Speech Recognition and its Applications to Computer-Assisted Language Learning

語音辨識與其在電腦輔助語言學習之應用

Berlin Chen (陳柏琳)
Professor, Department of Computer Science & Information Engineering
National Taiwan Normal University

2013/06/08

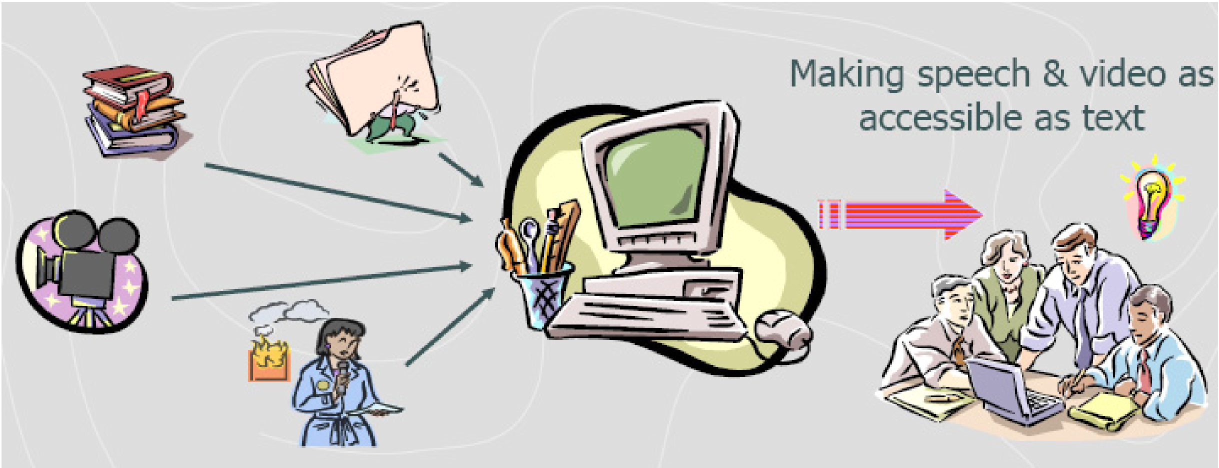
Introduction (1/3)

- Communication and search are by far the most popular activities in our daily lives
  - Speech is the most natural and convenient means for communications among humans (and between humans and machines in the future)
    - A spoken language interface could be more convenient than a visual interface on a small (or hand-held) 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 cues (e.g., transcripts, speakers, emotions, and scenes) for the data

Tur and Mori, *Spoken language understanding – systems for extracting Semantic Information from speech*, Wiley 2011.

Introduction (2/3)

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

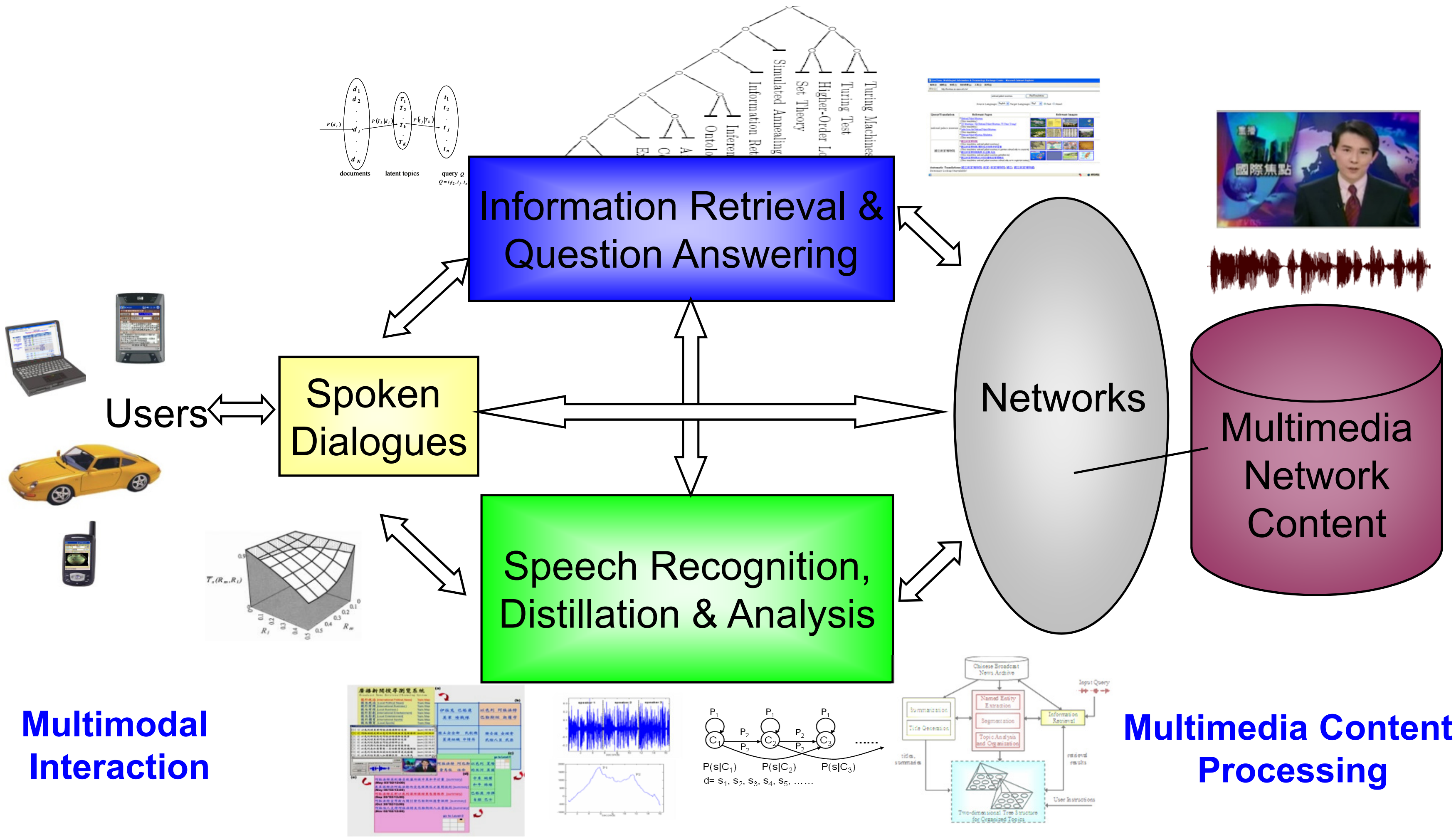


The figure is adapted from the presentation slides of Prof. Ostendorf at *Interspeech 2009.*

