音訊與語音辨識

Berlin Chen, 陳柏琳

berlin@csie.ntnu.edu.tw
http://berlin.csie.ntnu.edu.tw

About the Instructor

- Berlin Chen, 陳柏琳
  - Education:
    - Ph.D. Computer Science and Information Engineering
      National Taiwan University, Sept 1998 - May 2001
  - Professional Experiences
    - Aug 2002 ~ Assistant Professor,
      Graduate Institute of Computer Science
      and Information Engineering,
      National Taiwan Normal University
    - Dec 2000- July 2002 Postdoctoral Researcher,
      Graduate Institute of Communication Engineering,
      National Taiwan University
    - Oct 1996 - Nov 2001 Research Assistant,
      Institute of Information Science,
      Academia Sinica

About the Instructor (cont.)

- Research Interests
  - **Speech Signal Processing**
    - Large Vocabulary Continuous Speech Recognition
    - Discriminative Acoustic Feature Extraction
    - Supervised/Unsupervised Acoustic Modeling and Language Modeling
    - Utterance Verification and Confidence Measure
    - Speaker Adaptation
    - Spoken Dialogue Systems
  - **Information Retrieval**
    - Retrieval Modeling
    - Query/Document Representation, Robust Audio Indexing
    - Speech-based Multimedia Information Retrieval Systems
    - Keyword/Topic-word Extraction
  - **Natural Language Processing**
    - Part-of-Speech Tagging, Syntactic/Semantic Parsing
    - Speech Summarization using Heterogeneous Information Sources
    - Automatic Title Words Generation
  - **Artificial Intelligence and Neural Networks**
    - Search Algorithms/Machine Learning Techniques

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
  – 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.

Tentative Schedule

Date	Tentative Topic List
7/6	Introduction & Spoken Language Structure
7/13	Hidden Markov Models
7/20	Statistical Language Modeling
7/27	Search Algorithms (Digit Recognition、Word Recognition、Keyword Spotting、LVCSR)
8/3	Speech Signal Processing & Acoustic Modeling
8/10	Speech Enhancement & Robustness
8/17	Language and Acoustic Model Adaptation
8/24	Speech Information Retrieval & Spoken Dialogues
8/31	Tagging and Parsing of Natural Languages
9/1	Speaker Recognition & Speech Synthesis

Textbook and References

- Textbook
  - X. Huang, A. Acero, H. Hon. Spoken Language Processing, Prentice Hall, 2001
  - C. Manning and H. Schutze. Foundations of Statistical Natural Language Processing. MIT Press, 1999

- References books
  - 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
  - S. Young et al.. The HTK Book. Version 3.0, 2000 "http://htk.eng.cam.ac.uk"
  - 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

Textbook and References (cont.)

- Reference papers
  - Lawrence Rabiner. The Power of Speech. Science, Vol. 301, pp. 1494-1495, Sep. 2003
  - 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
  - ….

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, Member, A. Medl, And J. Flanagan, “Natural Communication with Information 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

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

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

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

SRI

BBN Technologies

Speech at CMU

LIMSI

MIT SLS

Cambridge HTK

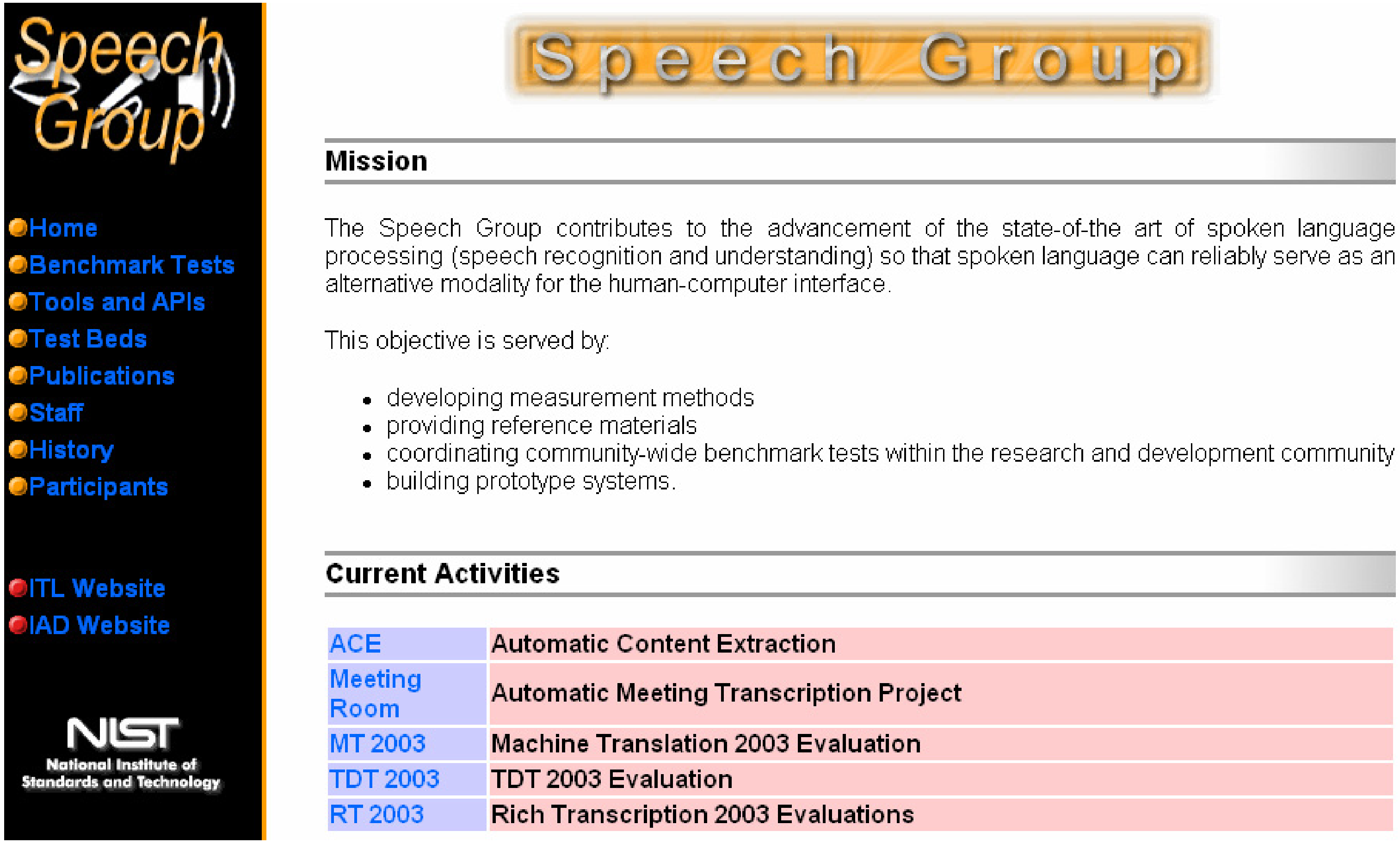
Philips

JHU CLSP

Microsoft

Progress of Technology

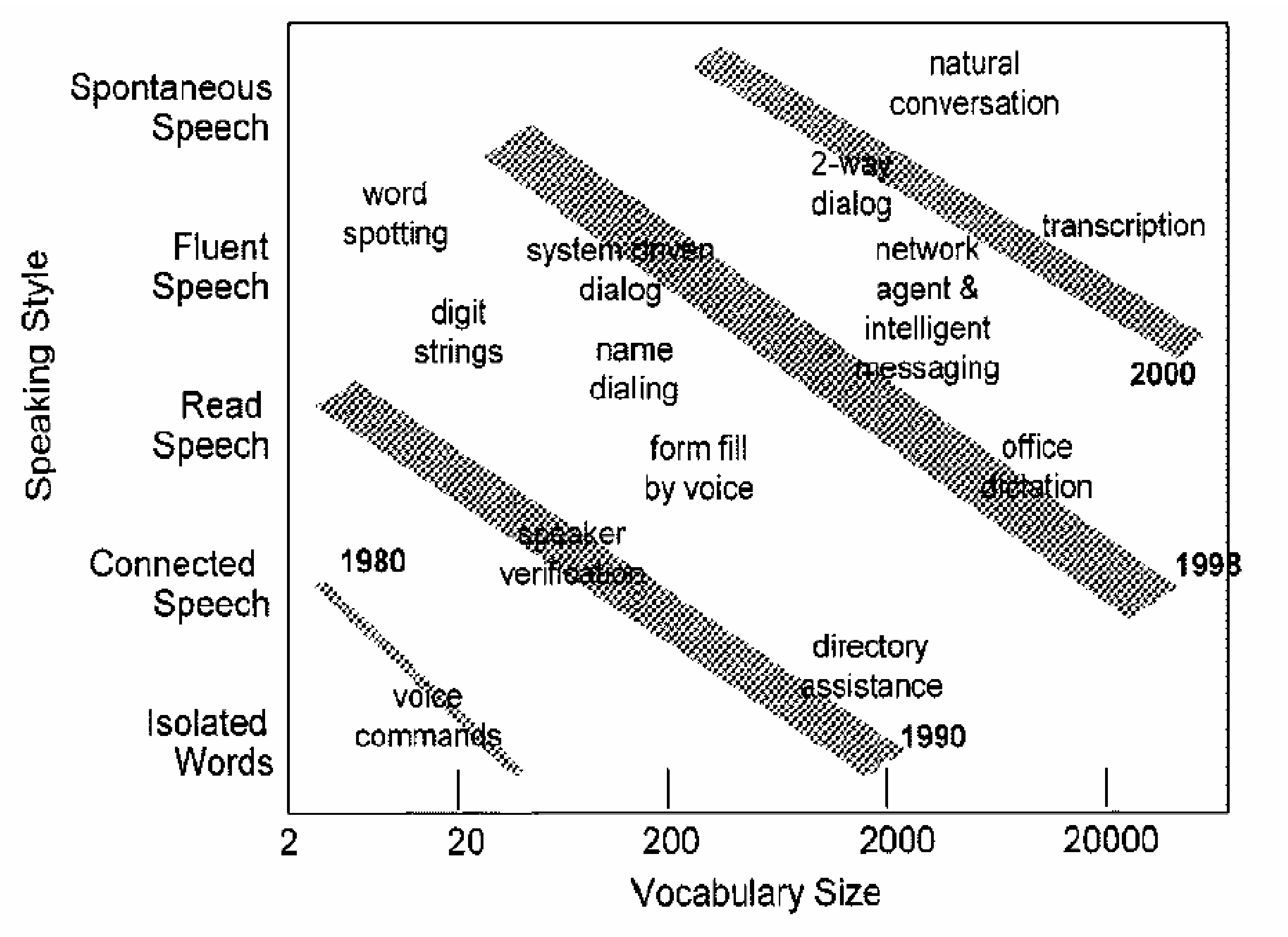
- US. National Institute of Standards and Technology (NIST)



