Recent Developments in Speech Retrieval and Summarization

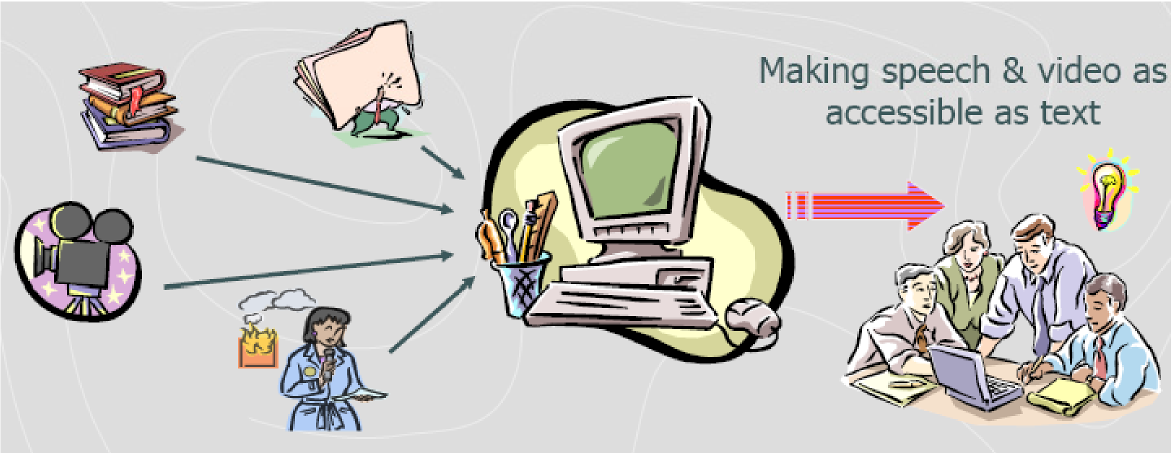
Berlin Chen (陳柏琳)
Department of Computer Science & Information Engineering
National Taiwan Normal University
2010/05/04

Introduction (1/2)

- Communication and search are by far the most popular activities in our daily lives
  - Speech is the most nature and convenient means of communication between humans, and between humans and machines
    - A spoken language interface could be more convenient than a visual interface on a small device
    - Provide "anytime" and "anywhere" access to information
  - Already over half of the internet traffic consists of video data
    - Though visual cues are important for search, the associated spoken documents often provide a rich set of semantic descriptions (e.g., transcripts, speakers, emotions, and scenes) for the data

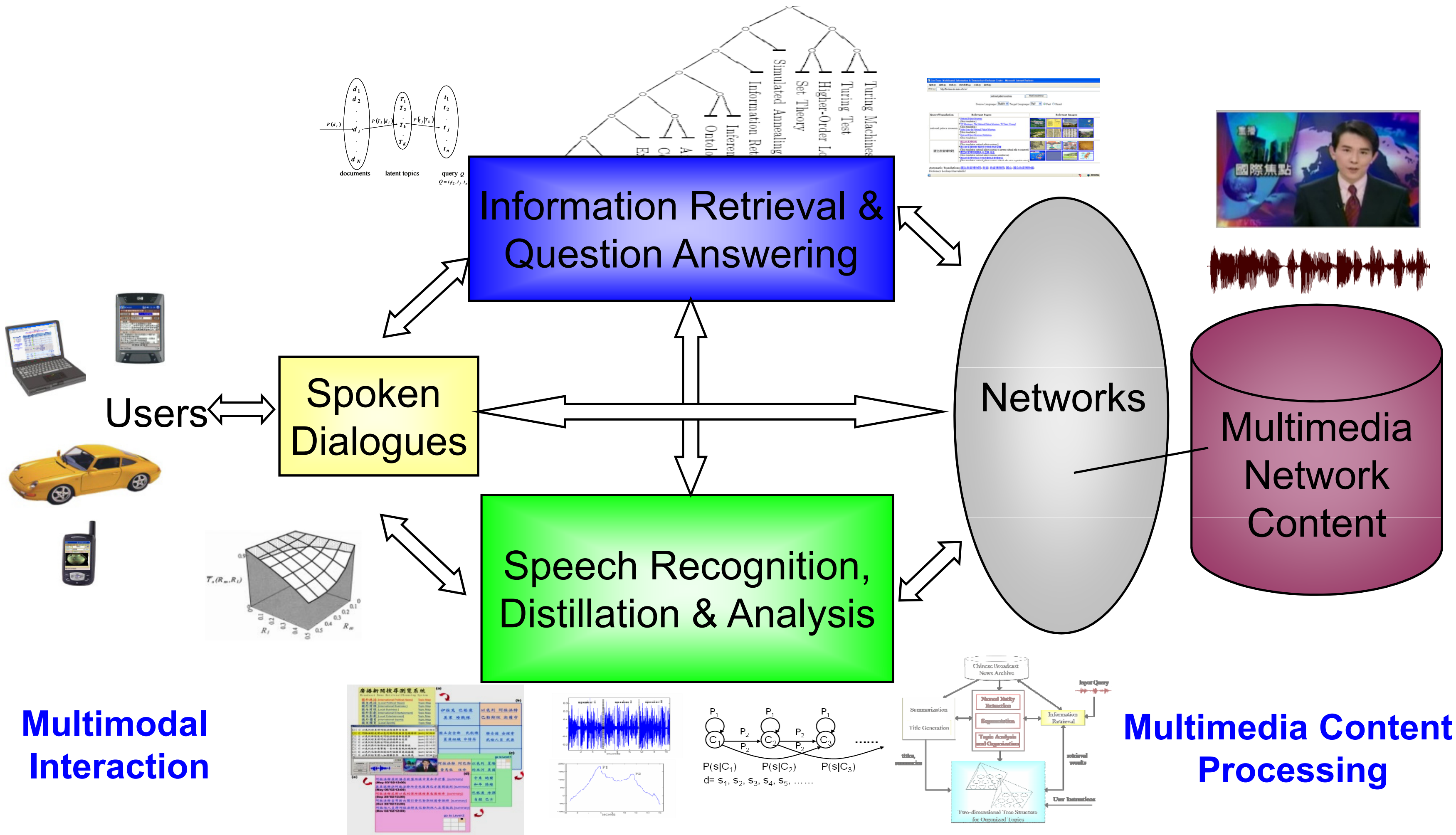
Introduction (1/2)

- Automatic speech recognition (ASR)
  - Transcribe the linguistic contents of speech utterances
  - Play a vital role in multimedia information retrieval, summarization and mining
    - Such as transcription of spoken documents and recognition of spoken queries



The figure is adapted from the presentation slides of Prof. Ostendorf at *Interspeech 2009.*

Multimodal Access to Multimedia in the Future



Automatic Speech Recognition (ASR)

- Bayes Decision Rule (Risk Minimization)

$$
\begin{aligned}
W_{opt} &= \arg\min_{W \in \mathbf{W}} Risk\left(W \middle| O\right) \\
&= \arg\min_{W \in \mathbf{W}} \sum_{W' \in \mathbf{W}} Loss\left(W, W'\right) P\left(W' \middle| O\right) \\
&\approx \arg\max_{W \in \mathbf{W}} P\left(W \middle| O\right) \quad \text{Assumption of Using the “0-1” Loss Function} \\
&= \arg\max_{W \in \mathbf{W}} \frac{p(O \mid W) P(W)}{p(O)} \\
&= \boxed{\arg\max_{W \in \mathbf{W}} p(O \mid W) P(W)} \quad \text{Linguistic Decoding}
\end{aligned}
$$

Feature Extraction & Acoustic Modeling Language Modeling

- There is an emerging trend of “direct modeling” the discriminant function $P\left(W \middle| O\right)$

Core Components of ASR

- Feature Extraction
  - Convert a speech signal into a sequence of feature vectors describing the inherent acoustic and phonetic properties
- Acoustic modeling
  - Construct a set of statistical models representing various sounds (or phonetic units) of the language
- Language modeling
  - Construct a set of statistical models to constrain the acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the final output from a speech recognizer
- Robustness
  - Eliminate varying sources of environmental (e.g., channel and background) variations

M.J.F. Gales and S.J. Young. *The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing*, 2008

Applications of ASR

- Multimedia (spoken document) retrieval and organization
  - Speech-driven Interface and multimedia content processing
  - Work in association with information retrieval techniques
  - A wild variety of potential applications (to be introduced later)
- Computer-Aided Language Learning (CALL)
  - Speech-driven Interface and multimedia content processing
  - Work in in association with natural language processing techniques
  - Applications
    - Synchronization of audio/video learning materials
    - Automatic pronunciation assessment/scoring
    - Read student essays and grade them
    - Automated reading tutor
    - Automated test assembly
- Others

IEEE Signal Processing Magazine 25(3), 2008 (*Spoken Language Technology*)
IEEE Signal Processing Magazine 22(5), 2005 (*Speech Technology and Systems in Human-Machine Communication*)