Introduction (3/3)

- Text Processing vs. Speech Processing
  - Recognition, Analysis and Understanding
    - **Text**: analyze and understand text
    - **Speech**: recognize speech (i.e., ASR), and subsequently analyze and understand the recognized text (propagations of ASR errors)
  - Variability
    - **Text**: different synonyms to refer to the same semantic object or meaning, such as 台灣師範大學, 師大, 教育界龍頭, etc.
    - **Speech**: an infinite number of utterances with respect to the same word (e.g., 台灣師範大學)
      - Manifested by a wide variety of oral phenomena such as disfluences (hesitations), repetitions, restarts, and corrections
      - Gender, age, emotional and environmental variations further complicate ASR
      - No punctuation marks (delimiters) or/and structural information cues exist in speech

Multimodal Access to Multimedia in the Future



Automatic Speech Recognition (ASR)

- Bayes Decision Rule (Risk Minimization)

$$W_{opt} = \arg\min_{W \in \mathbf{W}} Risk(W|O)$$

$$= \arg\min_{W \in \mathbf{W}} \sum_{W' \in \mathbf{W}} Loss(W, W') P(W'|O)$$

$$\approx \arg\max_{W \in \mathbf{W}} P(W|O)$$ Assumption of Using the "0-1" Loss Function

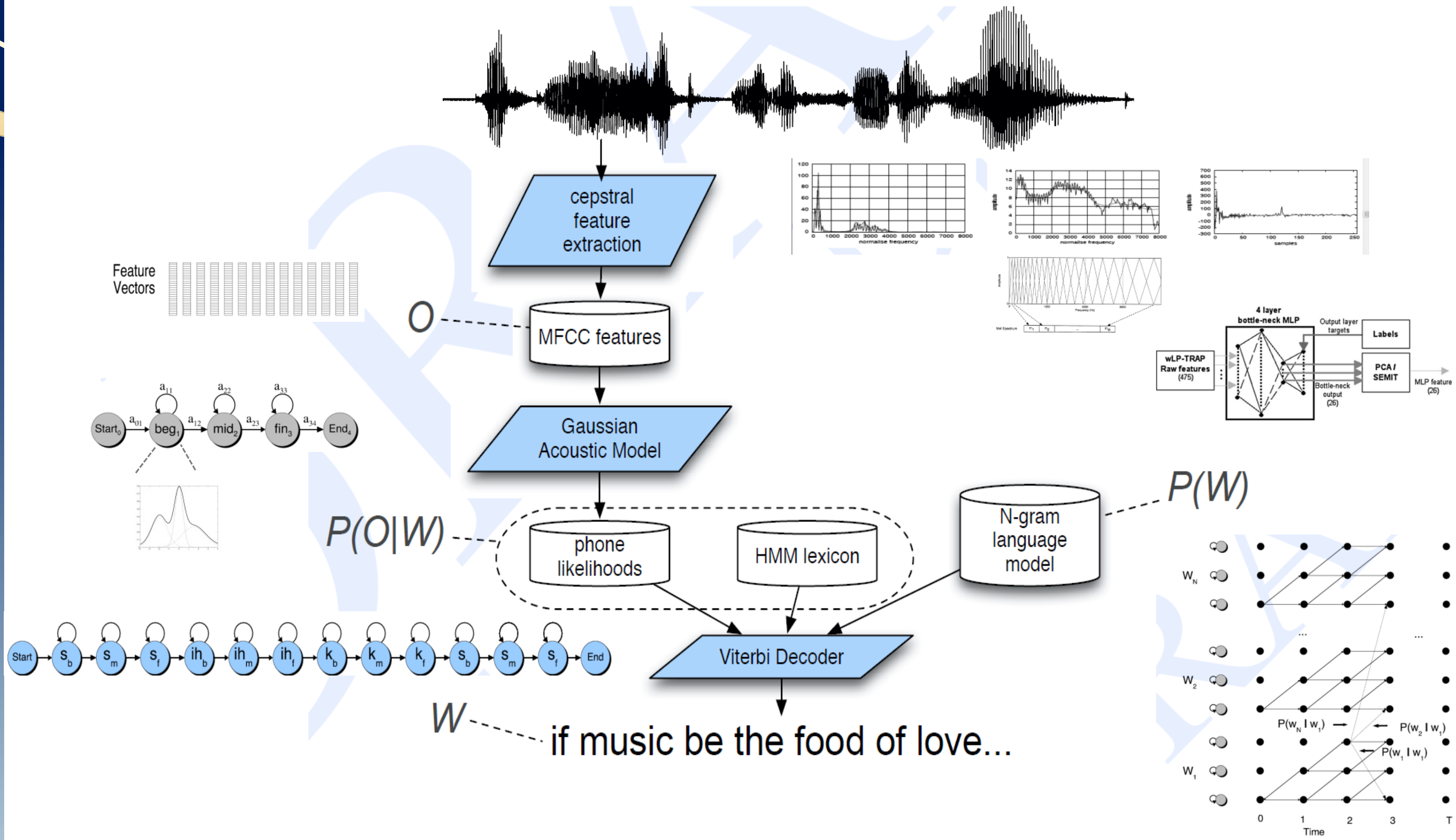
$$= \arg\max_{W \in \mathbf{W}} \frac{p(O|W)P(W)}{p(O)}$$

$$= \arg\max_{W \in \mathbf{W}} p(O|W)P(W)$$ Linguistic Decoding

Feature Extraction & Acoustic Modeling → $p(O|W)$; Language Modeling → $P(W)$

Possible variations: speaker, pronunciation, environment, context, etc. and OOV, domain, topic, style, etc.

