Speech Recognition

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
- Technology for **Automatic Speech Recognition** (ASR) 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
  - 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
  - T. F. Quatieri. Discrete-Time Speech Signal Processing - Principles and Practice. Prentice Hall, 2002
  - 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"

Website for This Course

- Visit http://berlin.csie.ntnu.edu.tw/ and then click the link "Fall 2010: Speech Recognition"

Speech Recognition

Fall 2010

2:10 ~5:00 pm, Mondays

Instructor: Dr. Berlin Chen (陳柏琳)

Topic List and Schedule:

09/13		Course Overview & Introduction	

Reference Books:

- X. Huang, A. Acero, H. Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development, Prentice Hall, 2001
- Jacob Benesty (ed.), M. Mohan Sondhi (ed.), Yiteng Huang (ed.), Springer Handbook of Speech Processing, Springer, 2007
- 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
- J. R. Deller, J. H. L. Hansen, J. G. Proakis, "Discrete-Time Processing of Speech Signals," IEEE Press, 2000
- C. Manning and H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999
- J. Bellegarda, Latent Semantic Mapping: Principles & Applications (Synthesis Lectures on Speech and Audio Processing), 2008
- T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch (eds.) , Handbook of Latent Semantic Analysis, Lawrence Erlbaum, 2007
- Ethem Alpaydin, Introduction to Machine Learning, MIT Press, 2004
- D. P. Bertsekas, J. N. Tsitsiklis, "Introduction to Probability," Athena Scientific, 2002

Reference Papers:

- L. Rabiner. The Power of Speech. Science, Vol. 301, pp. 1494-1495, Sep. 2003.
- Baker, J.M.et al., Research Developments and directions in speech recognition and understanding, part 1, IEEE Signal Processing Magazine 25(3), May 2009.
- Baker, J.M.et al., Research Developments and directions in speech recognition and understanding, part 2, IEEE Signal Processing Magazine 25(4), July 2009.
- M. Ostendorf, Speech Technology and Information Access, IEEE Signal Processing Magazine 25(3), May 2008.
- L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, No. 2, February 1989
- A. V. Oppenheim and R. W. Schafer, "From Frequency to Quefrency: A History of the Cepstrum," IEEE Signal Processing Magazine 21(5), September 2004.
- A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1, 1977
- J. 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
- J. W. Picone, "Signal modeling techniques in speech recognition," proceedings of the IEEE, September 1993, pp. 1215-1247
- R. Rosenfeld, "Two Decades of Statistical Language Modeling: Where Do We Go from Here?," Proceedings of IEEE, August, 2000
- H. Ney, "Progress in Dynamic Programming Search for LVCSR," Proceedings of the IEEE, August 2000
- Aubert, X. L., "An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition," Computer Speech and Language, vol. 16, 2002, pp. 89-114.
- H. Hermansky, "Should Recognizers Have Ears?", Speech Communication, 25(1-3), Speech Communication, 25(1-3), 1998.
- J. R. Bellegarda, "Statistical Language Model Adaptation: Review and Perspectives," Speech Communication, vol. 42, no.1, pp. 93-108, 2004.
- B. Roark, "A survey of discriminative language modeling approaches for large vocabulary continuous speech recognition," in Large Margin and Kernel Approaches to Speech and Speaker Recognition, J. Keshet and S. Bengio (Eds.), Wiley, 2009.
- L. Rabiner, B.H. Juang, "Speech Recognition: Statistical Methods," Encyclopedia of Language & Linguistics, pp. 1-18, 2006.
- P. Nguyen, "TechWare: Speech recognition software and resources on the web," IEEE Signal Processing Magazine 25(3), May 2009.
- J. B. Allen, F. Li, "Speech Perception and Cochlear Signal Processing," IEEE Signal Processing Magazine 25(4), July 2009.
- A. Orlitsky, N. P. Santhanam, J. Zhang, "Always Good Turing: Asymptotically Optimal Probability Estimation," Science, 17 October 2003.
  - Proceedings of IEEE 88(8), August, 2000 (Special Issue on Spoken Language Processing)
  - IEEE Signal Processing Magazine 22(5), September 2005 (Special Issue on Speech Technology and Systems in Human-Machine Communication)
  - IEEE Signal Processing Magazine 25(3), May 2008 (Special Issue on Spoken Language Technology)

Reference Presentations:

- J. Droppo, Noise Robust Automatic Speech Recognition, a comprehensive tutorial talk given at EUSIPCO 2008
- B. Chen, Latent Semantic Approaches for Information Retrieval and Language Modeling, a talk given at Telecommunication Laboratories, Chunghwa Telecom Co., Ltd., 2008
- B. Chen, Recent Developments in Chinese Spoken Document Search and Distillation, a talk given at Google Taipei, 2009

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

nuance

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 | 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)

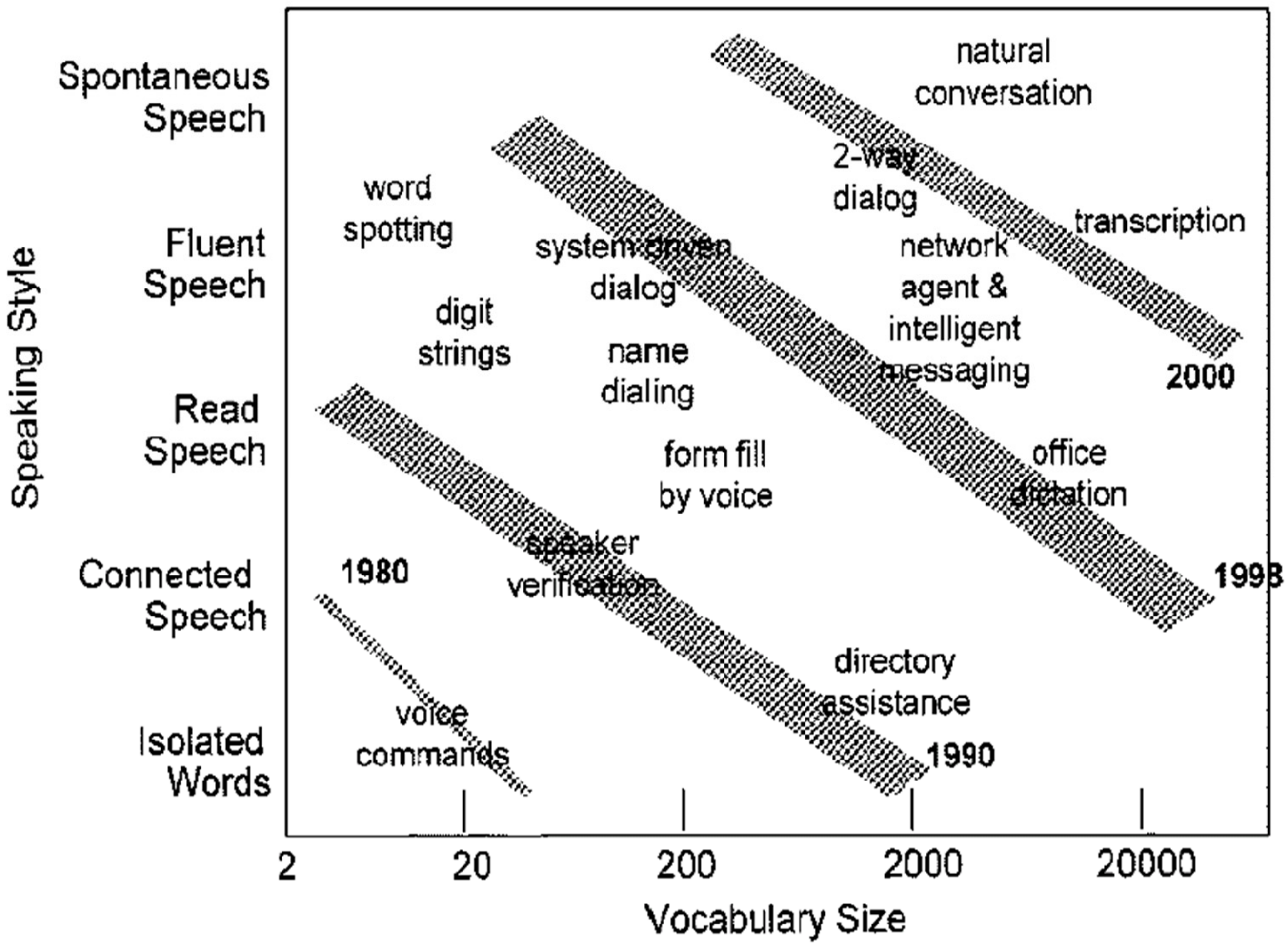
Past Benchmark Tests

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)