NTNU

Speech-driven Multimedia Retrieval & Organization

- Continuous and substantial efforts have been paid to speech-driven multimedia retrieval and organization 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 *Bing Mobile Voice Search*, Audio-Video Indexing System (*MAVIS*)
  - MIT Lecture Browser
  - NTT Speech Communication Technology for Contact Centers
  - Some Prototype Systems Developed in Taiwan

IEEE SLTC eNewsletter - Spring 2010 : *Following Global Events with IBM Translingual Automatic Language Exploration System (TALES)*

World-wide Speech Research Projects

- There also are several research projects conducted on a wild variety of 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

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", "烏普薩拉(Uppsala)" etc.
- Speech Retrieval
  - Robust representations of spoken queries and spoken documents
  - Matching between (spoken) queries and spoken documents

Lee and Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine* 22 (5), Sept. 2005

Key Technologies (2/2)

- Speech Summarization
  - 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, why and how for the information described by spoken documents

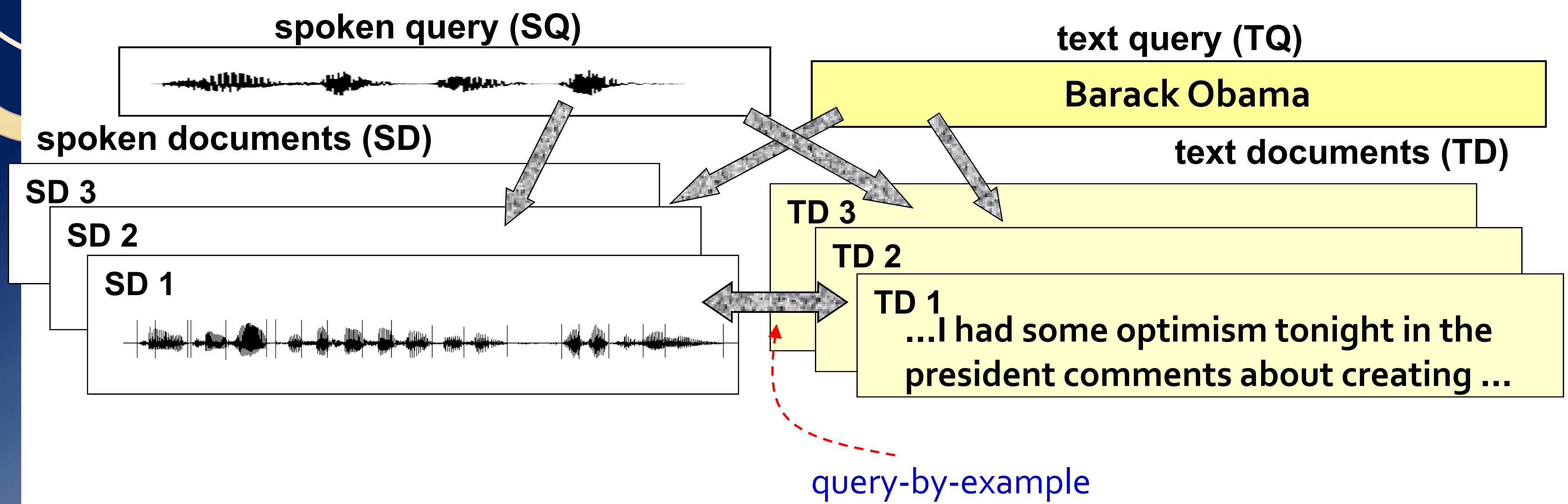
I. Speech Retrieval

Task Definition of Speech Retrieval

- Robustly Index spoken documents with speech recognition techniques
  - Explore better ways to represent the recognition hypotheses of spoken documents beyond the top scoring ones
  - Hybrid of words and subwords (phone/syllable/character *n*-grams) for indexing
- Retrieve relevant spoken documents in response to a user query
  - Spoken Document Retrieval (SDR)
    - Find spoken documents that are "topically related" to a given query
  - Spoken Term Detection (STD)
    - Find "literally matched" spoken documents where all/most query terms should be present (much like Web search)

N T N U

Scenarios of Spoken Document Retrieval

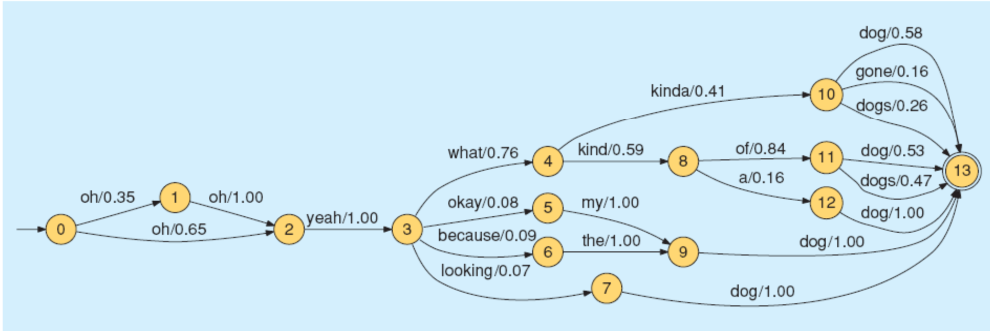


- SQ/SD is the most difficult
- TQ/SD is studied most of the time
  - "query-by-example": e.g., use text news documents to retrieve relevant broadcast news documents
    - Useful for news monitoring and tracking

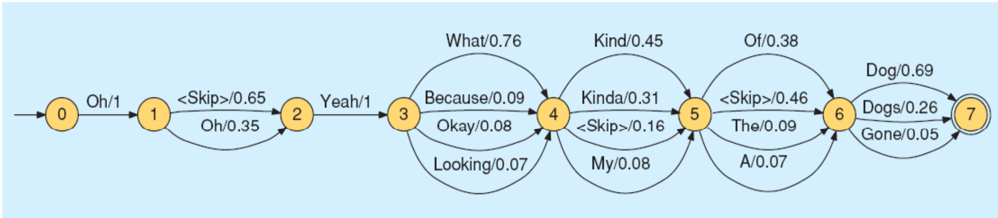
Representations of Spoken Queries and Documents

- Lattice/confusion network structures for retaining multiple recognition hypotheses

Lattice



Confusion Network



Position-Specific Posterior Probability Lattices

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				
						......	...	......	...	—					

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

Retrieval Models

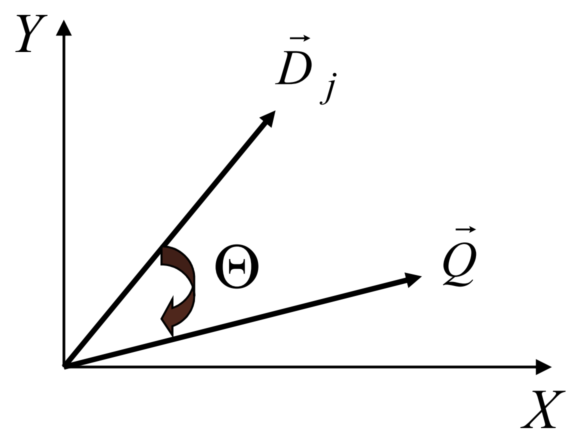
- Information retrieval (IR) models can be characterized by two different matching strategies
  - Literal term matching
    - Match queries and documents in an index term space
  - Concept matching
    - Match queries and documents in a latent semantic space

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。

Retrieval 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), describing the intra-/inter-document statistics between all terms and all documents (bag-of-words modeling)
  - Cosine measure for query-document relevance

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

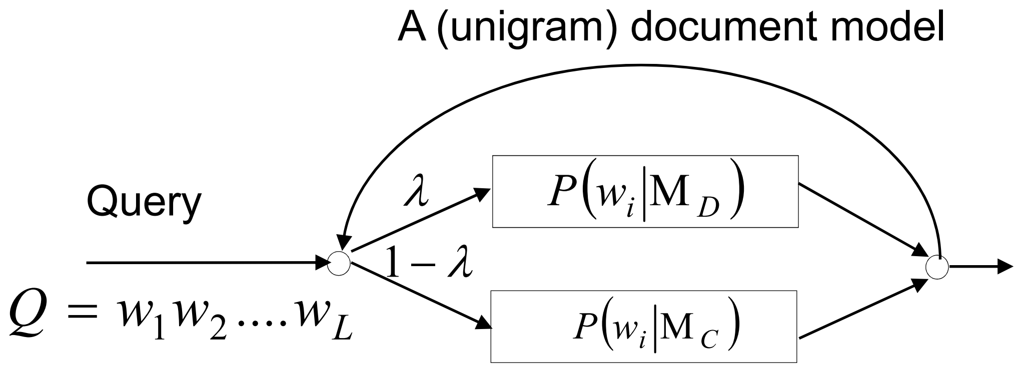


  - VSM can be implemented with an inverted file structure for efficient document search (instead of exhaustive search)

Manning, Raghavan, and Schütze, *Introduction to Information Retrieval,* Cambridge University Press. 2008.