http://www.nist.gov/speech/

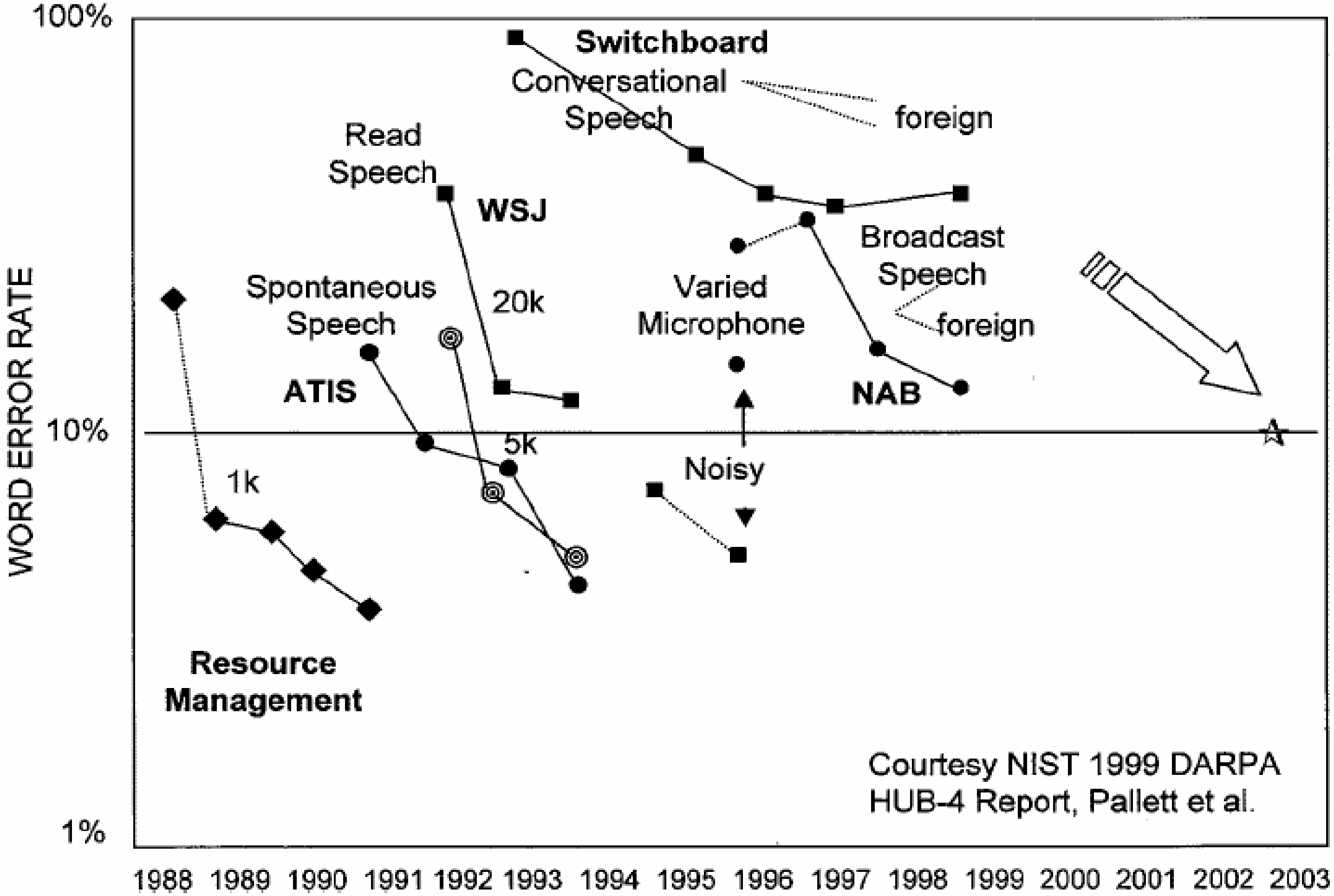
Progress of Technology (cont.)

- Generic Application Areas (vocabulary vs. speaking style)



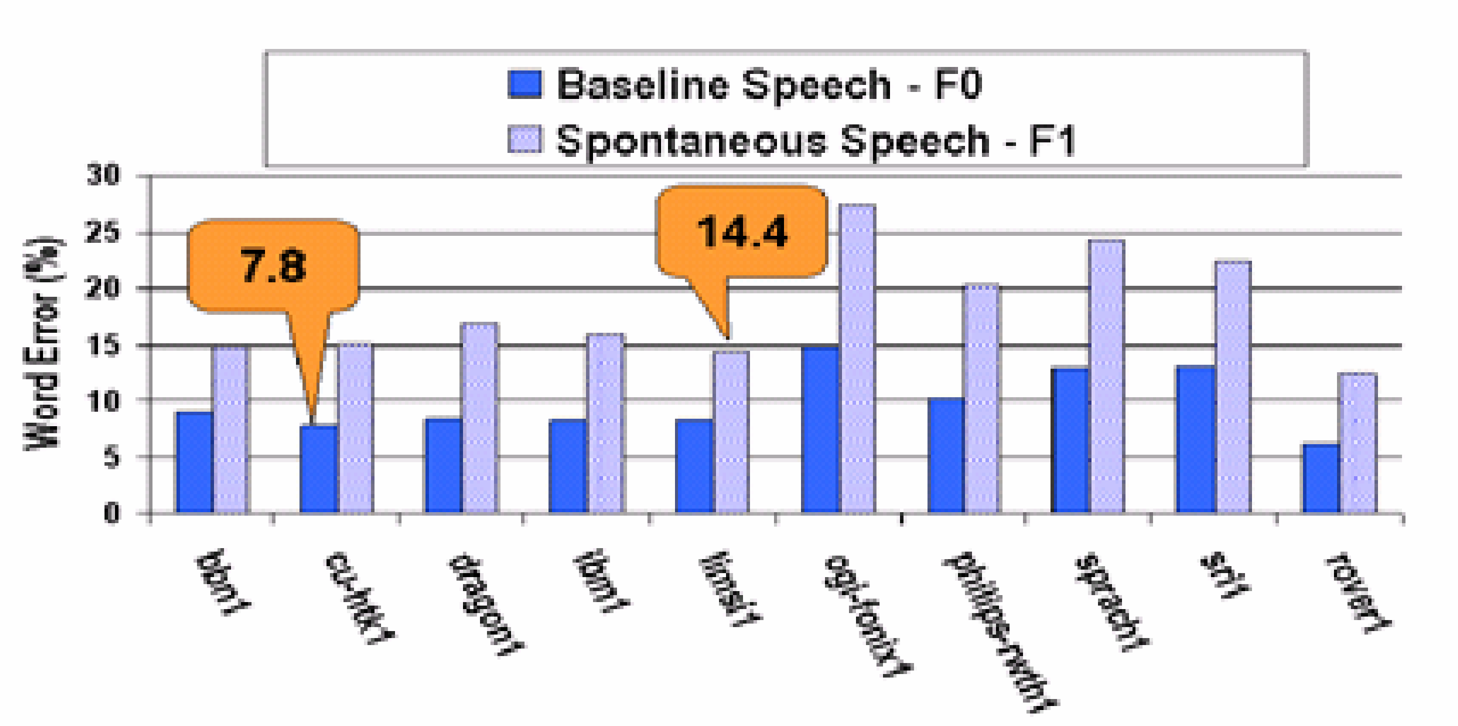
Progress of Technology (cont.)

- Benchmarks of ASR performance: Overview



Progress of Technology (cont.)

- Benchmarks of ASR performance: Broadcast News Speech



Progress of Technology (cont.)

- Benchmarks of ASR performance: Conversational Speech

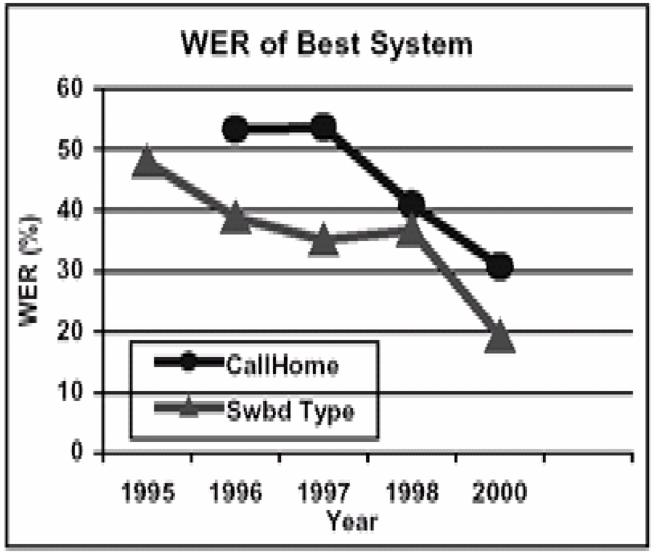


Figure 4 History of lowest word error rates (WER) obtained in NIST conversational speech evaluations on Switchboad and CallHome type conversations in English [26].

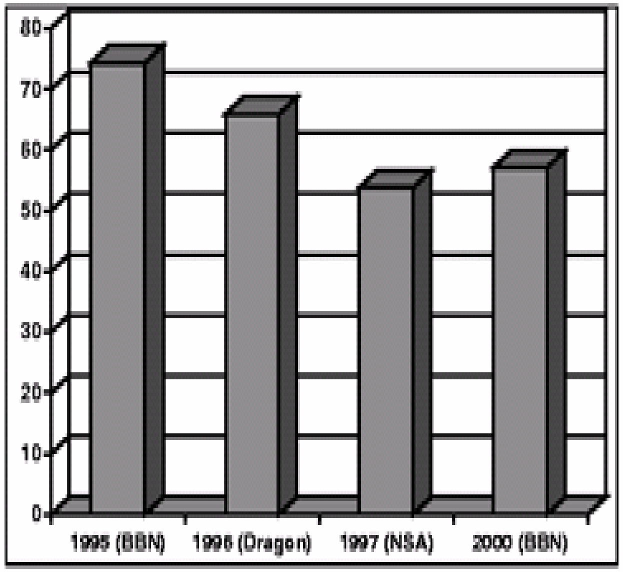


Figure 5 Chinese Character error rates of the best performing evaluation system in NIST Mandarin conversational speech evaluations 1995-2000 [26].