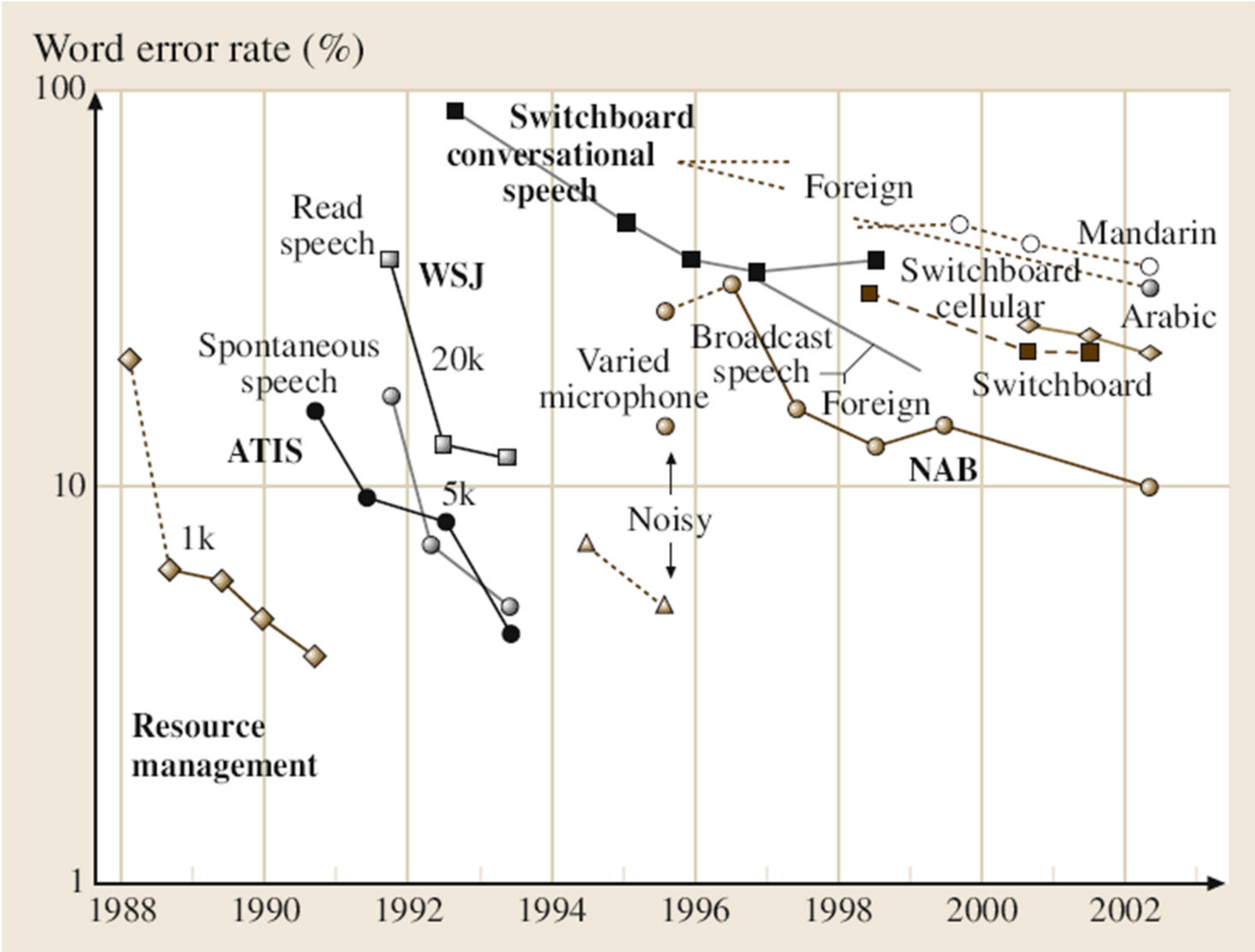
http://www.nist.gov/itl/iad/mig/bmt.cfm

Progress of Technology (2/6)

- Generic Application Areas (vocabulary vs. speaking style)



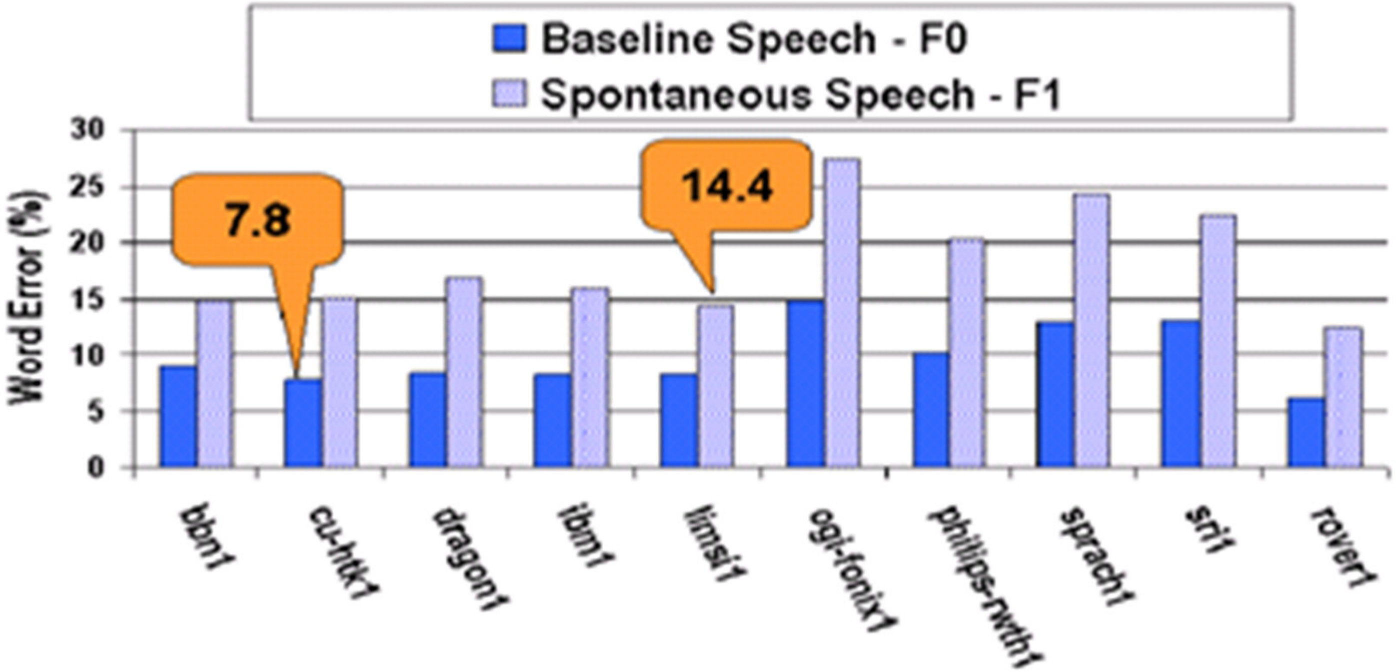
Progress of Technology (3/6)



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

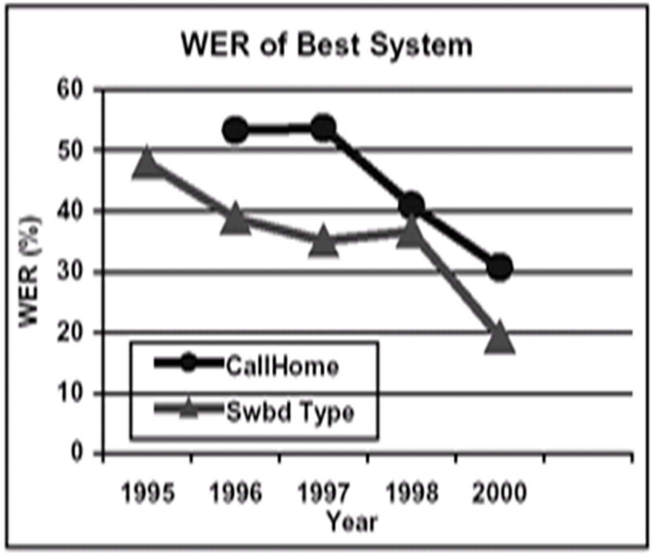


Figure 4 History of lowest word error rates (WER) obtained in NIST conversational speech evaluations on Switchboad and CallHome type conversations in English [26].

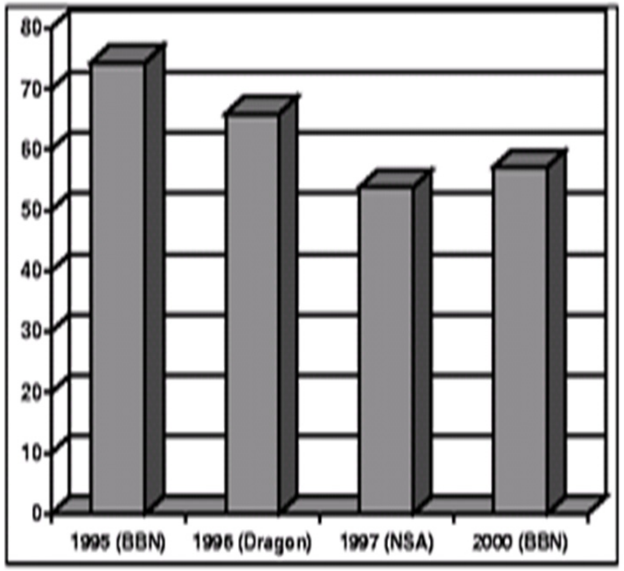


Figure 5 Chinese Character error rates of the best performing evaluation system in NIST Mandarin conversational speech evaluations 1995-2000 [26].