Retrieval Models: Literal Term Matching (2/2)

- Hidden Markov Model (HMM)
  - Also known as the Language Model (LM)
  - A language model consists of a set of *n*-gram distributions is established for each document for predicting the query



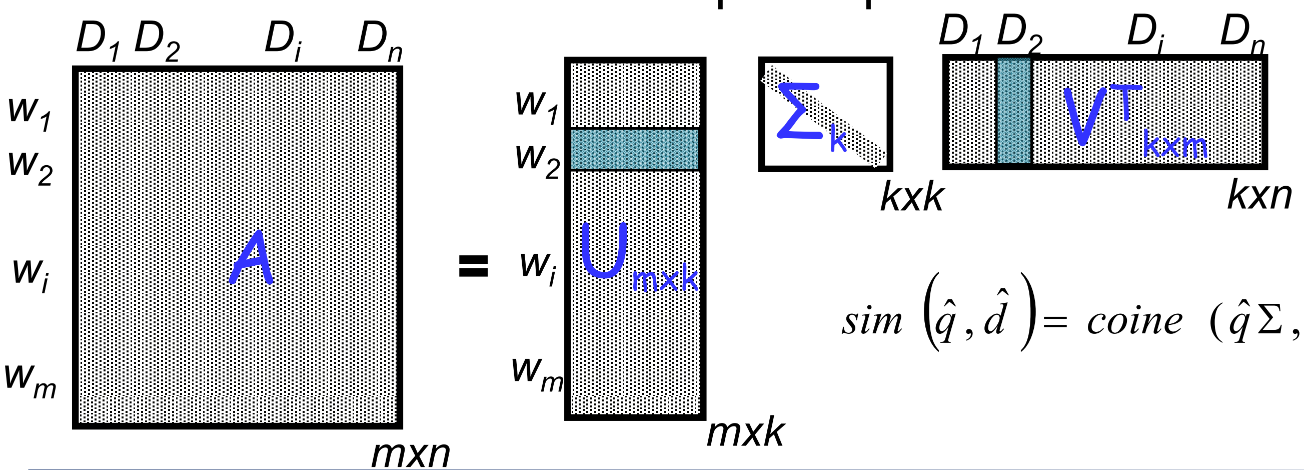
$$P_{\mathrm{HMM}}(Q|\mathrm{M}_D) = \prod_{i=1}^{L} \left[\lambda \cdot P(w_i|\mathrm{M}_D) + (1-\lambda) \cdot P(w_i|\mathrm{M}_C)\right]$$

  - Such models can be optimized with different criteria
  - Provide a potentially effective and theoretically attractive probabilistic framework for studying IR problems

Zhai, *Statistical Language Models for Information Retrieval* (Synthesis Lectures Series on Human Language Technologies), Morgan & Claypool Publishers, 2008

Retrieval Models: Concept Matching (1/2)

- Latent Semantic Analysis (LSA)
  - Start with a matrix (A) describing the intra-/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 $A \approx U \Sigma V^T$
  - Matching between words/queries and documents can be carried out in this latent topical space



$$sim\left(\hat{q}, \hat{d}\right) = coine\ (\hat{q}\Sigma, \hat{d}\Sigma) = \frac{\hat{q}\Sigma^2 \hat{d}^T}{\left|\hat{q}\Sigma\right|\left|\hat{d}\Sigma\right|}$$

Retrieval Models: Concept Matching (2/2)

- Recently, several probabilistic counterparts of LSA have proposed and demonstrated with good success
- 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

$$P_{\text{PLSA/LDA}}\left(Q \mid \mathrm{M}_D\right) = \prod_{i=1}^{L}\left[\sum_{k=1}^{K} P\left(w_i \mid T_k\right) P\left(T_k \mid \mathrm{M}_D\right)\right]$$

  - 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

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

- The relevance measure between a query and a document can be expressed by

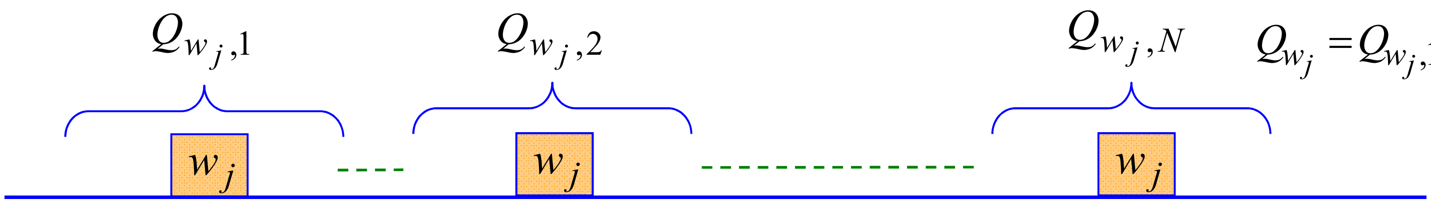
$$P_{\text{WTM}}(Q|D) = \prod_{i=1}^{L} \left[\sum_{w_j \in D} P_{\text{WTM}}\left(w_i \middle| \mathrm{M}_{w_j}\right) P\left(w_j \middle| D\right) \right]$$

  - 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
  - $P\left(w_j \middle| D\right)$ is estimated according to the frequency of w_j in D

Can we model topic al information using other units beyond "documents" ?

Unsupervised Training of WTM

- The WTM $P_{\text{WTM}}\left(w_i \mid \text{M}_{w_j}\right)$ 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



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

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

    - $\mathbf{w}$: the set of distinct words in the language

  - WTM was trained to optimize its prediction power over the observation

Comparison Between WTM and DTM
-- Probabilistic Matrix Decompositions

PLSA/LDA

$$\underset{\text{words}\times\text{documents}}{\mathbf{A}} \approx \underset{\text{words}\times\text{topics}}{\mathbf{G}}\ \underset{\text{topics}\times\text{documents}}{\mathbf{H}^{T}}$$

normalized "word-document" co-occurrence matrix

mixture components

mixture weights

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

WTM

$$\underset{\text{words}\times\text{vicinities of words}}{\mathbf{B}} \approx \underset{\text{words}\times\text{topics}}{\mathbf{Q}}\ \underset{\text{topics}\times\text{vicinities of words}}{\mathbf{Q}'^{T}}$$

normalized "word-word" co-occurrence matrix

mixture components

mixture weights

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

Chen, "Word topic models for spoken document retrieval and transcription," *ACM TALIP*, 8(1), March 2009

Example Topic Distributions of WTM

Topic 13	
word	**weight**
Vena (靜脈)	1.202
Resection (切除)	0.674
Myoma (肌瘤)	0.668
Cephalitis (腦炎)	0.618
Uterus (子宮)	0.501
Bronchus (支氣管)	0.500

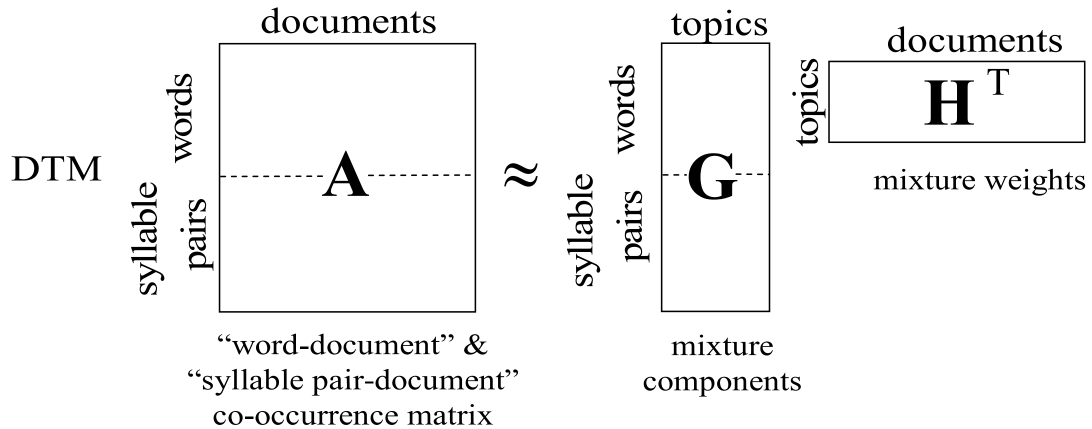
Topic 14	
word	**weight**
Land tax (土地稅)	0.704
Tobacco and alcohol tax law (菸酒稅法)	0.489
Tax (財稅)	0.457
Amend drafts (修正草案)	0.446
Acquisition (購併)	0.396
Insurance law (保險法)	0.373

Topic 23	
word	**weight**
Cholera (霍亂)	0.752
Colorectal cancer (大腸直腸癌)	0.681
Salmonella enterica (沙門氏菌)	0.471
Aphtae epizooticae (口蹄疫)	0.337
Thyroid (甲狀腺)	0.303
Gastric cancer (胃癌)	0.298