Schematic Diagram of ASR



F. Valente et al., "Transcribing Mandarin broadcast speech using multi-layer perceptron acoustic features," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011.

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, pronunciation, speaker and context) 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 concert with natural language processing (NLP) and information retrieval (IR) techniques
  - A wild variety of potential applications
- Computer-Aided Language Learning (CALL)
  - Speech-driven Interface and multimedia content processing
  - Work in in association with natural language processing techniques
  - Applications
    - Automatic pronunciation assessment/scoring (CAPT)
    - Synchronization of audio/video learning materials
    - Estimation of document (writing) readability
    - Automated reading tutor (with spoken dialogues)
- 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*)

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.
  - *Rough'n'Ready* System at BBN Technologies
  - 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*
  - Apple's *Siri* (QA)
  - MIT Lecture Browser

We are witnessing the golden age of ASR!

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

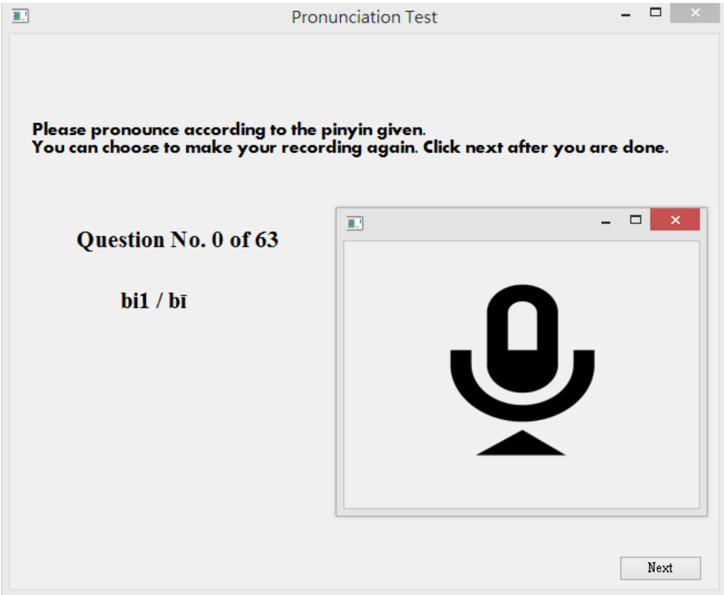
ASR for CALL Applications

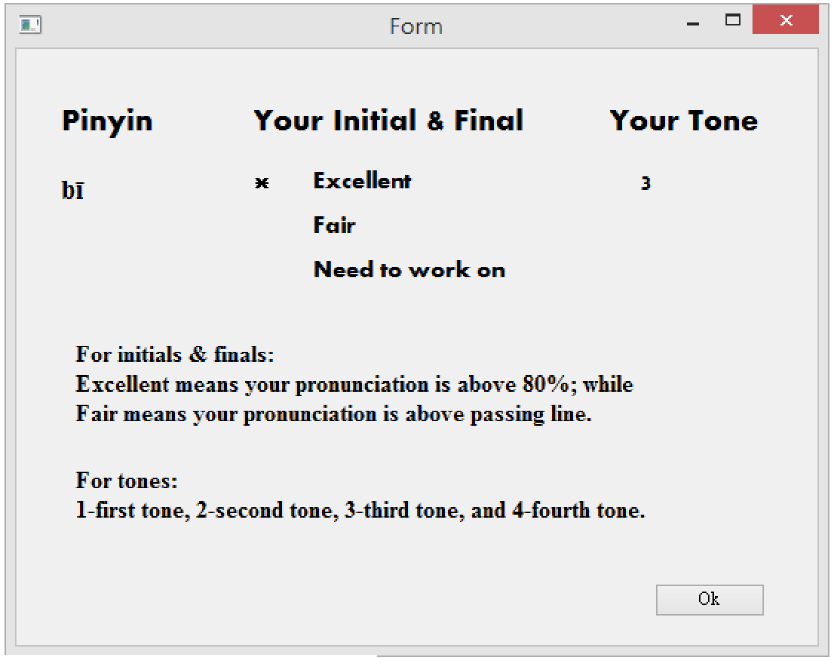
- The major task of applying ASR for CAPT (Computer Assisted Pronunciation Training) is to automatically detect pronunciation errors and evaluate pronunciation quality
- Facets for ASR-based CAPT in Mandarin Chinese
  - **Pronunciation of Lexical Tones**: Detection and Assessment
  - **Pronunciation of Sub-word (Syllable, INITIAL/FINAL) Units**: Detection and Assessment
  - **Duration/ Speaking Rate (Fluency/Proficiency):** Detection and Assessment
  - **Overall Scoring** (word-, phrase-, sentence-levels)

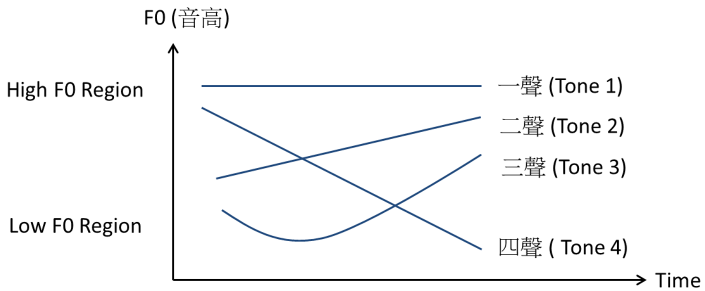
NTNU

Detection of Tone/Phone Mispronunciations (1/4)

- Detect possible mispronunciations and corresponding error patterns for a Chinese-language learner





