Speech Processing

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Course Contents

- Both the theoretical and practical issues for spoken language processing will be considered
- Technologies for Automatic Speech Recognition (ASR) and associated applications will be further emphasized
- Topics to be covered
  - Fundamentals and Statistical Modeling Paradigms
    - Spoken Language Structure
    - Hidden Markov Models
    - Speech Signal Analysis and Feature Extraction
    - Acoustic and Language Modeling
    - Search/Decoding Algorithms
  - Systems and Applications
    - Keyword Spotting, Dictation, Speaker Recognition, Spoken Dialogue, Speech-based Information Retrieval, etc.

Some Textbooks and References (1/3)

- References books
  - X. Huang, A. Acero, H. Hon. Spoken Language Processing, Prentice Hall, 2001
  - L. Rabiner, R. Schafer, Theory and Applications of Digital Speech Processing, Pearson, 2011
  - D. Yu and L. Deng, Automatic Speech Recognition: A Deep Learning Approach, Springer, 2015
  - Jacob Benesty (ed.), M. Mohan Sondhi (ed.), Yiteng Huang (ed.), Springer Handbook of Speech Processing, Springer, 2007
  - M.J.F. Gales and S.J. Young. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing, 2008
  - C. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 1999
  - J. R. Deller, J. H. L. Hansen, J. G. Proakis. Discrete-Time Processing of Speech Signals. IEEE Press, 2000
  - F. Jelinek. Statistical Methods for Speech Recognition. MIT Press, 1999
  - L. Rabiner, B.H. Juang. Fundamentals of Speech Recognition. Prentice Hall, 1993
  - 王小川教授， 語音訊號處理， 全華圖書 2004

Some Textbooks and References (2/3)

- Reference papers

1. L. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proceedings of the IEEE, vol. 77, No. 2, February 1989
2. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," J. Royal Star. Soc., Series B, vol. 39, pp. 1-38, 1977
3. Jeff A. Bilmes "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," U.C. Berkeley TR-97-021
4. J. W. Picone, “Signal modeling techniques in speech recognition,” proceedings of the IEEE, September 1993, pp. 1215-1247
5. R. Rosenfeld, ”Two Decades of Statistical Language Modeling: Where Do We Go from Here?,” Proceedings of IEEE, August, 2000
6. H. Ney, “Progress in Dynamic Programming Search for LVCSR,” Proceedings of the IEEE, August 2000
7. H. Hermansky, "Should Recognizers Have Ears?", Speech Communication, 25(1-3), 1998

Some Textbooks and References(3/3)

8. Frederick Jelinek, "The Dawn of Statistical ASR and MT," Computational Linguistics, Vol. 35, No. 4. (1 December 2009), pp. 483-494

9. L.S. Lee and B. Chen, “Spoken document understanding and organization,” *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 42-60, Sept. 2005

10. M. Gilbert and J. Feng, “Speech and Language Processing over the Web,” *IEEE Signal Processing Magazine* 25 (3), May 2008

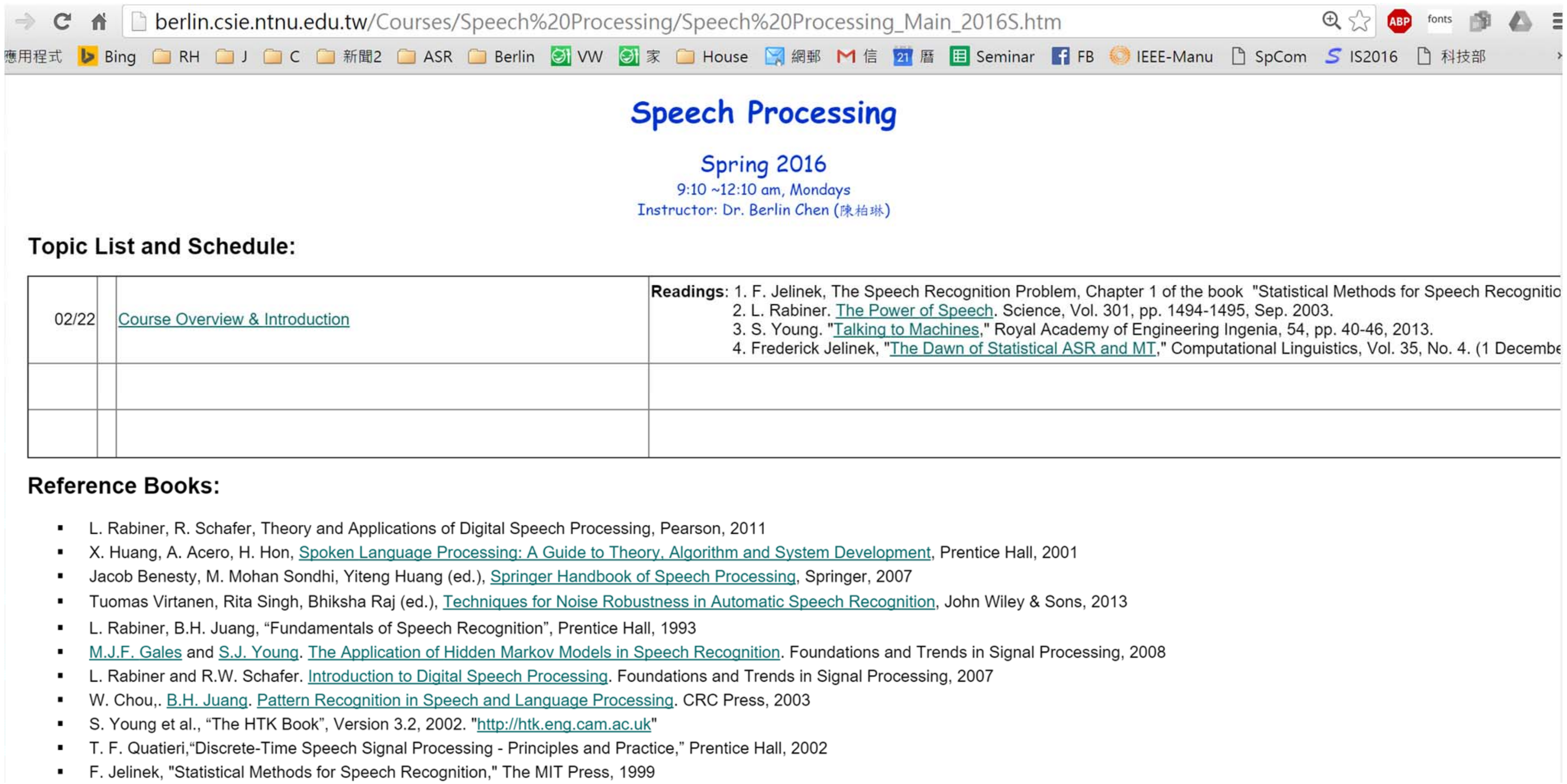
11. C. Chelba, T.J. Hazen, and M. Saraclar. Retrieval and Browsing of Spoken Content. *IEEE Signal Processing Magazine* 25 (3), May 2008

12. S. Young et al., The HTK Book. Version 3.4: http://htk.eng.cam.ac.uk

13. J. Schalkwyk et al., "Google Search by Voice: A case study," 2010

Website for This Course

- Visit http://berlin.csie.ntnu.edu.tw/ and then click the link "Spring 2016: Speech Processing"



berlin.csie.ntnu.edu.tw/Courses/Speech%20Processing/Speech%20Processing_Main_2016S.htm

應用程式 Bing RH J C 新聞2 ASR Berlin VW 家 House 網郵 信 曆 Seminar FB IEEE-Manu SpCom IS2016 科技部

Speech Processing

Spring 2016

9:10 ~12:10 am, Mondays

Instructor: Dr. Berlin Chen (陳柏琳)

Topic List and Schedule:

02/22		Course Overview & Introduction	**Readings**: 1. F. Jelinek, The Speech Recognition Problem, Chapter 1 of the book "Statistical Methods for Speech Recognitio 2. L. Rabiner. The Power of Speech. Science, Vol. 301, pp. 1494-1495, Sep. 2003. 3. S. Young. "Talking to Machines," Royal Academy of Engineering Ingenia, 54, pp. 40-46, 2013. 4. Frederick Jelinek, "The Dawn of Statistical ASR and MT," Computational Linguistics, Vol. 35, No. 4. (1 Decembe

Reference Books:

- L. Rabiner, R. Schafer, Theory and Applications of Digital Speech Processing, Pearson, 2011
- X. Huang, A. Acero, H. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001
- Jacob Benesty, M. Mohan Sondhi, Yiteng Huang (ed.), Springer Handbook of Speech Processing, Springer, 2007
- Tuomas Virtanen, Rita Singh, Bhiksha Raj (ed.), Techniques for Noise Robustness in Automatic Speech Recognition, John Wiley & Sons, 2013
- L. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", Prentice Hall, 1993
- M.J.F. Gales and S.J. Young. The Application of Hidden Markov Models in Speech Recognition. Foundations and Trends in Signal Processing, 2008
- L. Rabiner and R.W. Schafer. Introduction to Digital Speech Processing. Foundations and Trends in Signal Processing, 2007
- W. Chou,. B.H. Juang. Pattern Recognition in Speech and Language Processing. CRC Press, 2003
- S. Young et al., "The HTK Book", Version 3.2, 2002. "http://htk.eng.cam.ac.uk"
- T. F. Quatieri,"Discrete-Time Speech Signal Processing - Principles and Practice," Prentice Hall, 2002
- F. Jelinek, "Statistical Methods for Speech Recognition," The MIT Press, 1999

Introduction

References:

1. B. H. Juang and S. Furui, "Automatic Recognition and Understanding of Spoken Language - A First Step Toward Natural Human-Machine Communication,“ *Proceedings of IEEE*, August, 2000
2. I. Marsic, A. Medl, and J. Flanagan, “Natural Communication with Informatio Systems,” *Proceedings of IEEE*, August, 2000

Historical Review

1952, Isolated-Digit Recognition, Bell Lab.

1956, Ten-Syllable Recognition, RCA

1959, Ten-Vowel Recognition, MIT Lincoln Lab

1959, Phoneme-sequence Recognition using Statistical Information of Context, Fry and Denes

1960s, Dynamic Time Warping to Compare Speech Events, Vintsyuk

1960s-1970s, Hidden Markov Models for Speech Recognition, Baum, Baker and Jelinek

Gestation of Foundations

1970s ~

Voice-Activated Typewriter (dictation machine, speaker-dependent), IBM

Telecommunication (keyword spotting, speaker-independent), Bell Lab

Philips

BBN Technologies

Cambridge (HTK)

CMU

LIMSI

MIT (SLS)

JHU CLSP

Microsoft

SRI

Google

Apple

nuance

Amazon

Facebook

Speech Processing vs. Text 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 a specific 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

Areas for Speech Processing

- Production, Perception, and Modeling of Speech (phonetics and phonology)
- Signal Processing for Speech
- Speech Coding
- Speech Synthesis (Text-to-Speech)
- Speech Recognition (Speech-to-Text) and Understanding
- Speaker Recognition
- Language Recognition
- Speech Enhancement
- ….

C.f. Jacob Benesty (ed.), M. Mohan Sondhi (ed.), Yiteng Huang (ed.), Springer Handbook of Speech Processing, Springer, 2007

Progress of Technology (1/6)

- US. National Institute of Standards and Technology (NIST)

NIST Time | NIST Home | About NIST | Contact Us | A-Z Site Index

Information Technology Laboratory

About ITL ▼ Publications Topic/Subject Areas ▼ Products/Services ▼ News/Multimedia Programs/Projects

NIST Home > ITL > Information Access Division > Multimodal Information Group > Benchmark Tests

Ongoing Benchmark Tests

GALE Translation (2006 - present)
Language Recognition (1996 - present)
Machine Translation (2001 - present)
Metrics for Machine Translation (2008 - present)
Rich Transcription (2003 - present)
Speaker Recognition (1996 - present)
TRECVid Event Detection (2008-present)
MADCAT (2008-present)
Multiple Camera Single Person Tracking (2009-present)