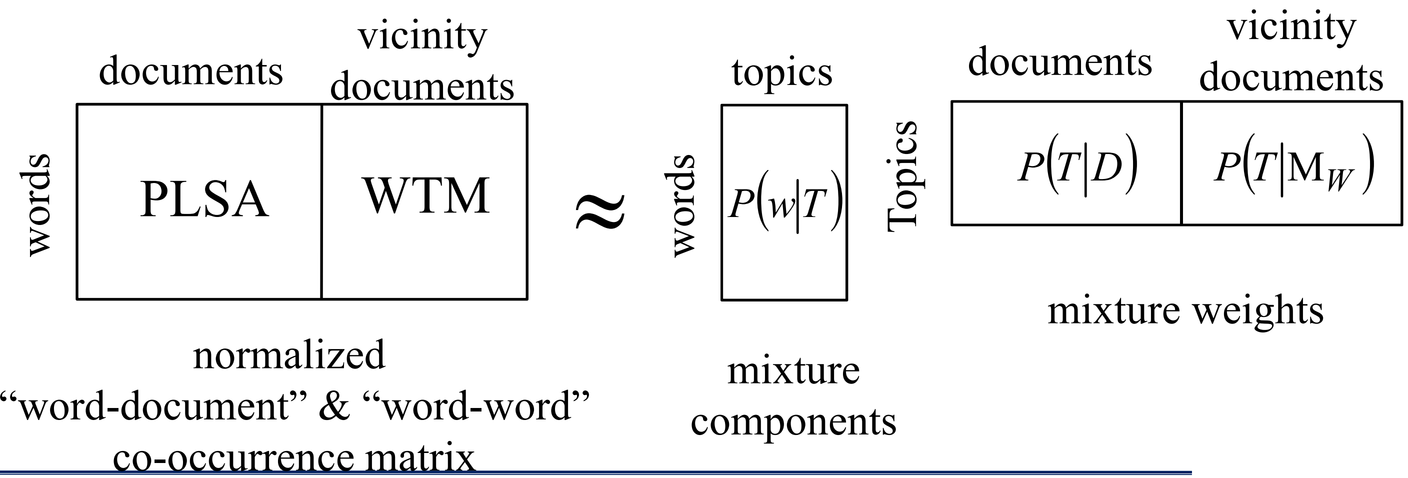
N T N U

Some Extensions of DTM and WTM

- Hybrid of Word- and Syllable-level Features by using DTM/WTM



- Hybrid of DTM and WTM by Sharing the Same Latent Topics



Lin and Chen, "Topic modeling for spoken document retrieval using word- and syllable-level information," *SSCS 2009*
Chen et al. "Latent topic modeling of word vicinity information for speech recognition," *ICASSP 2010*

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

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

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

relevant document

irrelevant document

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

Other irrelevant documents for the training query can be taken into consideration

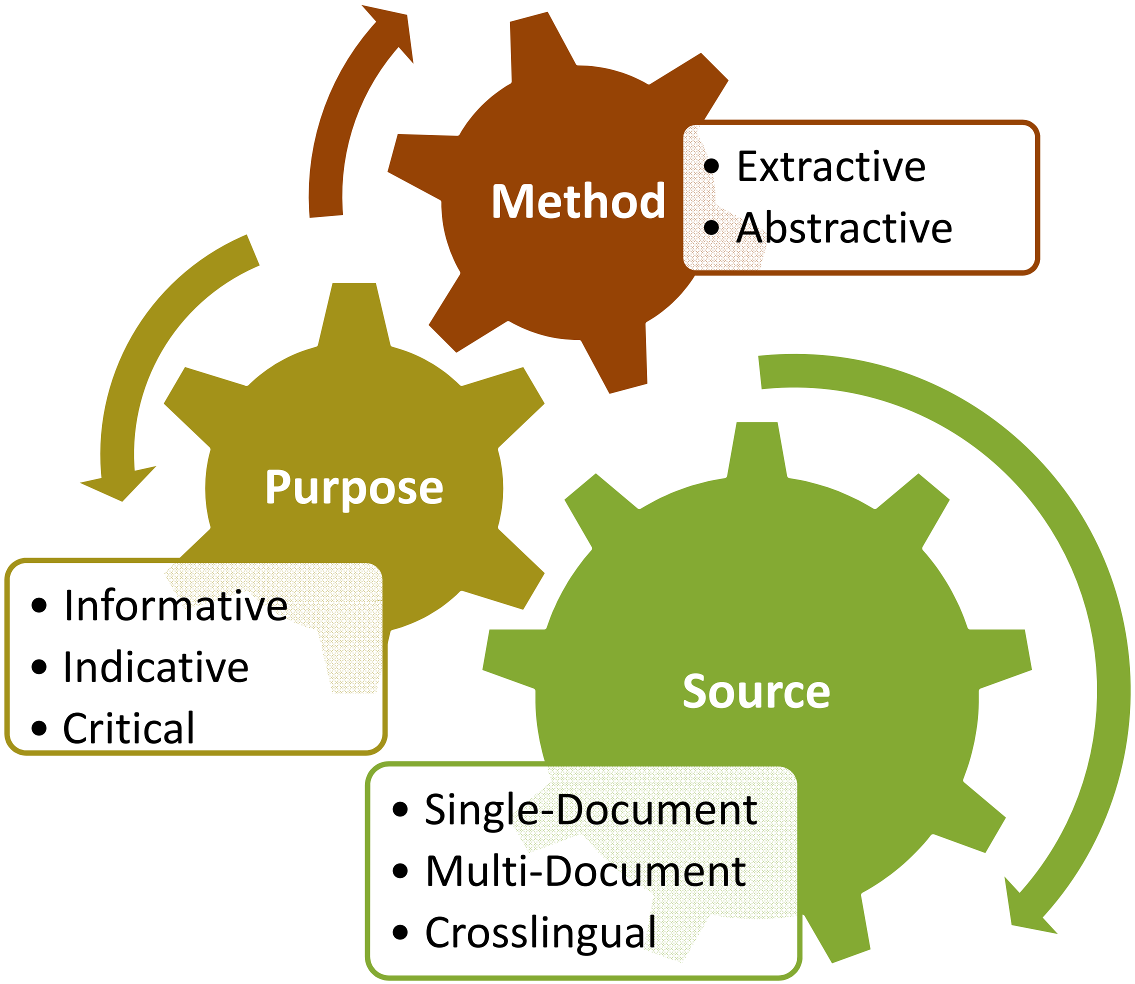
- Further converted to a loss function with a Sigmoid operator
- Corresponding parameters of WTM then are updated with a generalized probabilistic descent (GPD) procedure
- *Learning to rank* !

Associate documents with queries even if they do not share any of the query words!

II. Speech Summarization

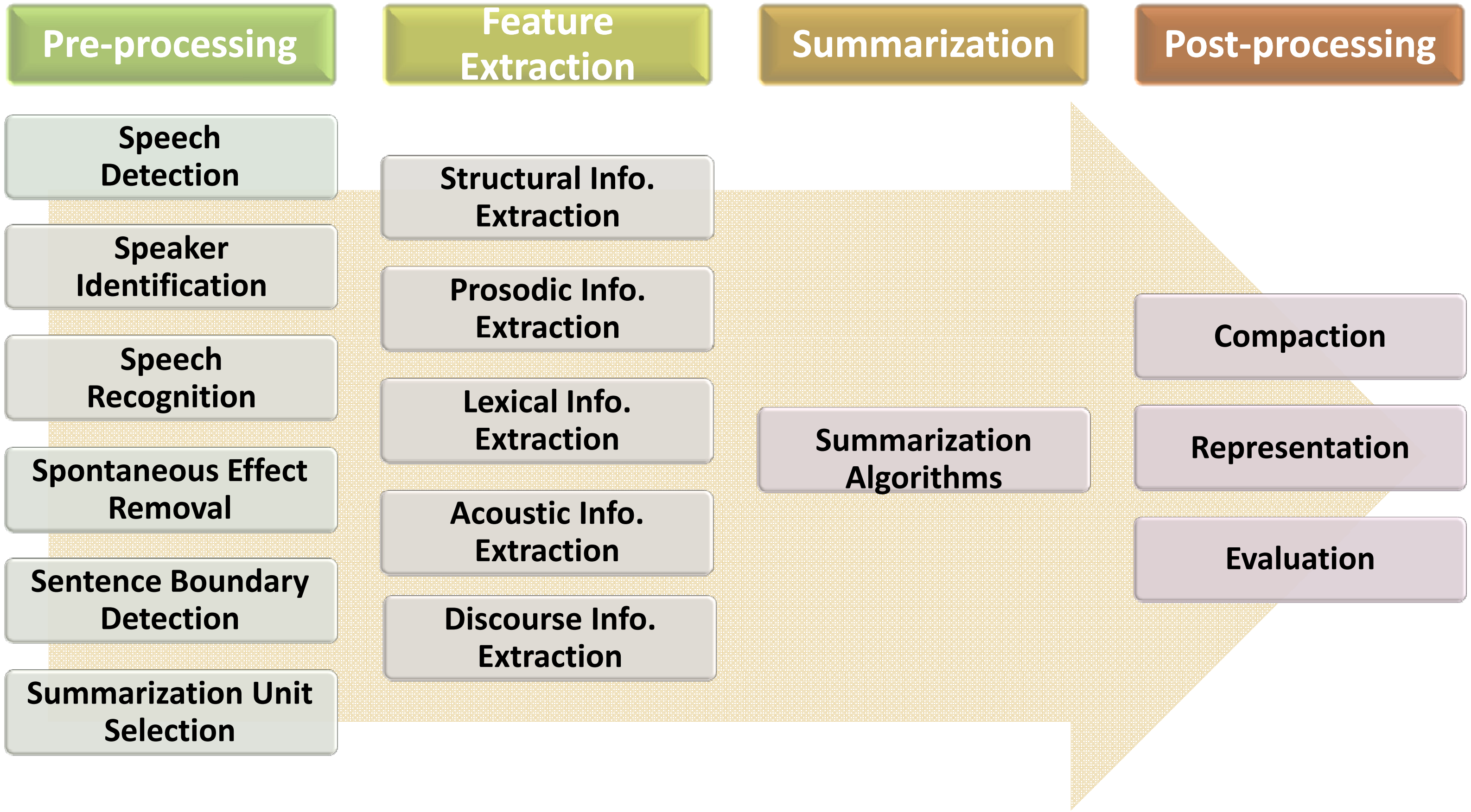
We proposed a "risk minimization" framework for speech summarization.
Co-contributor: Mr. Shih-Hsiang Lin