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

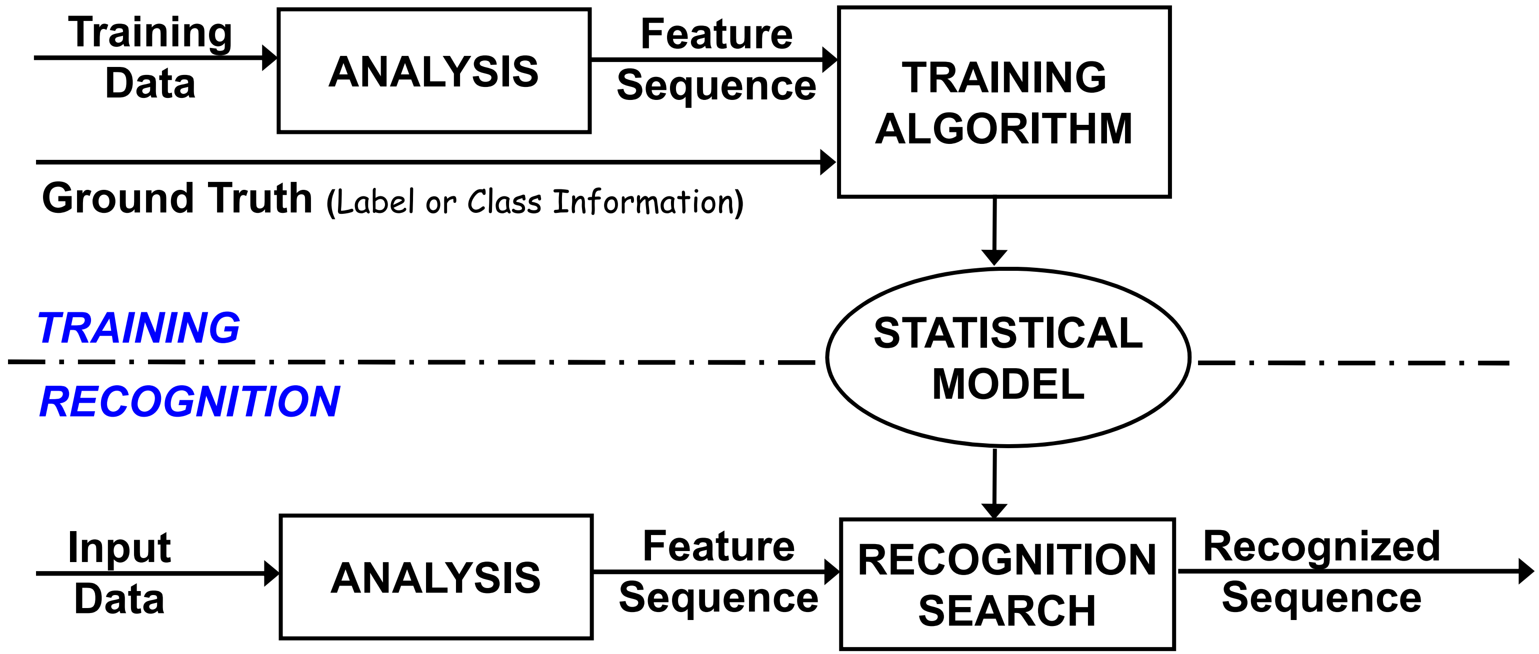
  - Adopted from

Cambridge University
Engineering Department

Rich Transcription Workshop 2003

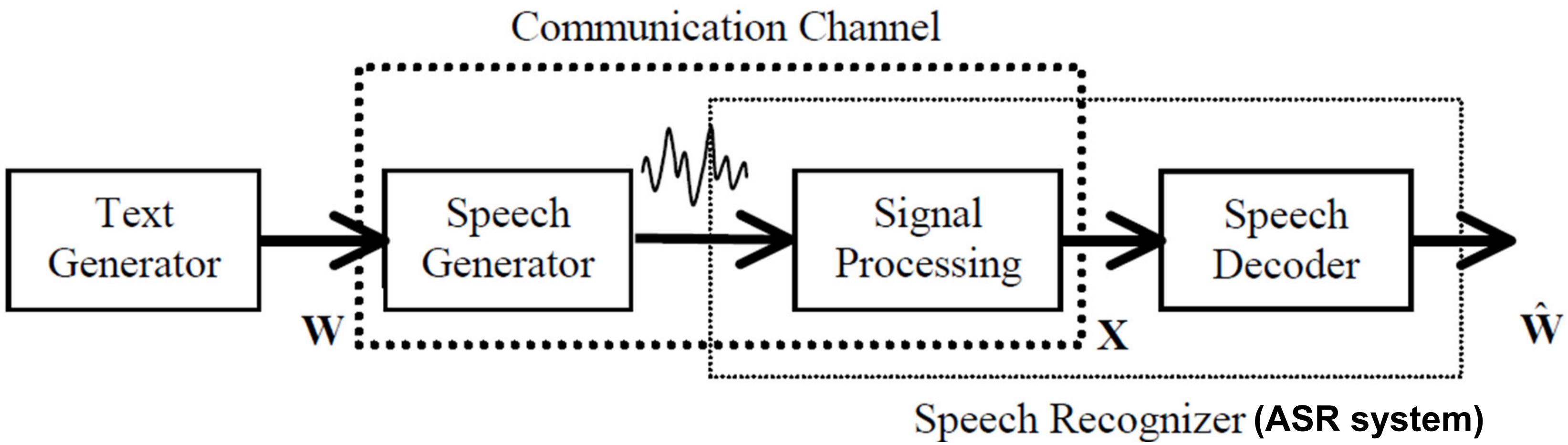
Statistical Modeling Paradigm

- 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

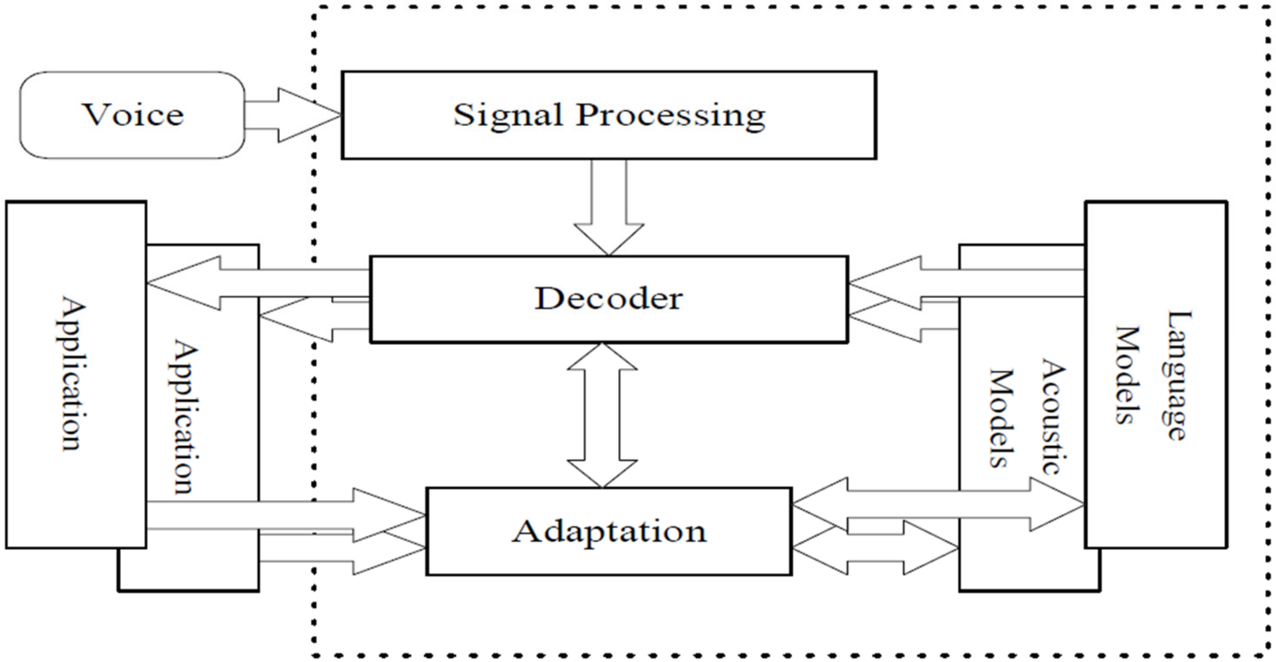
A Source-Channel Model for ASR



– 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

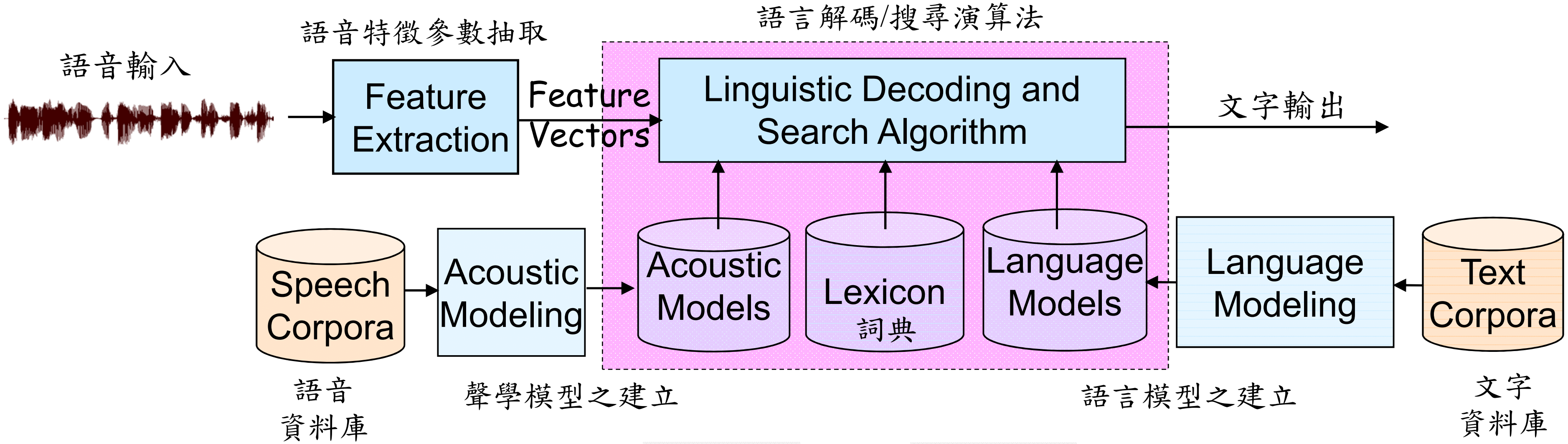


- Signal processing: extract salient features for the decoder
- Decoder: use both acoustic and language models to generate the “best” word sequence in response to the input voice
- Adaptation: modify either acoustic or language models so that improved performance can be obtained

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

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

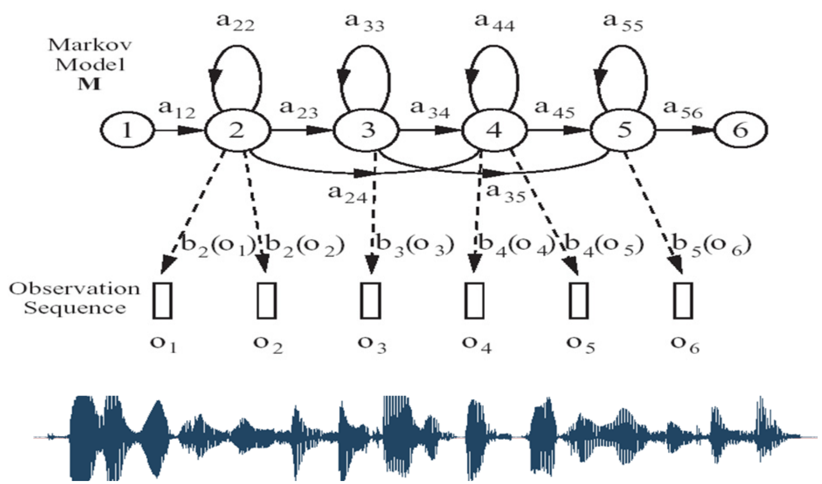
$$= \arg\max_{\mathbf{W}} \boxed{p(\mathbf{X}|\mathbf{W})P(\mathbf{W})}$$