Progress of Technology (cont.)

- 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

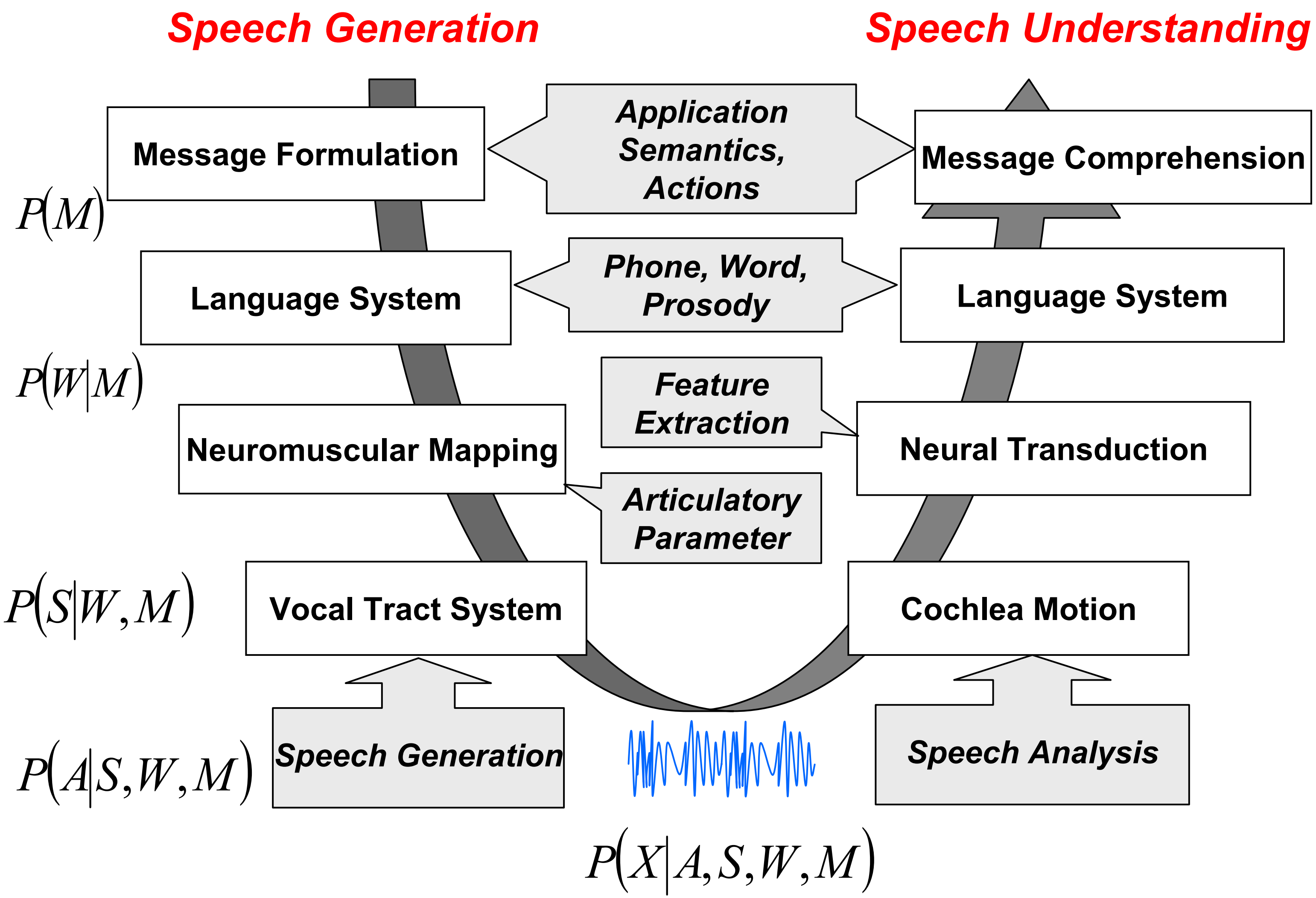
%CER on dev02 and eval03 for all stages of 2003 system

