



Recent Developments in Speech Recognition Technologies and Their Applications

Berlin Chen (陳柏琳)

Department of Computer Science & Information Engineering

National Taiwan Normal University

2012/01/03

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
- Variability
 - Text: 台灣師範大學, 師大, 教育界龍頭, etc.
 - Speech: an infinite number of utterances pertain to the same word (e.g., 台灣師範大學)
 - Gender, age, emotional and environmental variations further complicate ASR
 - No punctuation marks (delimiters) or/and structural information cues exist in speech

Speech Processing

- Speech Production, Perception, and Modeling
 - Phonetics and phonology
- Speech Coding
- Speech Synthesis
 - Text-to-speech: speech synthesis & natural language generation
- Speech Recognition and Understanding

Reference: D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2008.

Phonetics and Phonology

- Phonetics
 - The study of speech sounds and their production, classification, and transcription
 - Specifically, how sounds are produced by their articulators of the human vocal tract, how they are realized acoustically, and how they can be digitized and processed
 - IPA (International Phonetic Alphabet) is widely used to describe the phones (or transcribe the sounds) of different languages
- Phonology
 - The study of the distribution and patterning of speech sounds in a language and of the tacit rules governing the speech pronunciation
 - E.g. in Mandarin Chinese: the **combinations** of syllables / characters (“他呀” and “天哪”) and the **variations** of tone realization (“好酒” pronounced as “毫酒”, while “；”起碼” as “騎馬”)

Vowel Height and Formants

- Schematic of "vowel space" for English vowels

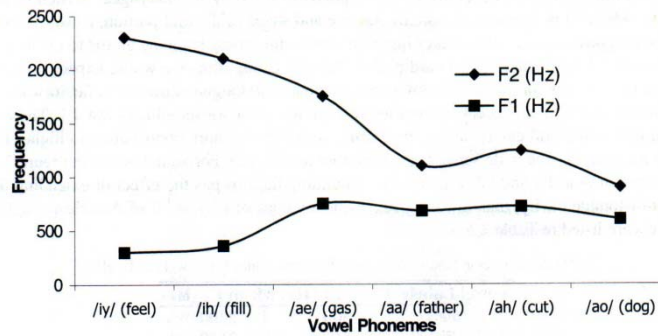
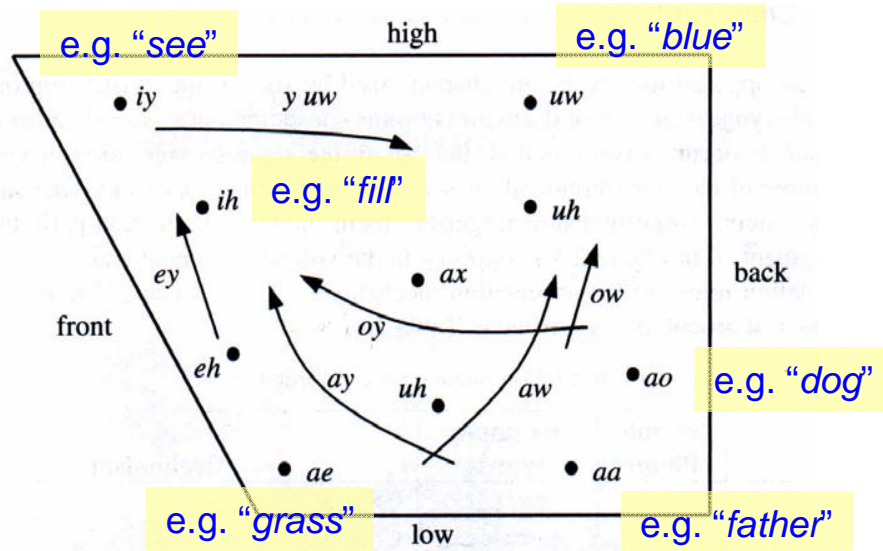


Figure 2.17 F1 and F2 values for articulations of some English vowels.

The major articulator for English vowels is the middle to rear portion of the tongue.

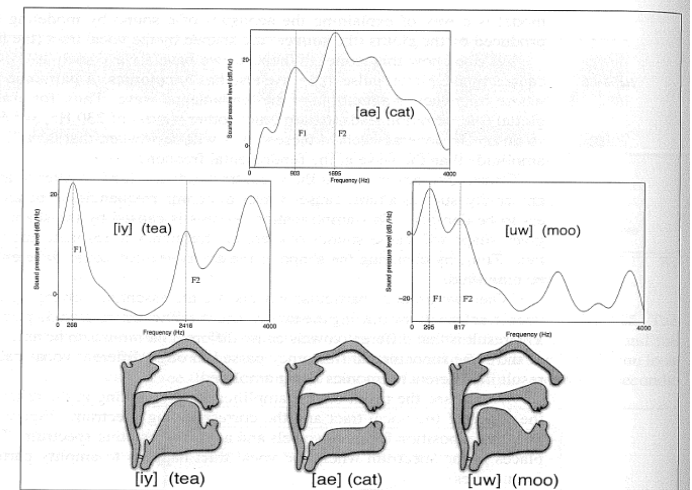


Figure 7.26 Visualizing the vocal tract position as a filter: the tongue positions for three English vowels and the resulting smoothed spectra showing F1 and F2.

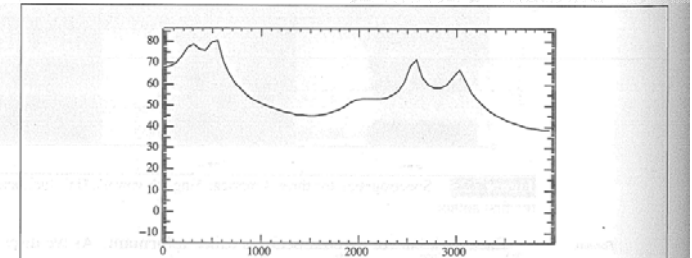


Figure 7.25 A smoothed (LPC) spectrum for the vowel [iy] at the start of *She just had a baby*. Note that the first formant (540 Hz) is much lower than the first formant for [ae] shown in Fig. 7.22, and the second formant (2581 Hz) is much higher than the second formant for [ae].

The location of the first two formants (called F1 and F2) plays a large role in determining vowel identity, although the formants still differ from speaker to speaker. Higher formants tend to be caused more by general characteristics of a speaker's vocal tract rather than by individual vowels. Formants also can be used to identify the nasal phonemes [n], [m], and [ŋ] and the liquids [l] and [r].