Spectrum of Summarization Research



Lin and Chen, "A survey on speech summarization techniques," *ACLCLP Newsletter* 21(1), 2010 (in Chinese)

Flowchart of Extractive Speech Summarization



Generic, Extractive Speech Summarization

- A summary is formed by selecting some salient sentences from the original spoken document
- A sentence to be selected as part of the summary is usually being considered by its
  - Significance
    - How importance of the sentence itself
  - Relevance
    - The degree of the similarity between the sentence and other sentences in the document
  - Redundancy
    - The information carried by the candidate summary sentence and the already selected summary sentences should be as dissimilar as possible

Related Work (1/3)

- Supervised Machine Learning Approaches (Significance)
  - The summarization task is usually cast as a two-class sentence-classification problem
  - A sentence is characterized by a set of indicative features
    - Acoustic cues, lexical cues, structural cues or discourse cues
  - Bayesian classifier (BC), support vector machine (SVM), conditional random fields (CRF)
  - Drawbacks
    - "*Bag-of-sentences*" assumption
    - Require manually labeled training data
    - Less generalization capability

Related Work (2/3)

- Unsupervised Machine Learning Methods (Relevance)
  - Based on the concept of sentence *centrality*
    - Sentences more similar to others are deemed more salient to the main theme of the document
  - Get around the need for manually labeled training data
  - Vector Space Model (VSM), Language Modeling (LM), Graph-based Algorithm (e.g., PageRank)
  - Drawbacks
    - The performance is usually worse than that of supervised summarizers
    - Most of methods constructed solely on the basis of the lexical information
      - Would be adversely affected by imperfect speech recognition

Related Work (3/3)

- Maximum Marginal Relevance (MMR) (Redundancy)
  - Perform sentence selection iteratively with the criteria of topic relevance and coverage
  - A summary sentence is selected according to
    - Whether it is more similar to the whole document than the other sentences (Relevance)
    - Whether it less similar to the set of sentences selected so far than the other sentences (Redundancy)

$$S_{MMR} = \arg\max_{S_i} \left[\beta \cdot Sim(S_i, D) - (1-\beta) \cdot \max_{S' \in \mathbf{Summ}} Sim(S_i, S') \right]$$

- None of the abovementioned methods fully address these three (Significance, Relevance, Redundancy) factors

A Risk Minimization Framework (1/4)

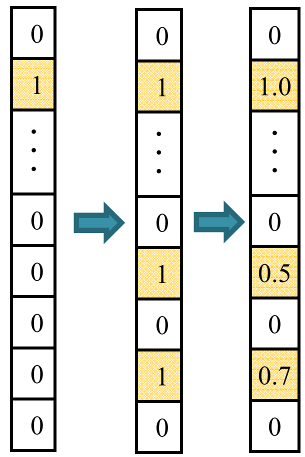
- Extractive summarization can be alternatively viewed as a decision making process
  - Select a representative subset of sentences or paragraphs from the original documents ➔ action

- Bayes decision theory can be employed to guide the summarizer in choosing a course of action
  - It quantifies the tradeoff between
    - Various decisions and the potential cost that accompanies each decision
  - The optimum decision can be made by contemplating each action
    - Choose the action that has the minimum expected risk

Lin and Chen, "A risk minimization framework for extractive speech summarization," *ACL 2010*

A Risk Minimization Framework (2/4)

- Without loss of generality, let us denote $\pi \in \mathbf{\Pi}$ as a selection strategy
  - It comprises a set of indicators to address the importance of each sentence S_i in a document D to be summarized
  - The feasible selection strategy can be fairly arbitrary according to the underlying principle
    - E.g., sentence-wise selection vs. list-wise selection



- Moreover, we refer to the k-th action a_k as choosing the k-th selection strategy π_k, and the observation O as the document D

A Risk Minimization Framework (3/4)

- The expected risk of a certain selection strategy π_k

$$R(\pi_k \mid D) = \int_{\pi} L(\pi_k, \pi) p(\pi \mid D) d\pi$$

- Therefore, the ultimate goal of extractive summarization could be stated as
  - The search of the best selection strategy π_{opt} from the space of all possible selection strategies that minimizes the expected risk

$$\begin{aligned}\pi_{opt} &= \arg\min_{\pi_k} R(\pi_k \mid D) \\ &= \arg\min_{\pi_k} \int_{\pi} L(\pi_k, \pi) p(\pi \mid D) d\pi\end{aligned}$$

A Risk Minimization Framework (4/4)

- Sentence-wise (iterative) selection

$$\begin{aligned} S^* &= \arg\min_{S_i \in \tilde{D}} R\left(S_i \mid \tilde{D}\right) \\ &= \arg\min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L\left(S_i, S_j\right) P\left(S_j \mid \tilde{D}\right) \end{aligned}$$

$$\pi_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

  - $\tilde{D}$ denotes the "residual" document

- By applying the Bayes' rule, the final selection strategy for extractive summarization is stated as

Relevance/Redundancy Relevance Significance

$$S^* = \arg\min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} L\left(S_i, S_j\right) \frac{P\left(\tilde{D} \mid S_j\right) P\left(S_j\right)}{\sum_{S_m \in \tilde{D}} P\left(\tilde{D} \mid S_m\right) P\left(S_m\right)}$$

Relation to Other Summarization Models

- 0-1 loss function

$$S^* = \arg\max_{S_i \in \tilde{D}} \frac{P\left(\tilde{D} \mid S_i\right) P\left(S_i\right)}{\sum_{S_m \in \tilde{D}} P\left(\tilde{D} \mid S_m\right) P\left(S_m\right)} = \arg\max_{S_i \in \tilde{D}} P\left(\tilde{D} \mid S_i\right) P\left(S_i\right)$$

  - A natural integration of the supervised and unsupervised summarizers
- Uniform prior distribution
  - Estimate the relevance between the document and sentence using $P\left(\tilde{D} \mid S_i\right)$
- Equal document-likelihood
  - Sentences are selected solely based on the prior probability $P\left(S_i\right)$

Implementation Details (1/4)

- Sentence Generative Model $P(\widetilde{D} \mid S_i)$
  - We explore the language modeling (LM) approach
    - Each sentence is simply regarded as a probabilistic generative model consisting of a unigram distribution for generating the document

$$P(\widetilde{D}|S_i) = \prod_{w \in \widetilde{D}} P(w|S_i)^{c(w,\widetilde{D})}$$

    - Maximum Likelihood Estimation (MLE) of $P(w|S_i)$
      - It may suffer from the problem of unreliable model estimation
      - Enhanced via topic modeling (PLSA, LDA, WTM, etc.)
      - Enhanced by incorporating relevance information

Chen et al., "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE TASLP*, 17(1), 2009

Implementation Details (2/4)

- Sentence Prior Model $P(S_i)$
  - We assume the sentence prior probability is in proportion to the posterior probability of a sentence being included in the summary class

$$P(S_i) \approx \frac{p(X_i \mid \mathbf{S})P(\mathbf{S})}{P(X_i \mid \mathbf{S})P(\mathbf{S}) + P(X_i \mid \overline{\mathbf{S}})P(\overline{\mathbf{S}})}$$

    - $\mathbf{S}$ and $\overline{\mathbf{S}}$: summary and non-summary classes
    - X_i : a set of indicative (prosodic/lexical/structural) features used for representing sentence S_i
  - Several popular supervised classifiers can be leveraged for this purpose
    - Bayesian Classifier (BC), Support Vector Machine (SVM), etc.