  - Adopted from

Cambridge University
Engineering Department

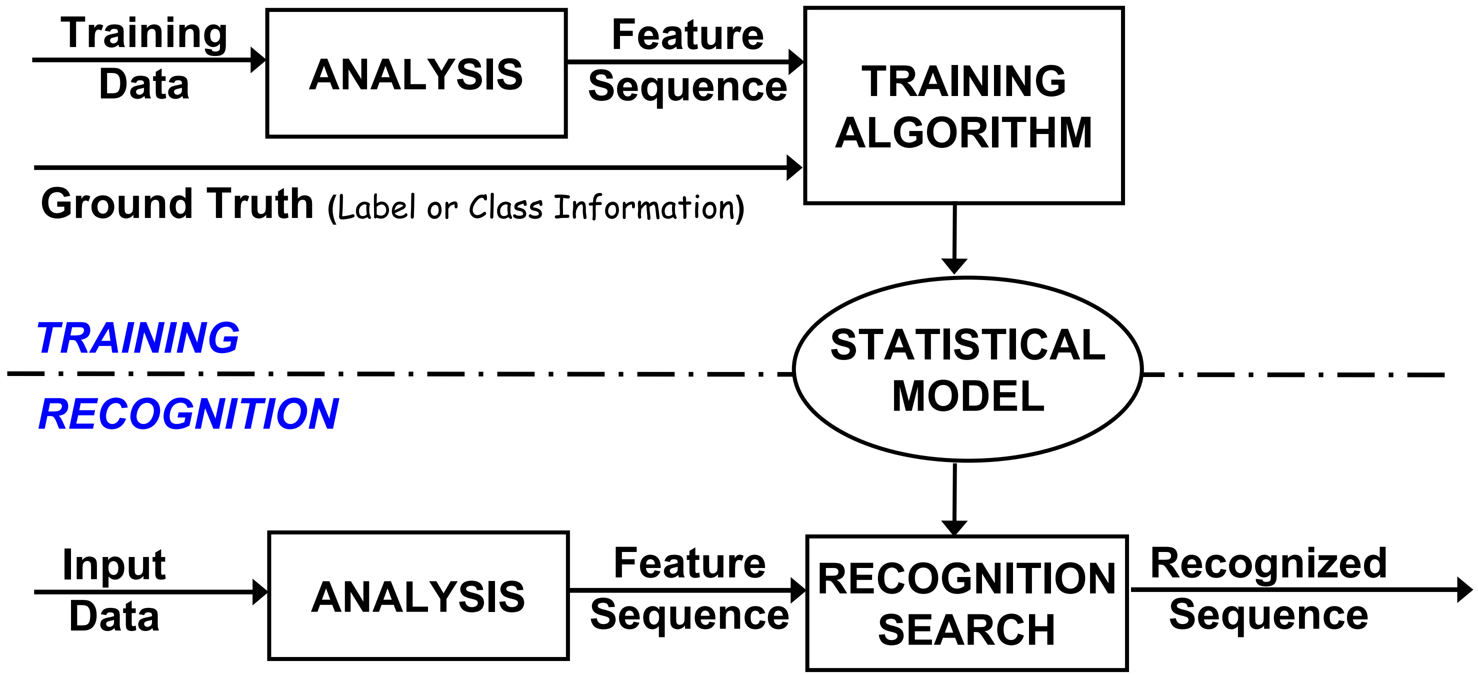
Rich Transcription Workshop 2003

Determinants of Speech Communication



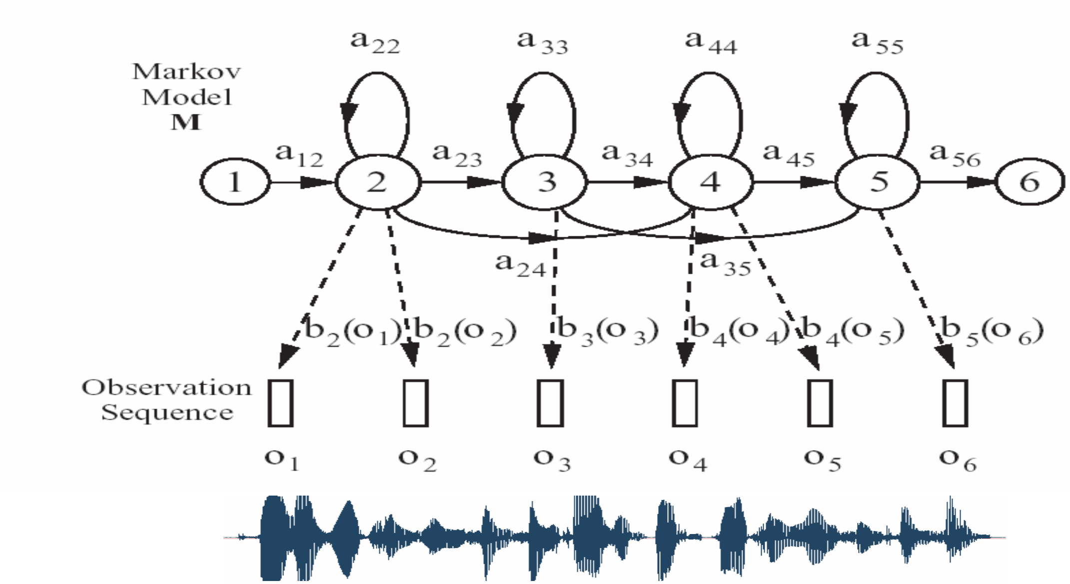
Statistical Modeling Paradigm

- The statistical modeling paradigm used in speech and language processing

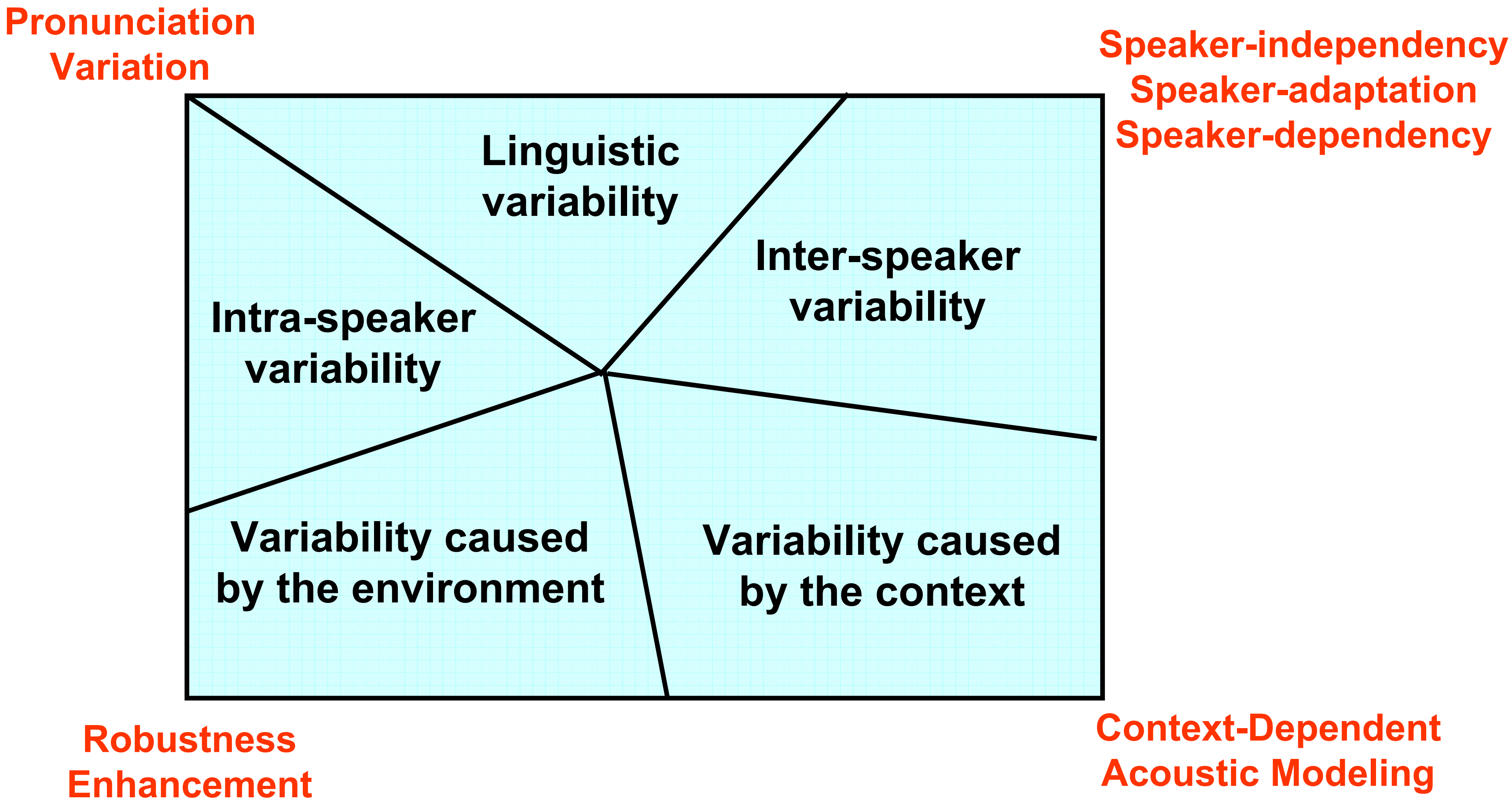


Statistical Modeling Paradigm

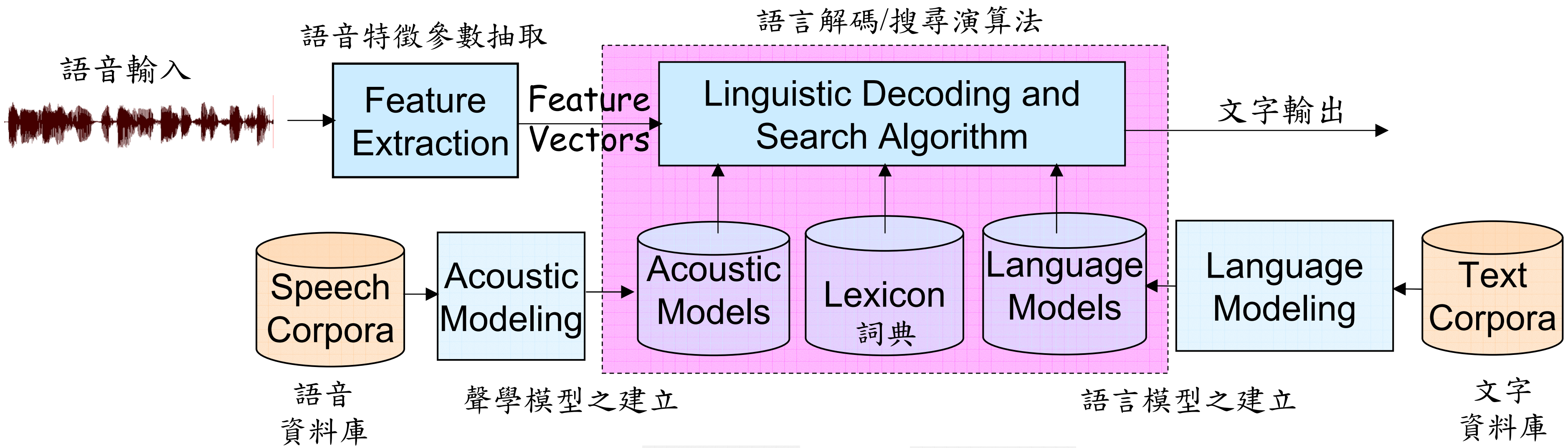
- Approaches based on Hidden Markov Models (HMMs) dominate the area of speech recognition
  - HMMs are based on rigorous mathematical theory built on several decades of mathematical results developed in other fields
  - HMMs are generated by the process of training on a large corpus of real speech data



Difficulties: Speech Variability



Large Vocabulary Continuous Speech Recognition (LVCSR)



可能詞句　　語音輸入

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

貝氏定理

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

詞彙網路搜尋

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

聲學模型機率　　語言模型機率

Large Vocabulary Continuous Speech Recognition (cont.)

- Transcription of Broadcast News Speech

0	SIL	71695	-1	35	1280.422	1.00000	1.00000
1	行政院	55302	35	80	720.973	1.00000	0.75715
2	秘書長	50877	80	118	459.867	0.56604	0.18618
3	劉	2406	118	137	371.101	0.26549	0.50987
4	世	6603	137	157	610.122	1.00000	1.00000
5	芳	1111	157	177	545.281	0.22222	1.00000
6	和	3407	177	196	374.724	0.15385	0.00000
7	蒙藏	66970	196	237	844.522	1.00000	0.53602
8	委員會	58282	237	281	776.631	1.00000	1.00000
9	委員長	58283	281	332	955.699	1.00000	0.83401
10	徐	5422	332	356	561.555	0.36598	0.54206
11	志	5919	356	372	420.553	0.40000	0.54860
12	修	5075	372	416	988.773	0.31579	0.84565
13	上午	40289	416	449	681.523	1.00000	0.75001
14	到	1302	449	463	337.270	0.33333	1.00000
15	立法院	52750	463	509	1077.581	1.00000	0.85865
16	報告	9234	509	550	1061.472	1.00000	1.00000
17	預算	49933	550	587	738.046	1.00000	0.82290
18	編列	9691	587	616	576.571	1.00000	0.60458
19	情況	31054	616	666	1020.239	0.75000	0.81394
20	SIL	71695	666	703	1341.544	1.00000	1.00000
21	好幾	24960	703	729	326.342	0.00760	0.73112
22	位	8111	729	741	273.841	0.18748	1.00000
23	在野	42491	741	767	605.460	0.99551	1.00000
24	立委	21015	767	792	518.366	0.98152	0.75214
25	認為	41950	792	842	957.432	0.96371	0.57802