Vowel Triangles

- Neutral Speech (German)

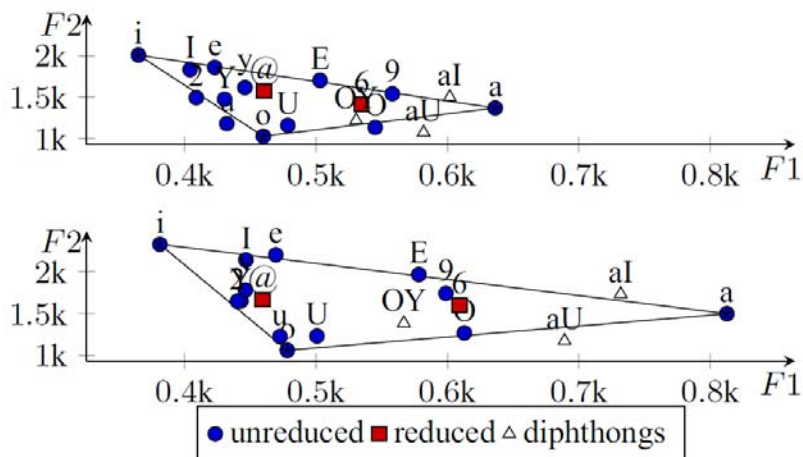


Figure 1: The German vowels triangle. Male (top), female (bottom).

- Affective speech (German)

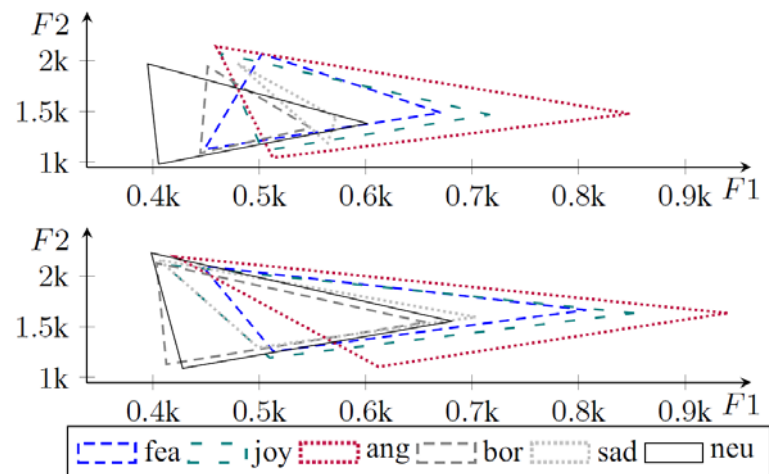
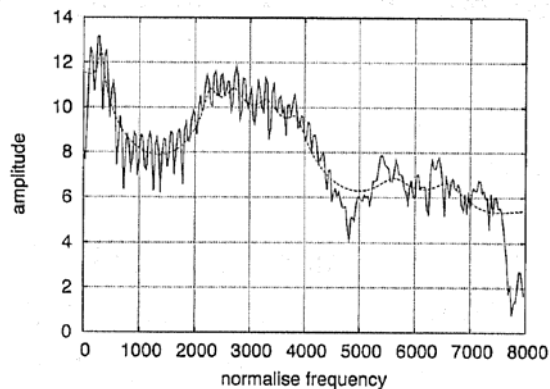


Figure 2: Classical vowel triangle form for different speaker's emotional states. Speakers: male (top), female (bottom).

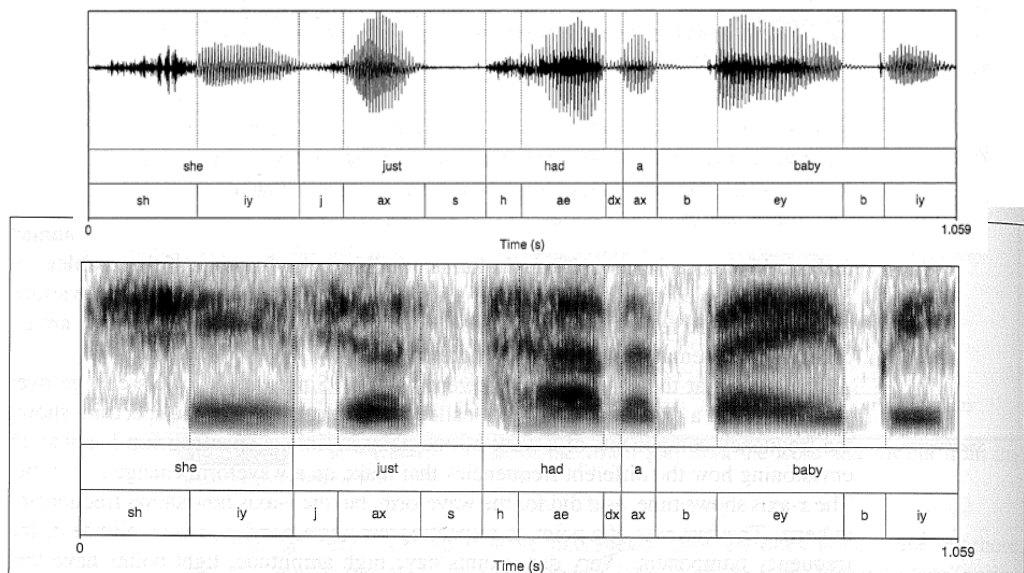
Spectrum, Spectrogram & Tone Patterns

- (log Magnitude) Spectrum



Effects of resonances
harmonics are evident.

- Spectrogram



- Tone Patterns

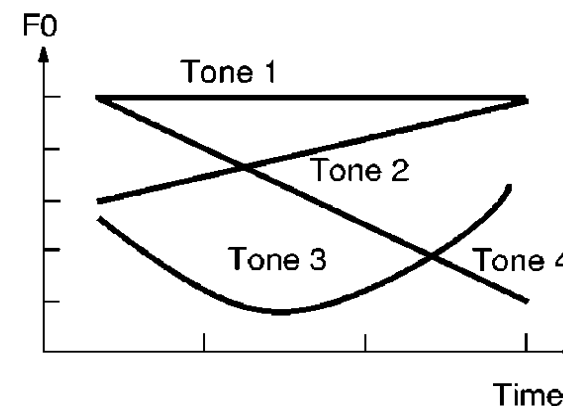
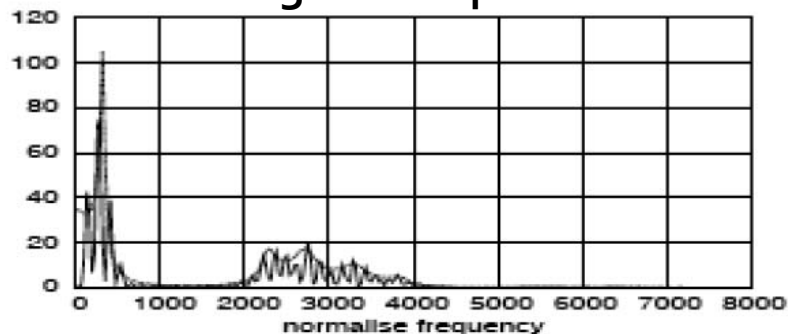


Figure 1: Pitch patterns of four lexical tones.

