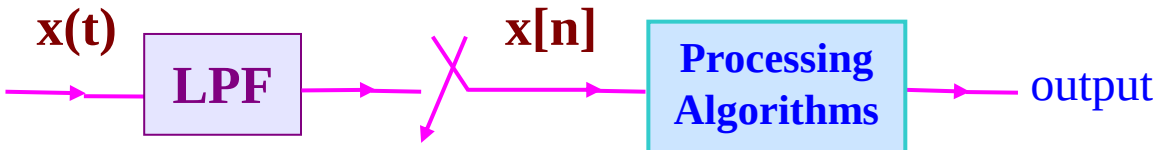
Digital Speech Processing

數位語音處理概論

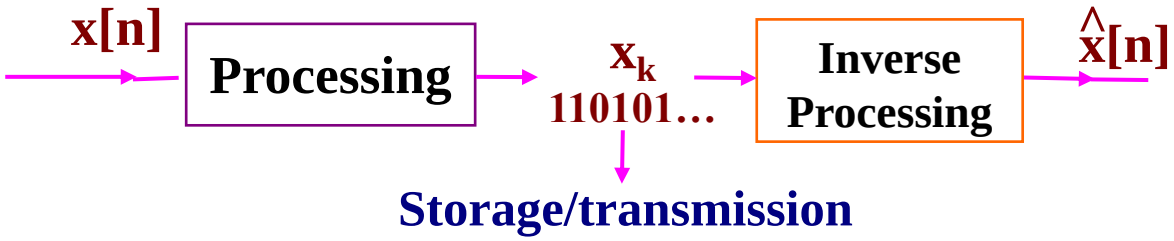
李琳山

Speech Signal Processing



- **Major Application Areas**

1. Speech Coding:Digitization and Compression



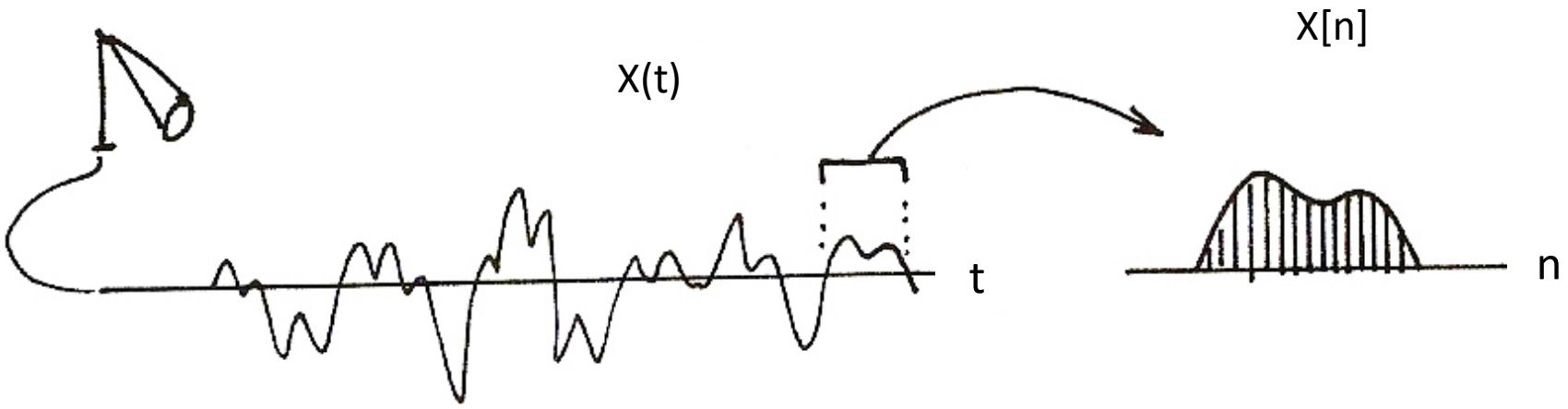
Considerations : 1) bit rate (bps)
2) recovered quality
3) computation complexity/feasibility