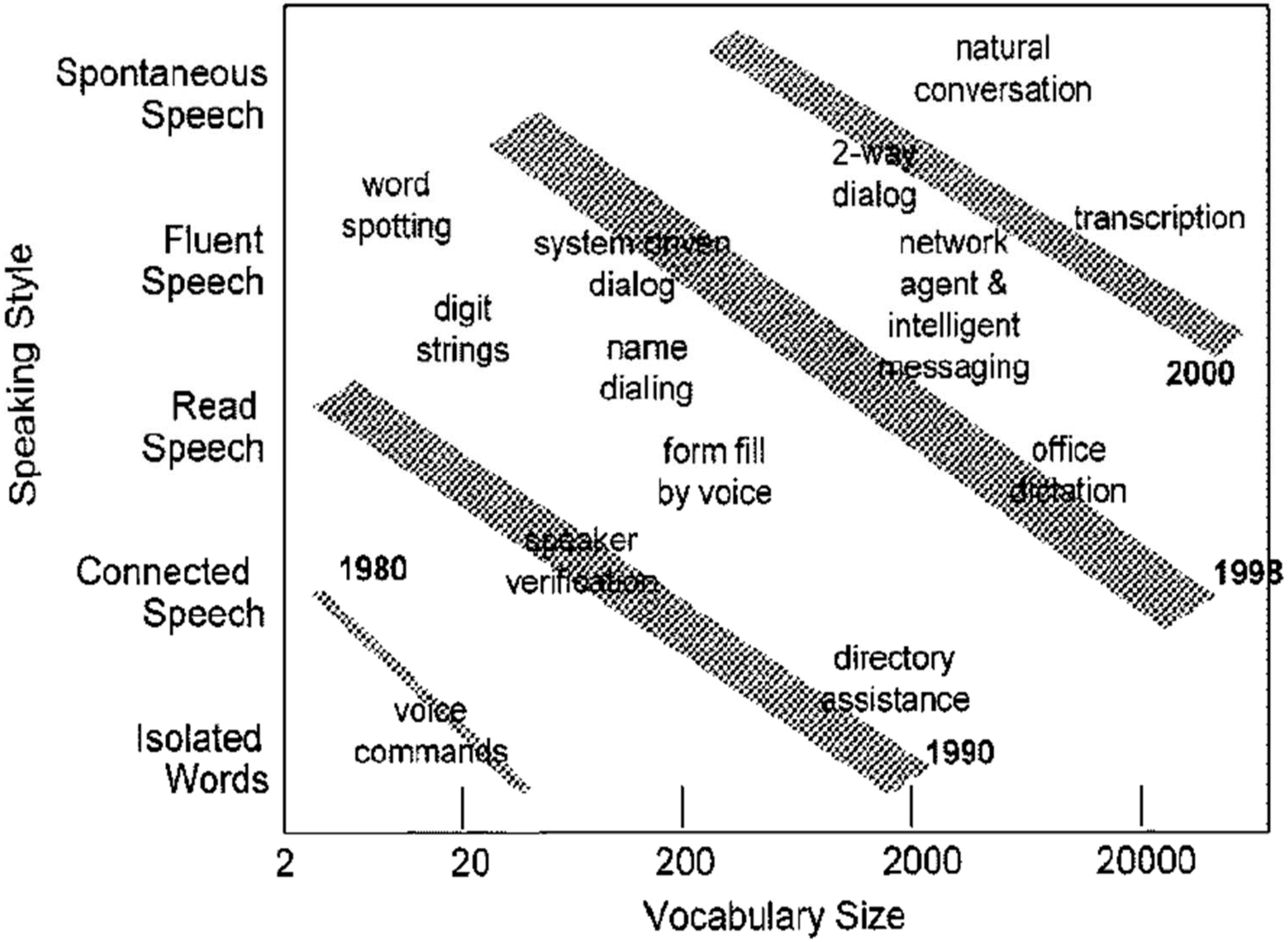
Past Benchmark Tests

CLEAR (2006 - 2007)
Spoken Term Detection (2006)
Broadcast News Recognition (1996 - 1999)
Conversational Telephone Recognition (1997 - 2001)
Spoken Document Retrieval (1997 - 2000)
Topic Detection and Tracking (1998 - 2004)
Automatic Content Extraction (1999 - 2008)

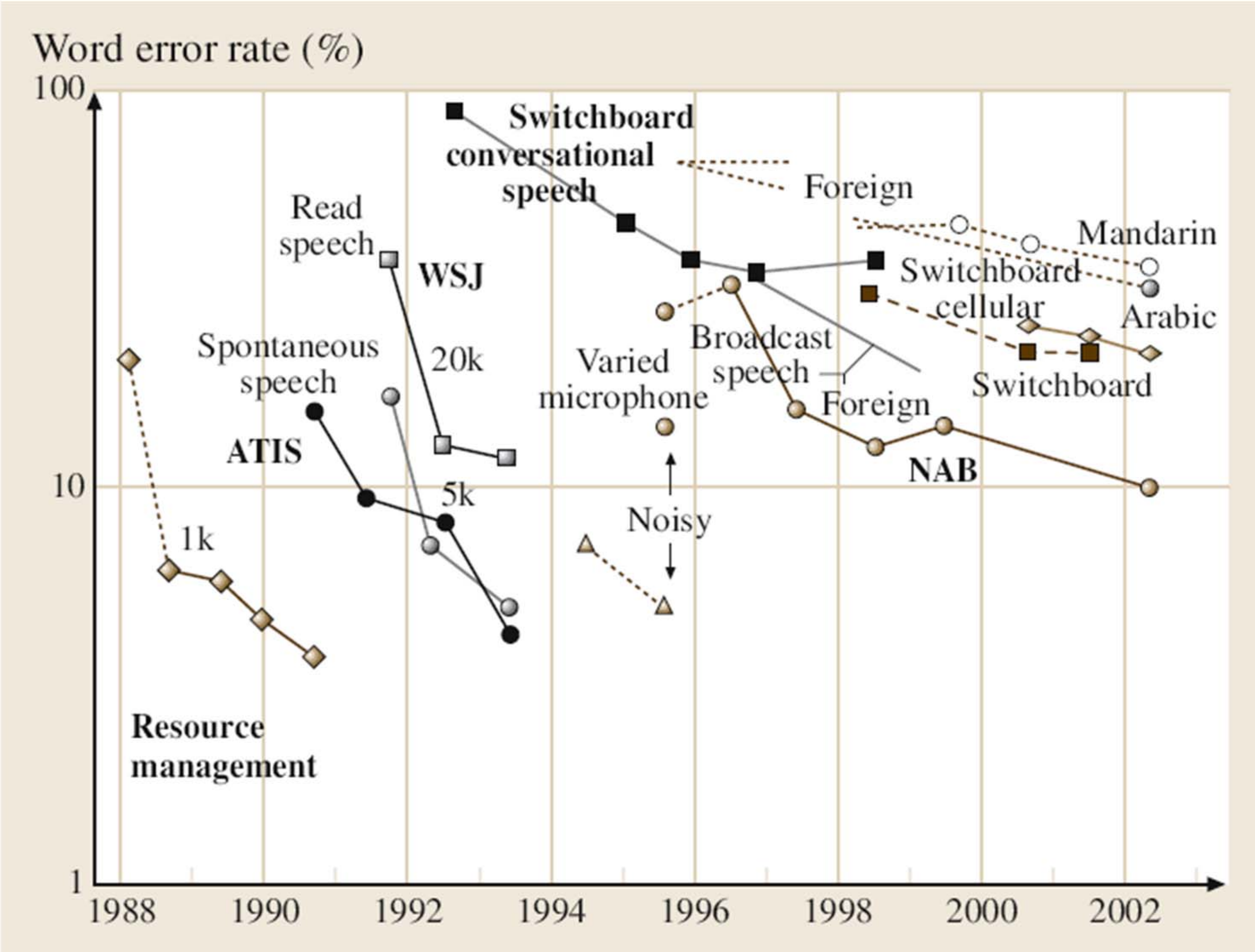
http://www.nist.gov/itl/iad/mig/bmt.cfm

Progress of Technology (2/6)

- Generic Application Areas (vocabulary vs. speaking style)



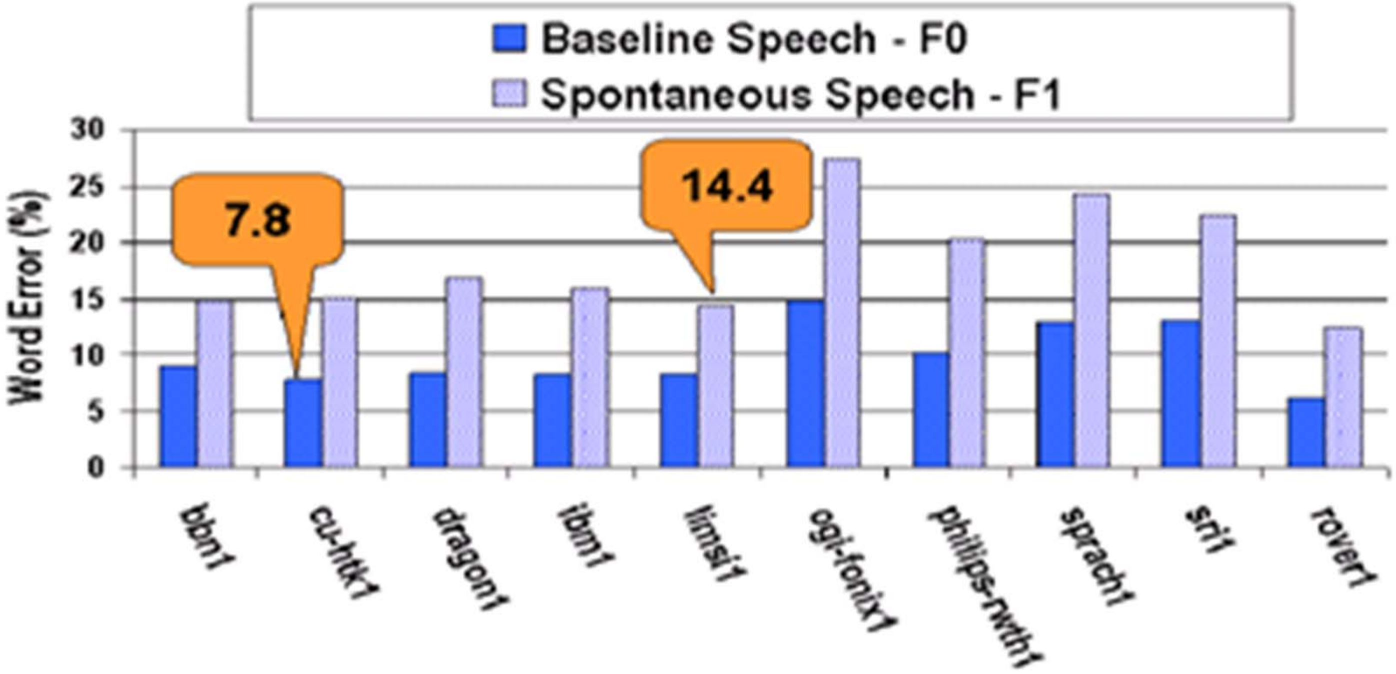
Progress of Technology (3/6)



L. Rabiner, B.-H. Juang, "Historical Perspective of the Field of ASR/NLU" Chapter 26 in the book " Springer Handbook of Speech Processing"

Progress of Technology (4/6)

- Benchmarks of ASR performance: Broadcast News Speech



FO: anchor speakers
F1: field reports and interviewees

Progress of Technology (5/6)

- Benchmarks of ASR performance: Conversational Speech

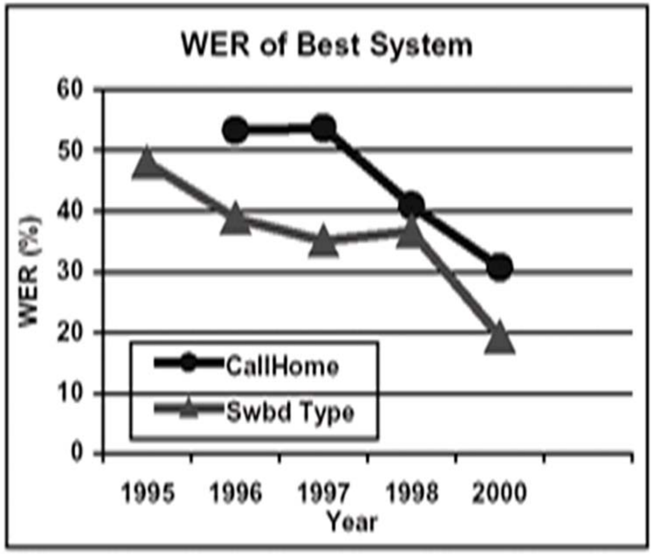


Figure 4 History of lowest word error rates (WER) obtained in NIST conversational speech evaluations on Switchboad and CallHome type conversations in English [26].

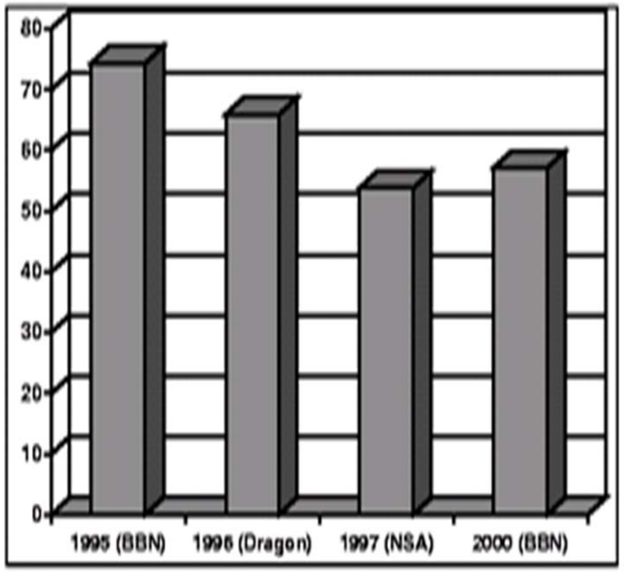


Figure 5 Chinese Character error rates of the best performing evaluation system in NIST Mandarin conversational speech evaluations 1995-2000 [26].