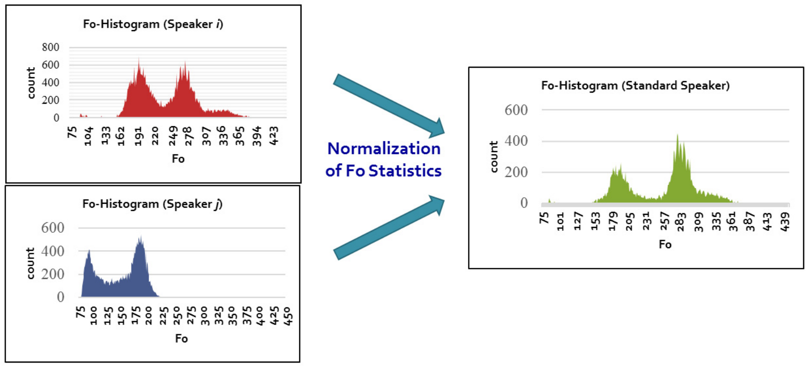


Detection of Tone/Phone Mispronunciations (2/4)

- Typical Steps for Lexical Tone Mispronunciation Detection

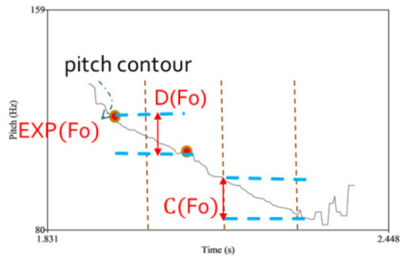
Pitch Extraction
↓
Pitch Normalization
↓
Segment-based Tonal Feature Extraction
↓
Detection

How to mitigate the negative effects caused by speaker and environmental variations?



How does the subtleness (granularity) of tonal features affect mispronunciation detection?

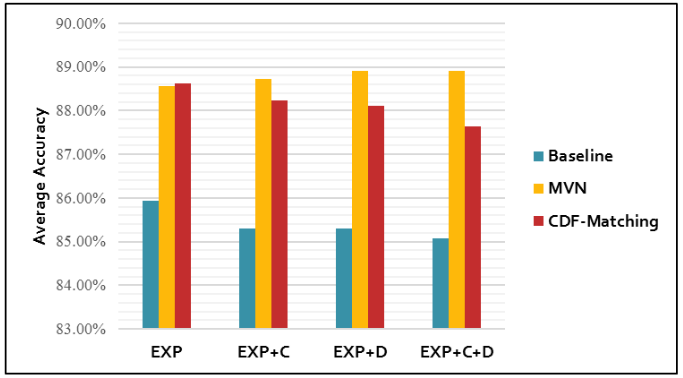
Types of Features	
EXP[Fo]	Mean Fo in each segment
C(Fo) (Within-Segment ΔFo)	Difference of beginning and ending Fo values within each segment
D(Fo) (Between-Segment ΔFo)	Difference of EXP[Fo] values between any pair of segments



The pitch contour is extracted with the **RAPT** (Robust Algorithm for Pitch Tracking) method.

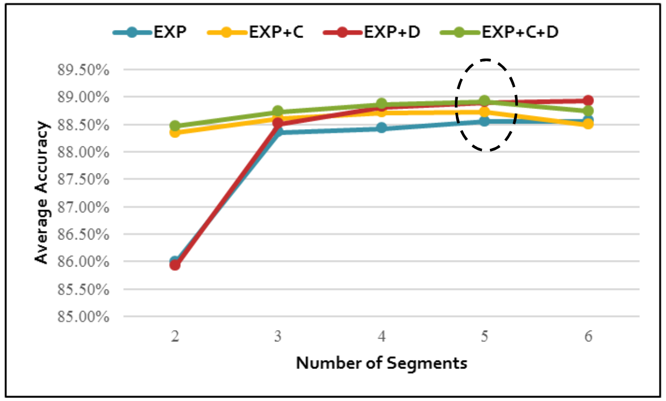
Detection of Tone/Phone Mispronunciations (3/4)

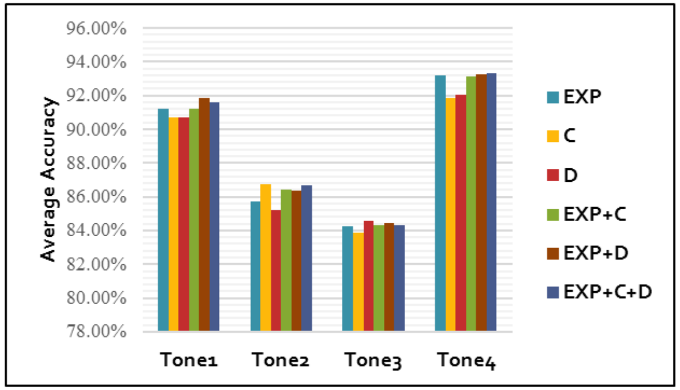
- Results on Automatic Detection of Tone Mispronunciations
  - Comparisons Among Different Normalization Methods



Both methods (MVN and CDF-matching) can offer significant performance boots compared to the baseline (without normalization)