Figure 7.24 A spectrogram of the sentence “she just had a baby” whose waveform was shown in Fig. 7.17. We can think of a spectrogram as a collection of spectra (time slices), like Fig. 7.22 placed end to end.

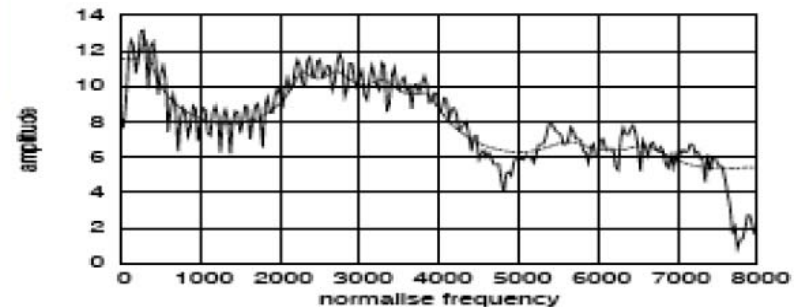
Cepstrum

magnitude spectrum



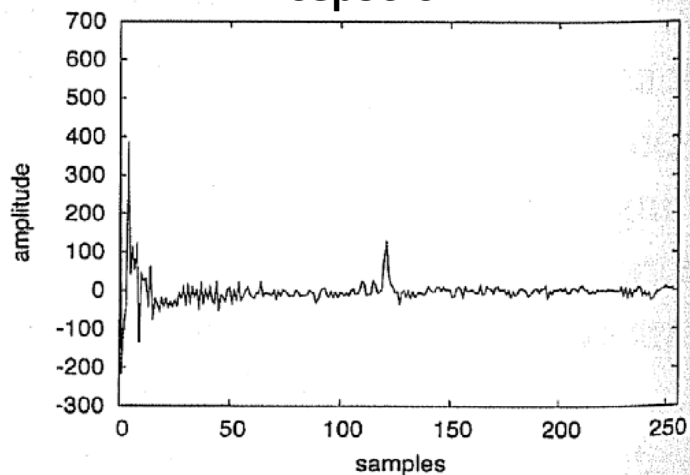
(a)

log magnitude spectrum



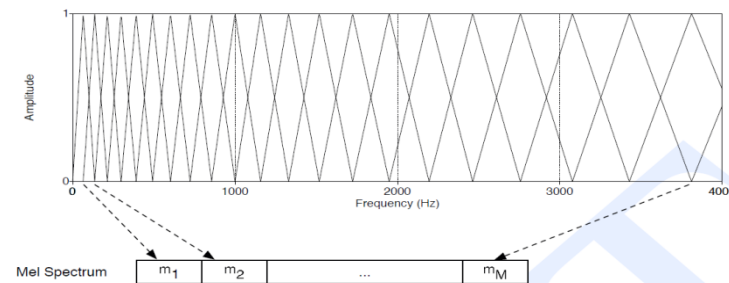
(b)

cepstrum



(c)