Progress of Technology (6/6)

- Mandarin Conversational Speech (2003 Evaluation)
  - Acoustic/Training Test Data:
    - training data: 34.9 hours, 379 sides, from LDC CallHome (22.4hrs) and CallFriend (12.5hrs), 451K Words (+7K English word), 628K Characters
    - development data: dev02 1.94 hours from CallFriend

		CER (%)	
		dev02	eval03
P1	trans for VTLN	55.1	54.7
P2	trans for MLLR	50.8	51.3
P3	lat gen (bg)	49.3	50.5
	tgintcat rescore	48.9	49.8
P4	lat MLLR	48.6	49.5
CN	P4	47.9	48.6

%CER on dev02 and eval03 for all stages of 2003 system

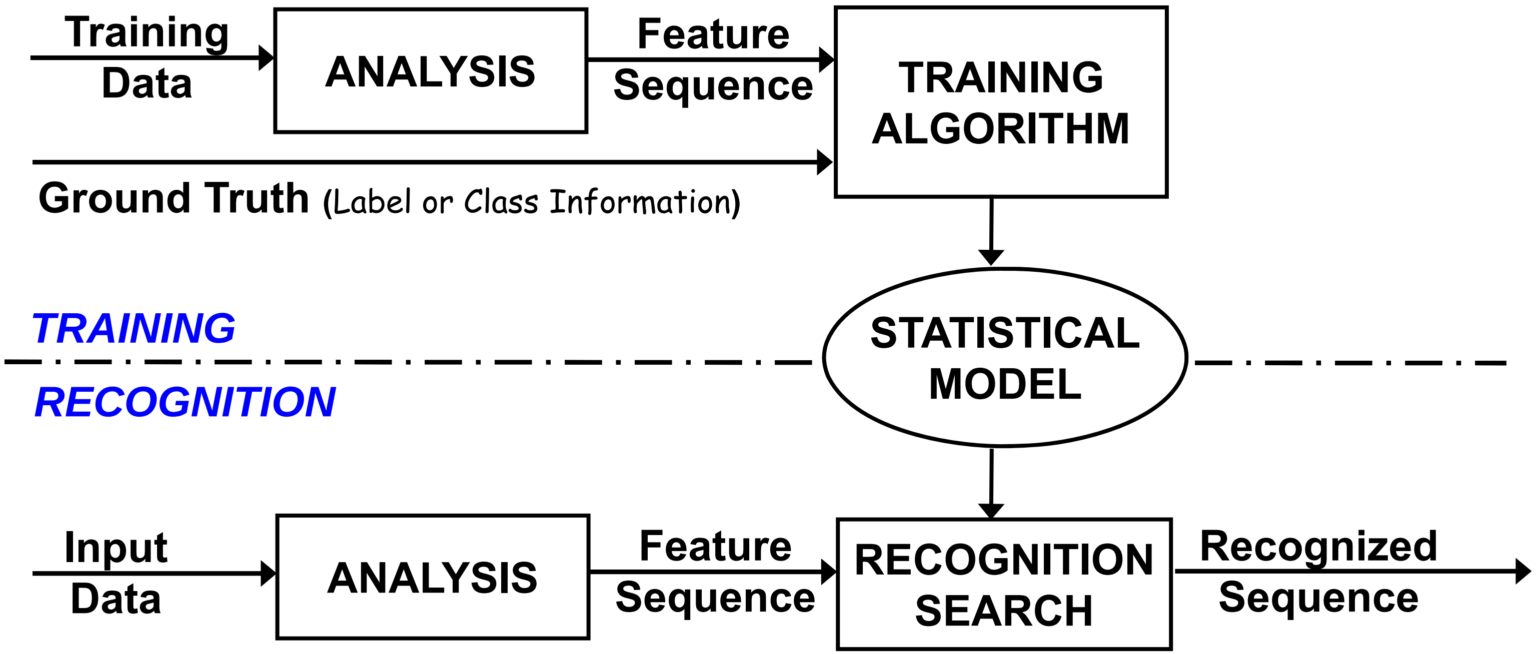
  - Adopted from

Cambridge University
Engineering Department

Rich Transcription Workshop 2003

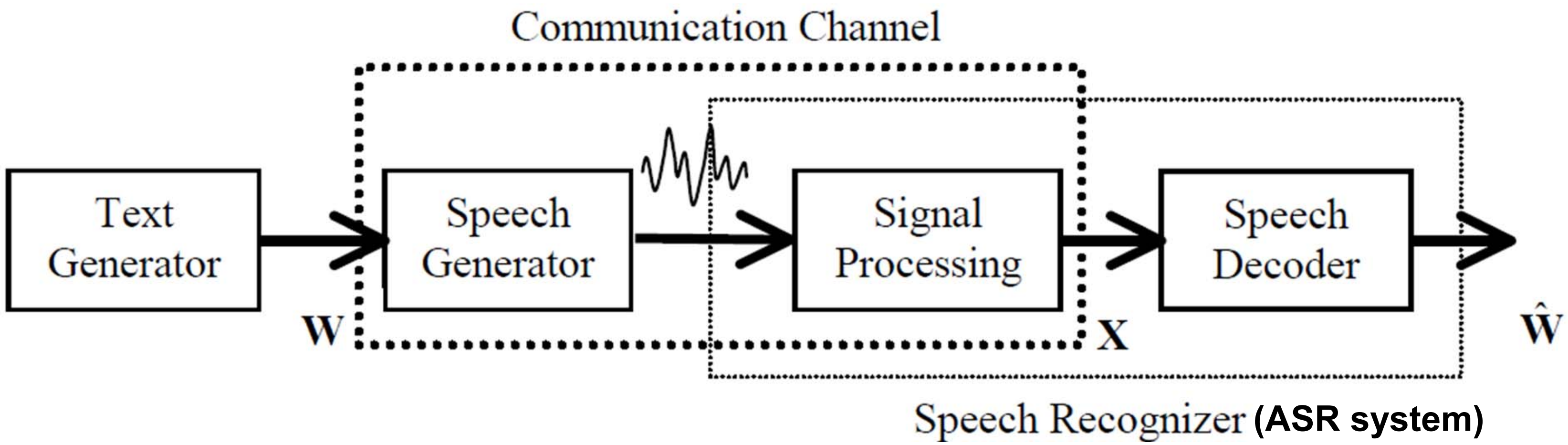
Statistical Modeling Paradigm

- Most approaches to speech and language processing generally follow the statistical modeling paradigm



  - Data-driven approaches: automatically extract "knowledge" from the data
  - It would be better to pair data-driven approaches with rule-based ones

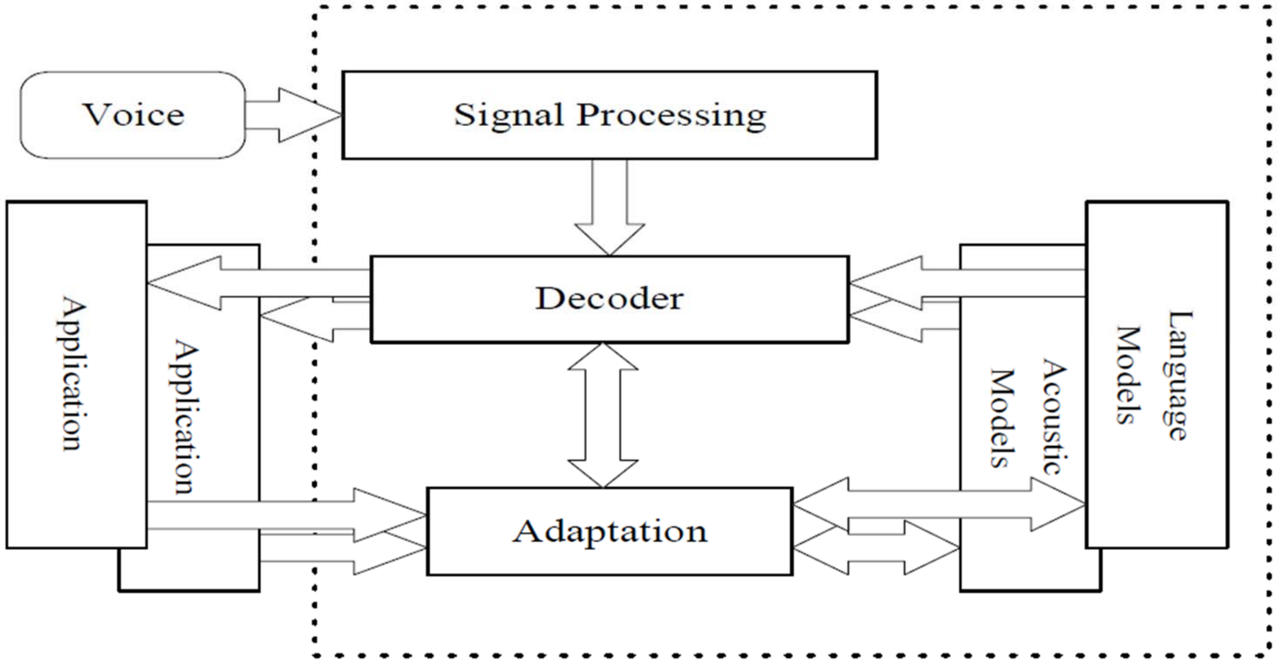
A Source-Channel Model for ASR



– Communication channel consists of speaker's vocal apparatus to produce speech (the waveform) and the signal processing component of the speech recognizer
– The speech decoder aims to decode the acoustic signal $\mathbf{X}$ into a word sequence $\hat{\mathbf{W}}$ (Hopefully, $\hat{\mathbf{W}} \approx \mathbf{W}$.)

Uncertainties to be contended with: unknown words, grammatical variation, noise interference, acoustic variation, to name a few

Basic Architecture of ASR System

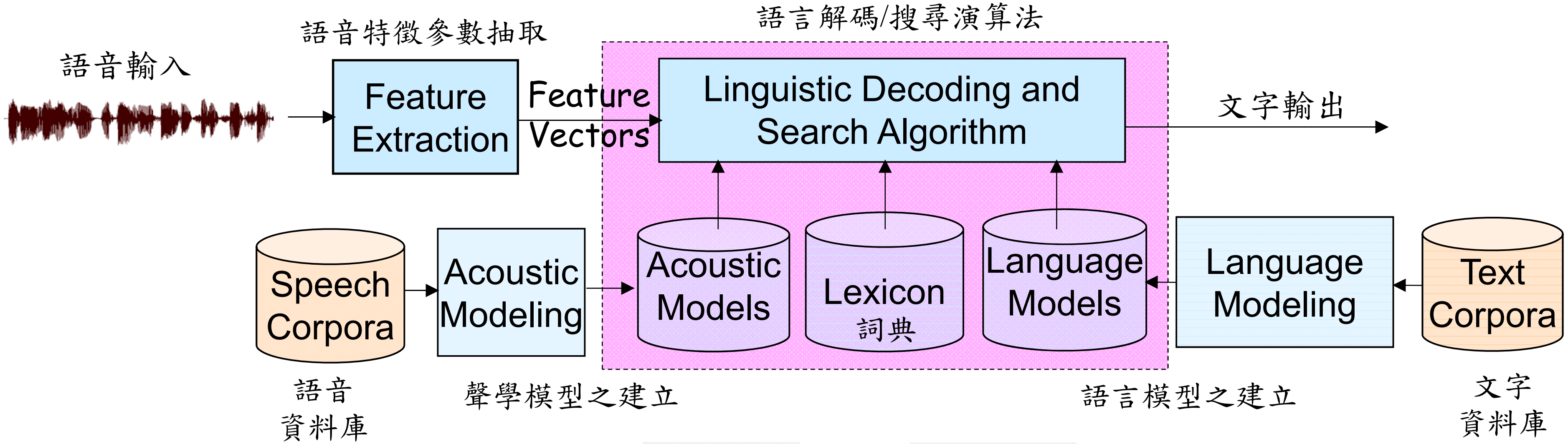


- **Signal processing:** extract salient features for the decoder
- **Decoder:** use both acoustic and language models to generate the “best” word sequence in response to the input voice
- **Adaptation:** modify either acoustic or language models so that improved performance can be obtained

ASR: Applications

- E.g., Transcription of Broadcast News Speech

ASR: A Bit of Terminology



可能詞句　　語音輸入

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{X})$$

Bayes Decision Theory

Bayes Rule

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

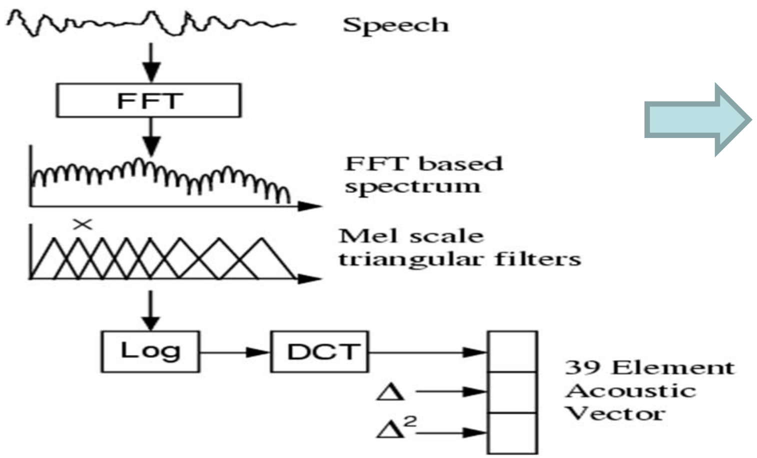
$$= \arg\max_{\mathbf{W}} p(\mathbf{X}|\mathbf{W})P(\mathbf{W})$$

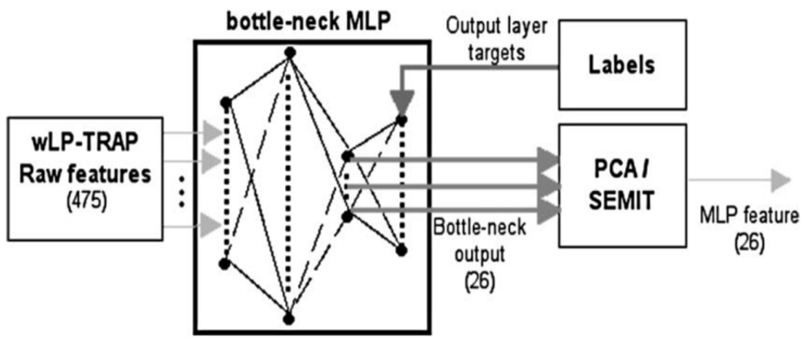
Decoding

Acoustic Modeling　　Language Modeling

Speech Feature Extraction

- The raw speech waveform is passed through feature extraction to generate relatively compact feature vectors at a frame rate of around 100 Hz
  - Parameterization: an acoustic speech feature is a simple compact representation of speech and can be modeled by cepstral features such as the Mel-frequency cepstral coefficient (MFCC)