Decoding

Acoustic Modeling　Language Modeling

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

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

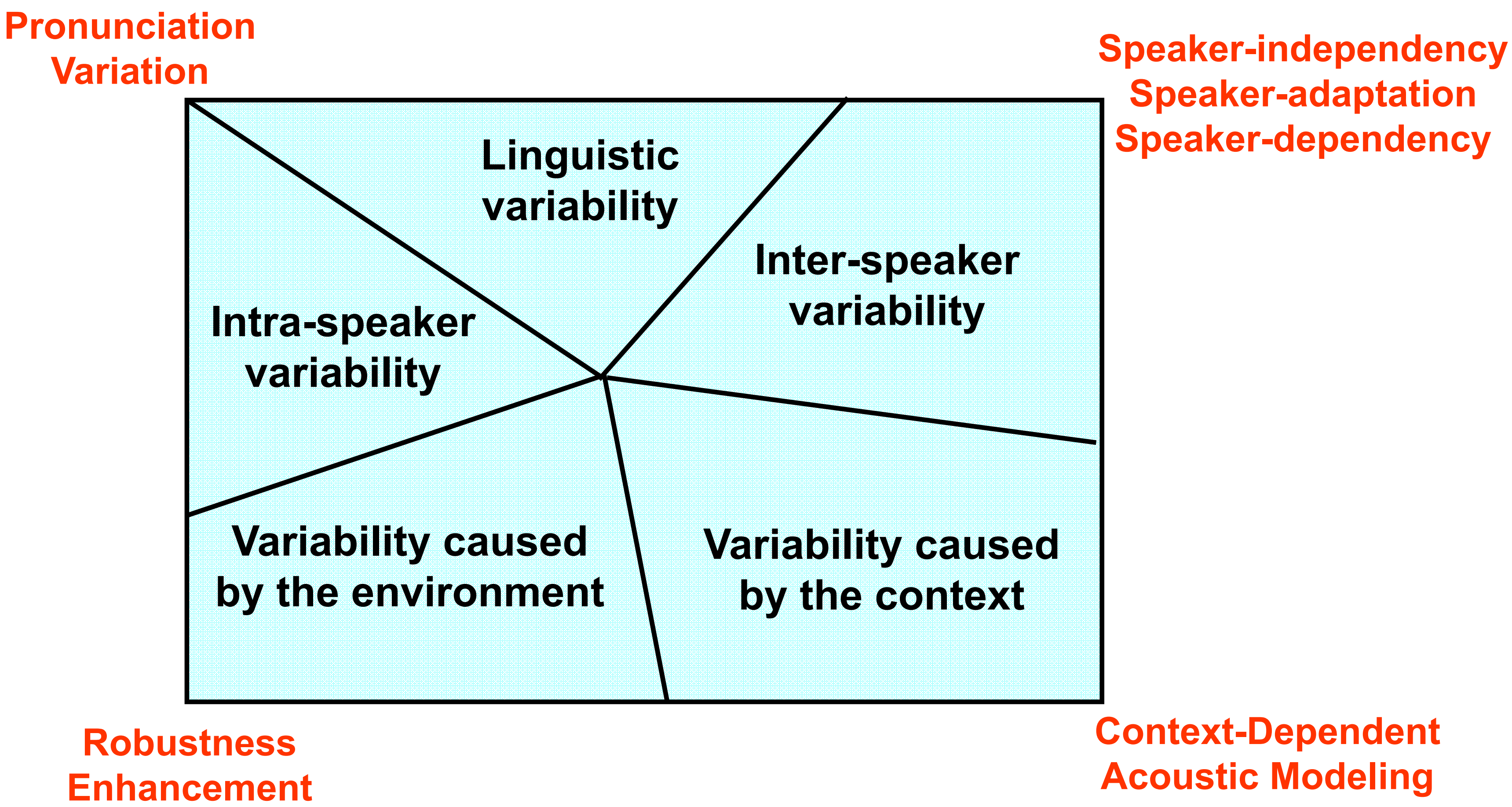
- The *n*-gram language model that follows a statistical modeling paradigm is the most prominently-used in ASR

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

bigram modeling

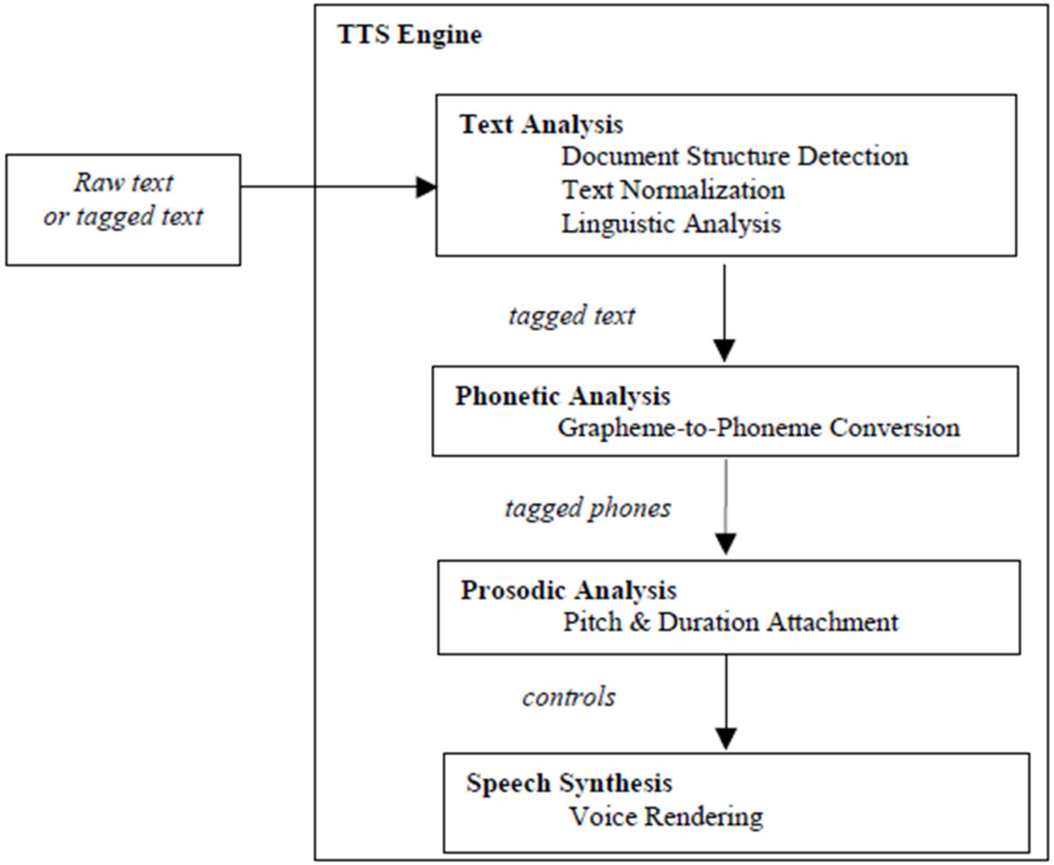
$$P(w_1 w_2 \ldots . w_L) = P(w_1) P(w_2|w_1) P(w_3|w_2) \cdots P(w_L|w_{L-1})$$

Difficulties: Speech Variability



Text to Speech (TTS)

- TTS can be viewed as ASR in reverse

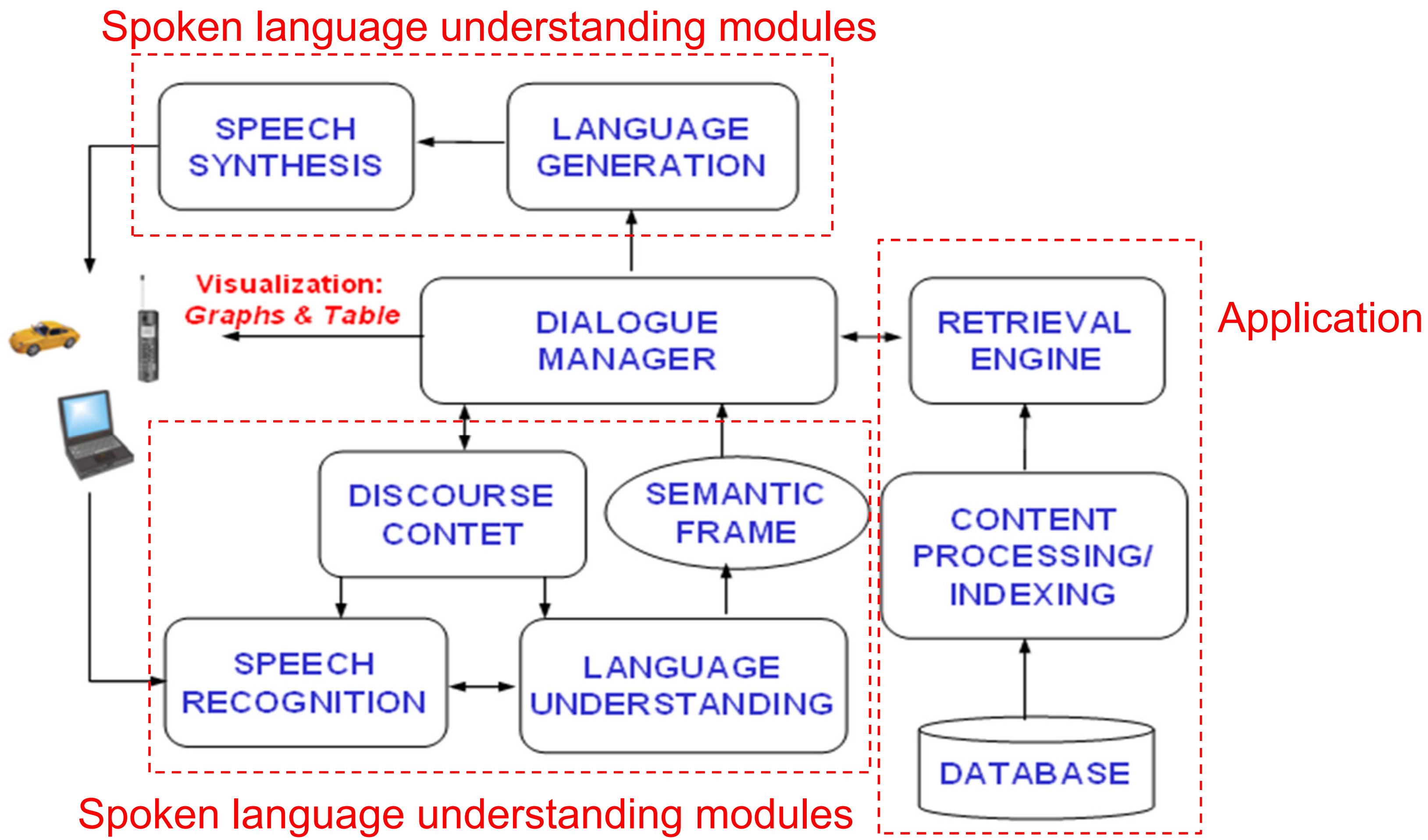


  – We are now able to general high-quality TTS systems, although the quality is inferior to human speech for general-purpose applications