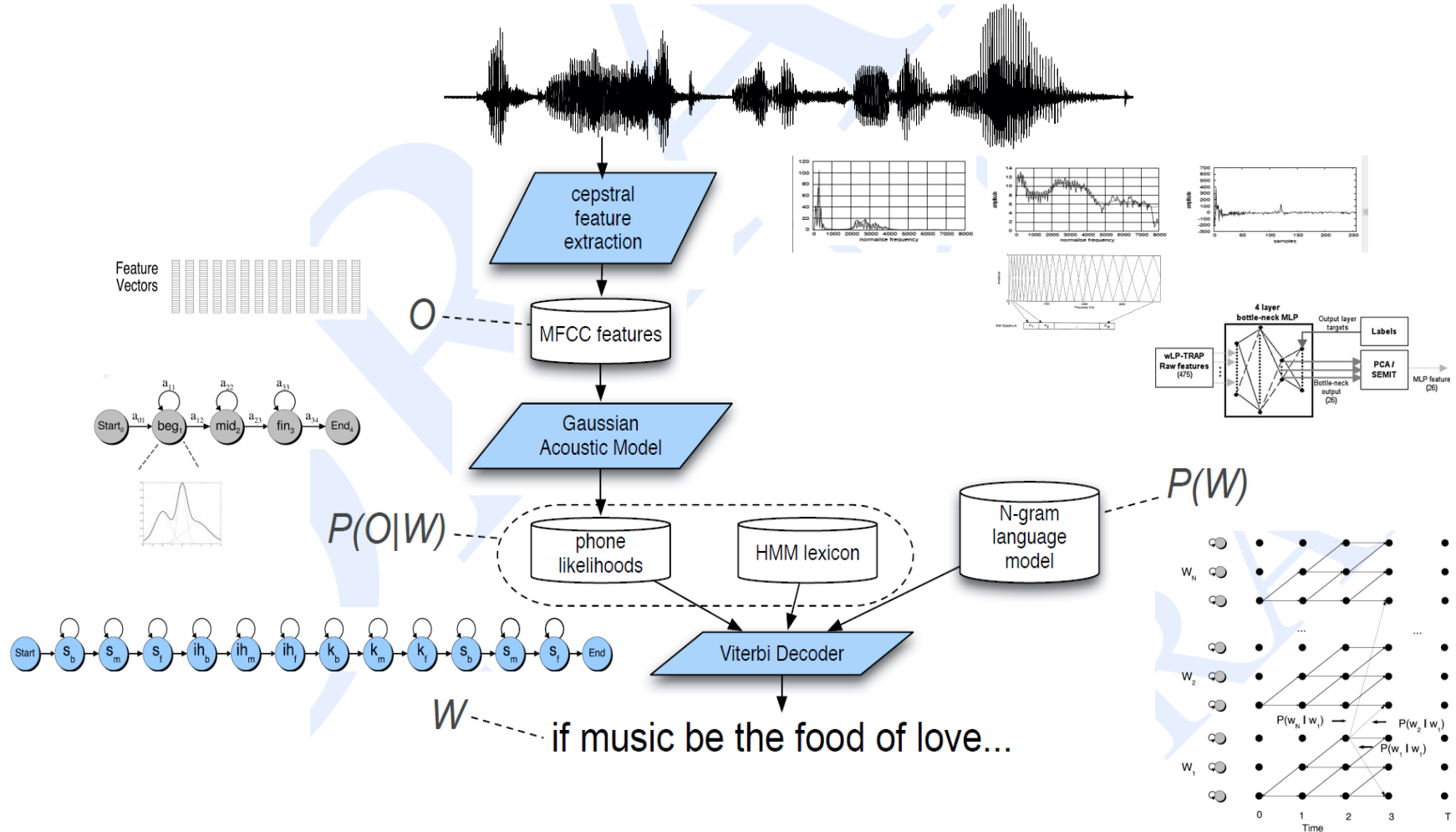
Mel filter bank



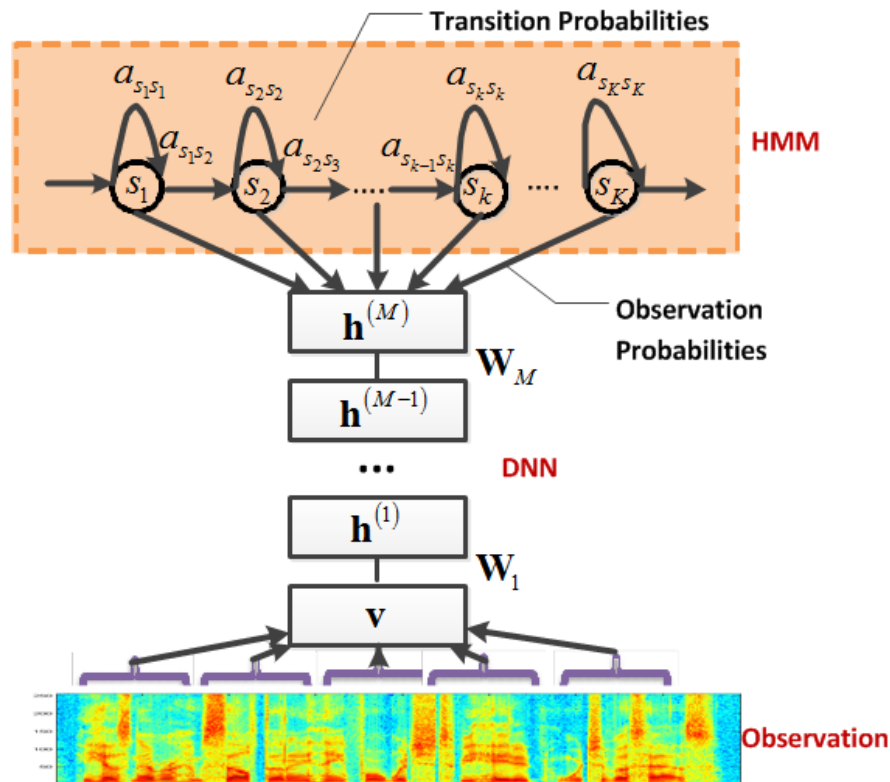
The Mel filter bank, after Davis and Mermelstein (1980). Each triangular filter collects energy from a given frequency range. Filters are spaced linearly below 1000 Hz, and logarithmically above 1000 Hz.

Automatic Speech Recognition (ASR)

- Schematic illustration



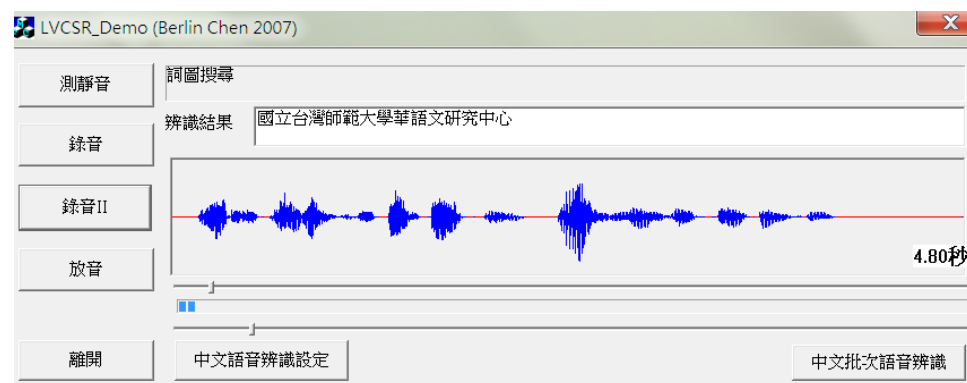
Context-Dependent Deep Neural Networks



- The HMM models the sequential property of the speech signal, and the DNN models the scaled observation likelihood of all the senones (tied tri-phone states). The same DNN is replicated over different points in time.

LVCSR (1/2)

- Large vocabulary continuous speech recognition (LVCSR)
大詞彙連續語音辨識



LVCSR (2/2)



Automatic

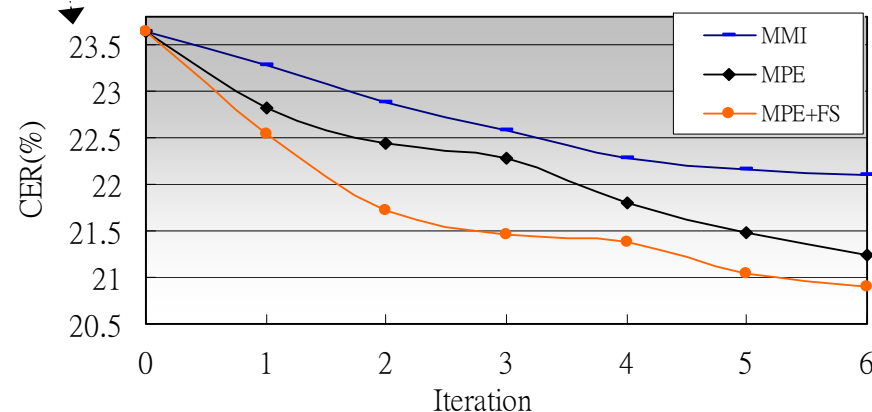
根據最新但雨量統計
 一整天下來
 費雪以試辦兩個水庫的雨量
 分別是五十三公里和二十九公厘
 對水位上升幫助不大
 不過就業機會期間也多在夜間
 氣象局也針對中部以北及東北部地區發佈豪雨特報
 因此還是有機會增加積水區的降雨量
 此外氣象局也預測
 華航又有另一道鋒面通過
 水利署估計如果這波鋒面能帶來跟著會差不多的雨水
 那個北台灣的第二階段限水時間
 渴望見到五月以後
 公視新聞當時匯率採訪報導

Manual

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

Maximum Likelihood (ML) Training

$$F_{ML}(\Lambda) = \sum_r P_{\Lambda}(O_r | W_r)$$



Discriminative training of acoustic models can further improve the ASR performance:

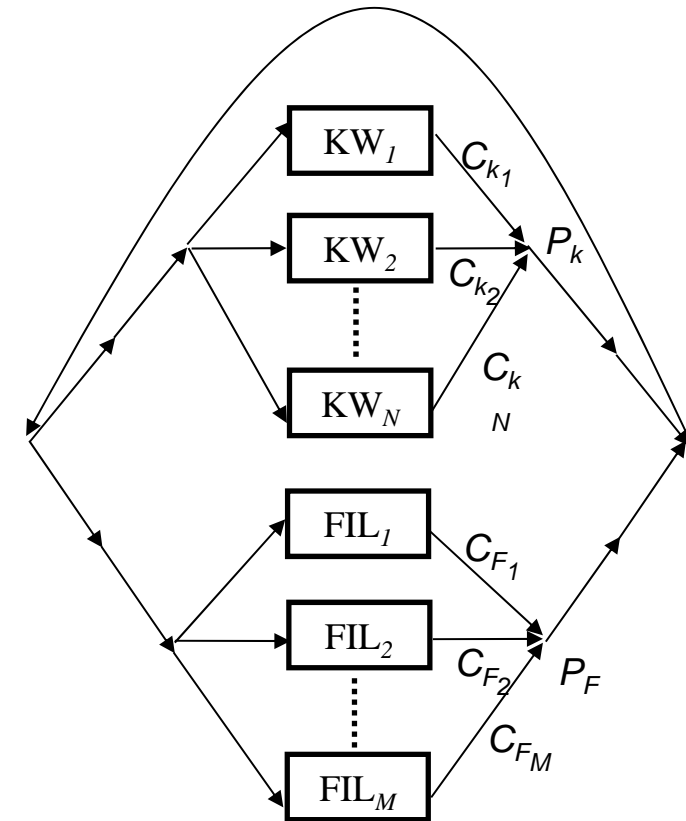
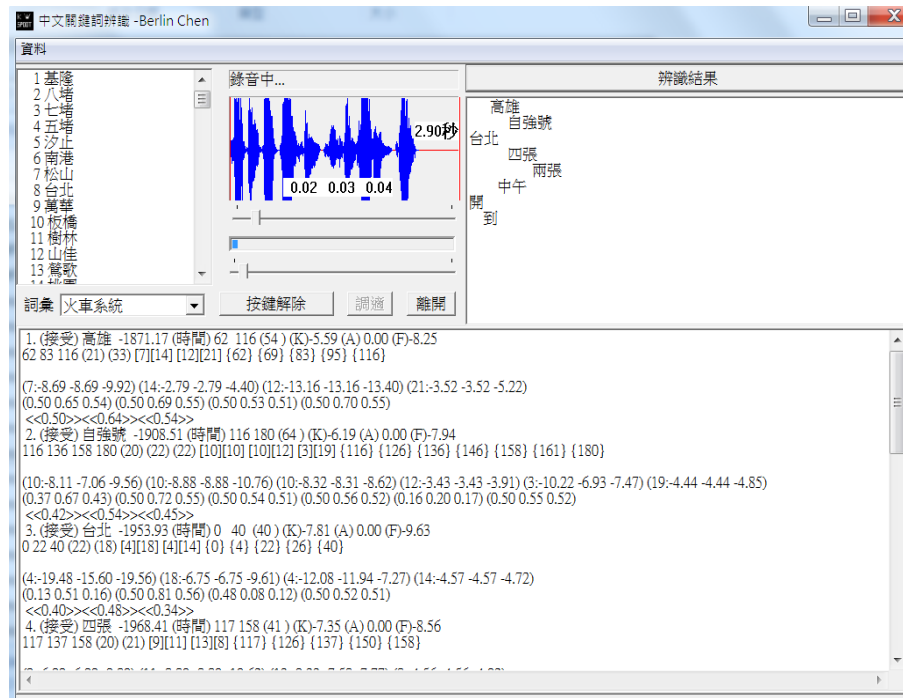
(Relative Character Error Rate Reduction)

- MMI: 6.5% , MPE: 10.1%

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

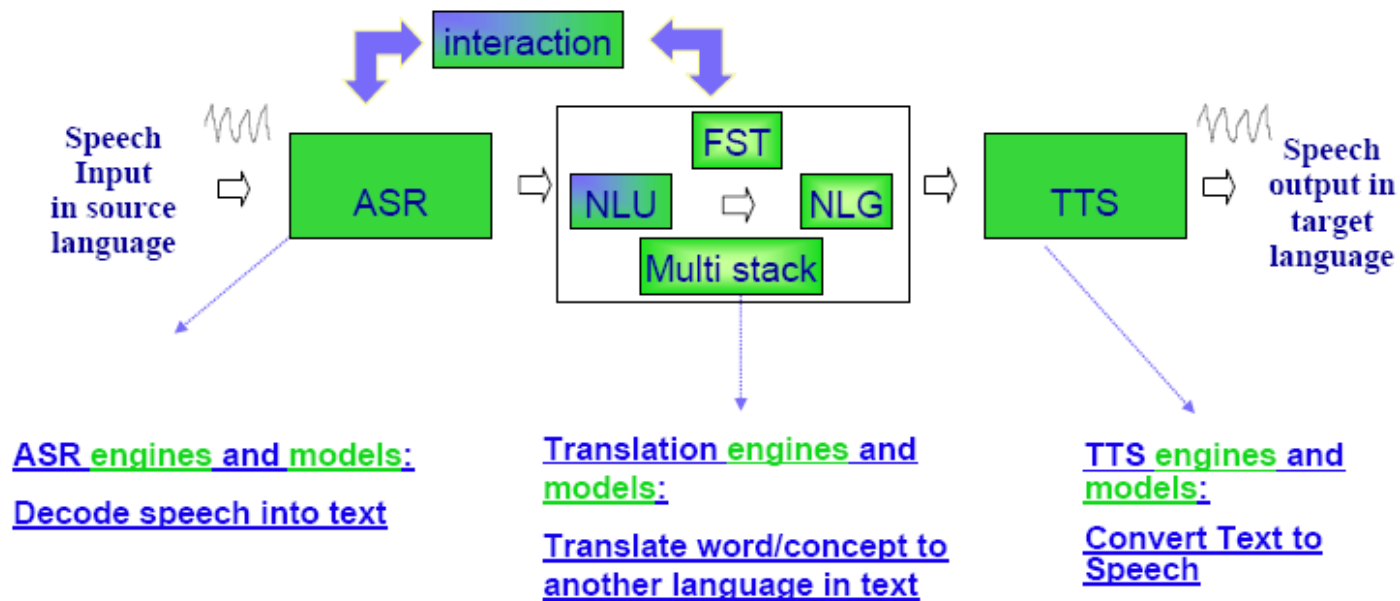
Keyword Spotting

- A relatively simple and efficient way to recognize salient semantic units from the speech utterances



Speech-to-Speech Translation (1/2)

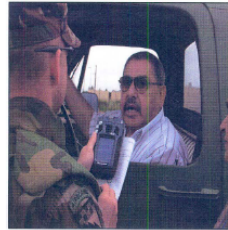
IBM Advanced Speech-to-Speech Translation Techniques



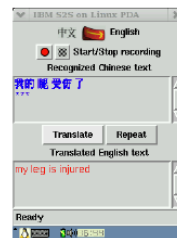
http://domino.research.ibm.com/comm/research_projects.nsf/pages/mastor.index.html

Speech-to-Speech Translation (2/2)

Handheld System



Laptop systems - hands-free, eyes-free function



Spoken Document Organization

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

(a) (b)

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

(c)

go to Level-1

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

(d)

go to Level-2

(e)

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)

Speech Retrieval



Google-411:
Finding and connecting to local business







Dial from any phone
1-800-GOOG-411
(1-800-466-4411)

About GOOG-411
Google's new 411 service is free, fast and easy to use. Give it a try now and see how simple it is to find and connect with local businesses for free.