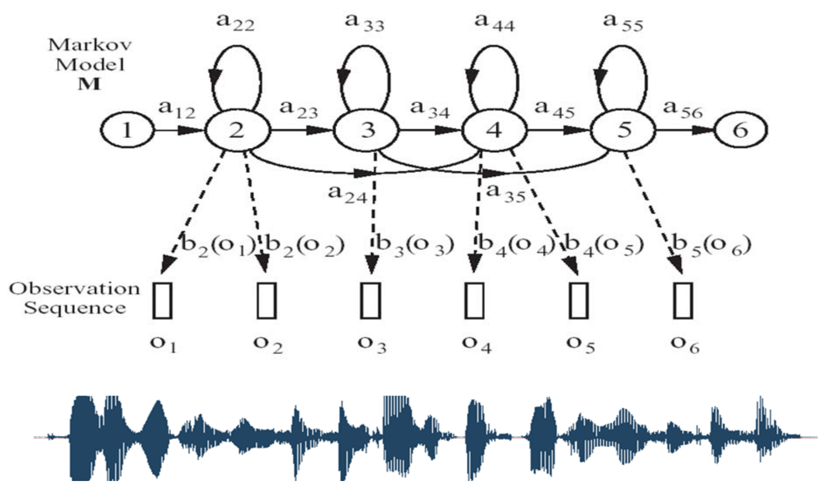




raw (perception-driven) features vs. discriminant (posterior) features

ASR: Acoustic Modeling

- Construct a set of statistical models representing various sounds (or phonetic units) of the language
  - Approaches based on Hidden Markov Models (HMMs) dominate the area of speech recognition
  - HMMs are based on rigorous mathematical theory built on several decades of mathematical results developed in other fields
  - HMMs are constructed by the process of training on a large corpus of real speech data



ASR: Language Modeling

- Constrain the acoustic analysis, guide the search through multiple candidate word strings, and quantify the acceptability of the final word string output from a speech recognizer

$$W = w_1 w_2 \ldots w_L \Longrightarrow P(W) = ?$$

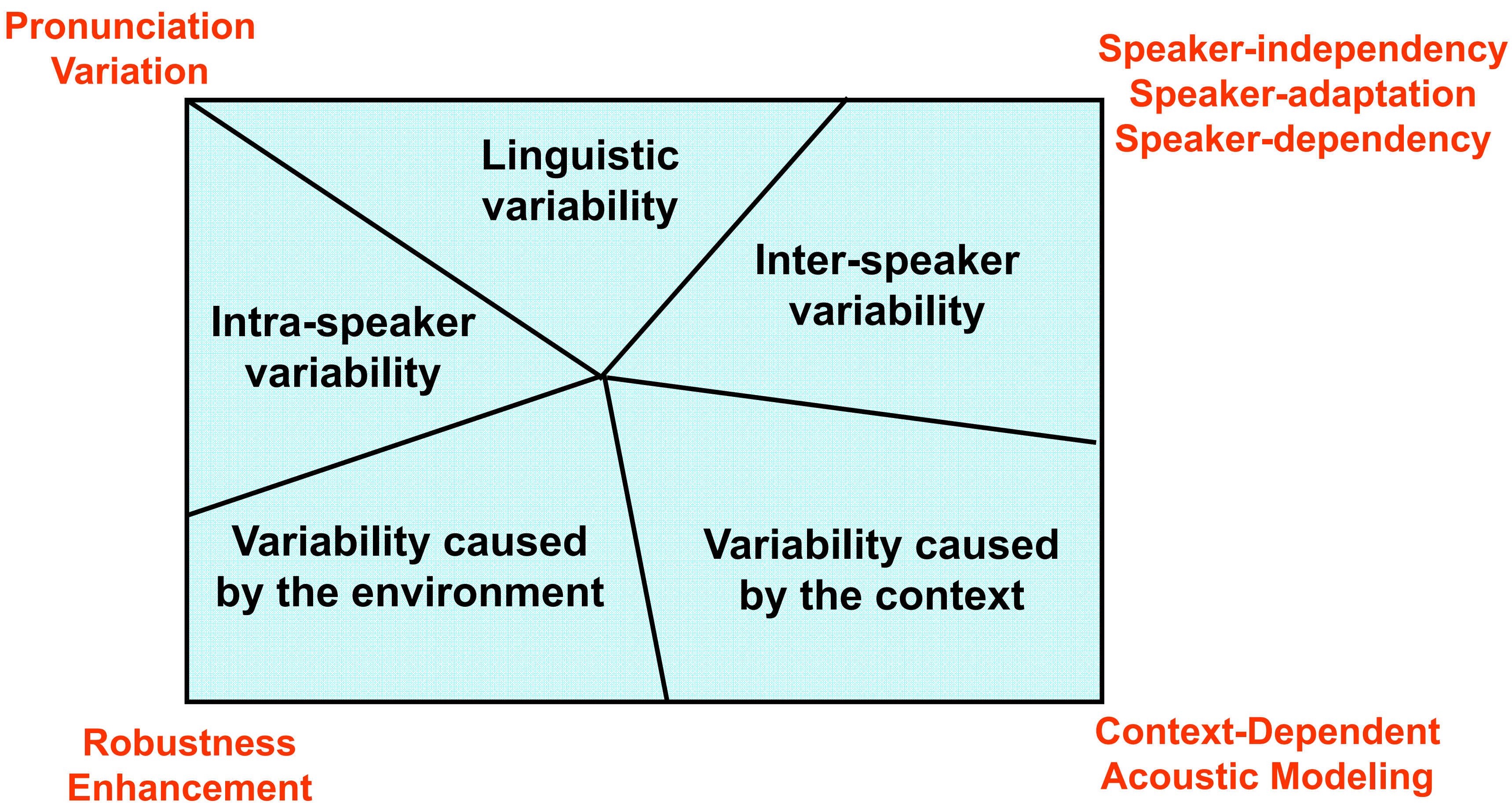
- The *n*-gram language model that follows a statistical modeling paradigm is the most prominently-used in ASR

$$P(w_1 w_2 \ldots. w_L) = P(w_1) P(w_2 | w_1) P(w_3 | w_1 w_2) \cdots P(w_L | w_1 w_2 \ldots w_{L-1})$$

bigram modeling

$$P(w_1 w_2 \ldots. w_L) = P(w_1) P(w_2 | w_1) P(w_3 | w_2) \cdots P(w_L | w_{L-1})$$

Difficulties: Speech Variability



What is Deep Learning?

WIKIPEDIA
The Free Encyclopedia

Deep learning

From Wikipedia, the free encyclopedia

Deep learning (*deep machine learning*, or *deep structured learning*, or *hierarchical learning*, or sometimes *DL*) is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures or otherwise, composed of multiple non-linear transformations.[1](p198)[2][3][4][5]

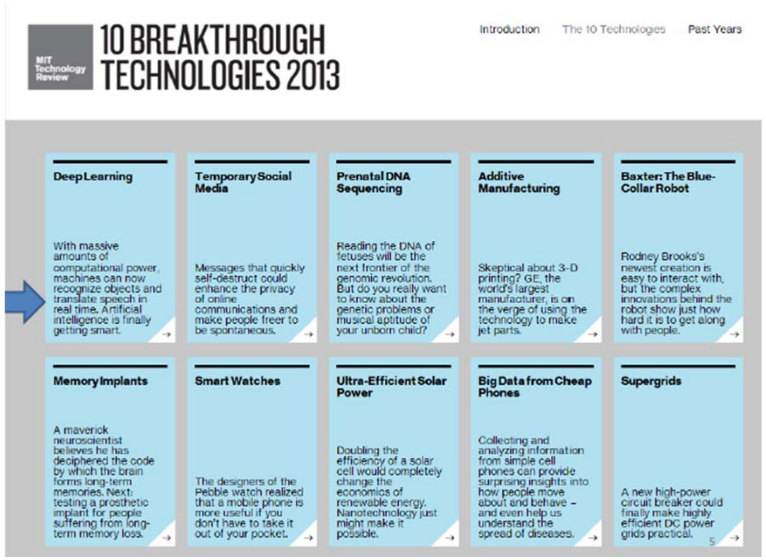
Representation Learning

- **Shallow Learning** GMM, SVM, CRF, NMF/PLSA/LDA, Perception, Boosting, etc.
- **Deep Learning** DNN, CNN, RNN, LSTM, etc.
  (help to discover intricate structure in large data sets)

Deeper is better? vs. Simple is elegant?

Deep Learning and its Applications to ASR (1/5)

- **Deep Learning** is concerned with learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text
- By virtue of **Deep Learning**
  - Our computers can learn and grow on their own
  - Our computers are able to understand complex, massive amount of data (deep learning is the holy grail of big data?)



http://www.technologyreview.com/lists/breakthrough-technologies/2013/

Deep Learning and its Applications to ASR (2/5)

MIT Technology Review

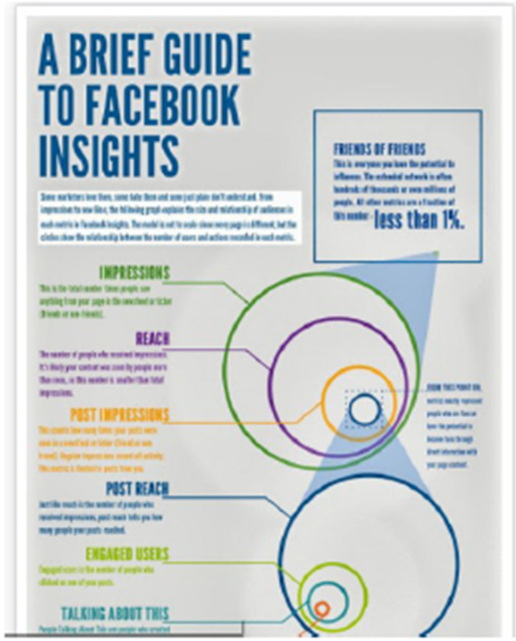
Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

September 20, 2013

A technique called deep learning could help Facebook understand its users and their data better.

By Tom Simonite on September 20, 2013

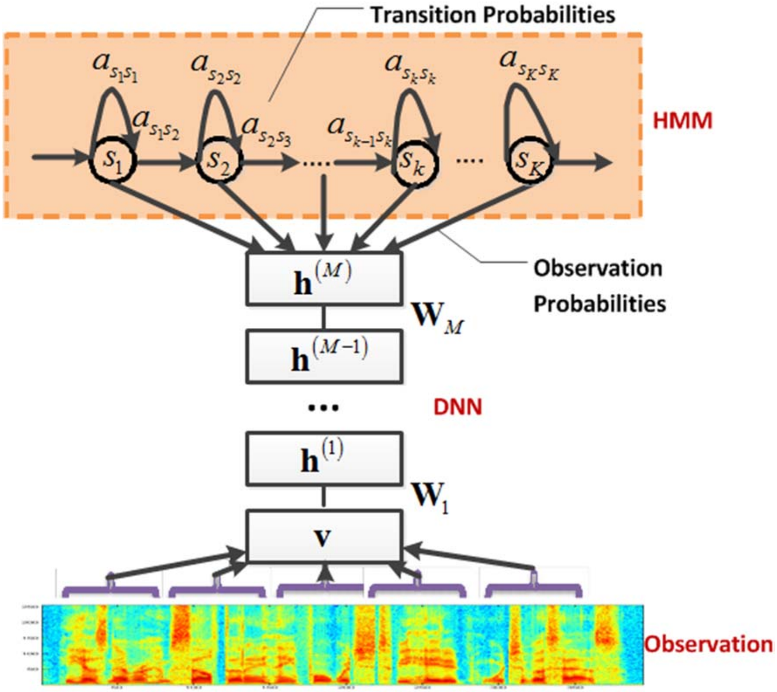
……Facebook's foray into deep learning sees it following its competitors Google and Microsoft, which have used the approach to impressive effect in the past year. Google has hired and acquired leading talent in the field (see "10 Breakthrough Technologies 2013: Deep Learning"), and last year created software that taught itself to recognize cats and other objects by reviewing stills from YouTube videos. The underlying deep learning technology was later used to slash the error rate of Google's voice recognition services (see "Google's Virtual Brain Goes to Work")….Researchers at Microsoft have used deep learning to build a system that translates speech from English to Mandarin Chinese in real time (see "Microsoft Brings Star Trek's Voice Translator to Life"). Chinese Web giant Baidu also recently established a Silicon Valley research lab to work on deep learning.