Implementation Details (3/4)

- Loss Function
  - VSM-based loss function $L(S_i, S_j)$
    - We use the TF-IDF weighting to calculate the cosine similarity
    - If a sentence is more dissimilar from most of the other sentences, it may incur higher loss

$$L(S_i, S_j) = 1 - Sim(S_i, S_j)$$

  - MMR-based loss function
    - Additionally address the "redundancy" issue

$$L(S_i, S_j) = 1 - \left[\beta \cdot Sim(S_i, S_j) - (1-\beta) \cdot \max_{S' \in \mathbf{Summ}} Sim(S_i, S') \right]$$

      - $\mathbf{Summ}$ the set of already selected summary sentences

Summarization Experiments (1/4)

- MATBN corpus
  - A subset of 205 broadcast news documents was reserved for the summarization experiments
  - The average Chinese character error rate (CER) is about 35%
  - Three subjects were asked to create summaries of the 205 spoken documents
  - The assessment of summarization performance is based on the widely-used ROUGE measure

	ROGUE-1	ROUGE-2	ROUGE-L
Agreement	0.600	0.532	0.527

*The agreement among the subjects for important sentence ranking for the evaluation set.

Summarization Experiments (2/4)

- Baseline experiments
  - Supervised summarizer – A Bayesian classifier (BC) with 28 indicative features determines the sentence prior probability $P(S_i)$
  - Unsupervised summarizer –A (unigram) language modeling approach determines the document-likelihood $P(D \mid S_i)$

	Text Document (TD)			Spoken Document (SD)		
	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
BC	0.445 (0.390 - 0.504)	0.346 (0.201 - 0.415)	0.404 (0.348 - 0.468)	0.369 (0.316 - 0.426)	0.241 (0.183 - 0.302)	0.321 (0.268 - 0.378)
LM	0.387 (0.302 - 0.474)	0.264 (0.168 - 0.366)	0.334 (0.251 - 0.415)	0.319 (0.274 - 0.367)	0.164 (0.115 - 0.224)	0.253 (0.215 - 0.301)

  - Erroneous transcripts cause significant performance degradation
  - BC outperforms LM
    - BC is trained with the handcrafted document-summary data
    - BC utilizes a rich set of features

Summarization Experiments (3/4)

- Experiments on proposed methods

		Text Document (TD)			Spoken Document (SD)		
Prior	Loss	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
BC	0-1	0.501	0.401	0.459	0.417	0.281	0.356
	SIM	0.524	0.425	0.473	0.475	0.351	0.420
	MMR	0.529	0.426	0.479	0.475	0.351	0.420

- Simple "0-1 Loss" gives about 4-5% absolute improvements as compared to the results of BC
- "SIM/MMR Loss" results in higher performance
  - MMR is slightly better than SIM
- The performance gaps between the TD and SD cases are reduced to a good extent

Summarization Experiments (4/4)

- Experiments on proposed methods

		Text Document (TD)			Spoken Document (SD)		
Prior	Loss	ROGUE-1	ROUGE-2	ROUGE-L	ROGUE-1	ROUGE-2	ROUGE-L
Uniform	SIM	0.405	0.281	0.348	0.365	0.209	0.305
	MMR	0.417	0.282	0.359	0.391	0.236	0.338

  - Assume the sentence prior probability $P(S_i)$ is uniformly distributed
    - The importance of a sentence is considered from two angles
      - Relationship between a sentence and the whole document
      - Relationship between the sentence and the other individual sentences
  - Additional consideration of the "sentence-sentence" relationship appears to be beneficial

Future Work on Speech Summarization

- Look for different selection strategies π_k
  - e.g., the listwise strategy
- Explore different modeling approaches and indicative features for the component models
- Investigate discriminative training criteria for training the component models
- Robustly represent the recognition hypotheses of spoken documents beyond the top scoring ones
- Extend and apply the proposed framework to multi-document summarization tasks
- …

Lin and Chen, "Improved speech summarization with multiple-hypothesis representations and Kullback-Leibler divergence measures," *Interspeech 2009*

III. Discriminative and Heteroscedastic Linear Feature Transformation

We proposed a new "heteroscedastic linear discriminant analysis" approach to speech recognition.
Co-contributors: Mr. Hung-Shin Lee and Dr. Hsin-Min Wang

Feature Dimension Reduction

- Two purposes of feature extraction by reducing the feature dimensionality :
  - Better discrimination
  - Less computational complexity

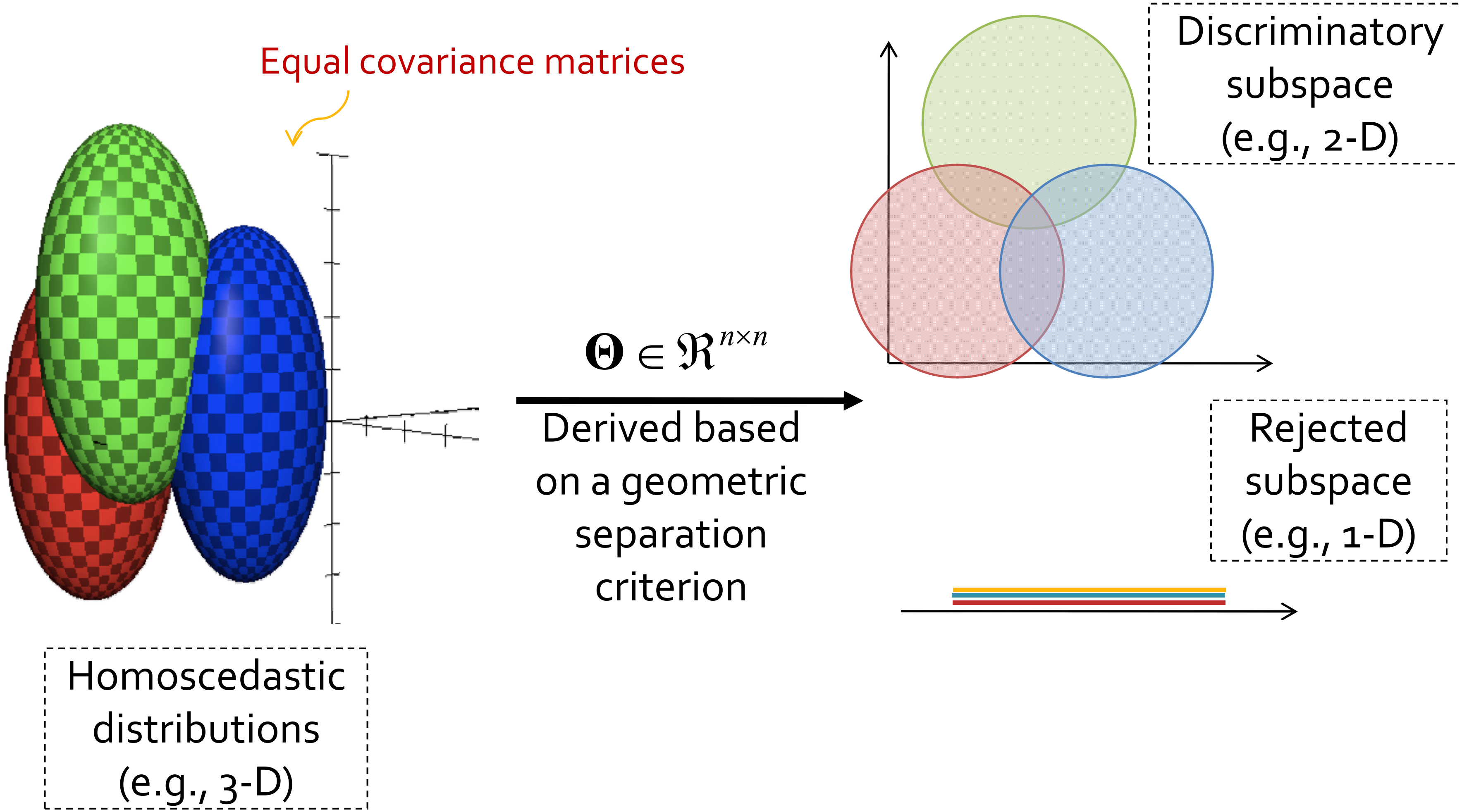
- Typically, we seek a linear transformation $\boldsymbol{\Theta} \in \Re^{n \times d}$
  - Project feature vectors from an ***n***-dimensional space to a ***d***-dimensional subspace
  - The resulting new features can possess good discriminatory power among classes

Previous Work (1/3)

- Linear discriminant analysis (LDA)
  - A procedure that maximizes the average squared Mahalanobis distance between each class-mean pair in the projective subspace
  - The derivation of the LDA transformation is equivalent to finding the parameters of multivariate Gaussian models by means of maximum-likelihood estimation (MLE), under the assumption that the whole class discrimination information resides in a d-dimensional subspace and that the within-class covariance matrices are equal for all classes

Previous Work (2/3)

- Pictorial Representation of LDA



Previous Work (3/3)

- Heteroscedastic linear discriminant analysis (HLDA)
  - Generalize LDA by dropping the homoscedastic assumption that all classes have equal within-class covariance matrices and maximizing the likelihood for these Gaussian models iteratively