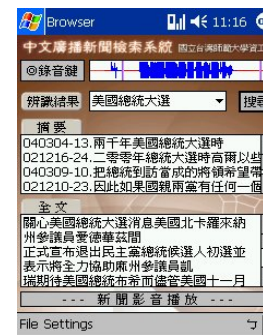
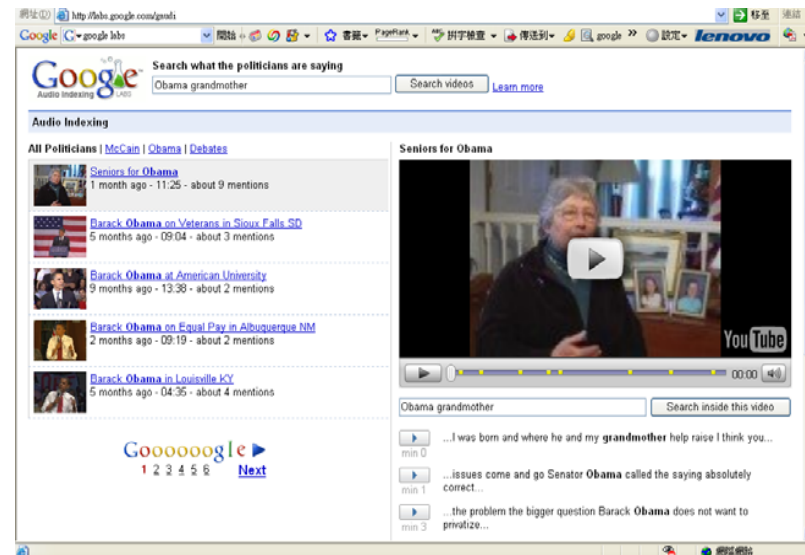
[Learn more - FAQ](#)

Liked the video? Want to comment or guess who the voice of GOOG-411 is? Post your opinion on our [YouTube page](#).

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

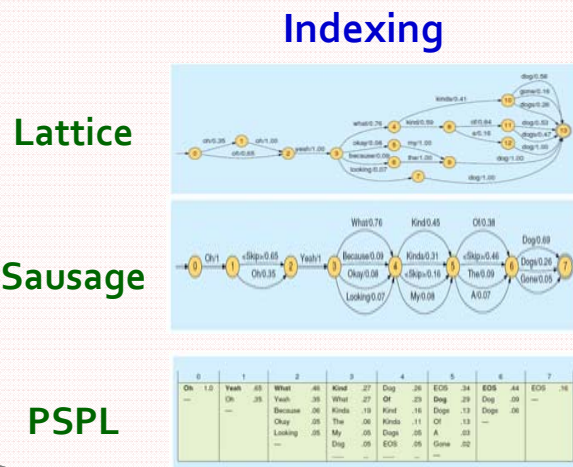
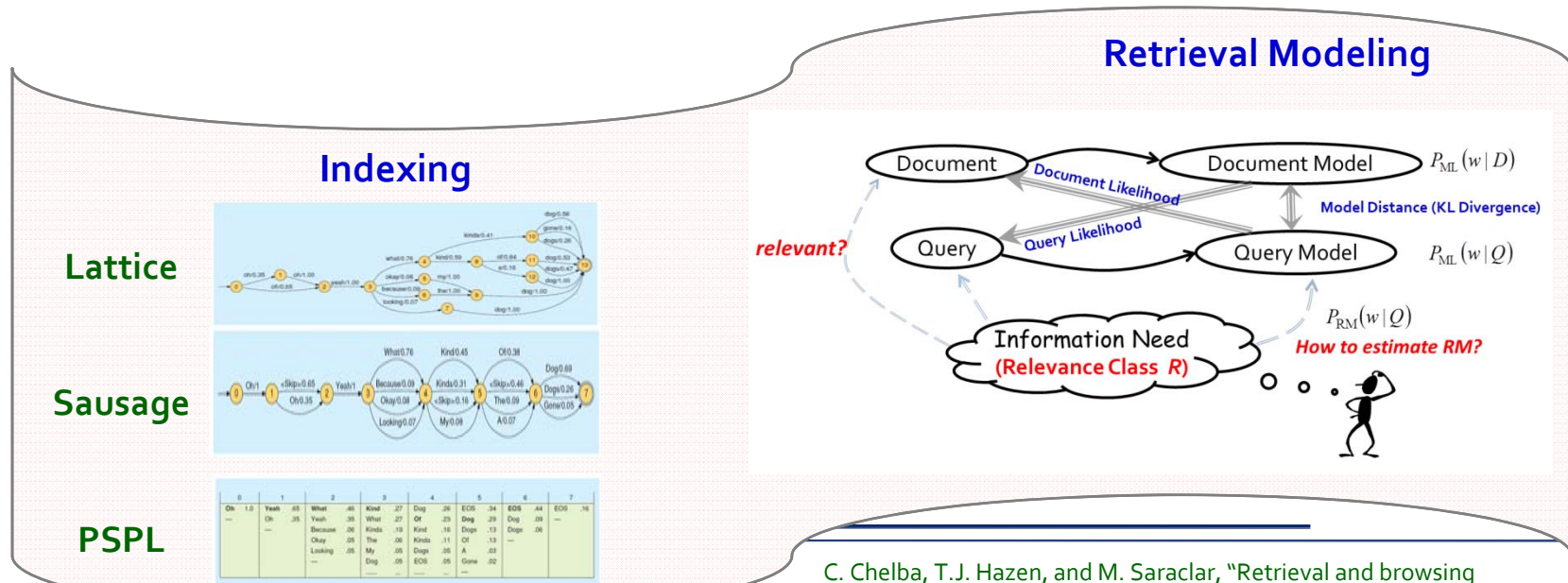
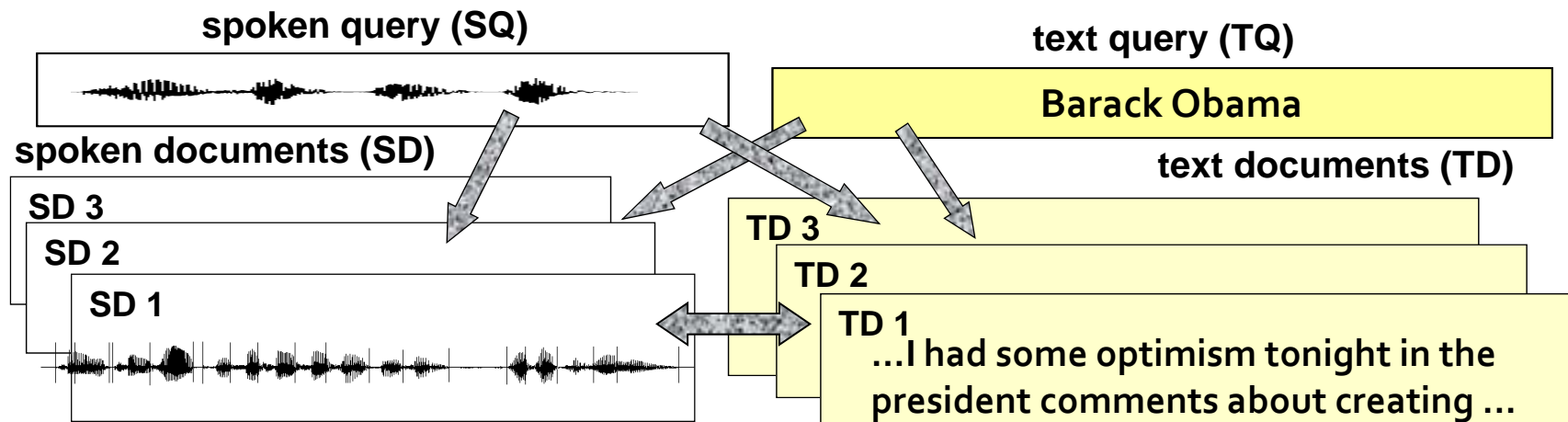
©2007 Google - [Terms of Service](#) - [Privacy Policy](#) - [Google Home](#) - [Mobile Home](#)