X. He, et al., "Deep learning for natural language processing and related applications," Tutorial given at ICASSP 2014.

Deep Learning and its Applications to ASR (3/5)

- **Deep Learning** is the cutting edge!
  - Use deep neural network hidden Markov model (DNN-HMM) hybrid architecture to train DNN to produce a distribution over senones (tied triphone states) as its output



deeper layers,
longer features &
wider temporal contexts

$$b_{s_i}(\mathbf{o}) = p(\mathbf{o} \mid s_i) = \frac{P_{\mathrm{DNN}}(s_i \mid \mathbf{o})\, p(\mathbf{o})}{P_{\mathrm{ML}}(s_i)} \propto \frac{P_{\mathrm{DNN}}(s_i \mid \mathbf{o})}{P_{\mathrm{ML}}(s_i)}$$

$$P_{\mathrm{DNN}}(s_i \mid \mathbf{o}) = v_i^L = \mathrm{softmax}_{\,i}\left(\mathbf{z}^L\right) = \frac{e^{z_i^L}}{\sum_j e^{z_j^L}}$$

$$\mathbf{v}^\ell = f\left(\mathbf{z}^\ell\right) = f\left(\mathbf{W}^\ell \mathbf{v}^{\ell-1} + \mathbf{b}^\ell\right), \text{ for } 0 < \ell < L$$

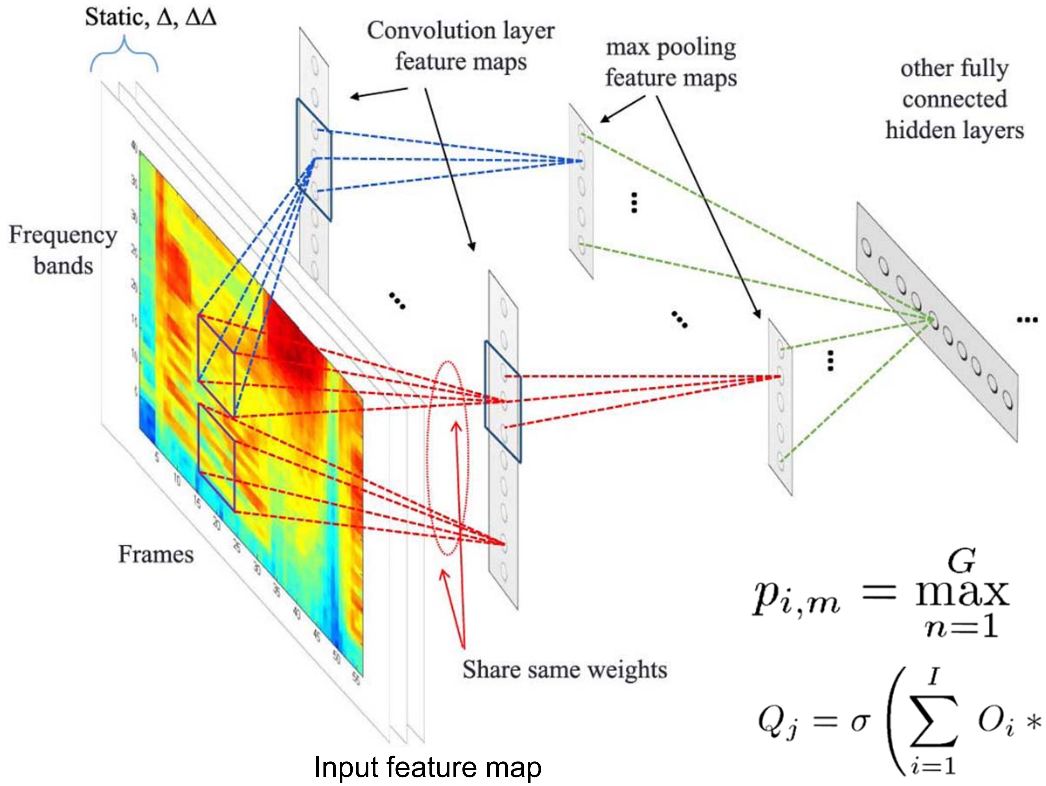
$f(\cdot)$: sigmoid, hyperbolic, or rectified linear unit (ReLU) functions

Model parameters of DNN can be estimated with **the error back-propagation** algorithm and **stochastic gradient decent** (SGD).

G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition," IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 1. pp. 30-42, 2012

Deep Learning and its Applications to ASR (4/5)

- **CNN-HMM**
  - CNN: Convolutional Neural Networks



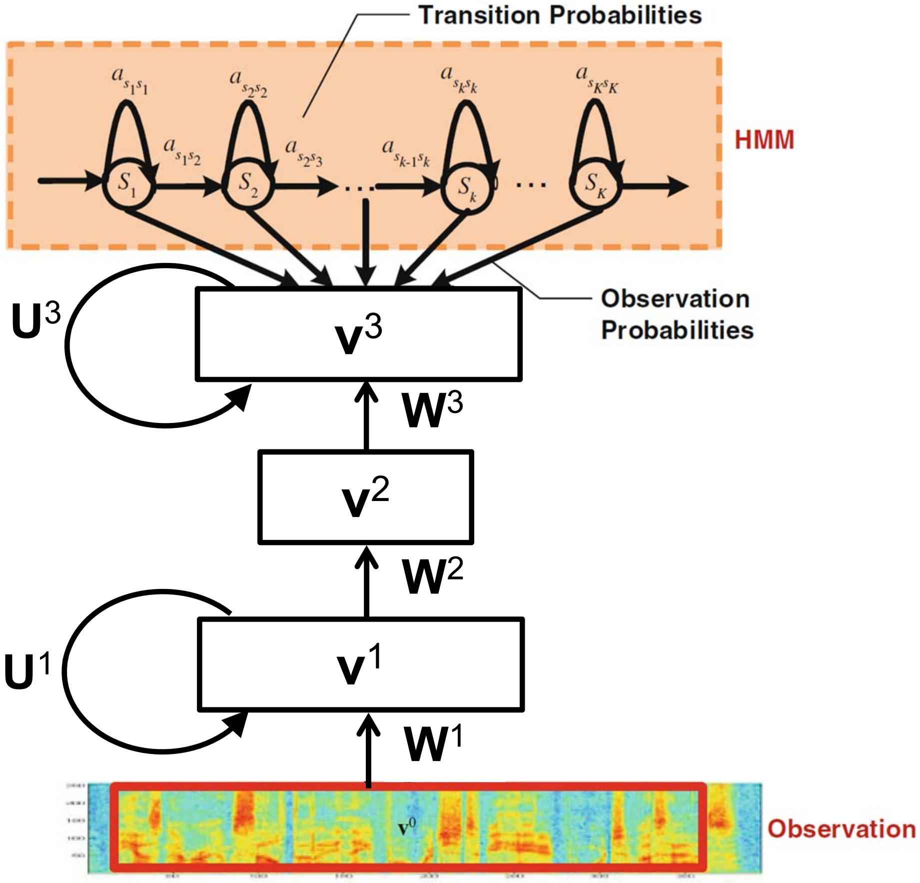
$$p_{i,m} = \max_{n=1}^{G} \quad q_{i,(m-1)\times s+n}$$

$$Q_j = \sigma\left(\sum_{i=1}^{I} O_i * \mathbf{w}_{i,j}\right) \quad (j = 1, \ldots, J)$$

Abdel-Hamid et al., Convolutional Neural Networks for Speech Recognition, IEEE/ACM Transactions on Adio, Speech, and Language Processing, Vol. 22, No. 10, 2014

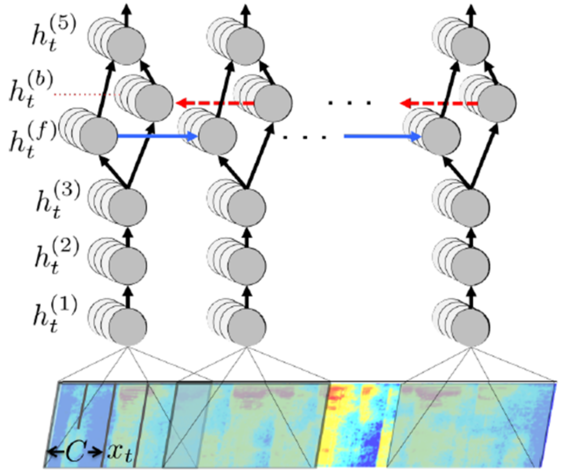
Deep Learning and its Applications to ASR (5/5)

- Recurrent Neural Networks (**RNN-HMM**)



Baidu Research Approach

Output: Text (Letters, Words, ….)



Structure of our RNN model and notation.

A. Hannun et al. (Lead by Andrew Ng), "Deep Speech: Scaling up end-to-end speech recognition," arXiv:1412.5567v2, December 2014.

Example: Automatic Meeting Transcription

Manual Transcripts

A: 那 會 在 二 a 那個 那 叫 什麼 二 b 啊 二 a
A: vip vip room
B: 欸
A: 就是 大家 開 all hands meeting 那裡
C: 錄音 的話 就 只 能 用 八爪魚 喔
A: 錄音 就 對 啊 那 場 就 反正 錄下 來 就 好了 對
A: 好 一 開始
D: 請問 一下
D: 上 次 二 a. 的 時候 那個 圓 方 不是 有 來 教 我們 怎麼 用 八爪魚 錄音 所以 那個 測試 設 定都 沒有 動
D: 就 直接 麥克風 可以 把 聲音 收進 來
A: 圓 圓形 會議 對 啊 圓形 會議 是 這樣
D: 好 好
A: 可是 我們 這 一 次 不是 在 圓形 我們 這 次 是 在 呃 vip
A: 就是 董事長 開會 的 地方

Automatic Transcripts

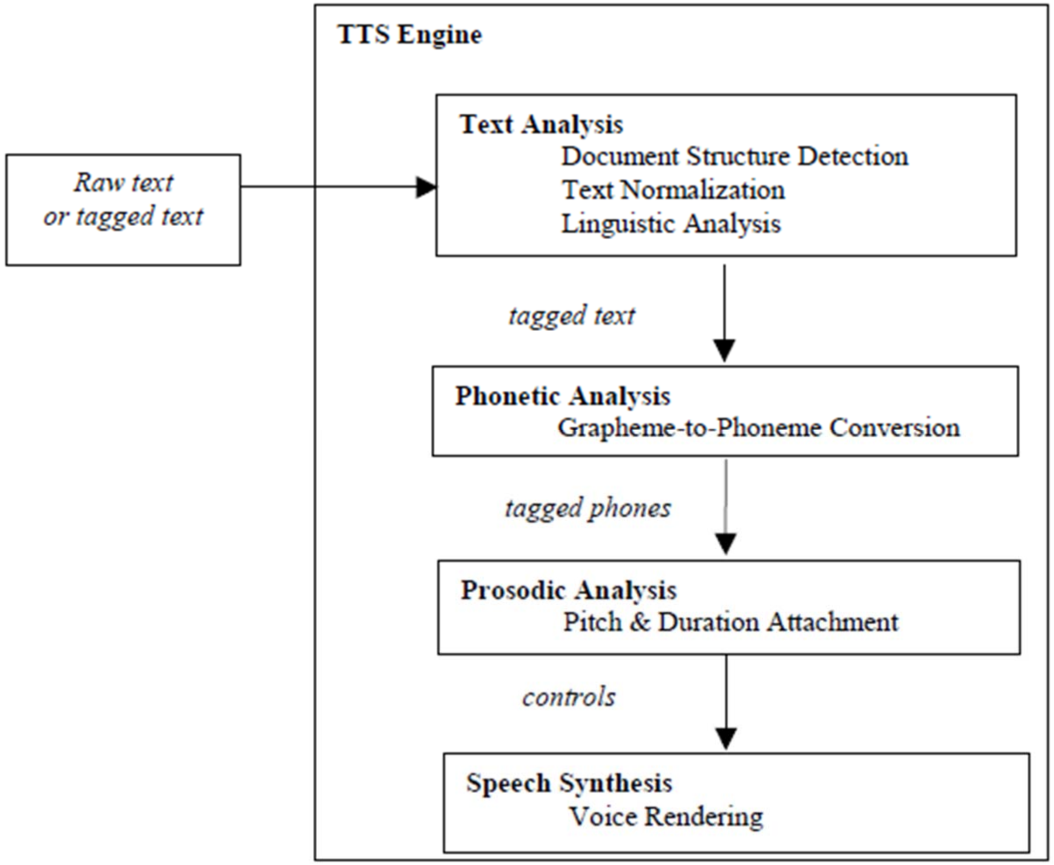
A: 那 會 在 二 a. h 那個 資料 怎麼 二 的 啊 把 二 a.
A: 七 vip 喔 vip vip room
B: 嘿
A: 可是 打開 過 hand meeting 那裡
C: 錄音 的話 是 怎麼 用 滑動 語料
A: 錄音 就 對 啊 那 一 場 就 反正 錄下 就 好了
A: 好 一 開始 了
D: 請問 一下
D:上 是 二月 的 時候 那個 員工 不是 來 教 我們 怎麼 跟 八爪魚 錄音 最 那個 測試 設定 檔 秒 鐘
D: 就 支 麥克風 可以 把 聲音 投 進來
A: 每 圓形 會議 對 啊 圓形 會議室 這樣
D: 好
A: 可是 我們 這 次 不是 在 圓形 我 edge vip
A: 是 董事會 開會 的 地方

Chinese Character Error Rate (CER%)

GMM-HMM	DNN-HMM	CNN-HMM	RNN(LSTM)-HMM	RNN(BLSTM)-HMM
49.74	40.95	35.41	51.63	41.87

Text to Speech (1/2)

- Text to speech (TTS) can be viewed as ASR in reverse



  – We are now able to general high-quality TTS systems, although the quality is inferior to human speech for general-purpose applications

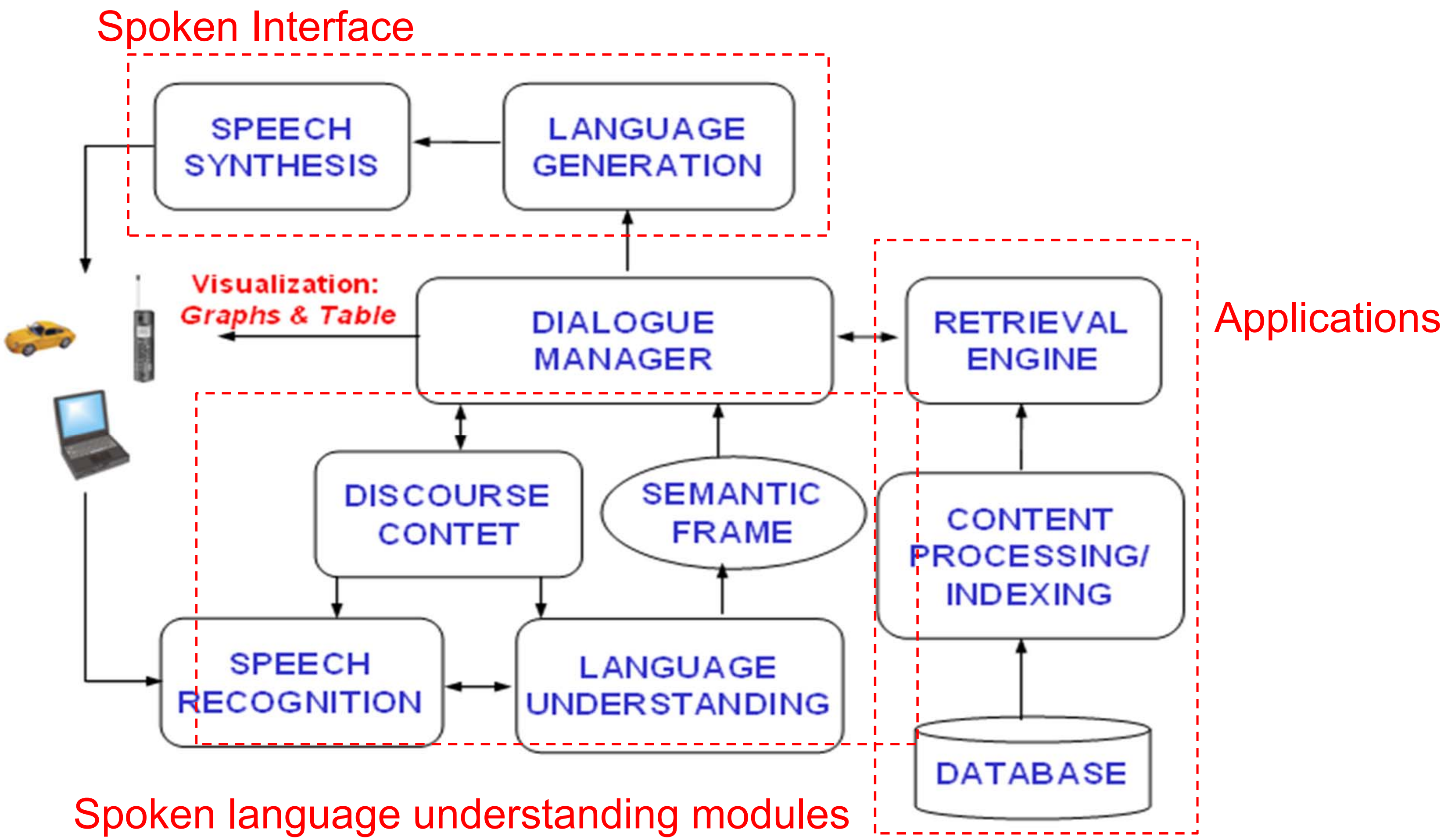
Text to Speech (2/2)