Spoken Dialogue: CMU's Systems

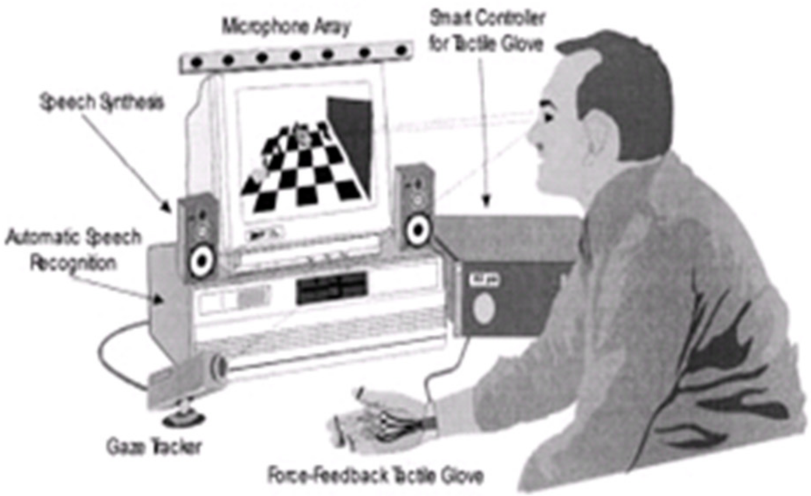
- Spoken language is attractive because it is the most natural, convenient and inexpensive means of exchanging information for humans
- In mobilizing situations, using keystrokes and mouse clicks could be impractical for rapid information access through small handheld devices like PDAs, cellular phones,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

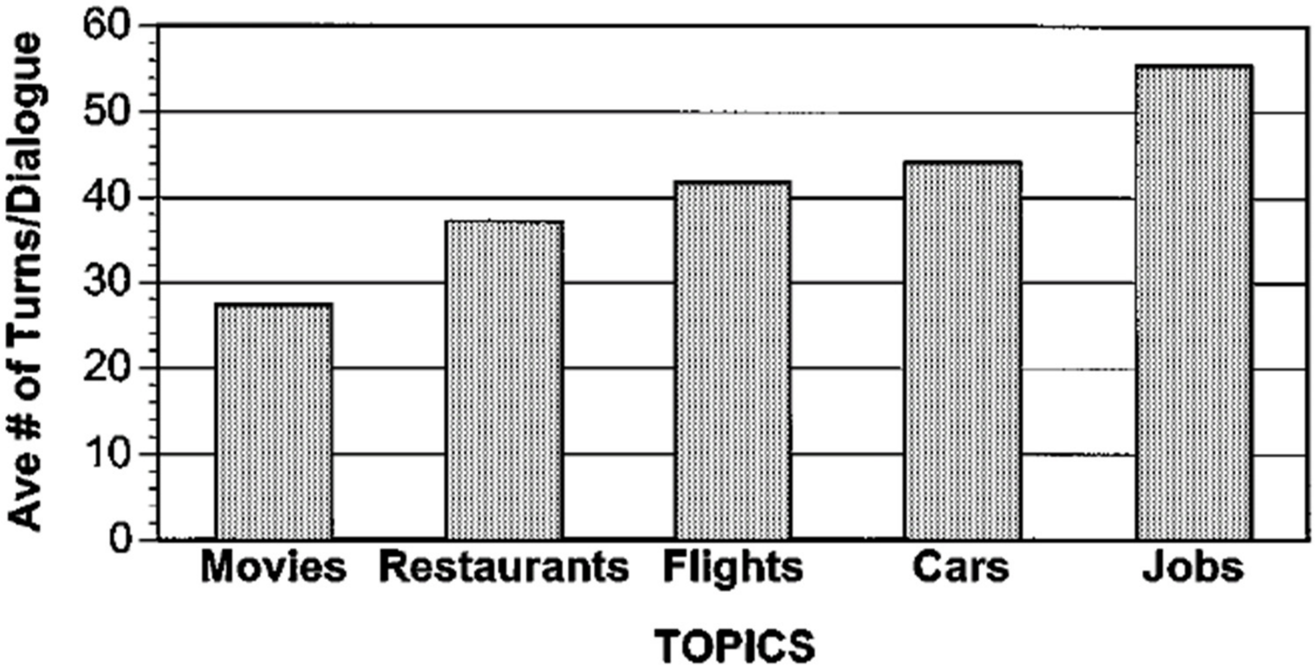
Spoken Dialogue: Some Deployed Systems

- 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

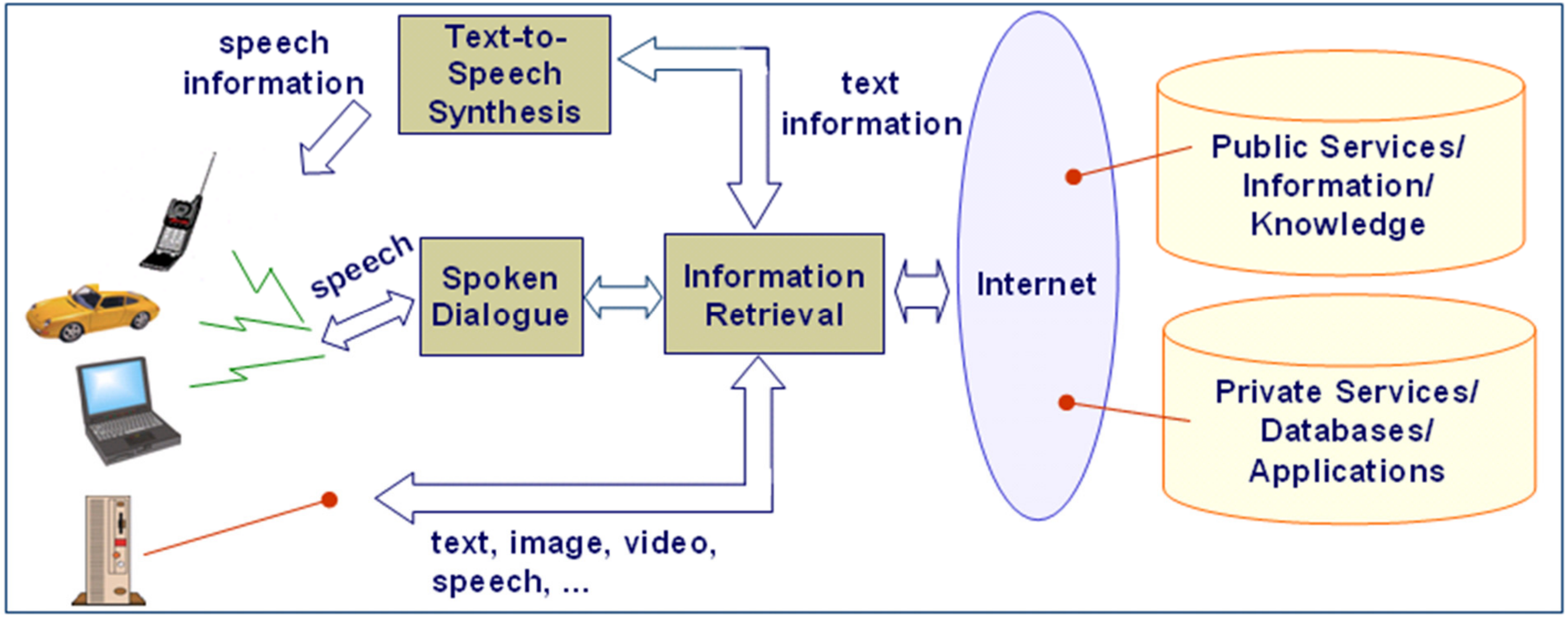
Spoken Dialogue: Some Statistics

- Topics vs. Dialogue Terms

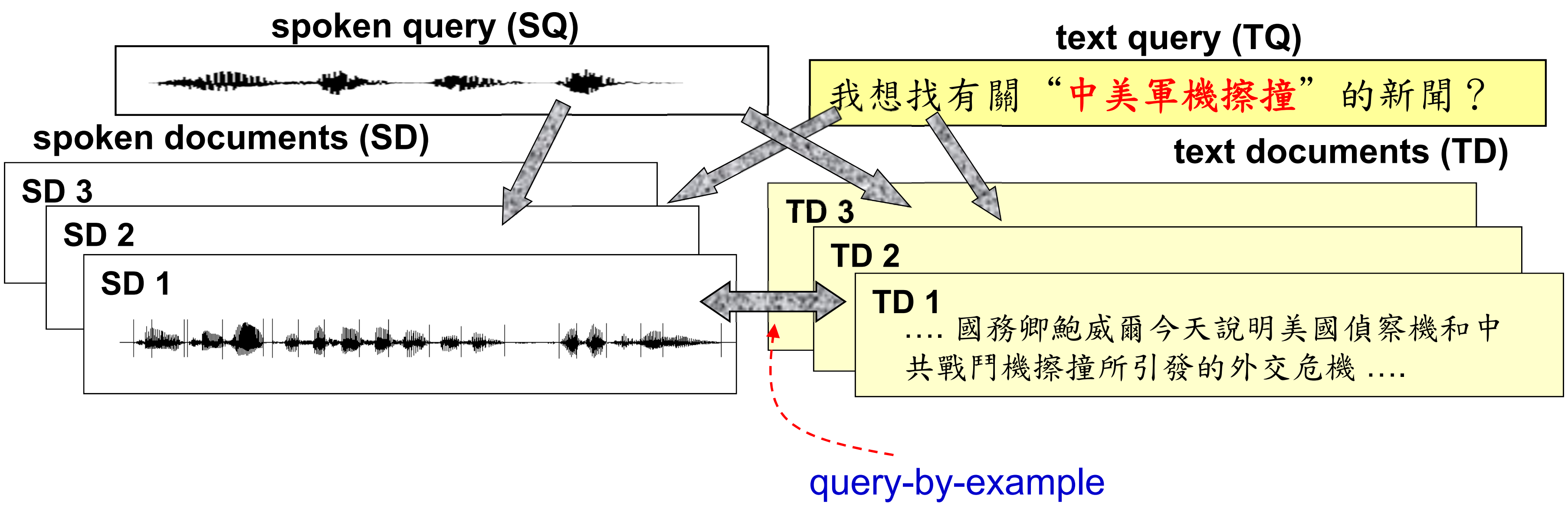


Speech-based Information Retrieval (1/5)

- 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)