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



<http://www.apple.com/iphone/features/siri.html>

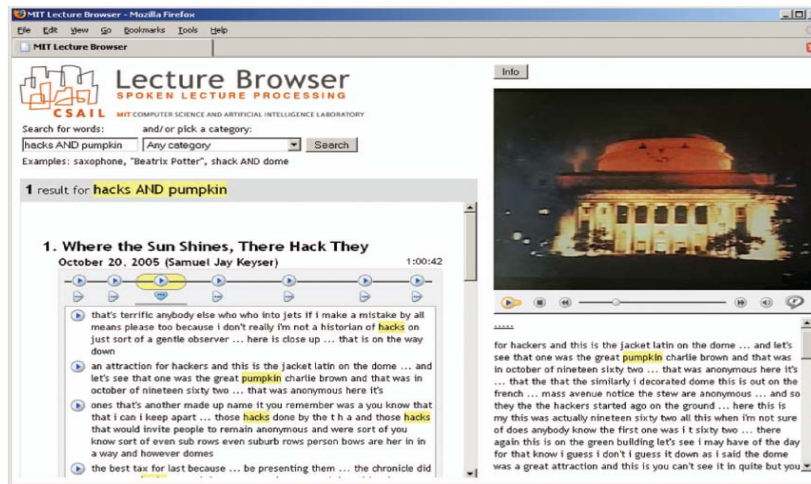
Speech Retrieval: Scenarios and Methodologies



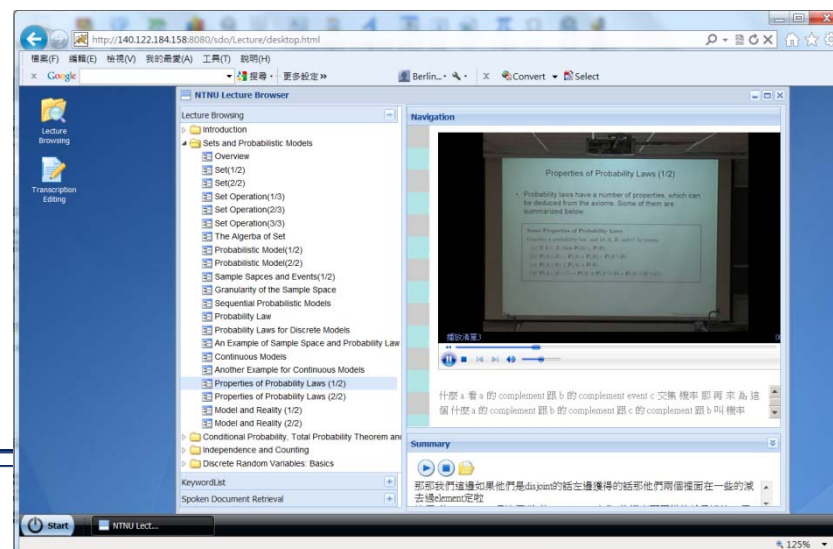
C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, 2008