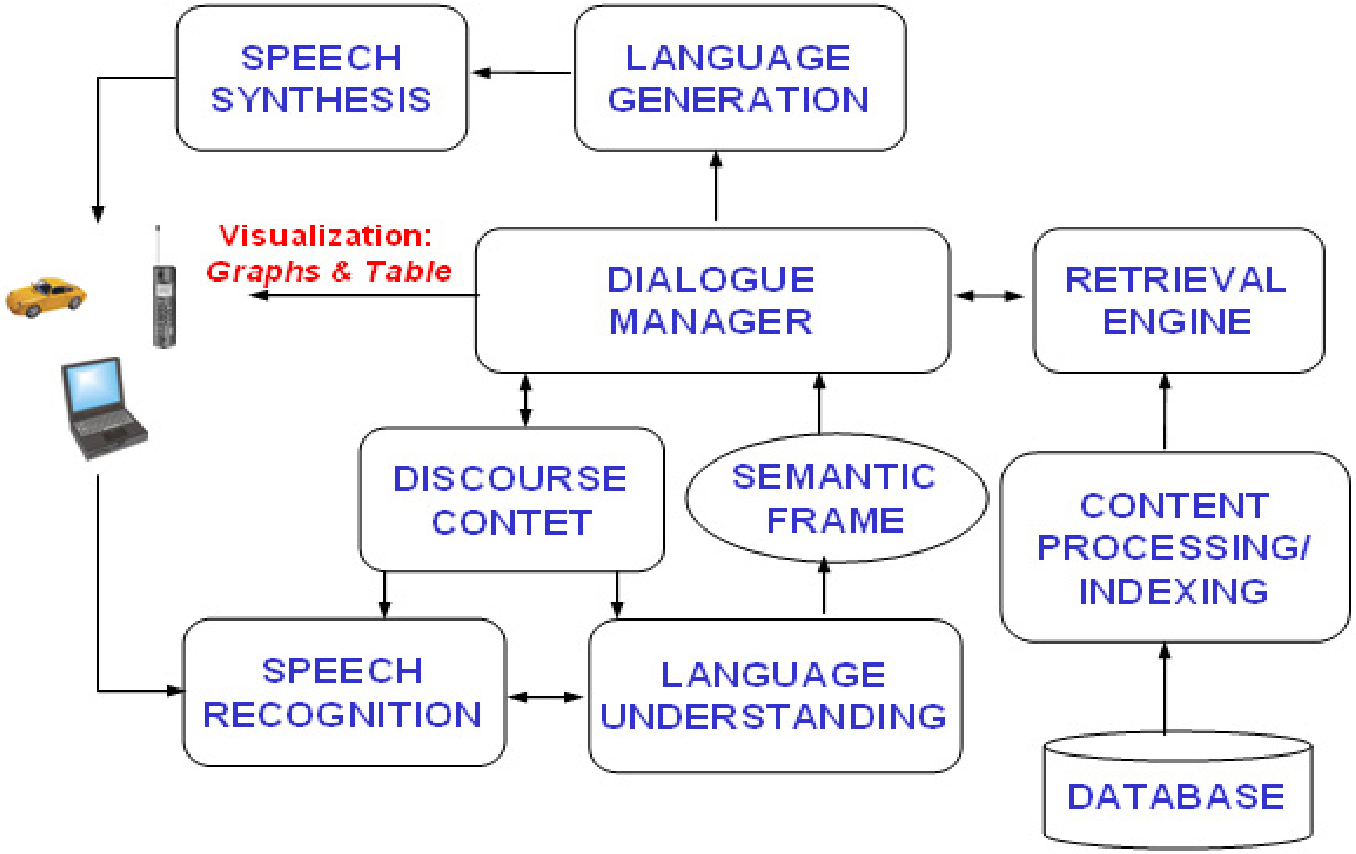
26	SIL	71695	842	872	1138.477	1.00000	1.00000
27	行政院	55302	872	934	1120.105	0.86107	0.87346
28	既然	29583	934	971	804.259	0.86107	0.95910
29	不	369	971	988	288.728	0.69917	1.00000
30	承認	38027	988	1043	931.888	0.46961	0.40323
31	外蒙	47896	1043	1084	786.448	1.00000	1.00000
32	為	8063	1084	1100	316.677	0.30057	1.00000
33	我國	47848	1100	1135	804.705	1.00000	1.00000
34	領土	20696	1135	1186	778.006	0.76186	0.96218
35	主張	36487	1186	1237	1003.320	0.07122	1.00000
36	全數	31649	1237	1304	1427.742	0.06937	1.00000
37	刪除	39728	1304	1349	818.702	1.00000	0.65401
38	蒙藏	66970	1349	1392	790.226	0.00928	0.51333
39	委員會	58282	1392	1432	870.207	1.00000	1.00000
40	的	1269	1432	1441	165.007	0.16667	1.00000
41	預算	49933	1441	1490	1304.056	0.23077	1.00000
42	SIL	71695	1490	1522	1101.760	1.00000	1.00000
43	從事	43981	1522	1566	1100.780	0.05556	0.76556
44	過	3023	1566	1580	279.248	0.07692	1.00000
45	院長	49392	1580	1613	632.123	0.10656	0.80456
46	許	3809	1613	1634	526.977	0.08333	1.00000
47	志	5919	1634	1650	222.692	0.05263	1.00000
48	雄	5420	1650	1685	762.830	0.33333	0.56287
49	也	7545	1685	1706	484.241	0.18462	1.00000
50	該	2847	1706	1721	403.345	0.18182	1.00000
51	下台	32060	1721	1781	1458.783	0.06522	1.00000
52	SIL	71695	1781	1843	2489.860	1.00000	1.00000

Spoken Dialogue

- 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.

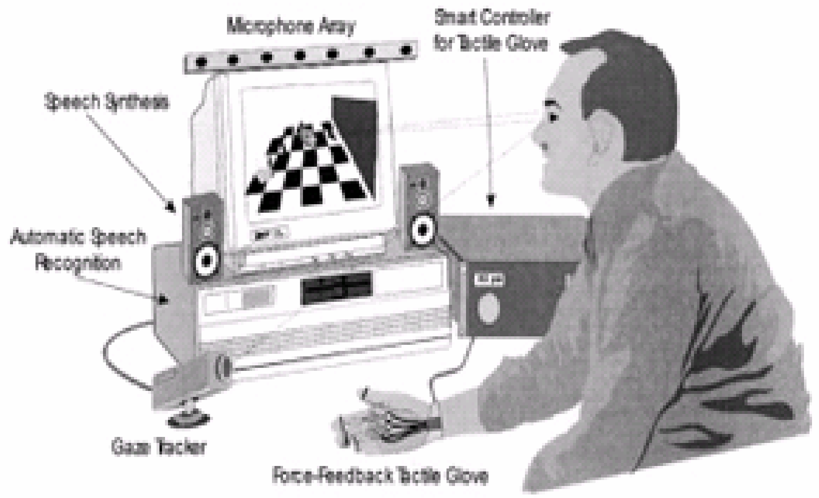
Spoken Dialogue (cont.)

- Flowchart



Spoken Dialogue (cont.)

- 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].

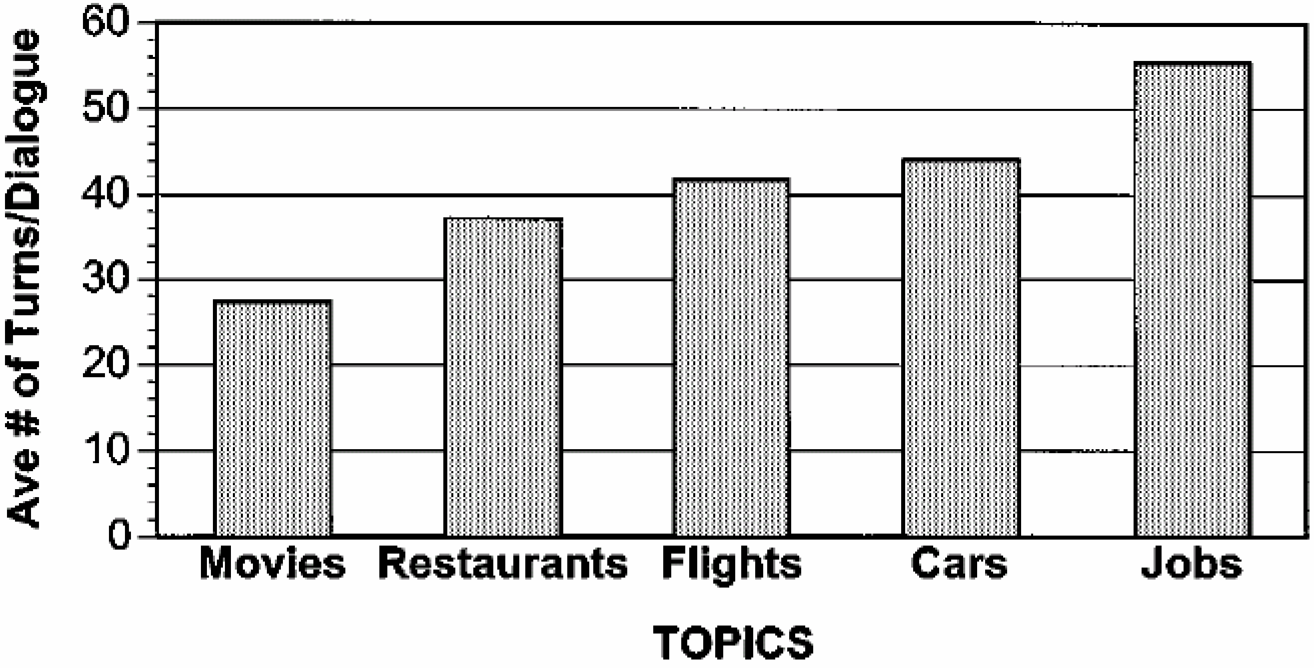
Spoken Dialogue (cont.)

- Deployed Dialogue Systems

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

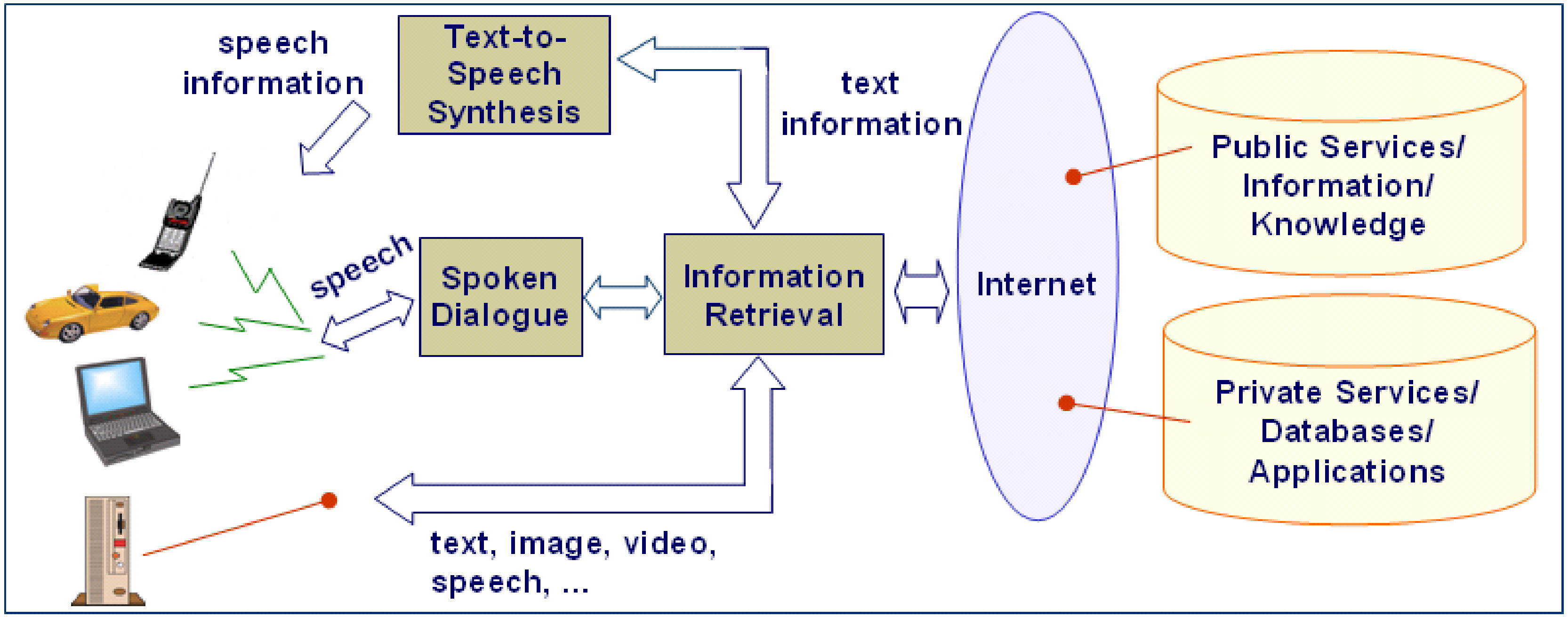
Spoken Dialogue (cont.)