  - Different Numbers of Segments for Tonal Feature Extraction



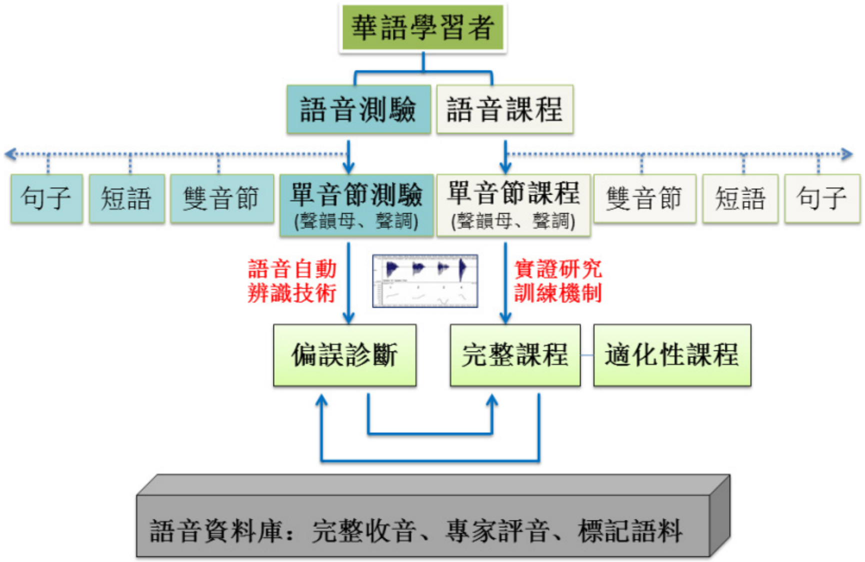


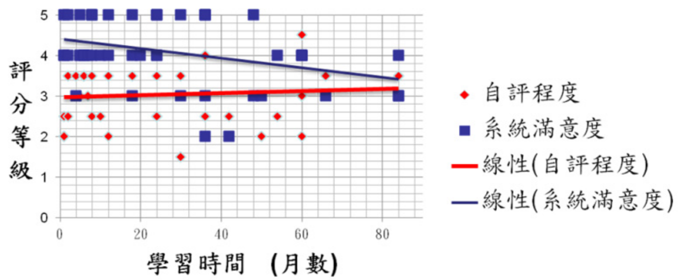
(with 5 segments)

H.-T. Hung et al., "An empirical study of tone mispronunciation detection methods for Mandarin CAPT applications," the 21st annual conference of the International Association of Chinese Linguistics (*IACL 2013*), 2013. (in Chinese)

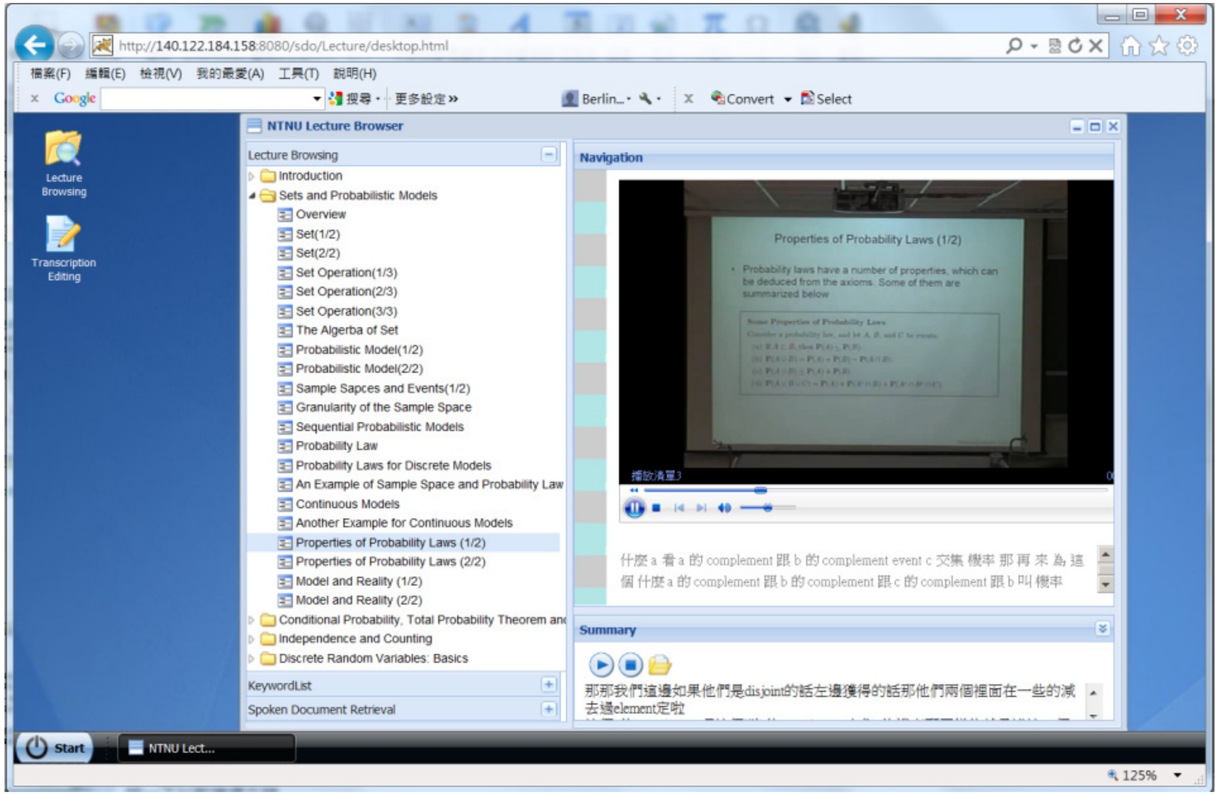
Detection of Tone/Phone Mispronunciations (4/4)

- We are now conducting a project to build a **Chinese Learning and Assessment System**