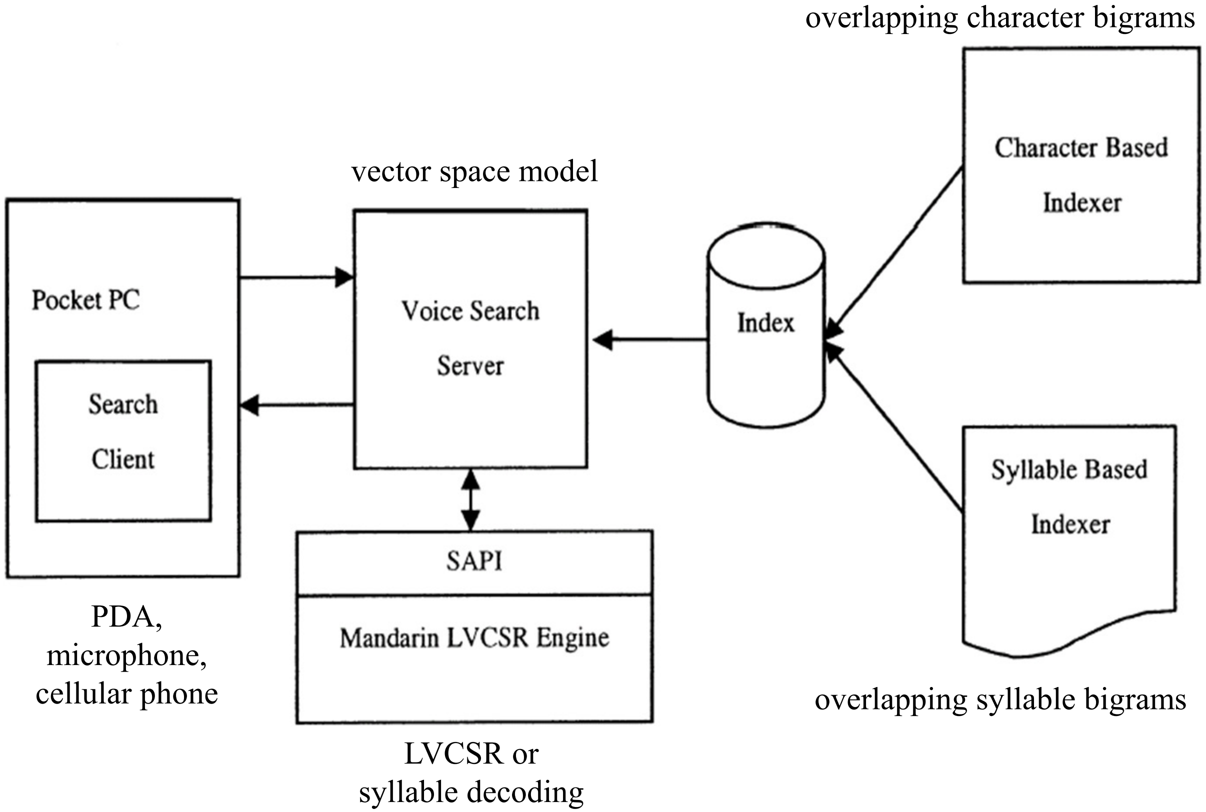
輸入聲音問句：“請幫我查總統府升旗典禮”

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)

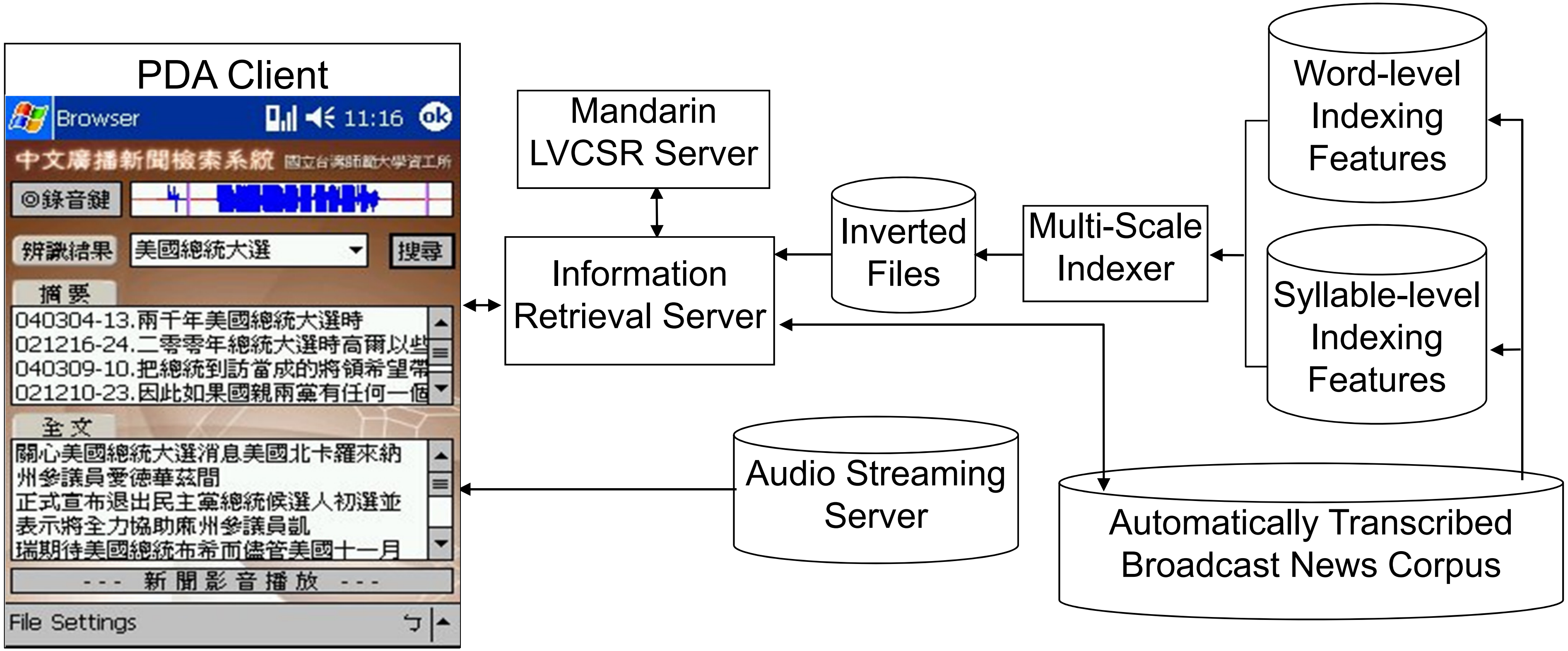
- Spoken queries retrieving text news documents via mobile devices



C.f. 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)

- Spoken queries retrieving text news documents via mobile devices

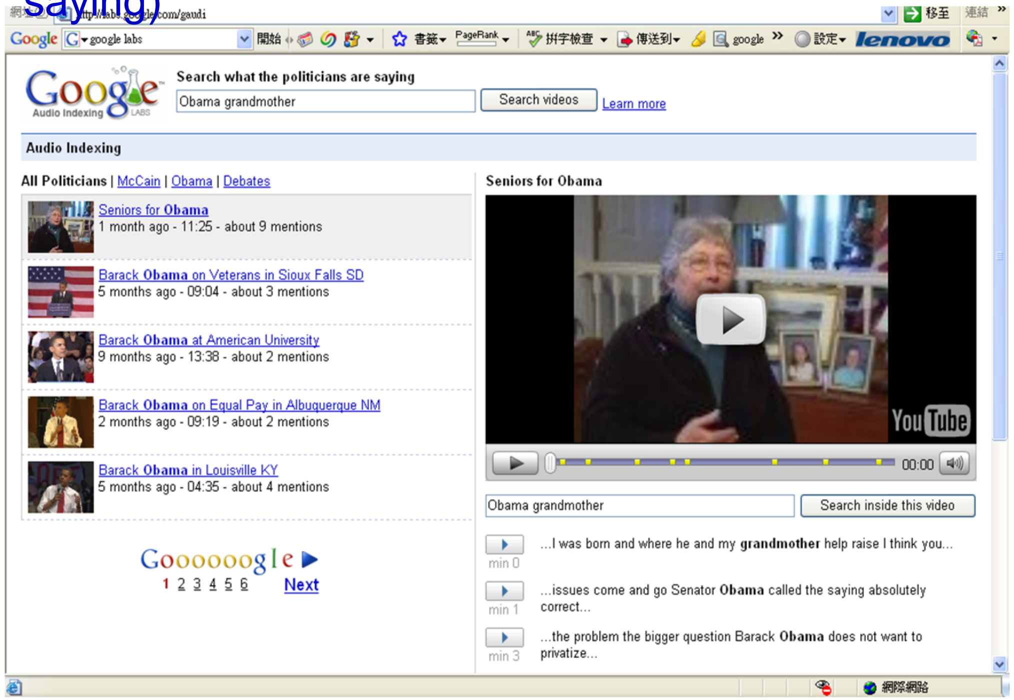


C.f. B. Chen, Y..T. Chen, C.H. Chang, H.B. Chen, "Speech Retrieval of Mandarin Broadcast News via Mobile Devices," Interspeech2005

Spoken Dialogue: Google Voice Search

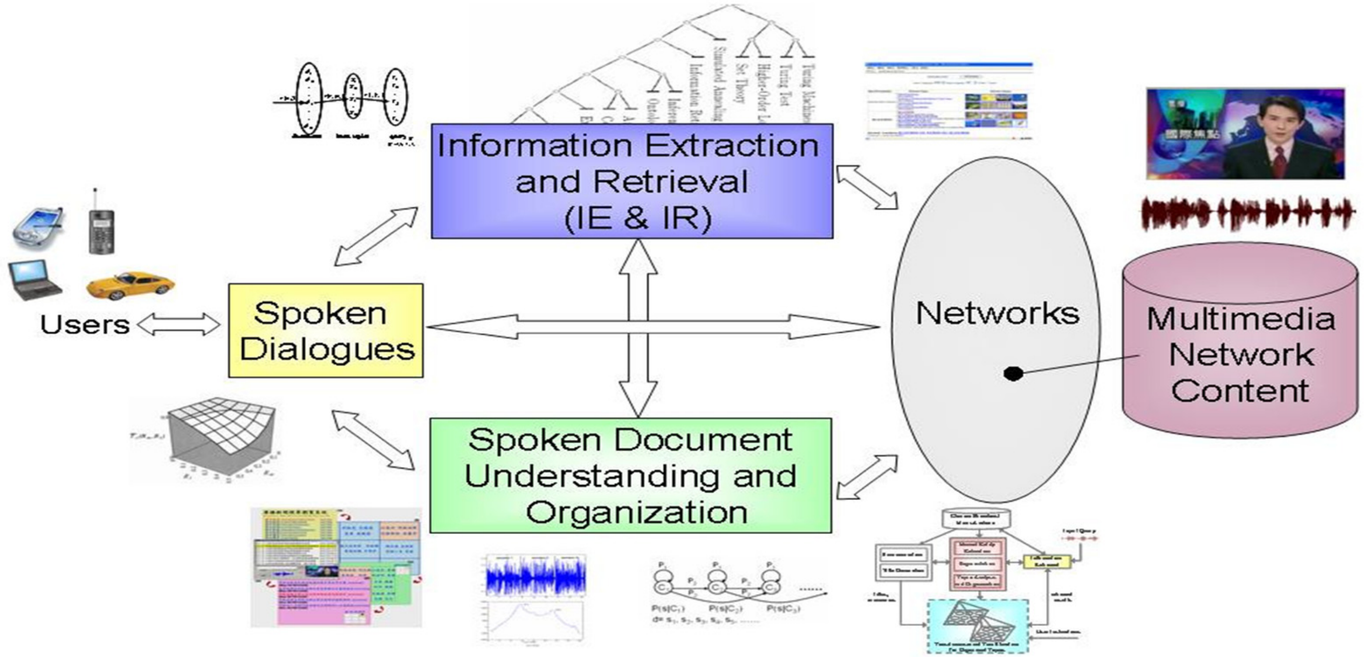
Google-411:
Finding and connecting to local business

