_{-\infty}^{F(x)} p_{Test}(F^{-1}(y')) \frac{dF^{-1}(y')}{dy'} dy' \\
 &= \int_{-\infty}^y p_{Train}(y') dy' \Big|_{y=F(x)} \\
 &= C_{Train}(y)
 \end{aligned}$$

, where  $F(x)$  is a transformation function





# Polynomial-Fit Histogram Equalization (PHEQ)

---

- We propose to use least squares regression for the fitting of the inverse function of CDFs of training speech
  - For each speech feature vector dimension of the training data, a polynomial function can be expressed as follows, given a pair of  $y_i$  and corresponding CDF  $C_{Train}(y_i)$

$$G(C_{Train}(y_i)) = \tilde{y}_i = \sum_{m=0}^M a_m (C_{Train}(y_i))^m$$

- The corresponding squares error

$$E^2 = \sum_{i=1}^N \left( y_i - \sum_{m=0}^M a_m (C_{Train}(y_i))^m \right)^2$$

- Coefficients  $a_m$  can be estimated by minimizing the squares error



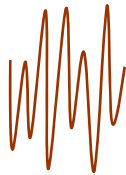
# PHEQ: Evaluation

- The speech recognition experiments were conducted under various noise conditions using the Aurora-2 database and task *Interspeech 2006, 2007; ICME 2007; ASRU 2007*

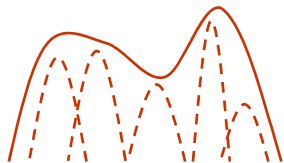
	Clean-Condition Training				Multi-Condition Training			
	Test A	Test B	Test C	Average	Test A	Test B	Test C	Average
ETSI	38.69	44.25	28.76	38.93	10.64	10.76	12.85	11.13
MFCC+CMVN	20.21	19.84	21.13	20.24	12.18	11.23	13.21	12.01
CMVN-TA	16.63	14.92	17.90	16.20	8.86	8.82	9.69	9.01
THEQ	18.13	16.41	19.51	17.71	11.97	11.47	13.44	12.06
GHEQ	17.69	15.59	18.70	17.05	9.00	8.73	9.60	9.01
QHEQ	23.74	21.73	23.11	22.81	8.91	10.03	11.75	9.93
SPLICE(1024)	17.03	17.12	26.90	19.04	--	--	--	--
PHEQ	15.91	14.43	16.80	15.49	9.23	8.89	10.38	9.32
PHEQ-TA	14.29	13.75	15.20	14.25	8.72	8.64	9.21	8.78
CPHEQ(1024)	13.07	12.24	17.90	13.70	--	--	--	--

– Proposed PHEQ and CPHEQ provides significant performance boosts

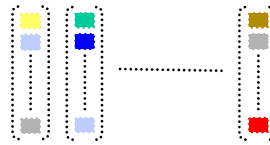




Noisy Speech



Statistical Distribution



Feature Domain

**Feature Normalization**

**Feature Transformation**

**Feature Compensation**

**Feature Reconstruction**

Cepstral Mean Normalization  
[Furui 1981]



Cepstral Mean/Variance Normalization  
[Vikki et al. 1998]



High Order Cepstral  
Moment Normalization  
[Hsu et al. 2006]



Histogram Equalization  
[Molau 2003;  
Torre et al. 2005]



Quantile-based  
Histogram Equalization  
[Hilger et al. 2006]

Principle Component Analysis  
Linear Discriminant Analysis  
[Duda et al. 1973]



Heteroscedastic Linear Discriminant Analysis  
[Kumar 1997]

Kernel Linear Discriminant Analysis  
[Mika 1999]

Heteroscedastic Discriminant Analysis  
[Saon et al. 2000]

Codeword Dependent  
Cepstral Normalization  
[Acero 1990]



Probabilistic Optimum Filtering  
[Neumeyer et al. 1994]



Stereo-based Piecewise  
Linear Compensation  
[Deng et al. 2000]



Discriminative Stochastic Vector Mapping

Maximum Mutual Information  
[Droppo et al. 2005]