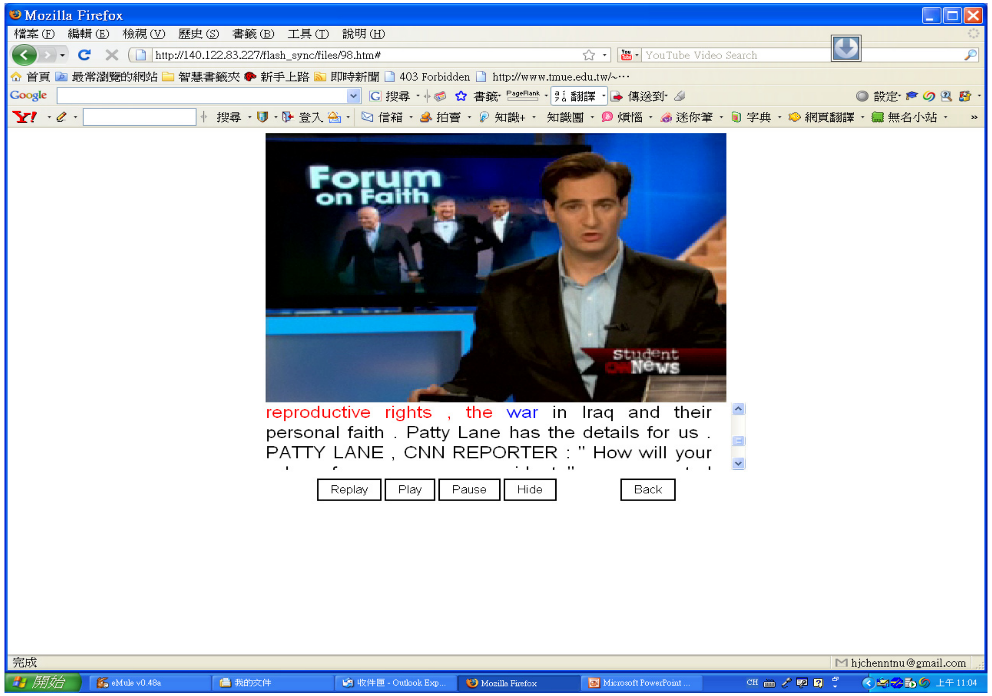


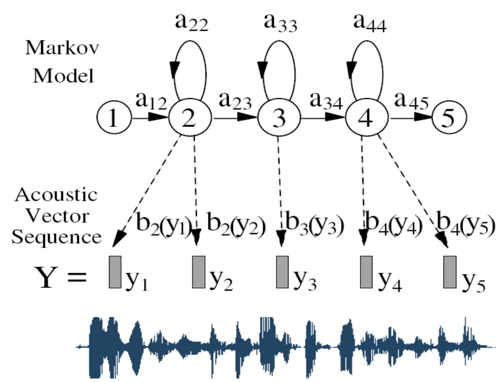
NTNU Lecture Browser for Self-Learning



- Featured with ASR-based automatic transcription and speech summarization and retrieval functionality

Video and Script Synchronization for Spoken English Learning





- This figure adapted from:

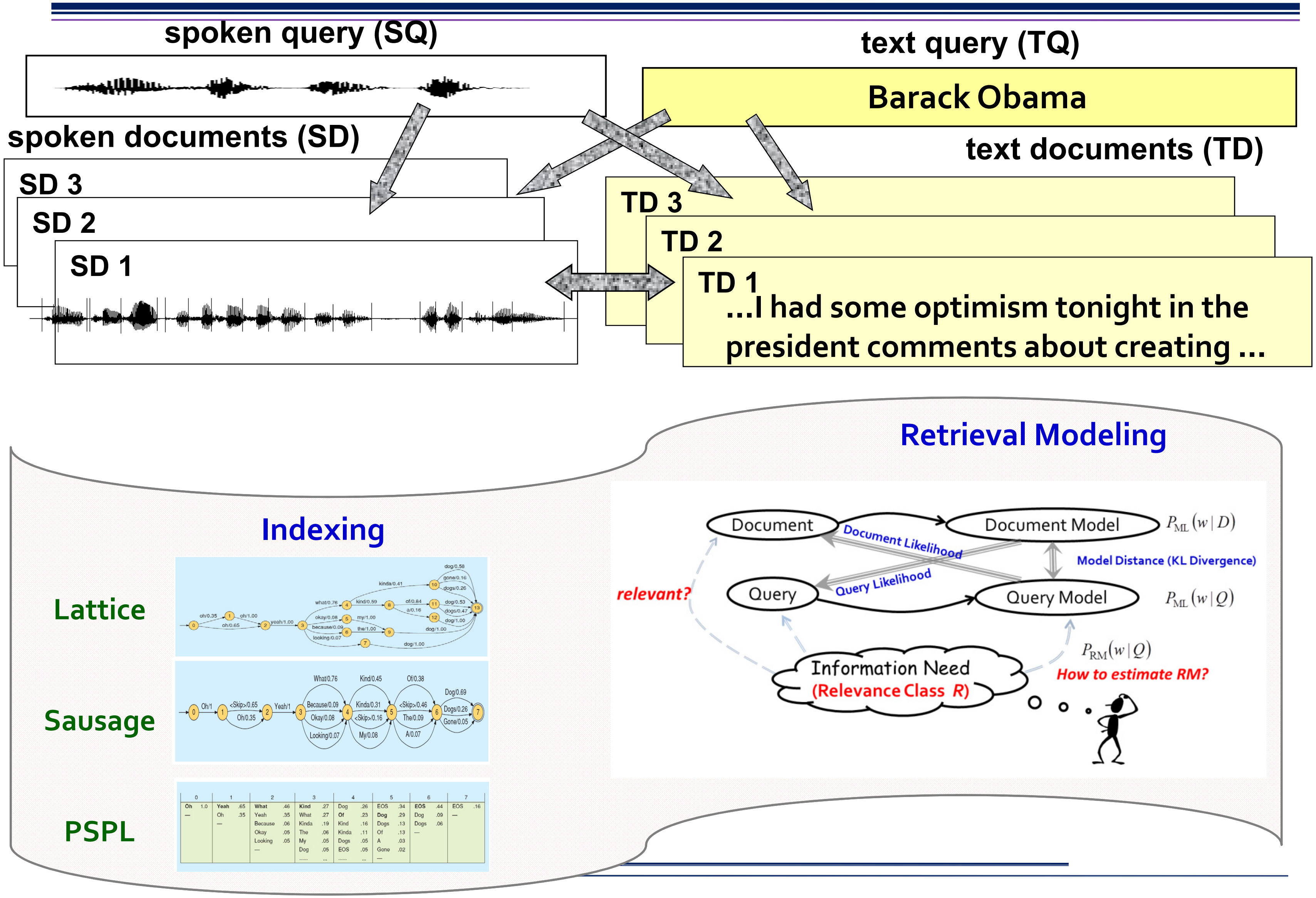
http://webh01.ua.ac.be/linguapolis/call2008/CALL%202008%20synctolearn.ppt

Howard Chen and Berlin Chen, "Developing SynctoLearn, an Automatic Video and Script Synchronization Tool, for Language Learners", *CALL2008*

Spoken Document Organization (NTU & NTNU)