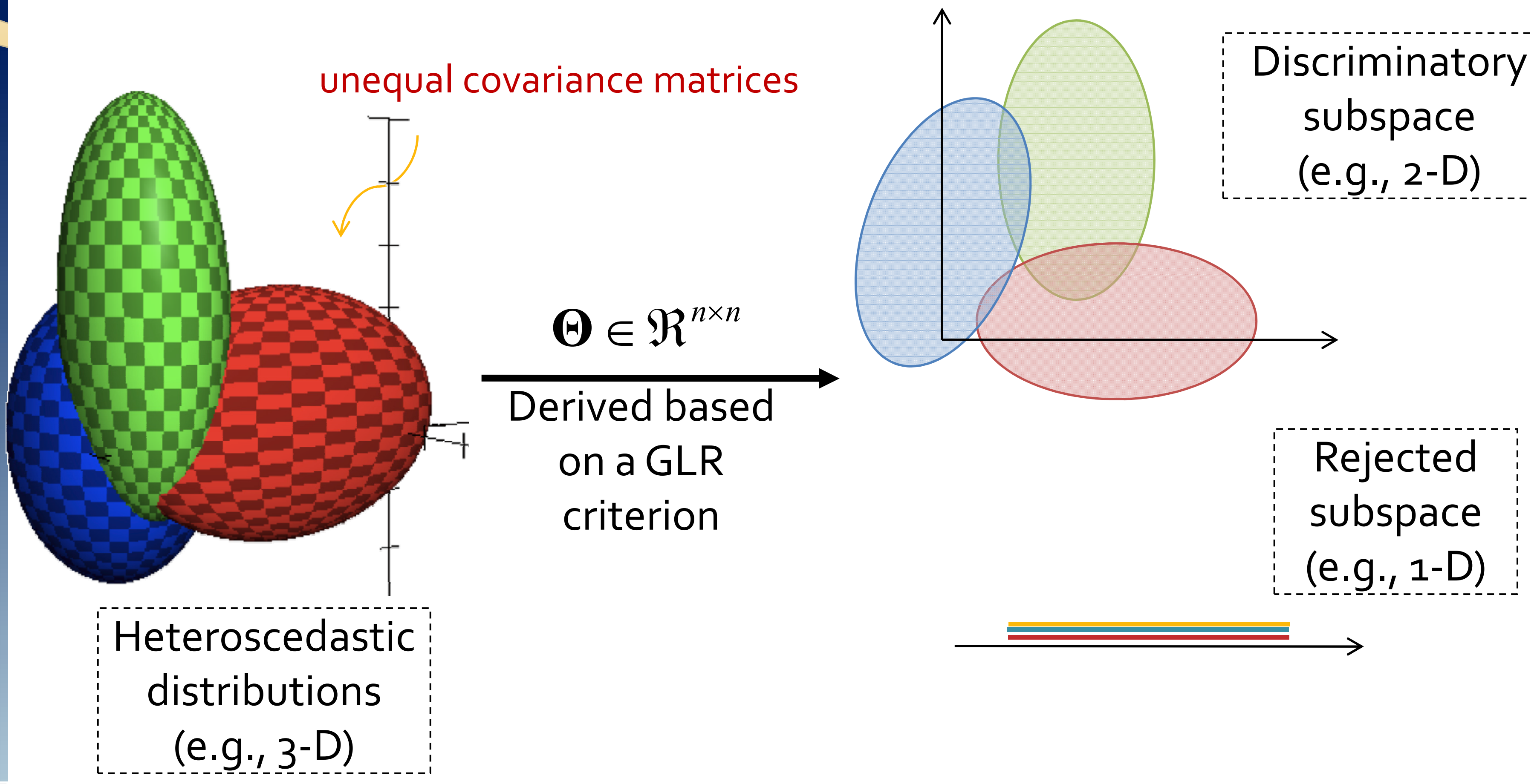
- Two common components in LDA and HLDA
  ① The transformation matrix is derived by maximizing the likelihood of all samples in the projective subspace
  ② The whole information for class discrimination resides in the ***d***-dimensional subspace, spanned by ***d*** column vectors of the transformation matrix

*f*GLRDA

- Full-rank generalized likelihood ratio discriminant analysis, *f*GLRDA
  - Seek a feature space, which is linearly isomorphic to the original ***n***-dimensional feature space and can be decomposed into a ***d***-dimensional discriminatory subspace and an (***n–d***)-dimensional non-discriminatory subspace
  - Make the most confusing situation, described by the null hypothesis, as unlikely as possible to happen without the homoscedastic assumption on underlying class distributions

Lee and Chen, "Generalized likelihood ratio discriminant analysis," *ASRU 2009*

Pictorial Representation of *f*GLRDA



Modified Likelihood Ratio Test (1/3)

- Likelihood Ratio Test (LRT):
  - A celebrated method of obtaining test statistics in any situation in which one wishes to test a null hypothesis $\boldsymbol{H_0}$ versus a completely general alternative $\boldsymbol{H_1}$
  - Criterion:

parameter space constrained by the null hypothesis

$$LR = \frac{\max_{\omega} L}{\max_{\Omega} L}$$

the likelihood of the sample

complete parameter space

Modified Likelihood Ratio Test (2/3)

- The logic behind LRT
  - If $\boldsymbol{H_0}$ is absolutely true with no confidence measure considered, the maximum likelihood over Ω should occur at a value of the parameter set consistent with ω
  - if $\boldsymbol{H_0}$ is false, the maximum likelihood will occur at a value of Ω that is not in ω. Thus $\max_{\omega} L$ will be smaller than $\max_{\Omega} L$

Modified Likelihood Ratio Test (3/3)

- Goal:
  - Seek a projecting subspace, where the null hypothesis unlikely to be true as far as possible
- *f*GLRDA:
  - The hypotheses should be designed for discriminating classes
    $\boldsymbol{H_0}$: All classes are **the same** in the projecting subspace
    $\boldsymbol{H_1}$: All classes are **different** in the projecting subspace

$$J_{\text{fGLRDA}}(\boldsymbol{\Theta}) = \frac{\max L_{\substack{\text{the parameter space that the} \\ \text{class populations are the same.}}}(\boldsymbol{\Theta})}{\max L_{\text{the complete paremeter space}}(\boldsymbol{\Theta})}$$

  - $\boldsymbol{\Theta}$ can be derived by minimizing $J_{\text{fGLRDA}}(\boldsymbol{\Theta})$

Derivational Details of *f*GLRDA (1/6)

- Model assumption: Gaussian distributions
  - The log-likelihood of the data in the transformed space

the Jacobian

$$\log L(\boldsymbol{\Theta}) = (-Nd/2)\log(2\pi) + N\log|\boldsymbol{\Theta}|$$

$$-\sum_{j=1}^{C}\frac{n_j}{2}\left((\tilde{\mathbf{m}}_j - \tilde{\boldsymbol{\mu}}_j)^T \tilde{\boldsymbol{\Sigma}}_j^{-1}(\tilde{\mathbf{m}}_j - \tilde{\boldsymbol{\mu}}_j) + \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_j^{-1}\tilde{\mathbf{S}}_j) + \log|\tilde{\boldsymbol{\Sigma}}_j|\right)$$

Means and covariance matrices
in the transformed space

Derivational Details of *f*GLRDA (2/6)

- The Gaussian parameters in the transformed space and the transformation matrix can be expressed by

$$\tilde{\boldsymbol{\mu}}_j = [\tilde{\boldsymbol{\mu}}_j^d, \tilde{\boldsymbol{\mu}}_0]^T = [\boldsymbol{\Theta}_d^T \boldsymbol{\mu}_j^d, \boldsymbol{\Theta}_{(n-d)}^T \boldsymbol{\mu}_0]^T,$$

$$\tilde{\boldsymbol{\Sigma}}_j = \begin{bmatrix} \boldsymbol{\Theta}_d^T \boldsymbol{\Sigma}_j^d \boldsymbol{\Theta}_d & 0 \\ 0 & \boldsymbol{\Theta}_{(n-d)}^T \boldsymbol{\Sigma}_0^{(n-d)} \boldsymbol{\Theta}_{(n-d)} \end{bmatrix}, \qquad \begin{matrix} \boldsymbol{\Sigma}_j^d \in \Re^{d \times d} \\ \boldsymbol{\Sigma}_0^{(n-d)} \in \Re^{(n-d)\times(n-d)} \end{matrix}$$

$$\boldsymbol{\Theta} = [\boldsymbol{\Theta}_d, \boldsymbol{\Theta}_{n-d}],$$

  - The discriminatory space $\boldsymbol{\Theta}_d$ and the rejected space $\boldsymbol{\Theta}_{n-d}$ are assumed to be independent

Derivational Details of *f*GLRDA (3/6)

- The heteroscedastic case
  - The hypotheses can be defined by
    H_0^{heter} : For each class, $\boldsymbol{\mu}_i = \boldsymbol{\mu}$, and $\boldsymbol{\Sigma}_i$ is unrestricted
    H_1^{heter} : For each class, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are unrestricted
  - The objective function of heteroscedastic *f*GLRDA, which needs to be minimized, can be logarithmically expressed as

$$J(\boldsymbol{\Theta}) = \max \log L_{H_0^{\text{hetero}}}(\boldsymbol{\Theta}) - \max \log L_{H_1^{\text{hetero}}}(\boldsymbol{\Theta}).$$