- Topics vs. Dialogue Terms

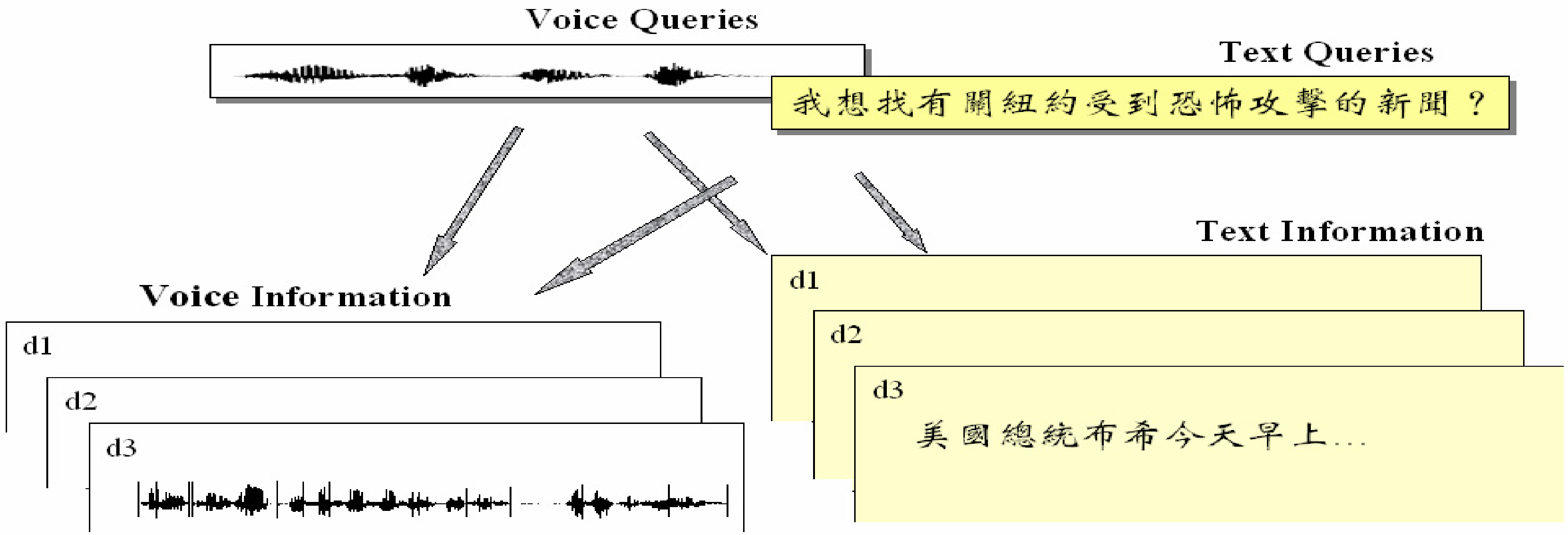


Speech-based Information Retrieval

- 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 (cont.)



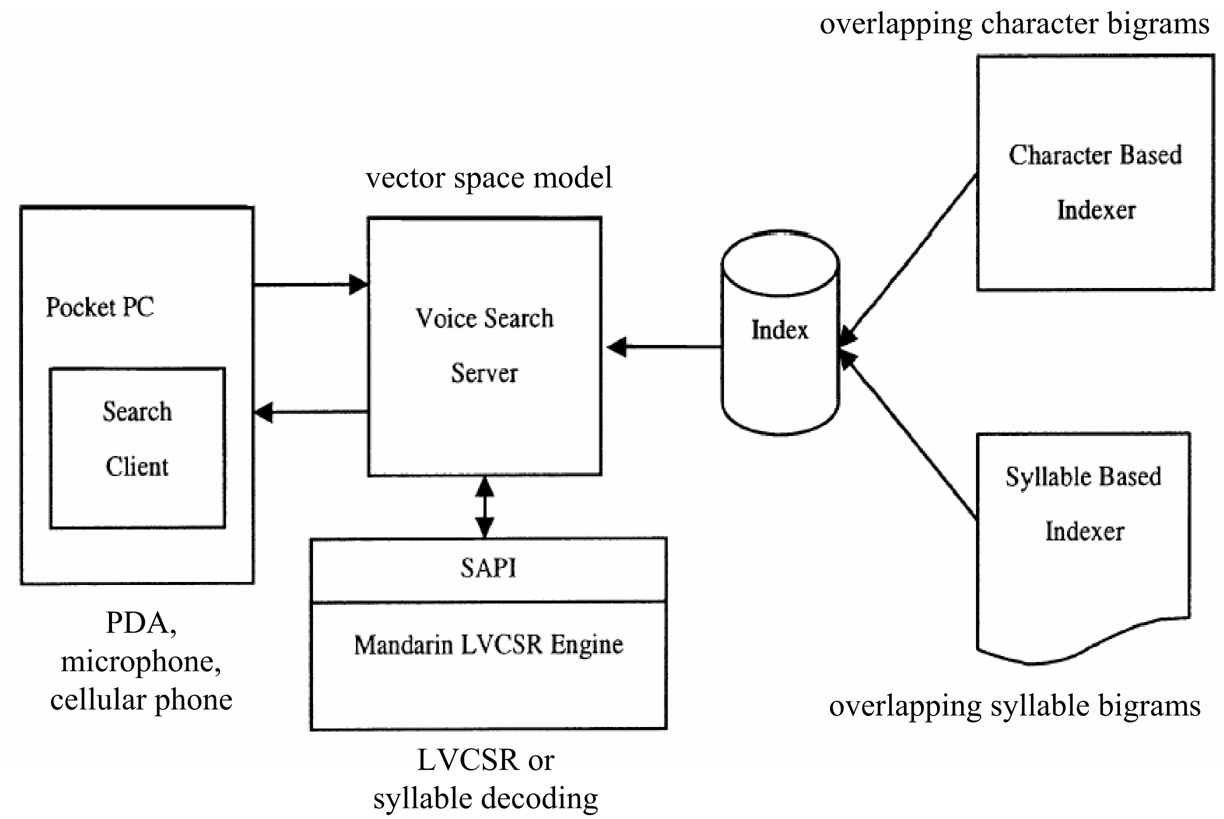
在四種不同時機下的資訊檢索過程。使用聲音問句(VQ，Voice Queries)或文字問句(TQ，Text Queries)去檢索聲音資訊(VI，Voice Information)或者是傳統的文字資訊(TI，Text Information)。

Speech-based Information Retrieval (cont.)

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。

Speech-based Information Retrieval (cont.)

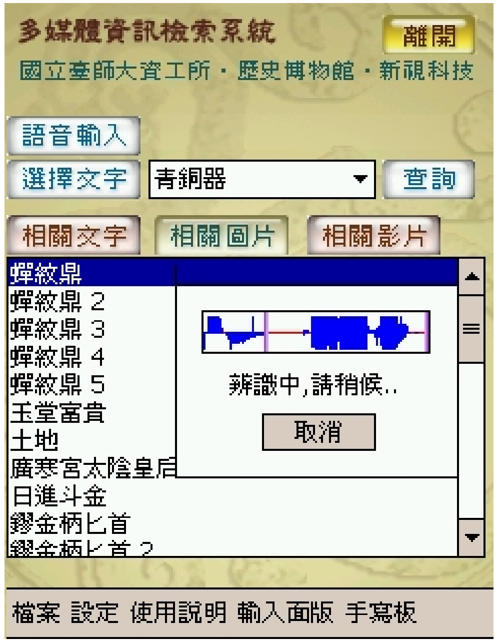


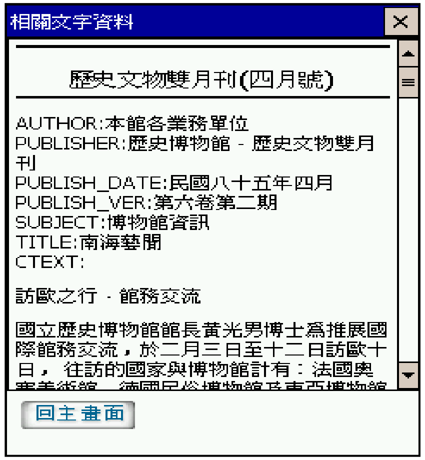
Speech-based Information Retrieval (cont.)

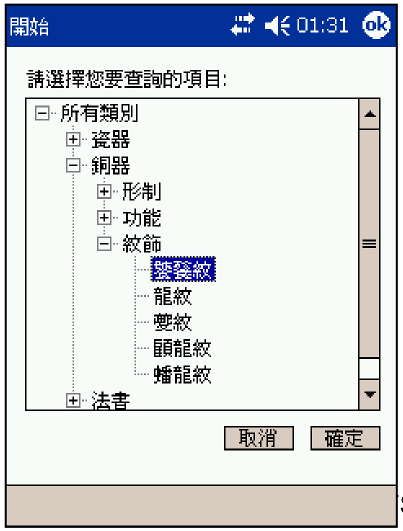
- PDA-based IR system for Mandarin broadcast news

Speech-based Information Retrieval (cont.)