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



Spoken Document Organization and Understanding (1/2)

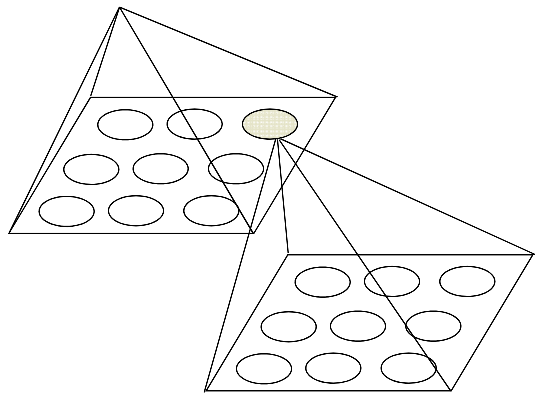
- Problems
  - The content of multimedia documents very often described by the associated speech information
  - Unlike text documents with paragraphs/titles easy to look through at a glance, multimedia/spoken documents are unstructured and difficult to retrieve/browse



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)

- For example, spoken documents can be clustered by the latent topics and organized in a two-dimensional tree structure, or a two-layer map



Two-dimensional Tree Structure for Organized Topics

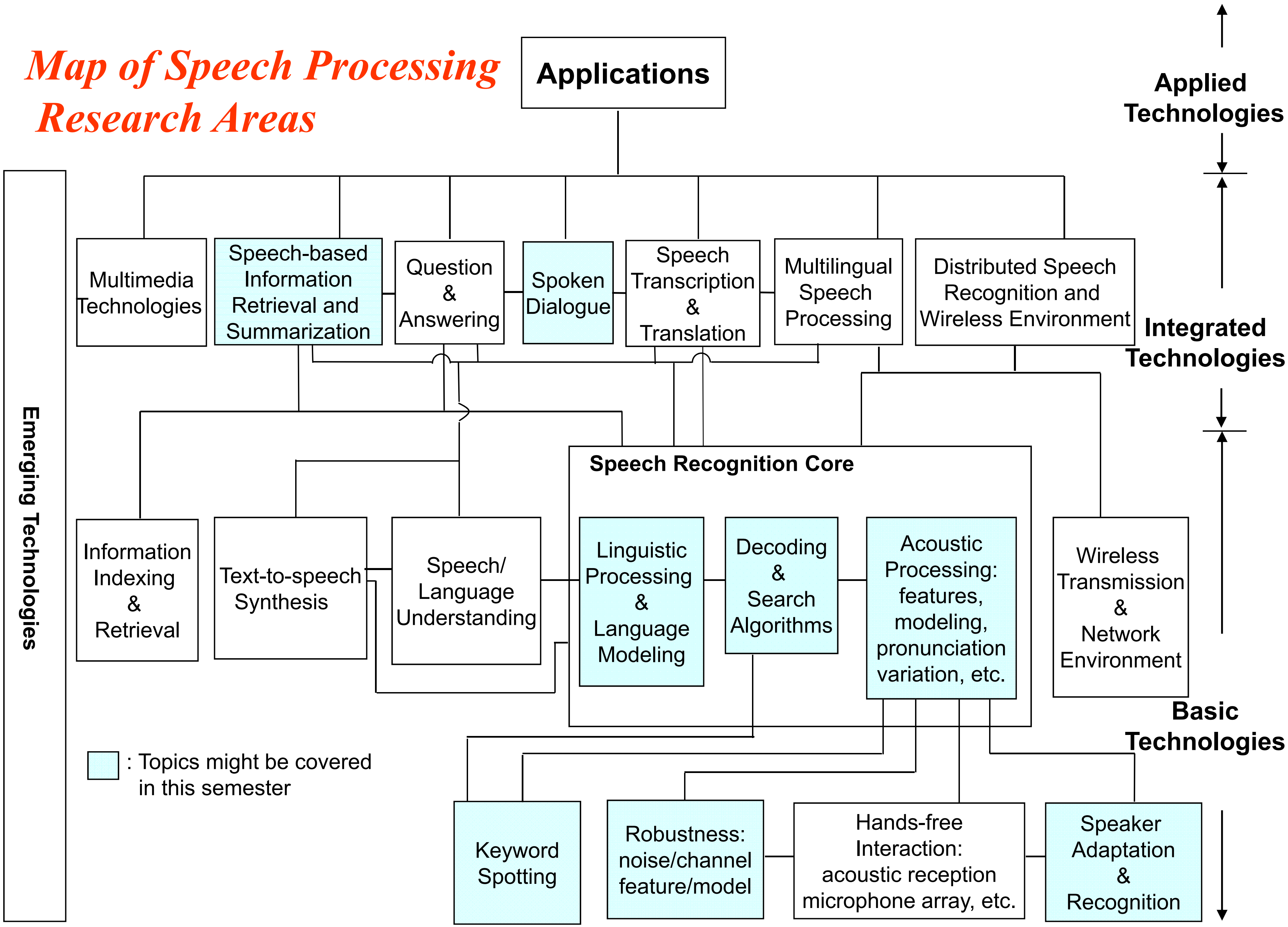
Speech-to-Speech Translation

- 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

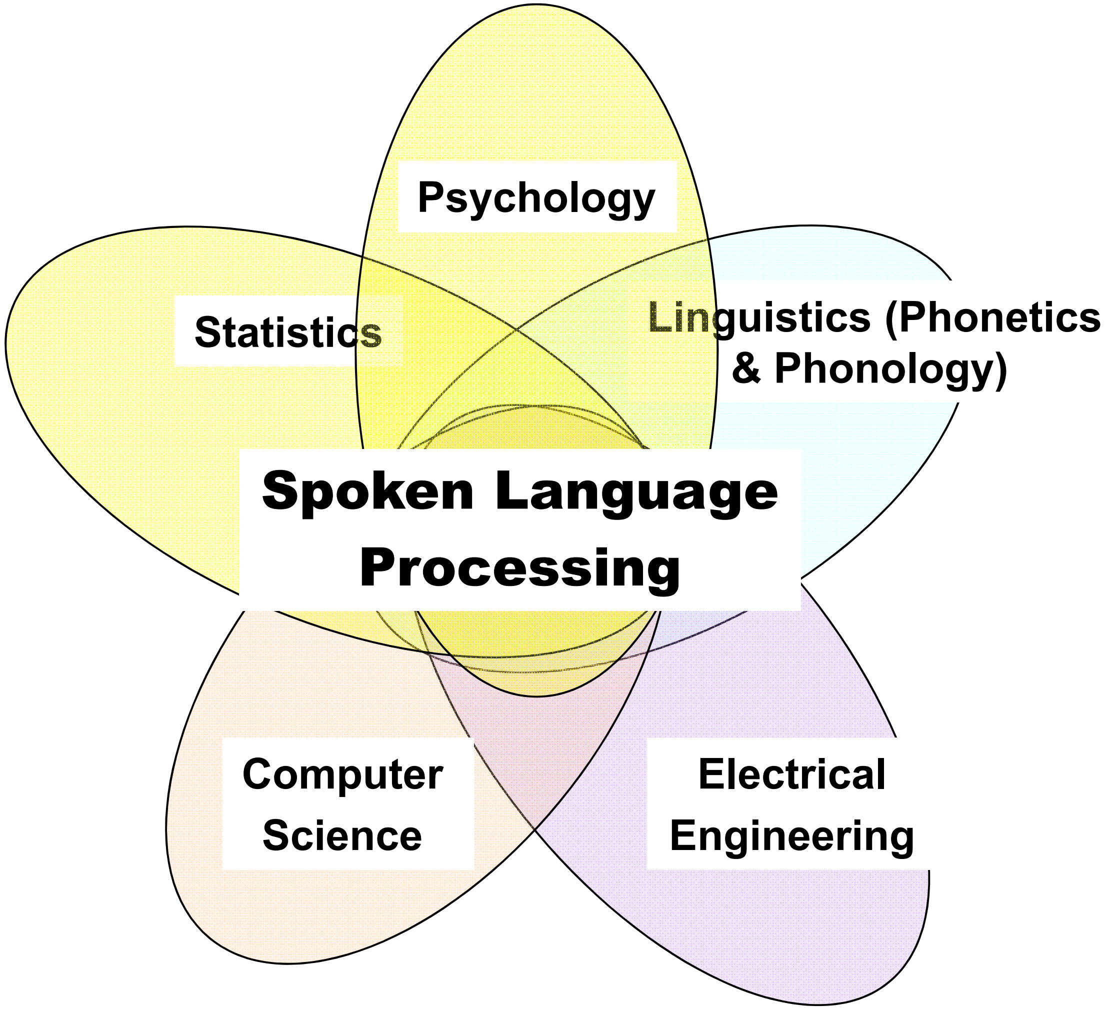
Map of Speech Processing Research Areas



Adapted from Prof. Lin-shan Lee

Different Academic Disciplines

- The foundations of spoken language processing lies in

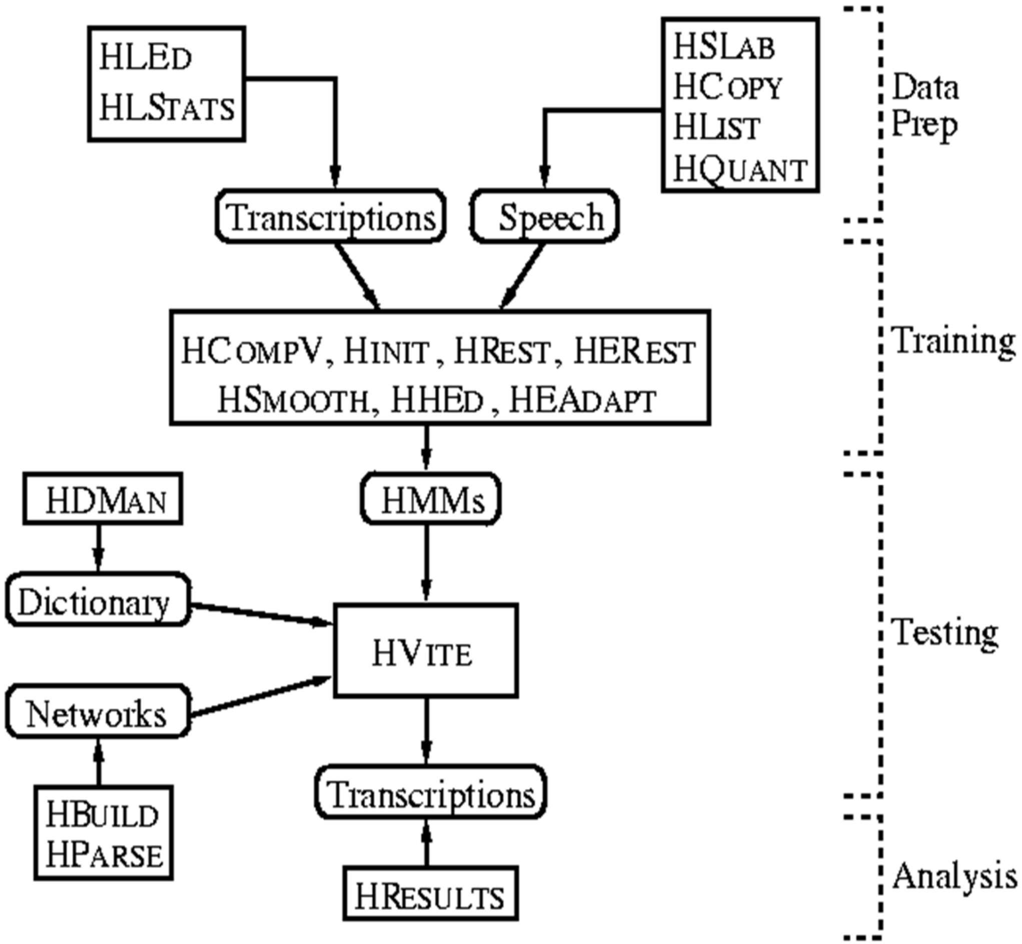


Speech Processing Toolkit (1/2)