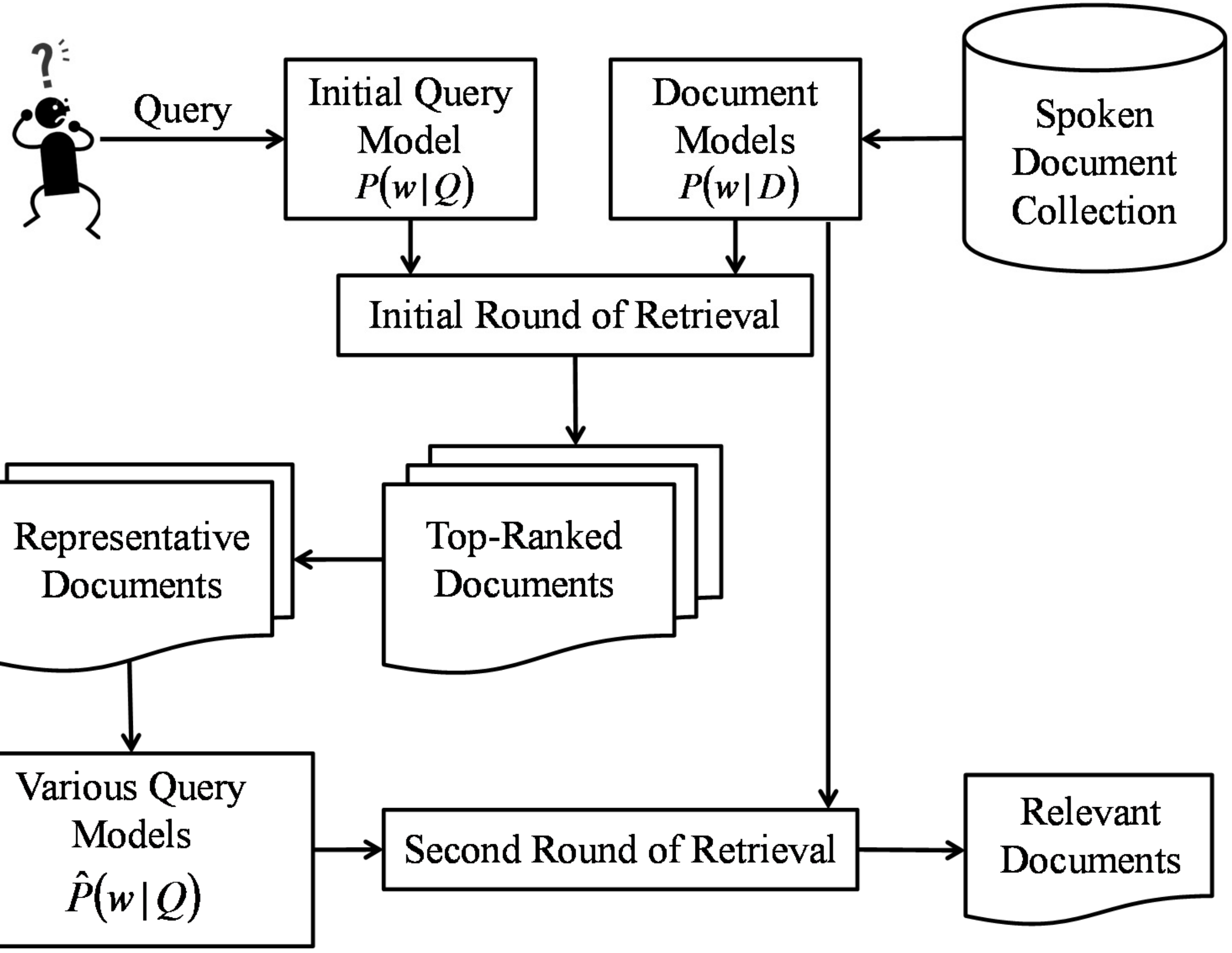
L.-S. Lee and B. Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, 2005.

Speech Retrieval: Scenarios and Methodologies



C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine* , 2008.

Speech Retrieval: Pseudo-relevance Feedback



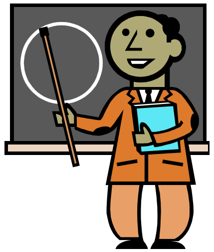
B. Chen et al., "Spoken document retrieval with unsupervised query modeling techniques," *IEEE Transactions on Audio, Speech and Language Processing*, 20(9), pp. 2602-2612, 2012.