- PDA-based IR system for digital archives
  - Current deployed at National Museum of History, Taipei

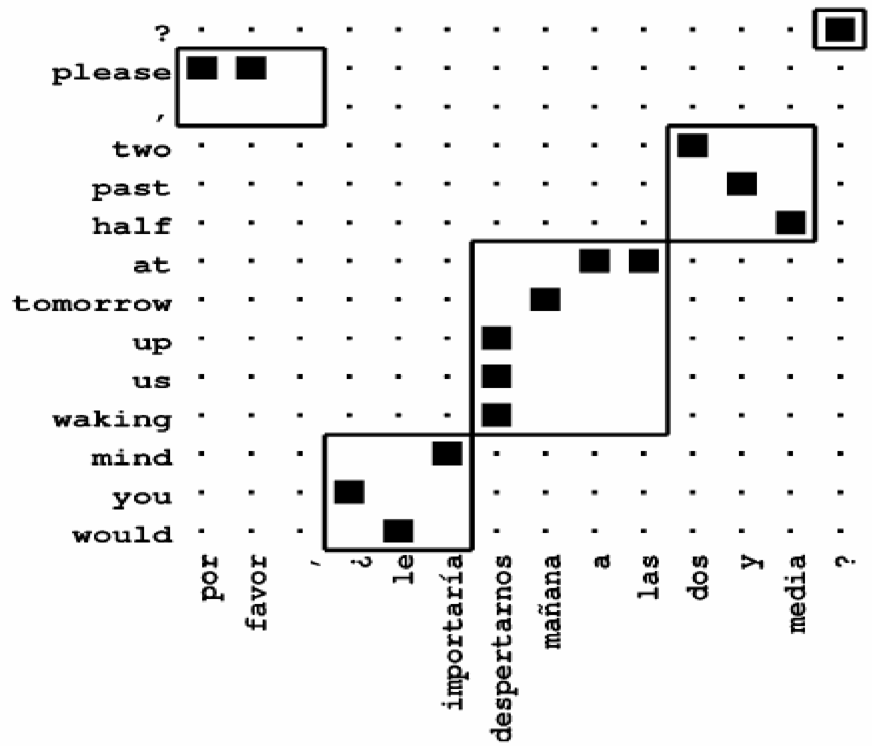






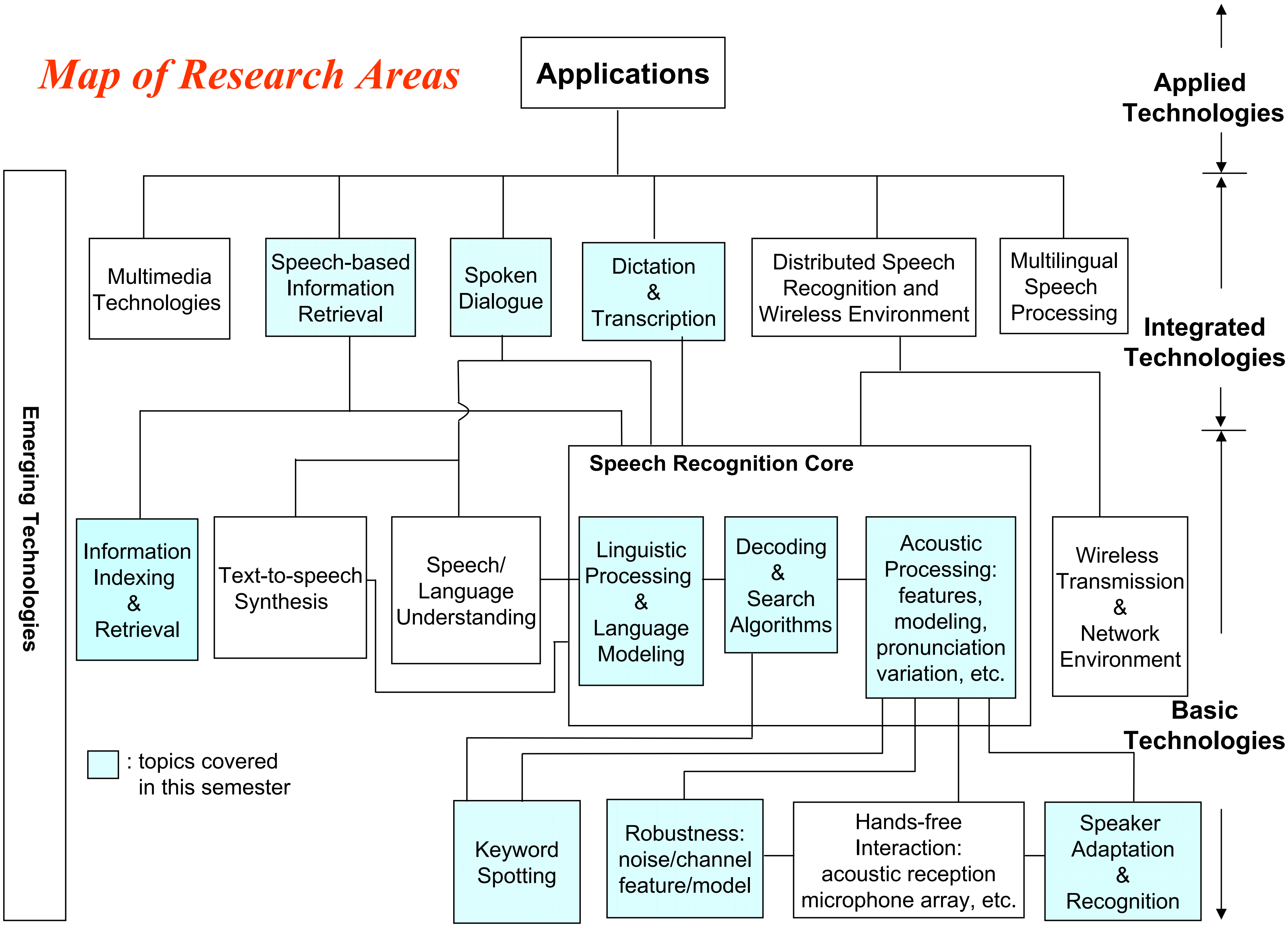
Speech-to-Speech Translation

- Multilingual interactive speech translation
  - Aims 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)



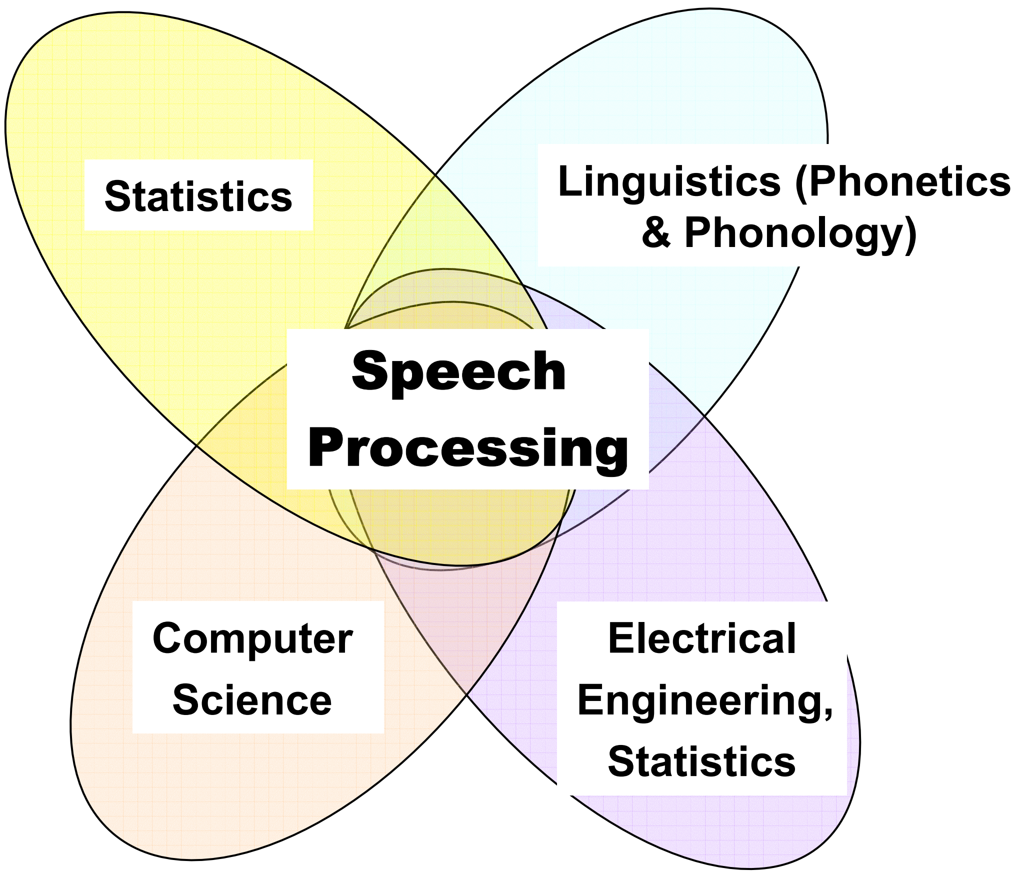
Example of a word alignment and of extracted alignment templates.