Minimum Classification Error  
[Wu et al. 2006]

Stochastic Matching  
[Sankar et al. 1994]



Maximum Likelihood  
Stochastic Vector Mapping  
[Wu et al. 2005]



Missing Feature  
- - Cluster-based  
[Raj et al. 2004]  
- - Covariance-based  
[Raj et al. 2004]



# Word Topical Mixture Model (1/4)

- In this research, each word of language are treated as a word topical mixture model (WTMM) for predicting the occurrences of other words *ICASSP 2007*

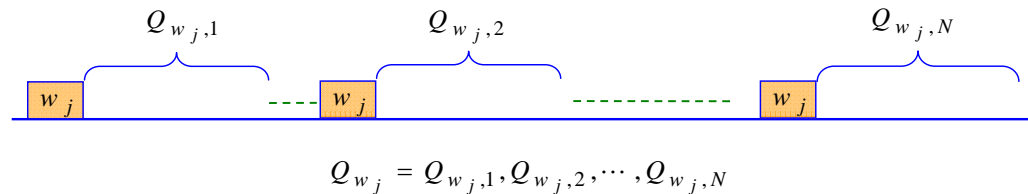
$$P(w_i | M_{w_j}) = \sum_{k=1}^K P(w_i | T_k) P(T_k | M_{w_j})$$

- WTMM in LM Adaptation
  - Each history consists of words
  - History model is treated as a composite word TMM
  - The history model of a decoded word can be dynamically constructed

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

# Word Topical Mixture Model (2/4)

- Exploration of Training Exemplars
  - Collect the words within a context window around each occurrence of word in the training corpus
  - Concatenate them to form the relevant observations for training the word TMM



- Maximize the sum of log-likelihoods of WTMM models generating their corresponding training exemplars

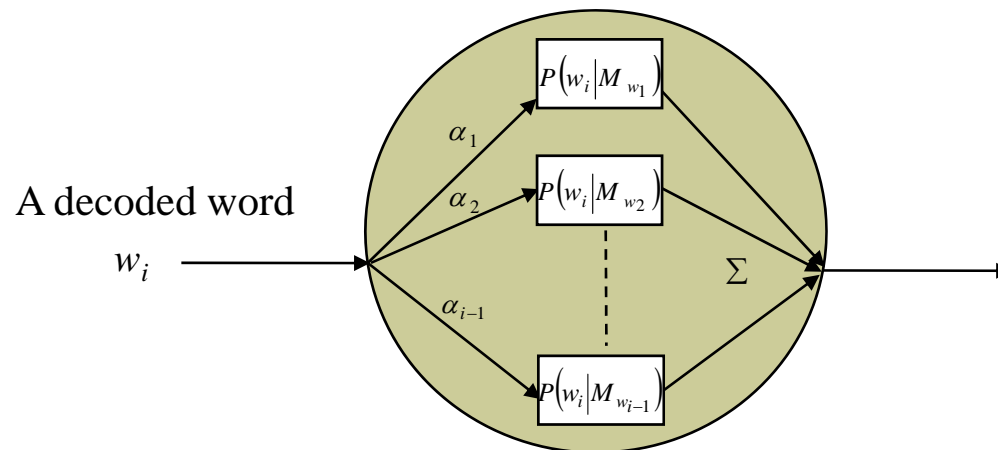
$$\log L_{\mathbf{Q}_{TrainSet}} = \sum_{Q_{w_j} \in \mathbf{Q}_{TrainSet}} \log P(Q_{w_j} | M_{w_j}) = \sum_{Q_{w_j} \in \mathbf{Q}_{TrainSet}} \sum_{w_n \in Q_{w_j}} n(w_n, Q_{w_j}) \log P(w_n | M_{w_j})$$



# Word Topical Mixture Model (3/4)

- Recognition using WTMM models
  - A simple linear combination of WTMM models of the words occurring in the search history

A composite word TMM model for the search history  $H_{w_i} = w_1, w_2, \dots, w_{i-1}$