Lecture Browsing

- MIT Lecture Browser <http://web.sls.csail.mit.edu/lectures>



- NTNU Lecture Browser <http://140.122.184.158:8080/sdo/Lecture/desktop.html>



Speech Summarization

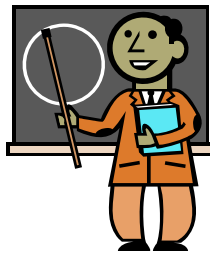
conversations



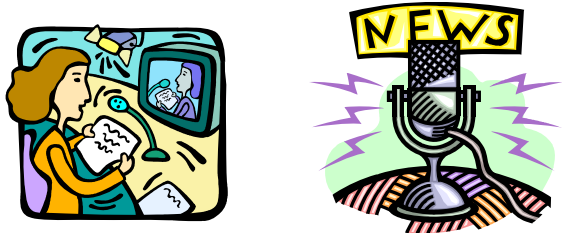
meetings



lectures



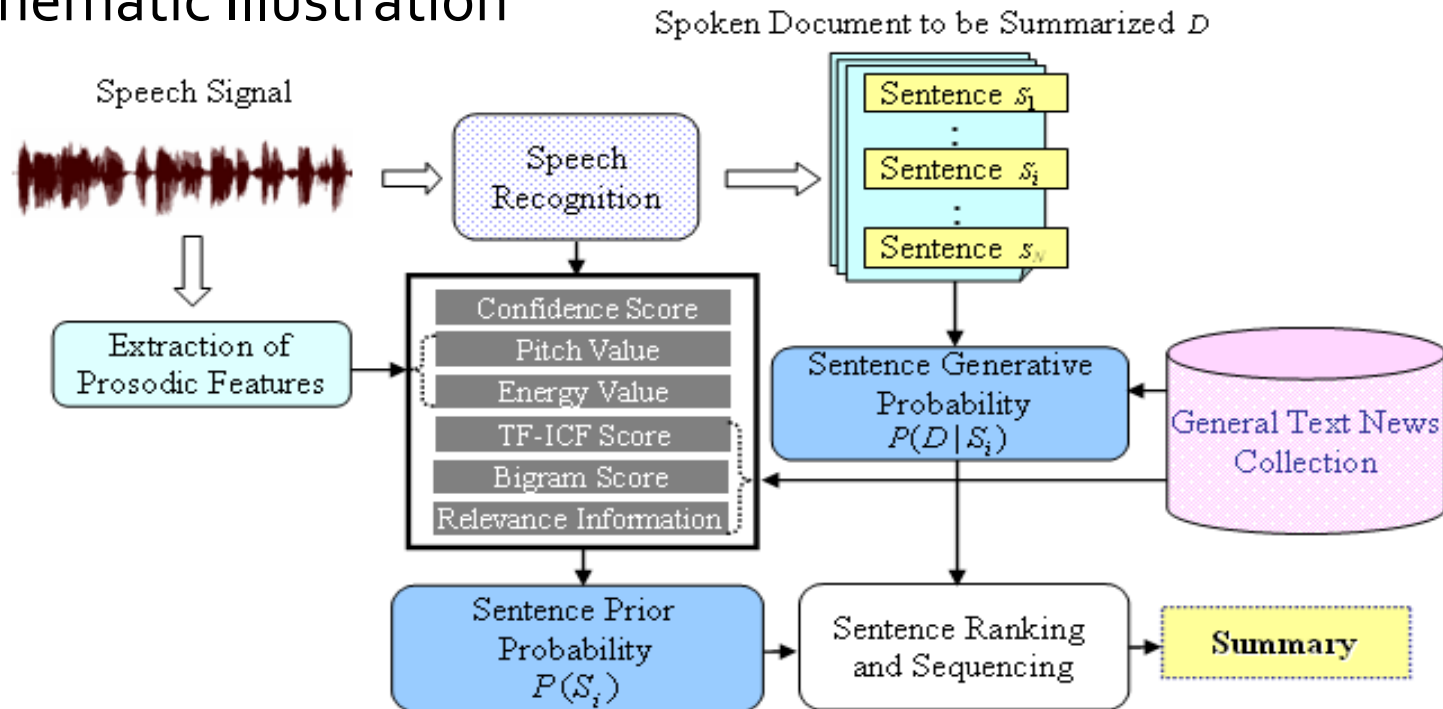
broadcast
and TV news



distilling
important information
abstractive vs. extractive
generic vs. query-oriented
single- vs. multi-documents

LM for Speech Summarization

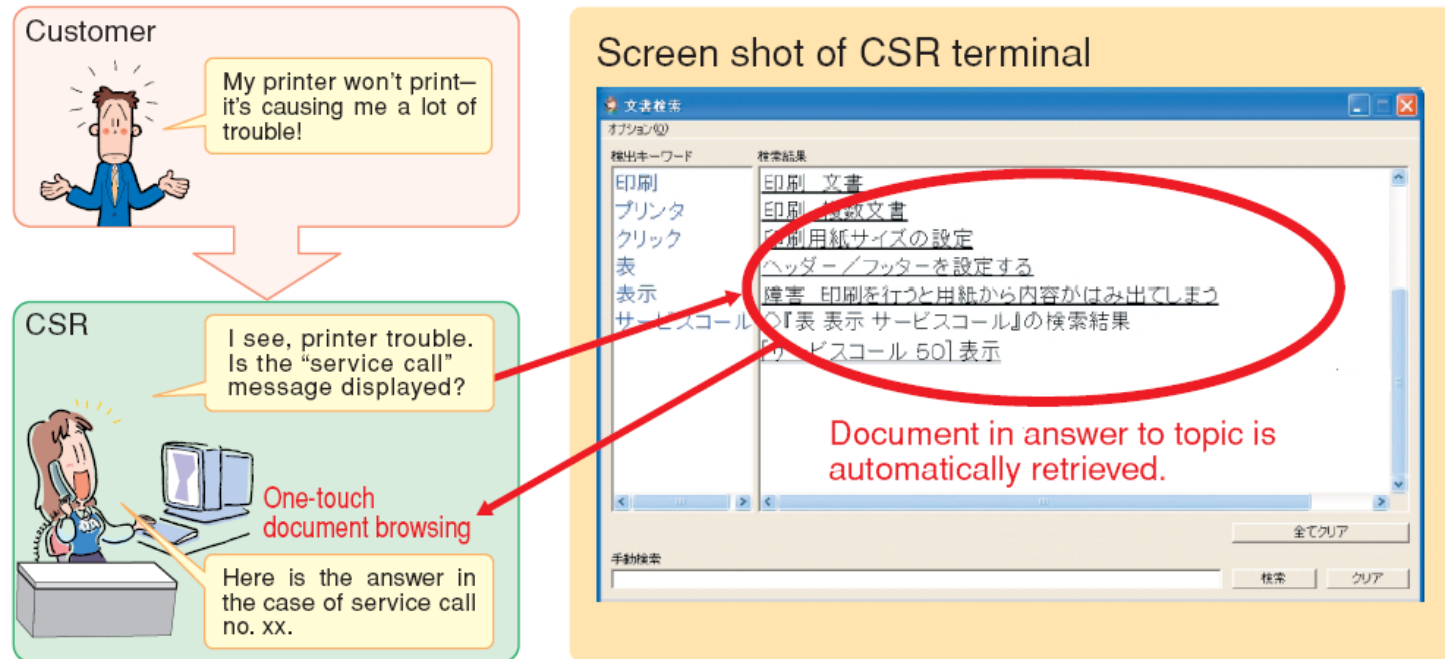
- Schematic Illustration



1. Y. T. Chen et al., "A probabilistic generative framework for extractive broadcast news speech summarization," *IEEE Transactions on Audio, Speech and Language Processing* 17(1), 2009
2. S.-H. Lin et al., "A comparative study of probabilistic ranking models for Chinese spoken document summarization," *ACM Transactions on Asian Language Information Processing*, 8(1), 2009
3. S.-H. Lin et al., "Leveraging Kullback-Leibler divergence measures and information-rich cues for speech summarization," *IEEE Transactions on Audio, Speech and Language Processing*, 19(4), 2011
4. 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
5. B. Chen et al., "Extractive speech summarization using evaluation metric-related training criteria," to appear in *Information Processing & Management*, 2012

Monitoring Contact (Call)-center conversations

Automatic document-retrieval by speech recognition



- CSR: Customer Service Representative
- Monitor agent conduct and customer satisfaction to increase service efficiency

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
 - Speech retrieval and summarization provide good assistance for companies, for instance, in
 - Contact (Call)-center conversations: 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 described program contents
 - Speech processing technologies are expected to play an essential role in computer-aided (language) learning
-