2. Voice-based Network Access —

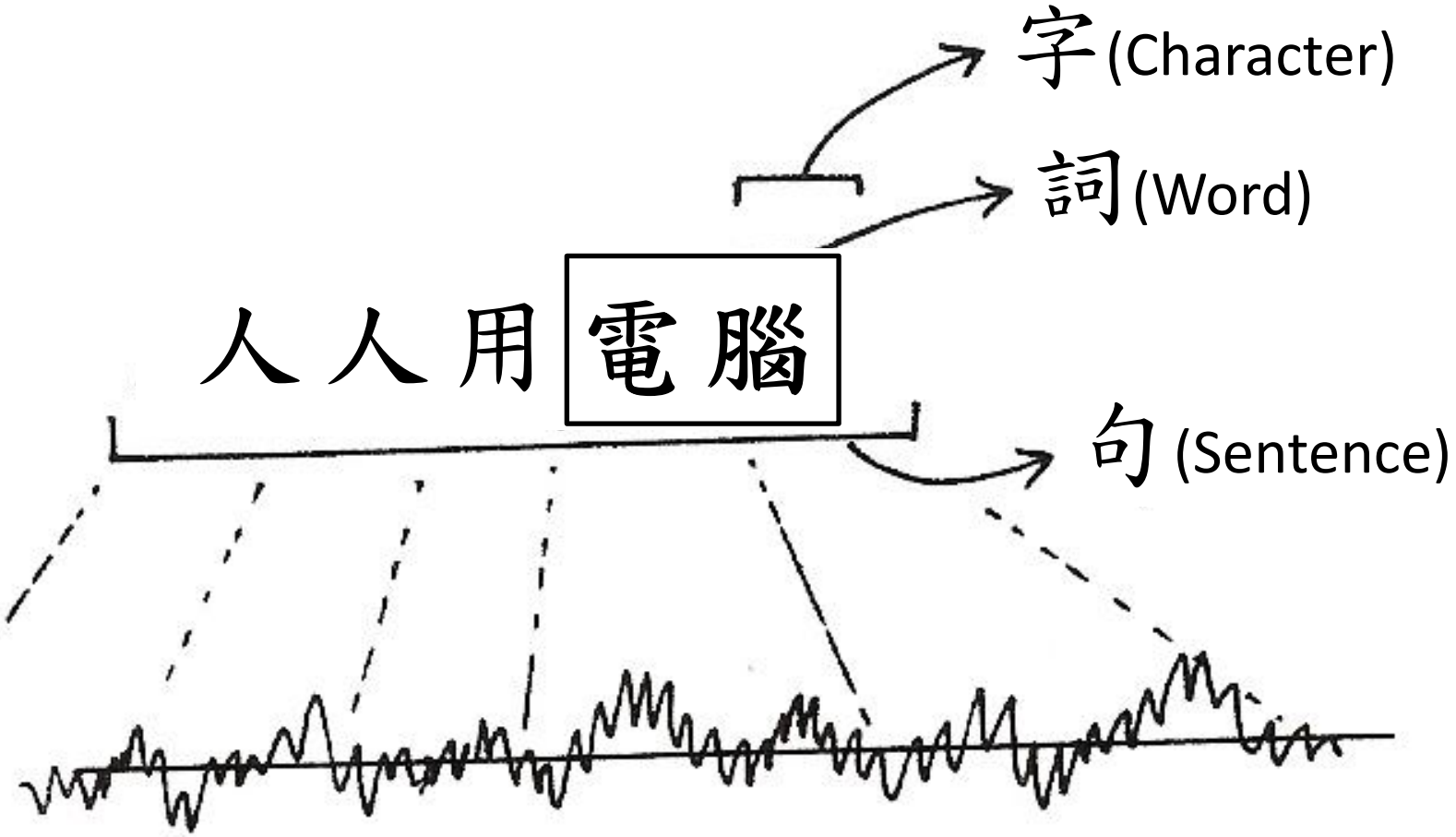
User Interface, Content Analysis, User-content Interaction

- **Speech Signals**
  - Carrying Linguistic Knowledge and Human Information: Characters, Words, Phrases, Sentences, Concepts, etc.
  - Double Levels of Information: Acoustic Signal Level/Symbolic or Linguistic Level
  - Processing and Interaction of the Double-level Information