- Example 1
  - 青少年在成長的過程中，非常需要角色模範的引導、族群的認同及自我的肯定，所以我一直在找這方面的好書來幫助孩子。
  - Original Speech: Synthesized Speech:
- Example 2
  - 新北市市長朱立倫昨天邀台北市市長柯文哲參加新北市天燈節第三場活動，兩人在廿呎高的剪紙天燈上寫下「雙北合作」「神采飛羊」，柯則寫下「天佑台灣」，大小天燈齊放升空，照亮平溪夜空。
  - Synthesized Speech:

Spoken Dialogue: CMU's Systems

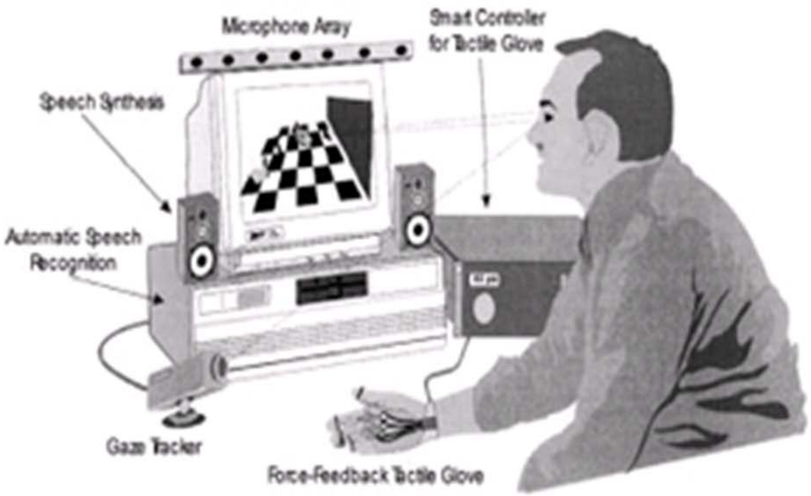
- Spoken language is attractive because it is the most natural, convenient and inexpensive means of exchanging information for humans
- In mobilizing situations, using keystrokes and mouse clicks could be impractical for rapid information access through small handheld devices like PDAs, cellular phones, etc.

Spoken Dialogue: Basic System Architecture



V. Zue, J.R. Glass, Conversational Interfaces: Advances and Challenges. Proceedings of the IEEE, Vol. 88, No. 8, August 2000

Spoken Dialogue: Multimodality of Input and Output



Experimental client workstation incorporating sight, sound, and touch modalities for human/machine communication. The eye tracker provides a gaze-controlled cursor for indicating objects in the display. The tactile force-feedback glove allows displayed objects to be grasped, "felt," and moved. Hands-free speech recognition and synthesis provides natural conversational interaction [7].

I. Marsic, A. Medl, and J. Flanagan, Natural Communication with Information Systems. Proceedings of the IEEE, Vol. 88, No. 8, August 2000

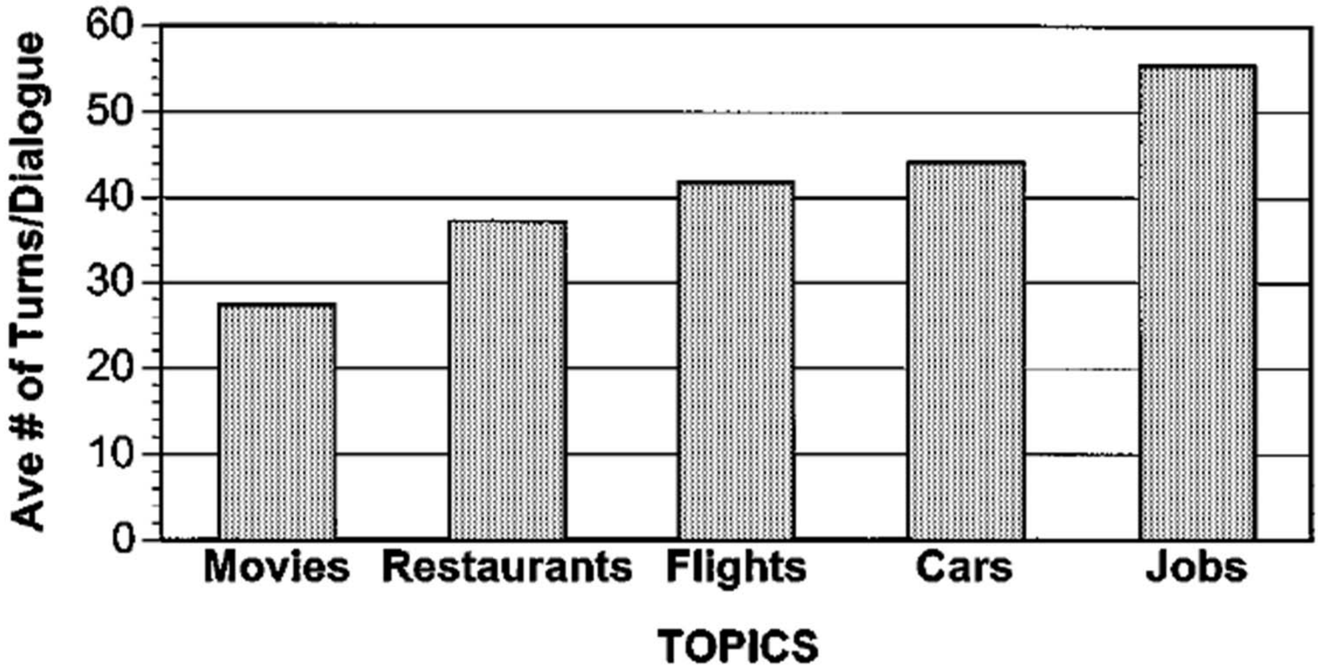
Spoken Dialogue: Some Deployed Systems

- Complexity Analysis

Domain	Language	Vocabulary Size	Average Words/Utt	Average Utts/Dialogue
CSELT Train Timetable Info	Italian	760	1.6	6.6
SpeechWorks Air Travel Reservation	English	1000	1.9	10.6
Philips Train Timetable Info	German	1850	2.7	7.0
CMU Movie Information	English	757	3.5	9.2
CMU Air Travel Reservation	English	2851	3.6	12.0
LIMSI Train Timetable Info	French	1800	4.4	14.6
MIT Weather Information	English	1963	5.2	5.6
MIT Air Travel Reservation	English	1100	5.3	14.1
AT&T Operator Assistance	English	4000	7.0	3.0
Air Travel Reservations (human)	English	?	8.0	27.5

Spoken Dialogue: Some Statistics

- Topics vs. Dialogue Terms



Current Deployed Speech Retrieval and Spoken Dialogue Systems

- Google, Apple, Microsoft and Amazon's Deployed Services

Google-411:
Finding and connecting to
local business

Google Voice Search

Apple Siri

Microsoft Cortana

http://www.google.com/mobile/voice-search/

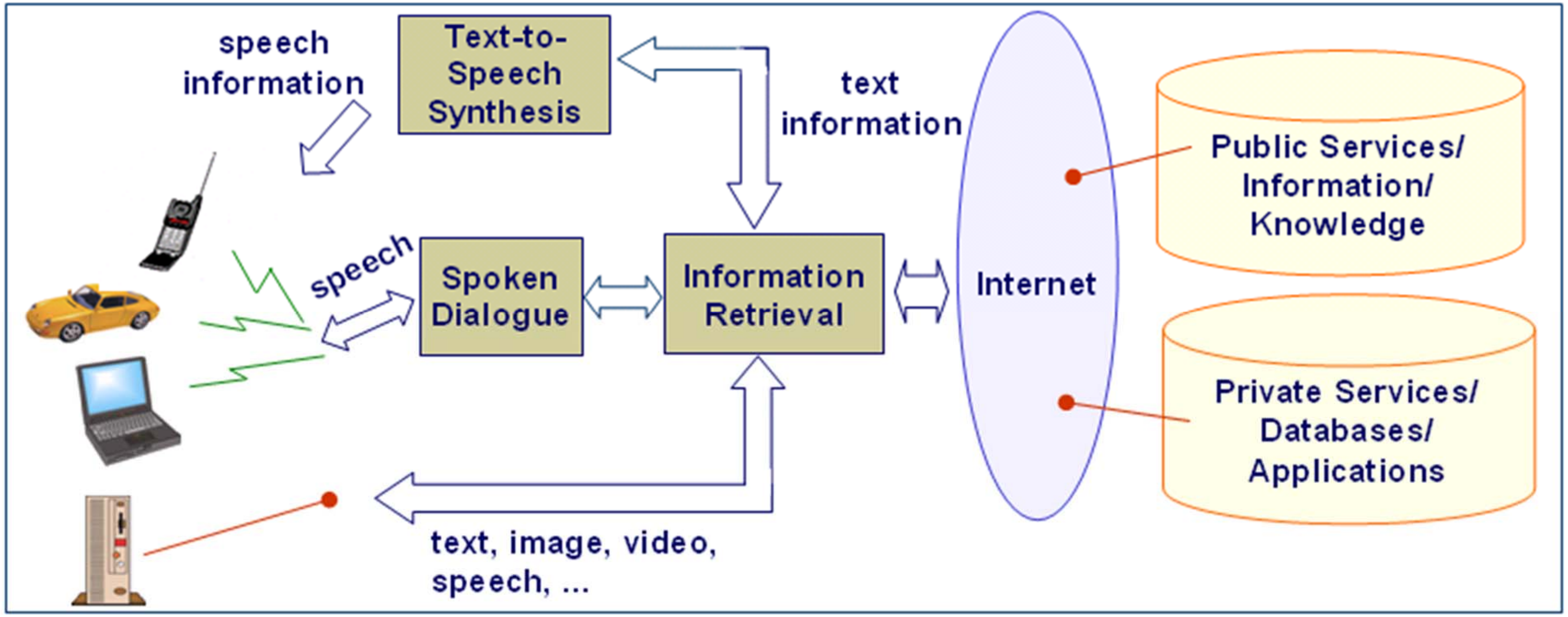
http://zh.wikipedia.org/wiki/Microsoft_Cortana

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

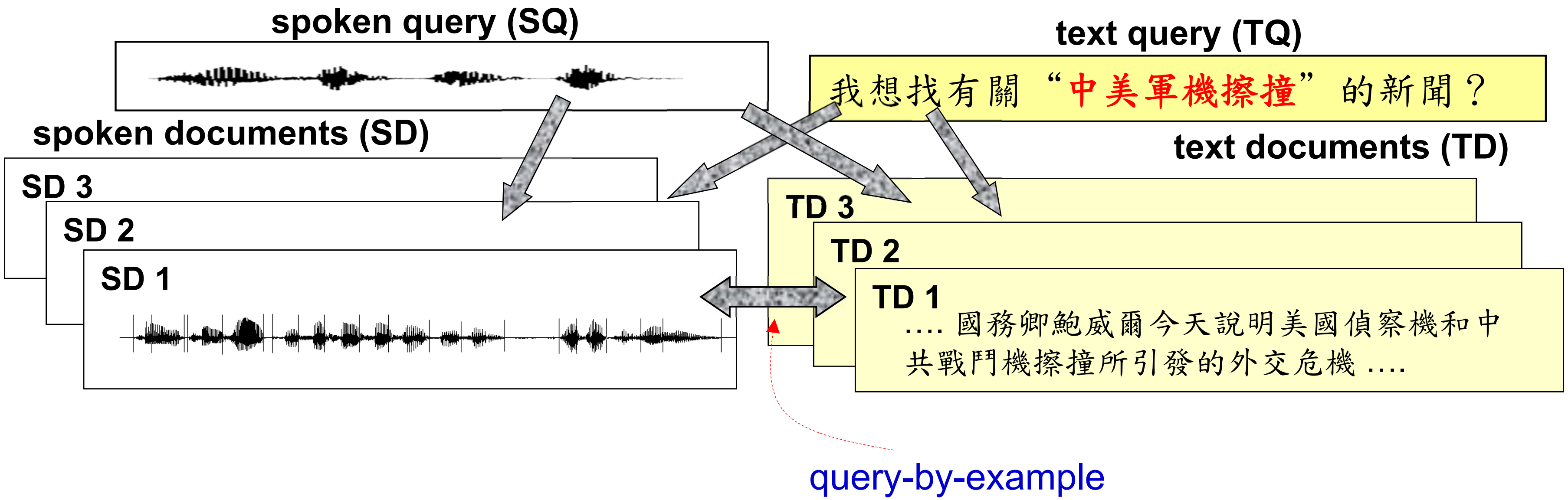
We are witnessing the golden age of ASR!