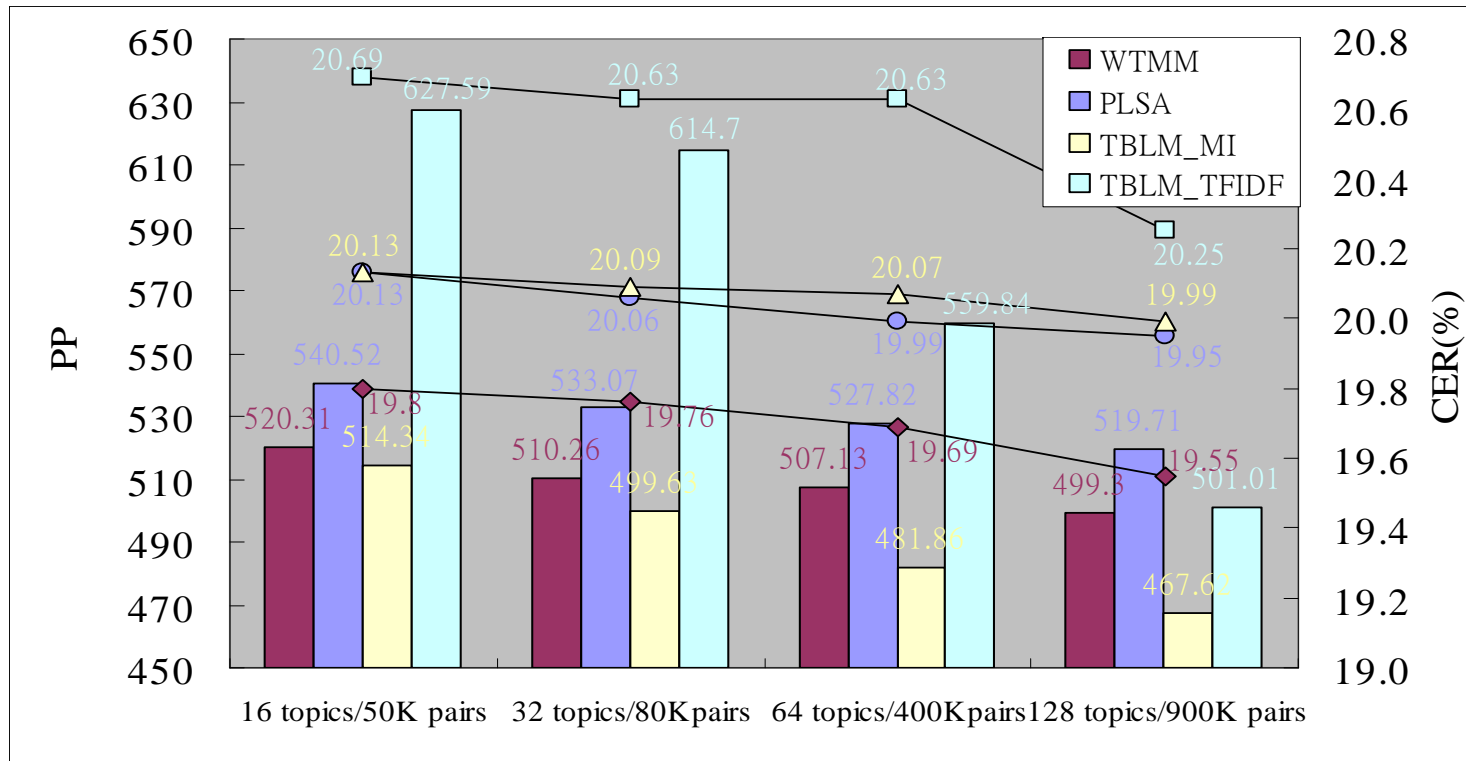


- Weights are empirically set to be exponentially decayed as the words in the history are apart from current decoded word



# Word Topical Mixture Model (4/4)

- Experiments: Comparison of WTMM, PLSALM, TBLM

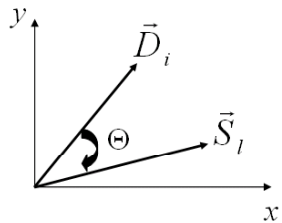


- WTMM performs slightly better than PLSALM and TBLM in CER measure
- TBLM trained with MI score performs better in PP (perplexity) measure