  - The model parameters and the transformation matrix can be derived by the **maximum likelihood estimation (MLE)**

Derivational Details of *f*GLRDA (4/6)

- The MLE statistics of *f*GLRDA under various hypotheses

Statistical Hypotheses	ML Estimates			
	Mean Vectors		Covariance Matrices	
	The first *d*-D Subspace	Rejected Subspace	The first *d*-D Subspace	Rejected Subspace
$H_0^{\text{hetero}}\begin{cases}\boldsymbol{\Sigma}_j : \text{unrestricted} \\ \boldsymbol{\mu}_j = \boldsymbol{\mu}\end{cases}$	$\boldsymbol{\Theta}_d^T\overline{\mathbf{w}}$	$\boldsymbol{\Theta}_{(n-d)}^T\overline{\mathbf{m}}$	$\boldsymbol{\Theta}_d^T\mathbf{W}_j\boldsymbol{\Theta}_d$	$\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}$
$H_1^{\text{hetero}}\begin{cases}\boldsymbol{\Sigma}_j : \text{unrestricted} \\ \boldsymbol{\mu}_j : \text{unrestricted}\end{cases}$	$\boldsymbol{\Theta}_d^T\mathbf{m}_j$	$\boldsymbol{\Theta}_{(n-d)}^T\overline{\mathbf{m}}$	$\boldsymbol{\Theta}_d^T\mathbf{S}_j\boldsymbol{\Theta}_d$	$\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}$

Derivational Details of *f*GLRDA (5/6)

- The MLE statistics of *f*GLRDA under various hypotheses

Statistical Hypotheses	(Relevant) Maximum Log-likelihood (not including the term $N\log\|\boldsymbol{\Theta}\|$)
$H_0^{\text{hetero}}\begin{cases}\boldsymbol{\Sigma}_j : \text{unrestricted} \\ \boldsymbol{\mu}_j = \boldsymbol{\mu}\end{cases}$	$-\sum_{j=1}^{C}\frac{n_j}{2}\log(\|\boldsymbol{\Theta}_d^T\mathbf{W}_j\boldsymbol{\Theta}_d\|\,\|\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}\|)$
$H_1^{\text{hetero}}\begin{cases}\boldsymbol{\Sigma}_j : \text{unrestricted} \\ \boldsymbol{\mu}_j : \text{unrestricted}\end{cases}$	$-\sum_{j=1}^{C}\frac{n_j}{2}\log(\|\boldsymbol{\Theta}_d^T\mathbf{S}_j\boldsymbol{\Theta}_d\|\,\|\boldsymbol{\Theta}_{(n-d)}^T\mathbf{S}_T\boldsymbol{\Theta}_{(n-d)}\|)$

Derivational Details of *f*GLRDA (6/6)

- The objective function of *f*GLRDA can be derived as

$$J(\boldsymbol{\Theta}) = \sum_{j=1}^{C} -\frac{n_j}{2}\log(|\boldsymbol{\Theta}_d^T \mathbf{W}_j \boldsymbol{\Theta}_d| \big/ |\boldsymbol{\Theta}_d^T \mathbf{S}_j \boldsymbol{\Theta}_d|)$$

$$\begin{cases} \mathbf{W}_j = (\mathbf{m}_j - \overline{\mathbf{w}})(\mathbf{m}_j - \overline{\mathbf{w}})^T + \mathbf{S}_j = \mathbf{B}_j + \mathbf{S}_j \\ \overline{\mathbf{w}} = \left(\sum_{j=1}^{C} n_j \boldsymbol{\Sigma}_j^{-1}\right)^{-1} \sum_{j=1}^{C} n_j \boldsymbol{\Sigma}_j^{-1} \mathbf{m}_j \end{cases}$$

- The derivative (for numerical optimization) is given by

$$\frac{\partial J(\boldsymbol{\Theta})}{\partial \boldsymbol{\Theta}} = -\sum_{j=1}^{C} n_j \frac{(-\mathbf{S}_j \boldsymbol{\Theta} \widetilde{\mathbf{S}}_j^{-1} \widetilde{\mathbf{B}}_j + \mathbf{B}_j \boldsymbol{\Theta}) \widetilde{\mathbf{S}}_j^{-1}}{1 + \mathrm{tr}(\widetilde{\mathbf{S}}_j^{-1} \widetilde{\mathbf{B}}_j)}$$

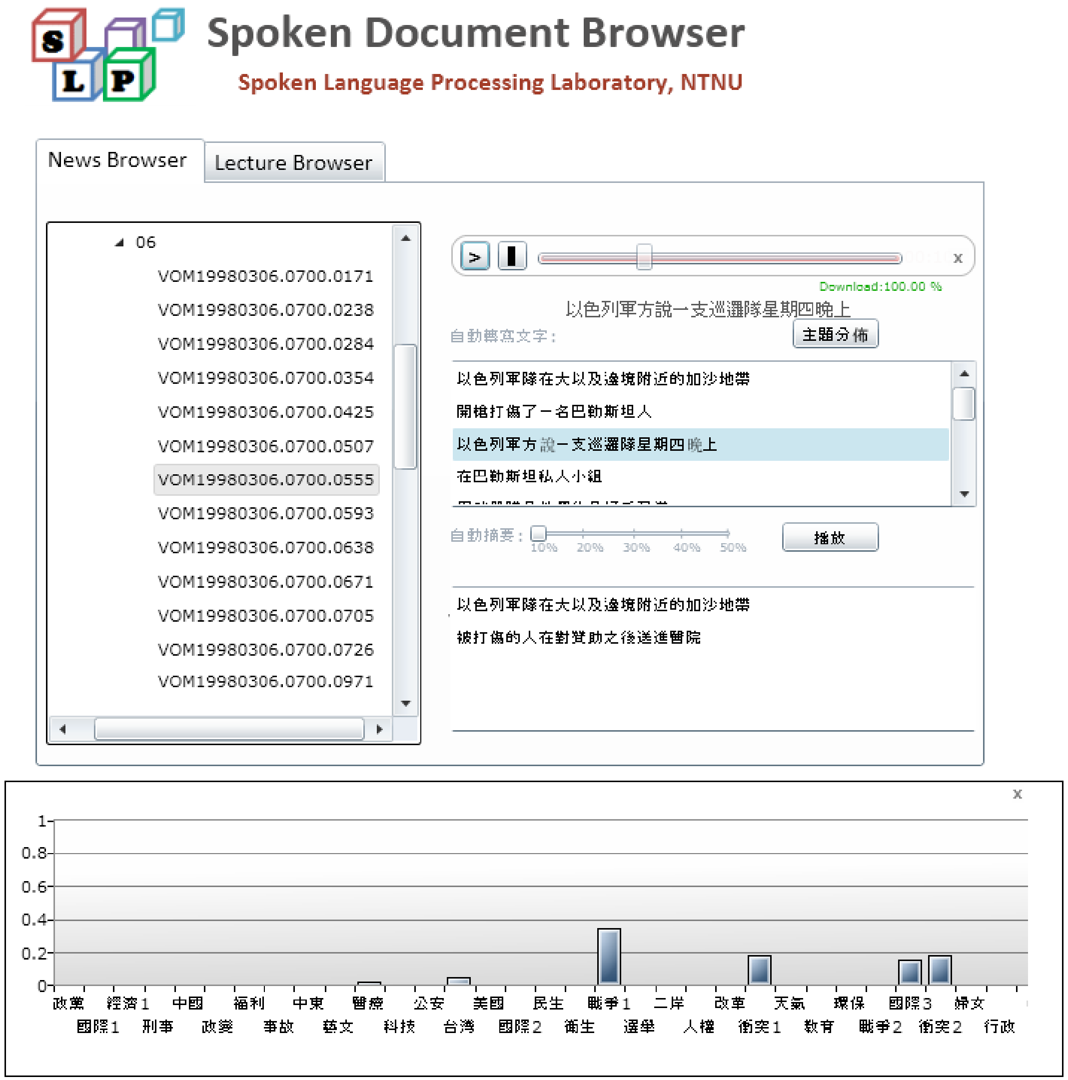
N T N U

Comments on *f*GLRDA

- A full-rank version of generalized likelihood ratio discriminant analysis (*f*GLRDA) for discriminative feature transformation on the basis of the likelihood ratio test was presented

- It is believed that methods designed along this vein would be more feasible when being applied to a wide array of pattern recognition tasks. As part of future work, the parameter settings in the rejected subspace can be designed in a more elaborate way

NTNU Lecture/News Browsing System



This system was designed and implemented by Mr. Shih-Hsiang Lin, 2010.
Currently available at: http://140.122.184.169/Browser/LectureAppTestPage.aspx

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
- Voice search provides good assistance for companies, for instance, in
  - Contact (Call)-center conservations: monitor agent conduct and customer satisfaction, increase service efficiency
  - Content-providing services: such as MOD (Multimedia on Demand): provide a better way to retrieve and browse descried program contents
- Speech processing technologies are expected to play an essential role in computer-aided (language) learning