Sampling of Signals

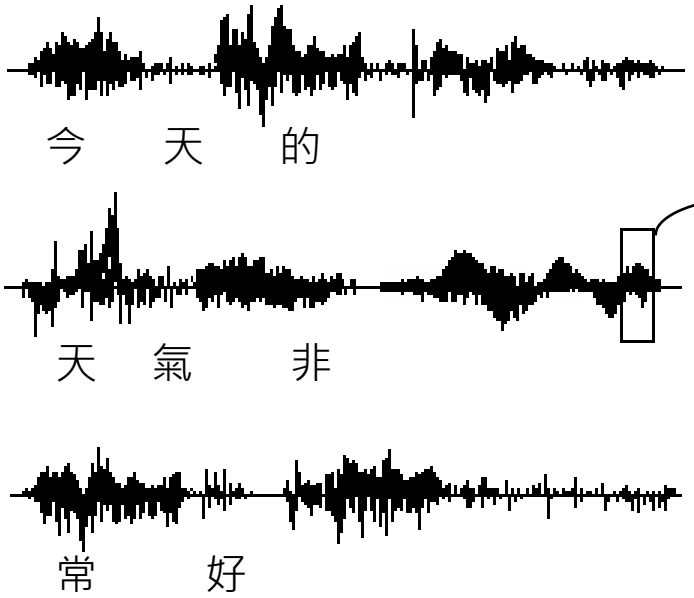


Double Levels of Information

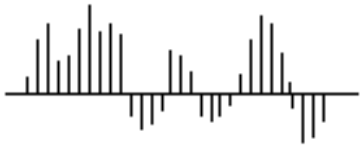


Speech Signal Processing – Processing of Double-Level Information

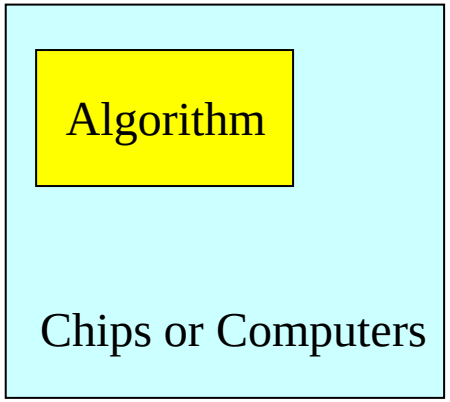
- Speech Signal



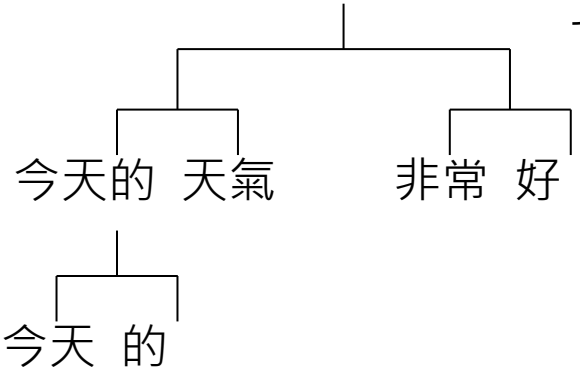
- Sampling



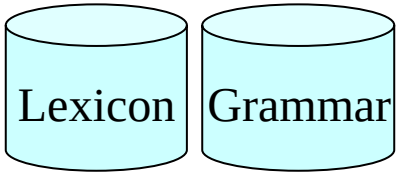
- Processing



- Linguistic Structure



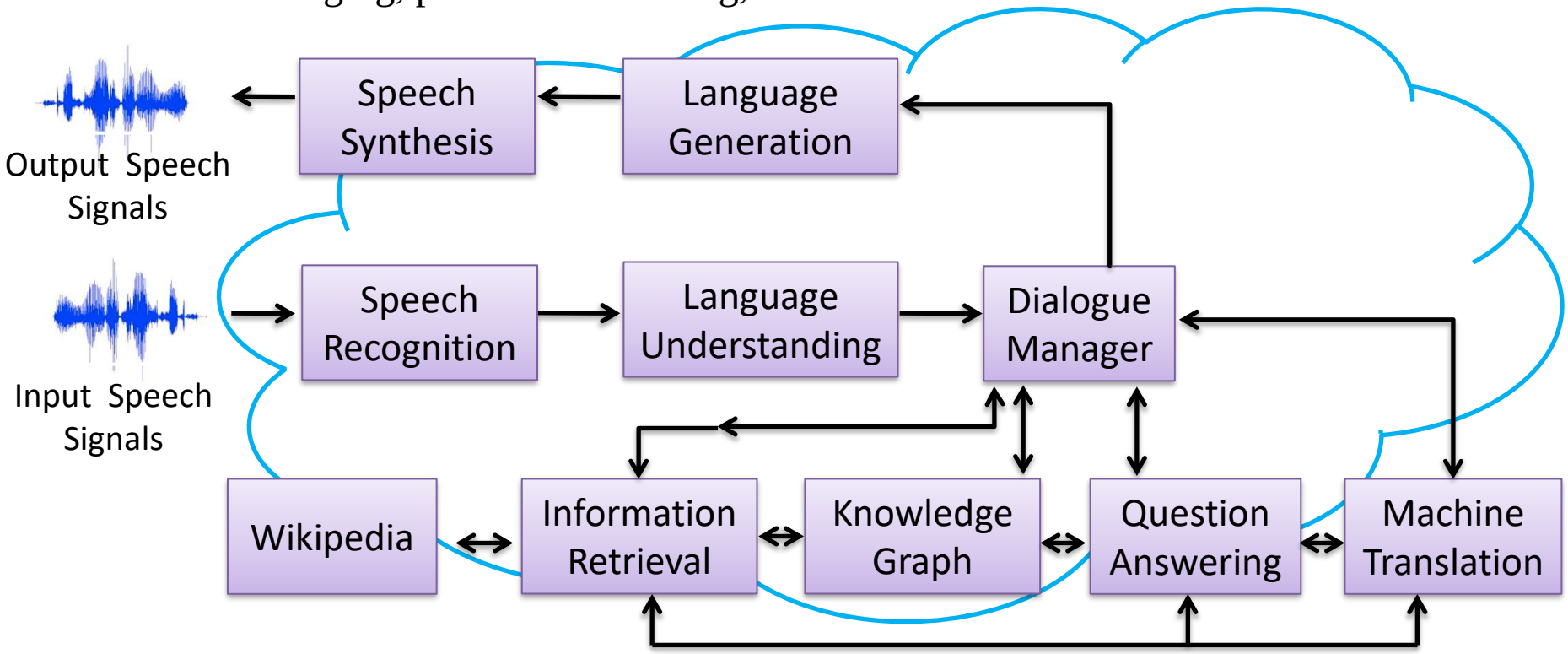
- Linguistic Knowledge