Speech-based Information Retrieval (1/5)

- Task :
  - Automatically indexing a collection of spoken documents with speech recognition techniques
  - Retrieving relevant documents in response to a text/speech query



Speech-based Information Retrieval (2/5)



- SQ/SD is the most difficult
- TQ/SD is studied most of the time

- Query-by-example
  - Attempt to retrieve relevant documents when users provide some specific query exemplars describing their information needs
  - Useful for news monitoring and tracking

Speech-based Information Retrieval (3/5)

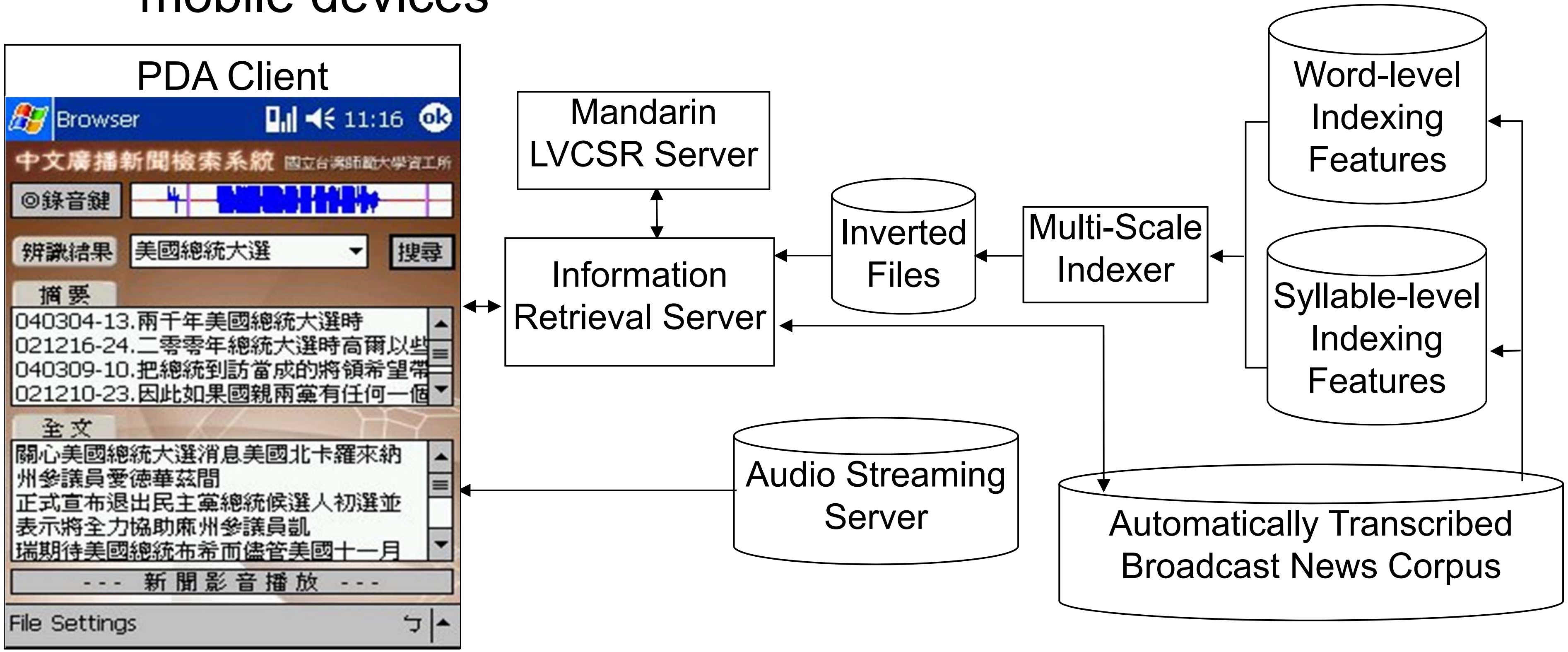
輸入聲音問句：“請幫我查總統府升旗典禮”

中文語音資訊檢索雛形展示系統。

C.f. B. Chen, H.M. Wang, Lin-shan 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 , Vol. 10, No. 5, pp. 303-314, July 2002.

Speech-based Information Retrieval (4/5)

- Spoken queries retrieving text news documents via mobile devices



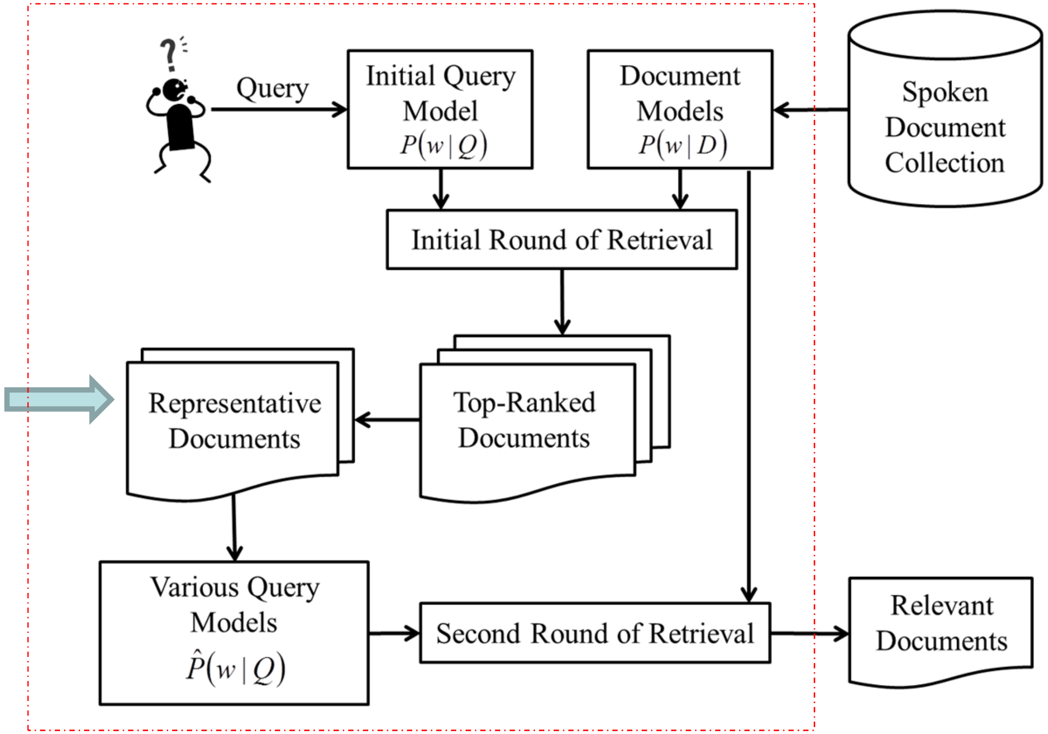
C.f. B. Chen, Y..T. Chen, C.H. Chang, H.B. Chen, "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," Interspeech2005

Chang, E., Seide, F., Meng, H., Chen, Z., Shi, Y., And Li, Y. C. 2002. A system for spoken query information retrieval on mobile devices. IEEE Trans. on Speech and Audio Processing 10, 8 (2002), 531-541.

Speech-based Information Retrieval (5/5)

- Query modeling for information retrieval

Considering **relevance**, **non-relevance**, **diversity** and **density** cues

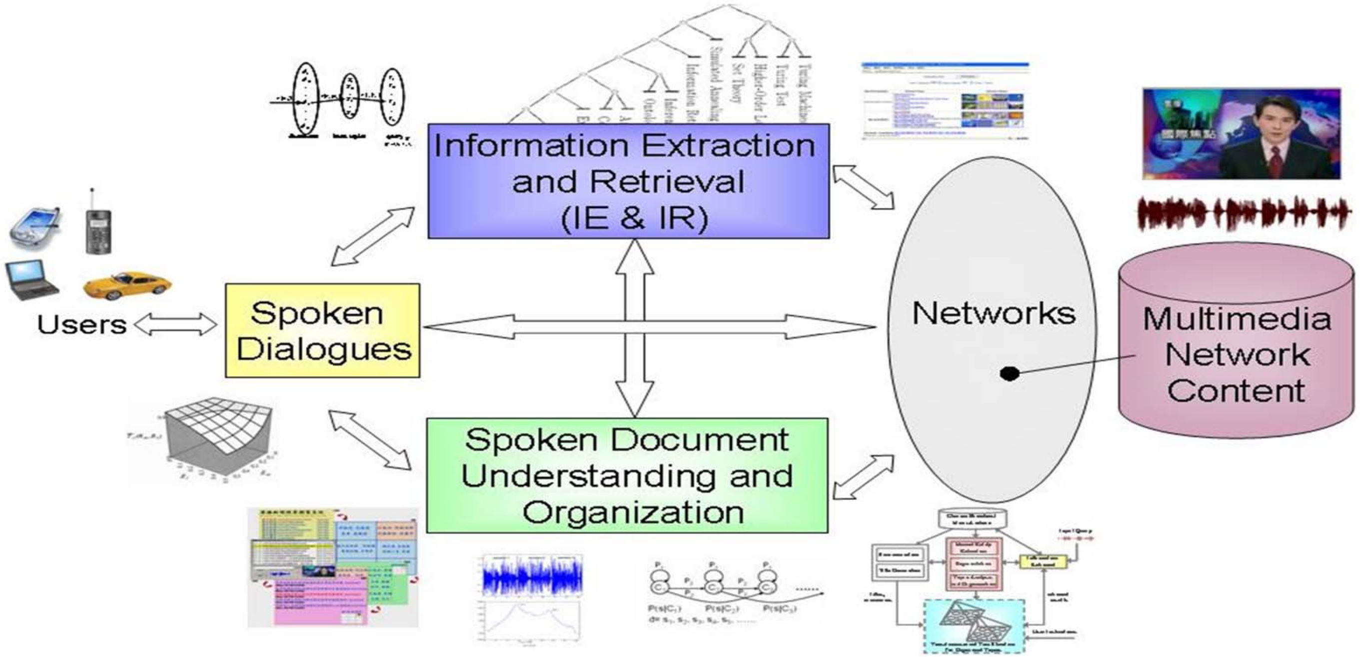


$$D^* = \underset{D \in \mathbf{D}_{\mathrm{Top}} - \mathbf{D}_{\mathrm{P}}}{\arg\max} \left[(1-\alpha-\beta-\gamma) \cdot M_{Rel}(Q,D) + \alpha \cdot M_{NR}(Q,D) + \beta \cdot M_{Diversity}(D) + \gamma \cdot M_{Density}(D) \right]$$

C.f. B. Chen, K.-Y. Chen, P.-N. Chen, Y.-W. Chen, "Spoken document retrieval with unsupervised query modeling techniques," IEEE Transactions on Audio, Speech and Language Processing, Vol. 20, No. 9, pp. 2602-2612, 2012

Spoken Document Organization and Understanding (1/2)

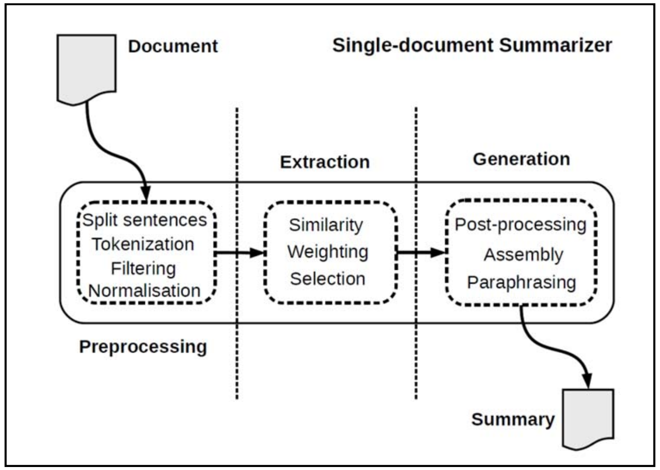
- Problems
  – The content of multimedia documents very often described by the associated speech information
  – Unlike text documents with paragraphs/titles easy to look through at a glance, multimedia/spoken documents are unstructured and difficult to retrieve/browse



C.f. L.S. Lee and B. Chen, "Spoken document understanding and organization," IEEE Signal Processing Magazine, vol. 22, no. 5, pp. 42-60, Sept. 2005

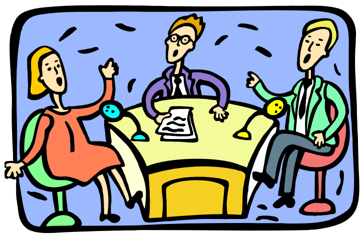
Spoken Document Organization and Understanding (2/2)

- Speech Summarization

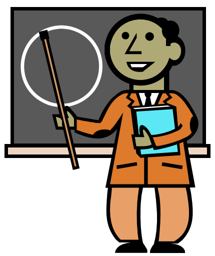


conversations

meetings



lectures



broadcast and TV news

distilling important information

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

C.f. Y. Liu and D. Hakkani-Tür, "Speech summarization," Chapter 13 in Spoken Language Understanding: Systems for Extracting Semantic Information from Speech, G. Tur and Renato D. Mori (eds.), Wiley, 2011.