- 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)

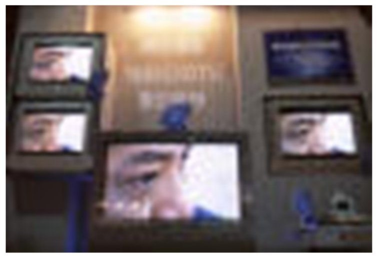


Journals & Conferences

- 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 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
- …..



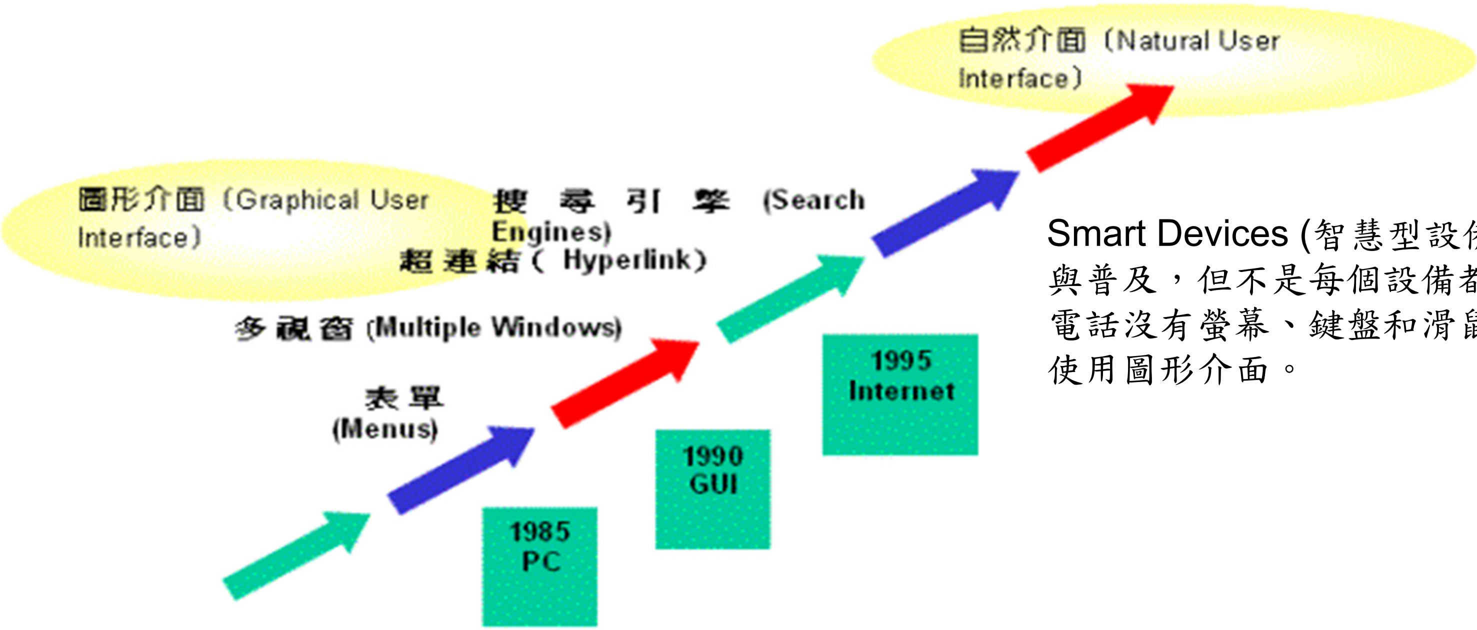
Tentative Schedule

Topics to be Covered
Overview & Introduction
Hidden Markov Models
Spoken Language Structure
Acoustic Modeling & HTK Toolkit
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

Speech Industry (2/3)

- Microsoft: Smart Device/Natural UI

使用介面的發展



Source：微軟自然互動服務產品部門 (NISD)副總裁李開複博士講稿， 2003/04

Smart Devices (智慧型設備)日益繁多與普及，但不是每個設備都有螢幕，例如：電話沒有螢幕、鍵盤和滑鼠，就無法使用圖形介面。

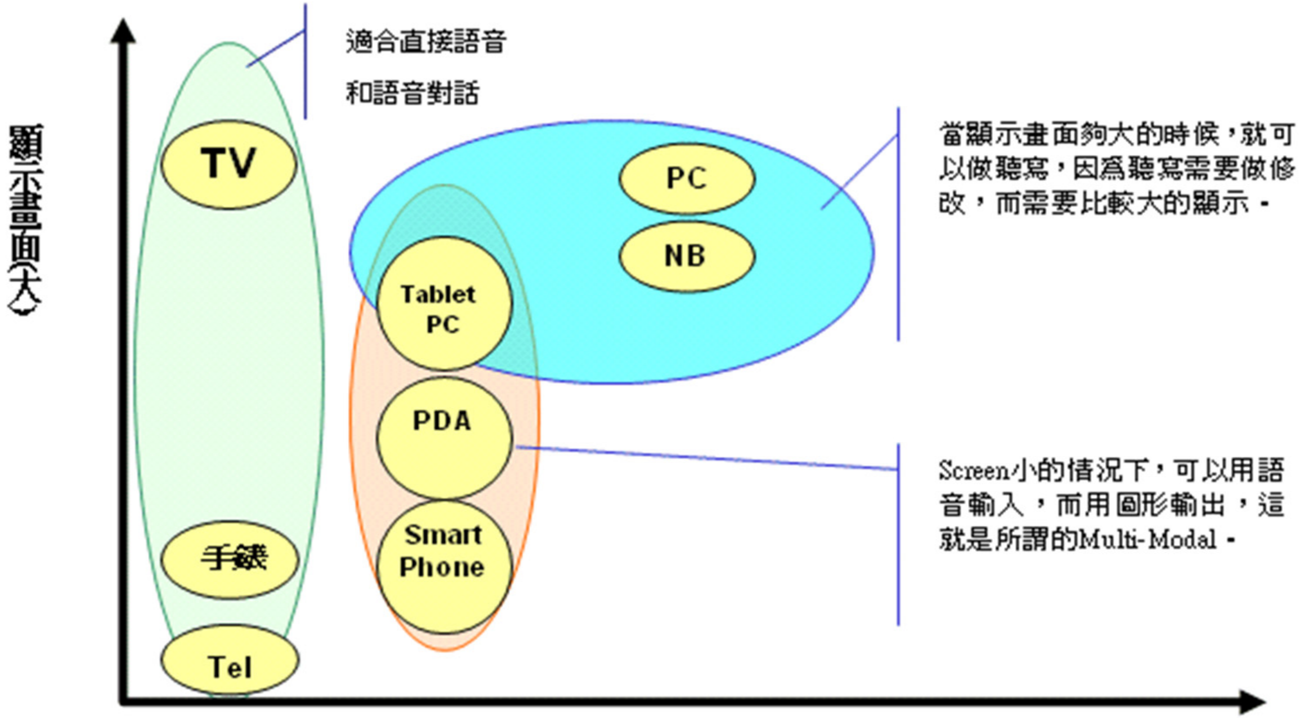
.NET 的最初構想，以符合人類需求的自然介面，其包括 –

- 語音合成
- 語音辨識技術
- 結合XML為基礎的網路服務

Speech Industry (3/3)

- Microsoft: Smart Device/Natural UI

Smart Device 與語音使用需求的關係



手寫	X	O	O
滑鼠	X	X	O
鍵盤	X	X	O

Good Words

- Your attitude determines your altitude.
- Stay Hungry; Stay Foolish
- Every job is a self-portrait of those who did it. Autograph your work with quality.
- ….