Map of Research Areas



Adapted from Prof. Lin-shan Lee

Different Academic Disciplines

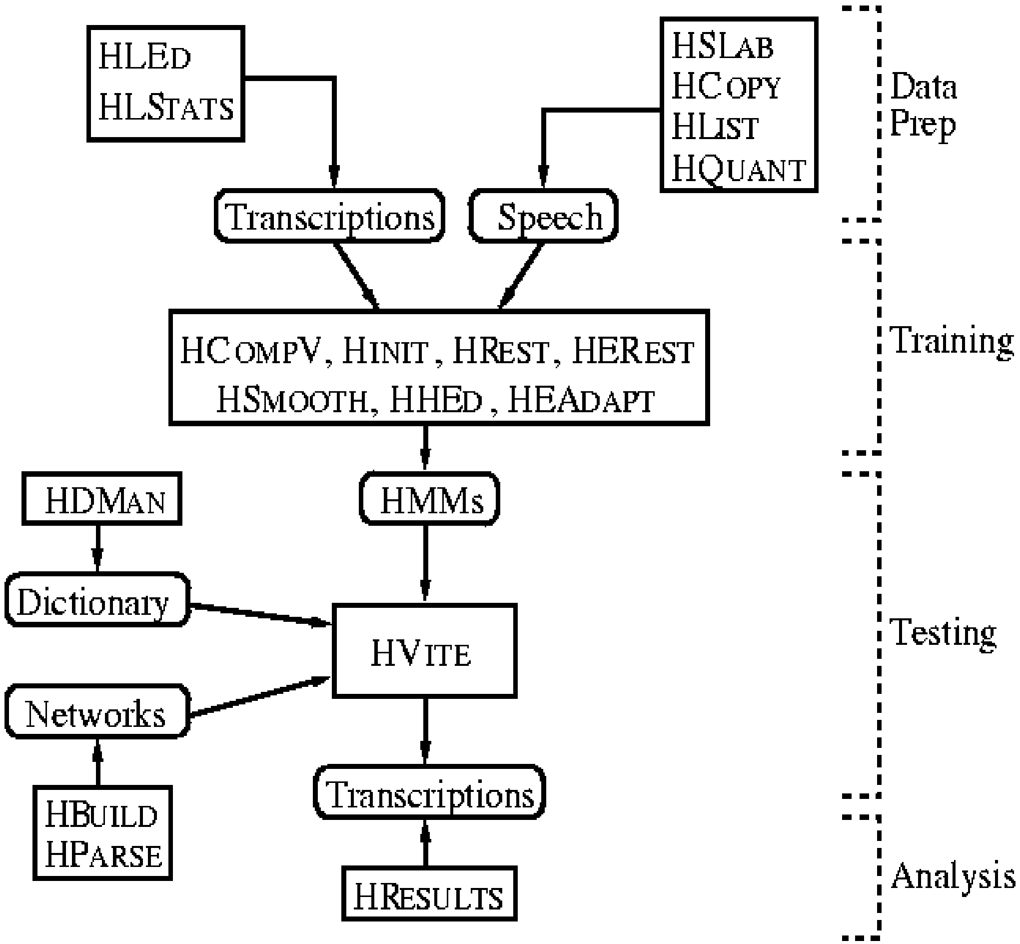


Speech Processing Toolkit

- 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

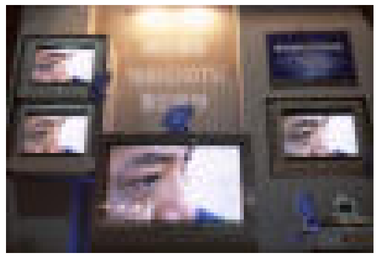
Speech Processing Toolkit

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



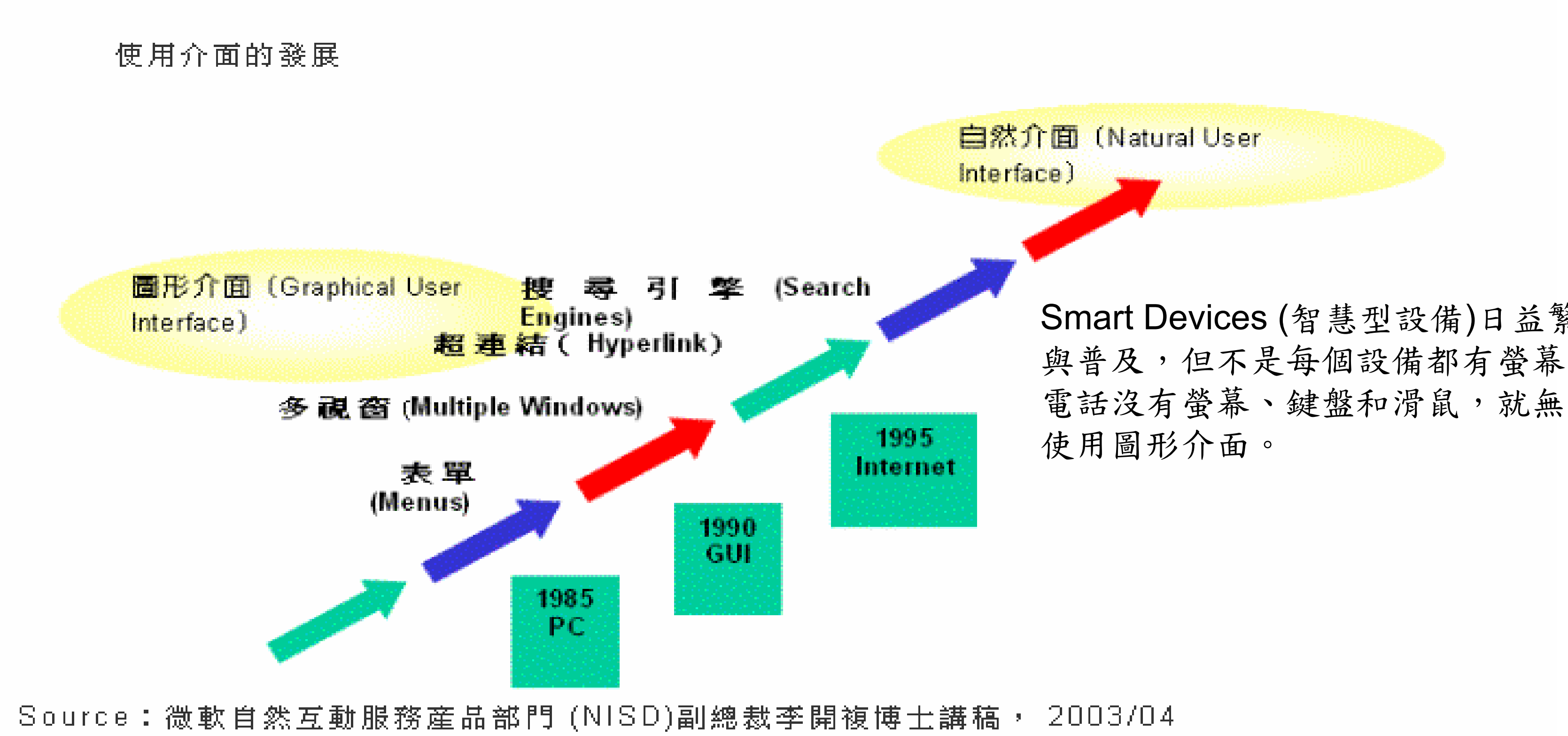
Speech Industry

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



Speech Industry

- Microsoft: Smart Device/Natural UI



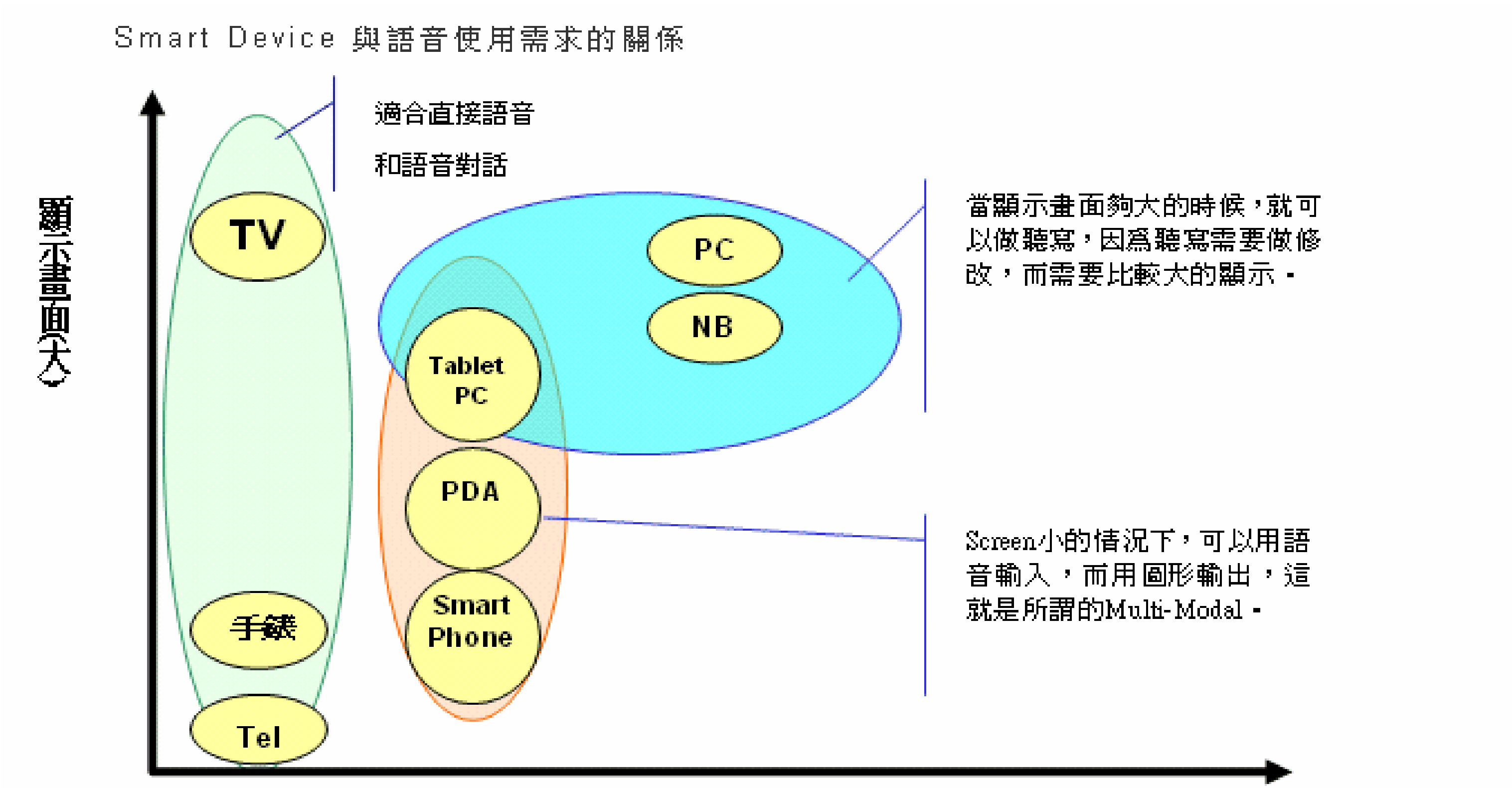
Smart Devices (智慧型設備)日益繁多與普及，但不是每個設備都有螢幕，例如：電話沒有螢幕、鍵盤和滑鼠，就無法使用圖形介面。

.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

- Microsoft: Smart Device/Natural UI



手寫	X	O	O
滑鼠	X	X	O
鍵盤	X	X	O

Source：微軟自然互動服務產品部門(NISD)副總裁李開複博士講稿，2003/4

Journals & Conferences

- Journals
  - IEEE Transactions on Speech and Audio Processing
  - Computer Speech and Language
  - Speech Communication
- Conferences
  - IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP)
  - Int. Conf. on Spoken Language Processing (ICSLP)
  - European Conference on Speech Communication and Technology (Eurospeech)
  - IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)
  - International Symposium on Chinese Spoken Language Processing (ISCSLP)
  - ROCLING Conference on Computational Linguistics and Speech Processing