Speech-to-Speech Translation (1/2)

- Multilingual interactive speech translation
  - Aim at the achievement of a communication system for precise recognition and translation of spoken utterances for several conversational topics and environments by using human language knowledge synthetically (adopted form ATR-SLT)

ATR-SLT

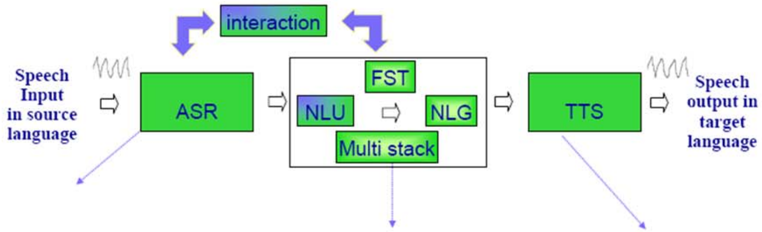
IBM Mastor Project

Speech-to-Speech Translation (2/2)

Handheld System

Laptop systems
- hands-free, eyes-free function

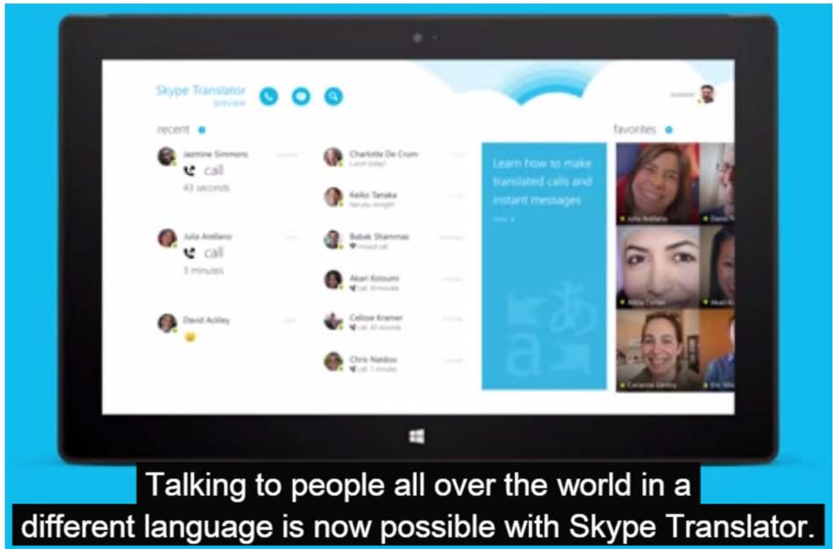
IBM Advanced Speech-to-Speech Translation Techniques



ASR engines and models:
Decode speech into text

Translation engines and models:
Translate word/concept to another language in text

TTS engines and models:
Convert Text to Speech



Car Play Systems

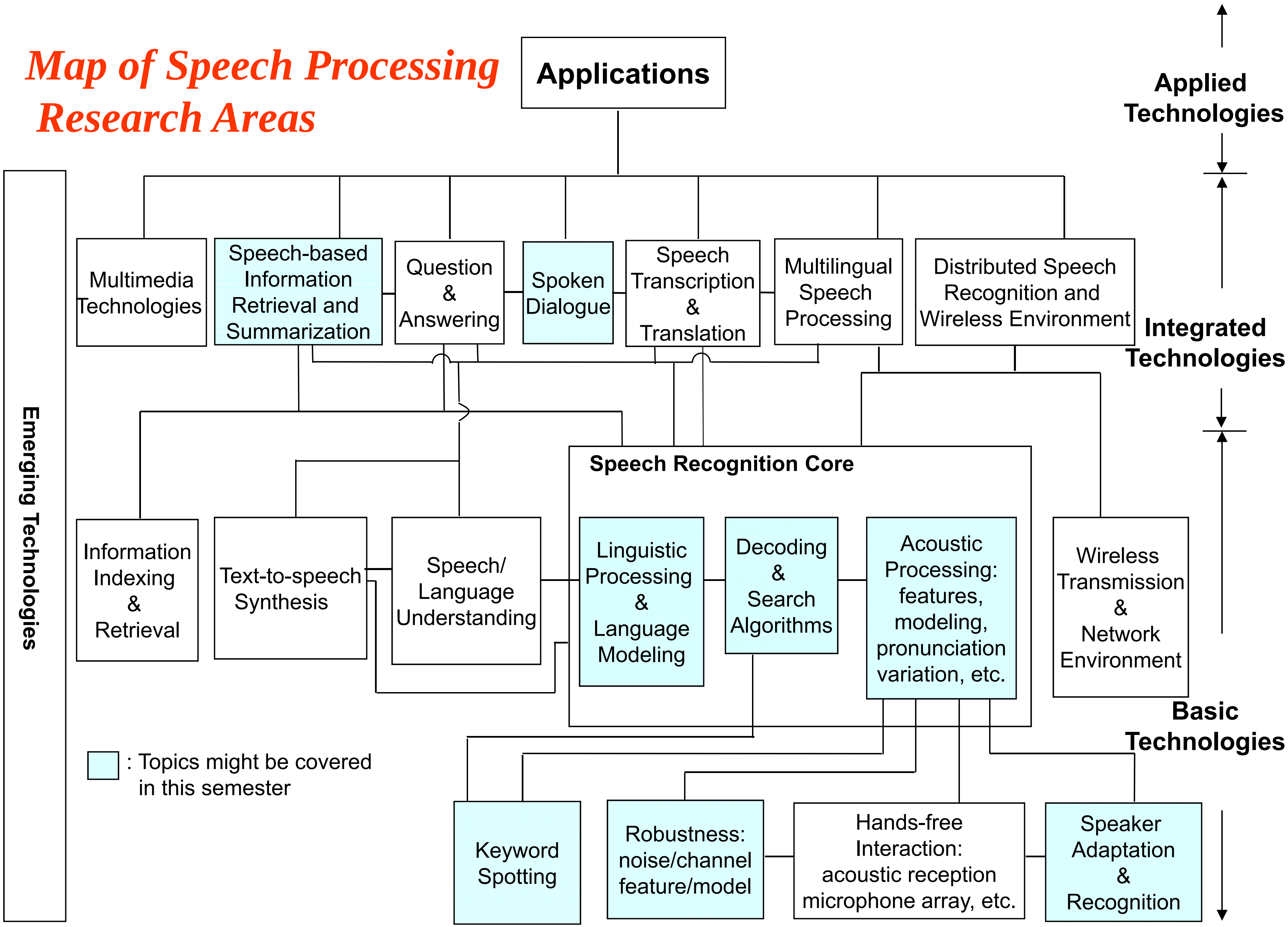
- Aim to s provide a smarter, safer way to use your communication devices in vehicle
- E.g. Apple Car Play

Computer-Assisted Pronunciation Training (CAPT)

- **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)

1. Mandarin Chinese CAPT: http://140.122.96.191/ALS/assessment.aspx
2. English CAPT: http://www.coolenglish.edu.tw/

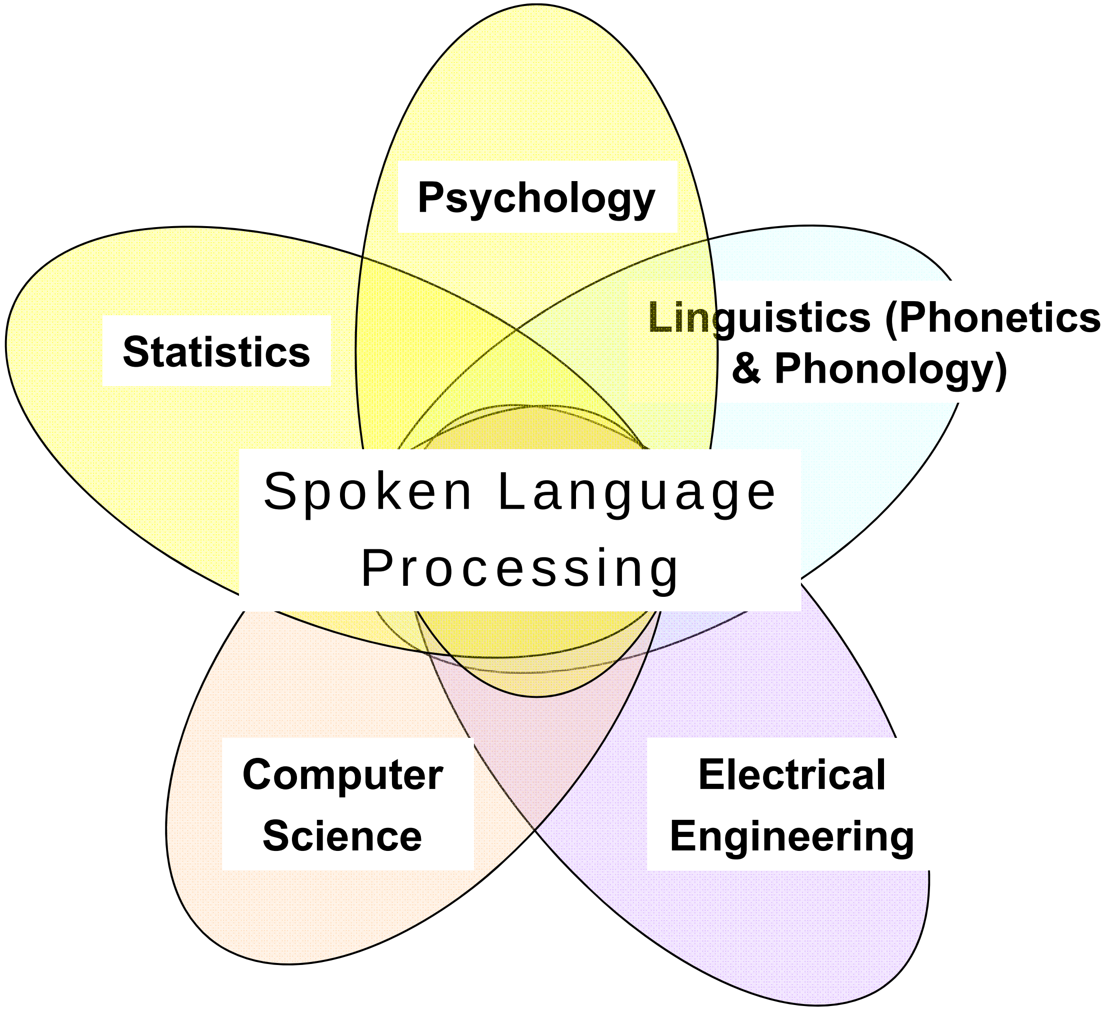
Map of Speech Processing Research Areas



Adapted from Prof. Lin-shan Lee

Different Academic Disciplines

- The foundations of spoken language processing lies in

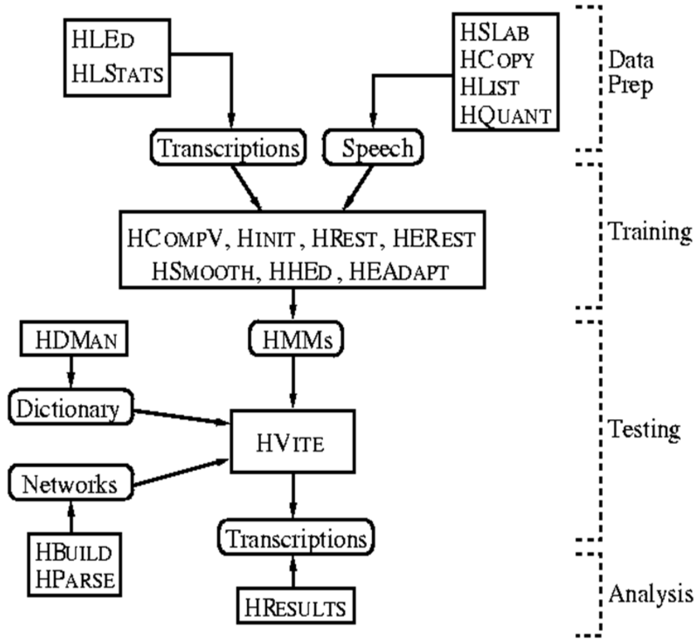


Speech Processing Toolkit (1/2)

- HTK (**H**idden Markov Model **T**ool**K**it)
  - A toolkit for building Hidden Markov Models (HMMs)
  - The HMM can be used to model any time series and the core of HTK is similarly general-purpose
  - In particular, for the acoustic feature extraction, HMM-based acoustic model training and HMM network decoding

Speech Processing Toolkit (2/2)

- HTK (**H**idden Markov Model **T**ool**K**it)



- Nowadays, **Kaldi** emerges as a cutting-edge toolkit for developing speech recognition tasks

http://kaldi.sourceforge.net/

Journals & Conferences

- Journals
  - IEEE Transactions on Audio, Speech and Language Processing
  - Computer Speech & Language
  - Speech Communication
  - Proceedings of the IEEE
  - IEEE Signal Processing Magazine
  - ACM Transactions on Speech and Language Processing
  - ACM Transactions on Asian and Low-Resource Language Information Processing
  - …
- Conferences
  - IEEE International Conference on Acoustics, Speech, Signal processing (ICASSP)
  - Annual Conference of the International Speech Communication Association (Interspeech)
  - IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)
  - IEEE Workshop on Spoken Language Technology (SLT)
  - International Symposium on Chinese Spoken Language Processing (ISCSLP)
  - ROCLING Conference on Computational Linguistics and Speech Processing
  - …

Speech Industry (1/3)

- Telecommunication
- Information Appliance
- Interactive Voice Response
- Voice Portal
- Multimedia Database
- Education
- …..

Tentative Schedule

Topics to be Covered
Overview & Introduction
Hidden Markov Models
Spoken Language Structure
Acoustic Modeling & HTK Toolkit & Kaldi Toolkit
Statistical Language Modeling & SRI LM Toolkit
Speech Signal Representations
Digit Recognition, Word Recognition and Keyword Spotting
Large Vocabulary Continuous Speech Recognition (LVCSR)
Speech Enhancement and Environment Robustness
Model Training and Adaptation Techniques
Utterance Verification and Confidence Measures