# SDS: Other Approaches (1/4)

- Vector Space Model (VSM) Y. Gong, SIGIR 2001
  - Vector representations of sentences and the document to be summarized using statistical weighting such as *TF-IDF*
  - Sentences are ranked based on their proximity to the document
  - To summarize more important and different concepts in a document



- The terms occurring in the sentence with the highest relevance score  $Sim(S_l, D_i)$  are removed from the document
- The document vector is then reconstructed and the ranking of the rest of the sentences is performed accordingly

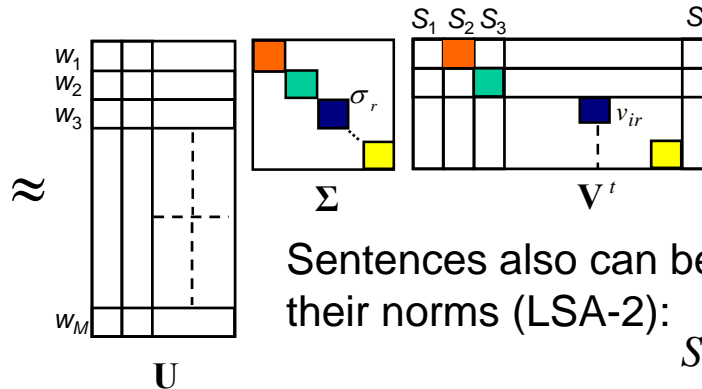
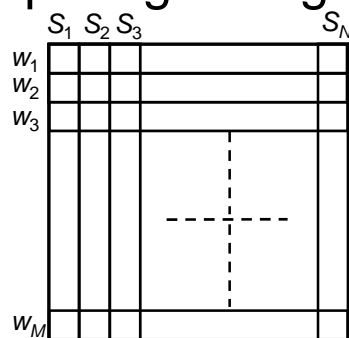
- Or, using the Maximum Marginal Relevance (MMR) model

$$NextSen = \max_{S_l} [\lambda \cdot Sim(S_l, D_i) - (1 - \lambda) Sim(S_l, Summ)]$$

- *Summ* : the set of already selected sentences

## SDS: Other Approaches (2/4)

- Latent Semantic Analysis (LSA) Y. Gong, SIGIR 2001
  - Construct a “term-sentence” matrix for a given document
  - Perform SVD on the “term-sentence” matrix
    - The **right singular vectors** with larger singular values represent the dimensions of the more important latent semantic concepts in the document
    - Represent each sentence of a document as a vector in the latent semantic space
  - Sentences with the largest index (element) values in each of the top  $L$  right singular vectors are included in the summary (LSA-1)



*Hirohata et al., ICASSP2005*

Sentences also can be selected based on their norms (LSA-2):

$$Score(S_i) = \sqrt{\sum_{r=1}^L (\sigma_r v_{ir})^2}$$

## SDS: Other Approaches (3/4)

---

- Sentence Significance Score (SenSig)
  - Sentences are ranked based on their significance which, for example, is defined by the average importance scores of words in the sentence

$$\text{SenSig}(S) = \frac{1}{N_s} \sum_{n=1}^{N_s} I(w_n)$$

$$I(w_n) = f_w \cdot icf = f_w \cdot \log \frac{F_c}{F_w}$$

similar to *TF-IDF* weighting

*S. Furui et al., IEEE SAP 12(4), 2004*

- Other features such as *word confidence*, *linguistic score*, or *prosodic information* also can be further integrated into this method





## SDS: Other Approaches (4/4)

---

- Sentence selection is formulated as a binary classification problem
  - A sentence can either be included in a summary or not
- A bulk of classification-based methods using statistical features also have been developed
  - Gaussian mixture models (GMM)
  - Bayesian network classifier (BN)
  - Support vector machine (SVM)
  - Logistic Regression (LR)
- However, the above methods need a set of training documents together with their corresponding handcrafted summaries (or labeled data) for training the classifiers