Speech Summarization

conversations

meetings

lectures



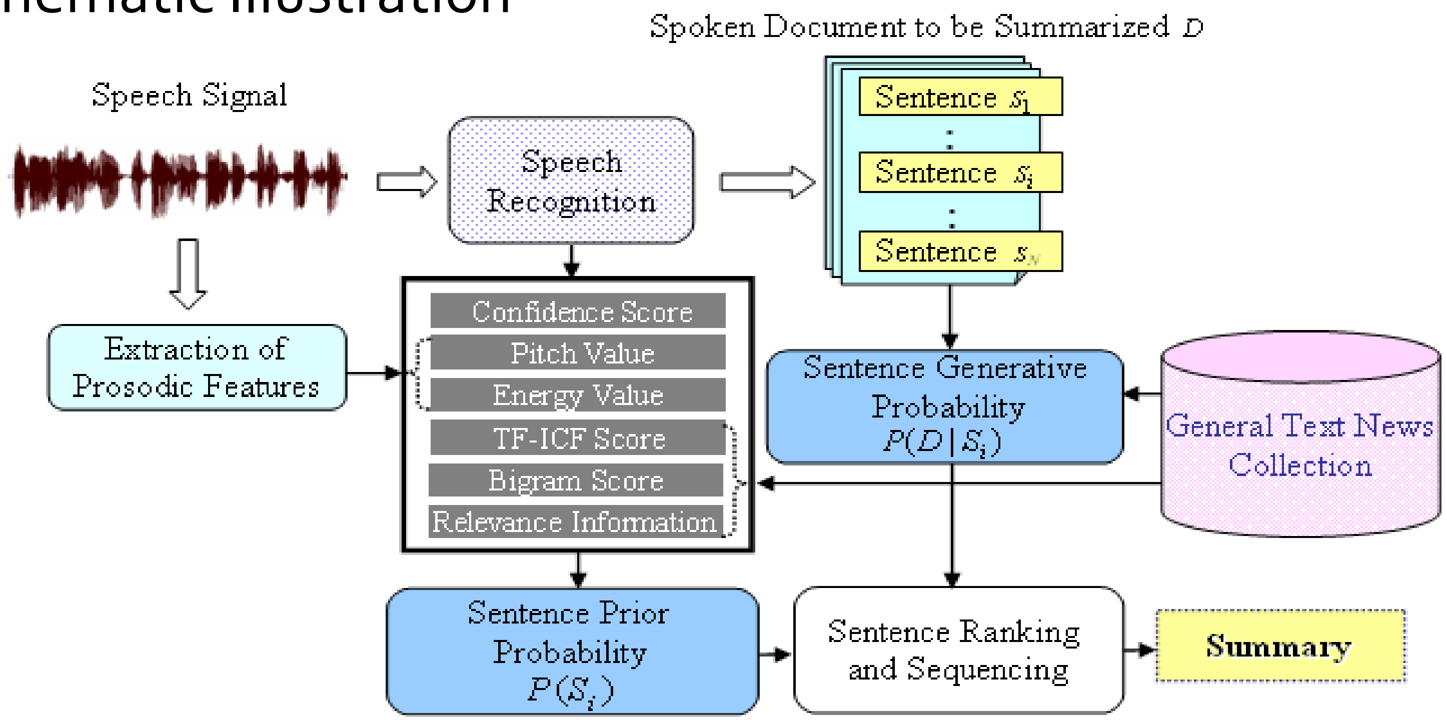
broadcast and TV news

distilling important information

abstractive vs. extractive
generic vs. query-oriented
single- vs. multi-documents

Risk Minimization-based Speech Summarization

- Schematic Illustration



$$S^* = \arg\min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} Loss(S_i, S_j) \cdot P(S_j | \tilde{D})$$

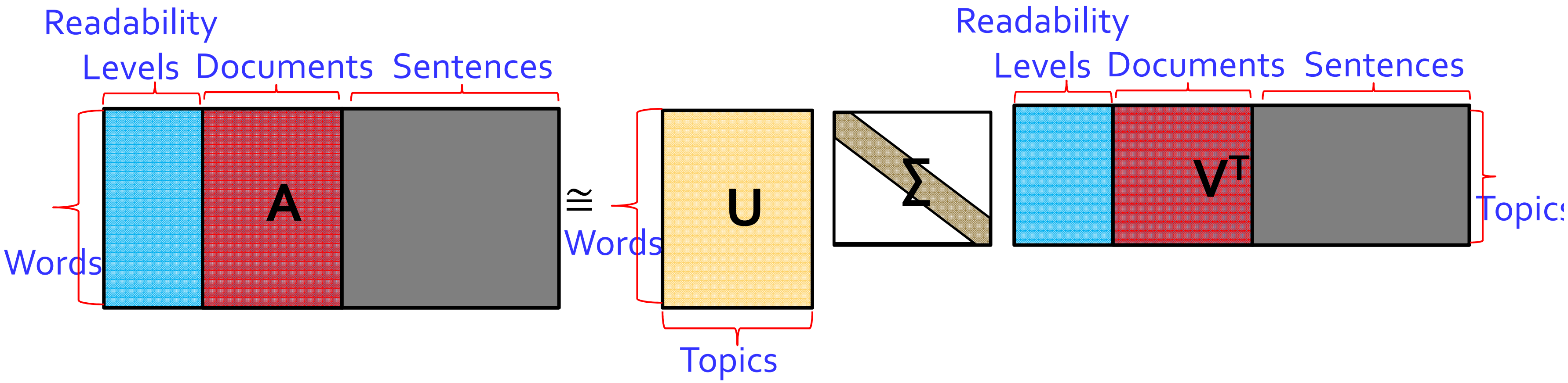
$$= \arg\min_{S_i \in \tilde{D}} \sum_{S_j \in \tilde{D}} Loss(S_i, S_j) \cdot \frac{P(\tilde{D}|S_j) P(S_j)}{\sum_{S_m \in \tilde{D}} P(\tilde{D}|S_m) P(S_m)}$$

Types	Description
Structural feature	1. Duration of the current sentence (S1)
Lexical features	1. Number of named entities (L1)
	2. Number of stop words (L2)
	3. Bigram language model scores (L3)
	4. Normalized bigram scores (L4)
Acoustic features	1. The 1st formant (F1-1 to F1-5)
	2. The 2nd formant (F2-1 to F2-5)
	3. The pitch value (P-1 to P-5)
	4. The peak normalized cross-correlation of pitch (C-1 to C-5)
Relevance features	1. Relevance score obtained by WTM
	2. Relevance score obtained by VSM
	3. Relevance score obtained by LSA
	4. Relevance score obtained by MRW

B. Chen and S.-H. Lin, "A risk-aware modeling framework for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 20(1), 2012

Readability Classification

- Leverage the **LSA (Latent Semantic Analysis)** based language modeling technique to extract "word-readability level", "word-document" and "word sentence" co-occurrence relationships



- Very Preliminary Results (10-fold tests; w.r.t. classification accuracy (%))

	NHK98 (410 documents)	國編版 (265 documents)
"word-readability level" relationship (dimensionality=6)	0.329	0.260
"word-readability level" & "word-document" relationships (dimensionality=20)	0.346	0.426

T. K. Landauer et al. (eds.) , *Handbook of Latent Semantic Analysis*, Lawrence Erlbaum, 2007.

Conclusion and Outlook

- Multimedia information (knowledge) 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 and warrant further investigation
- ASR technologies are expected to play an essential role in computer-aided (language) learning
  - As to future work, we would like to apply and extend our methods to automatic pronunciation scoring for sub-word (syllable, INITIAL/FINAL) units, and overall pronunciation quality evaluation
  - In addition, we are planning to leverage more state-of-the-art machine learning techniques for CAPT in Mandarin Chinese

NTNU