Well-Known Application Examples of Speech and Language Technologies – Speaking Personal Assistant

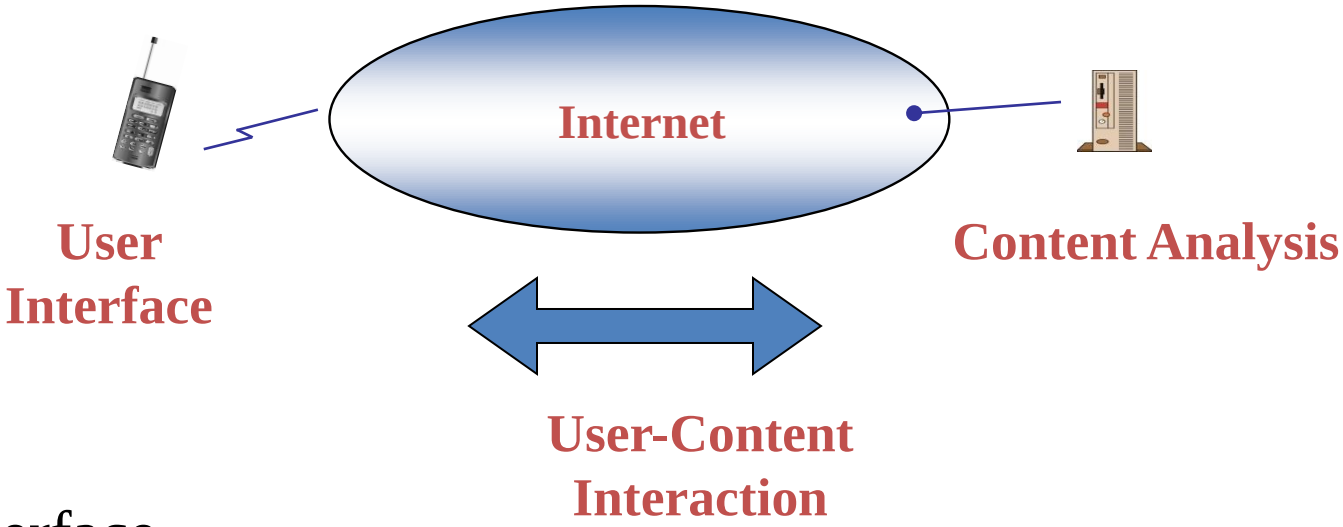
- **Examples**
  - Weather in New York next week ?
  - Who is the president of US ? What did he say today ?
  - How can I go to National Taiwan University ?
  - Short messaging, personal scheduling, etc.

- **Special Questions:**
  - 唐詩宋詞, 出師表…
  - 說個笑話…



- **Examples:**
  - Siri (Apple), Google Now (Google), Cortana (Microsoft)

Voice-based Network Access



- User Interface

  —when keyboards/mice inadequate

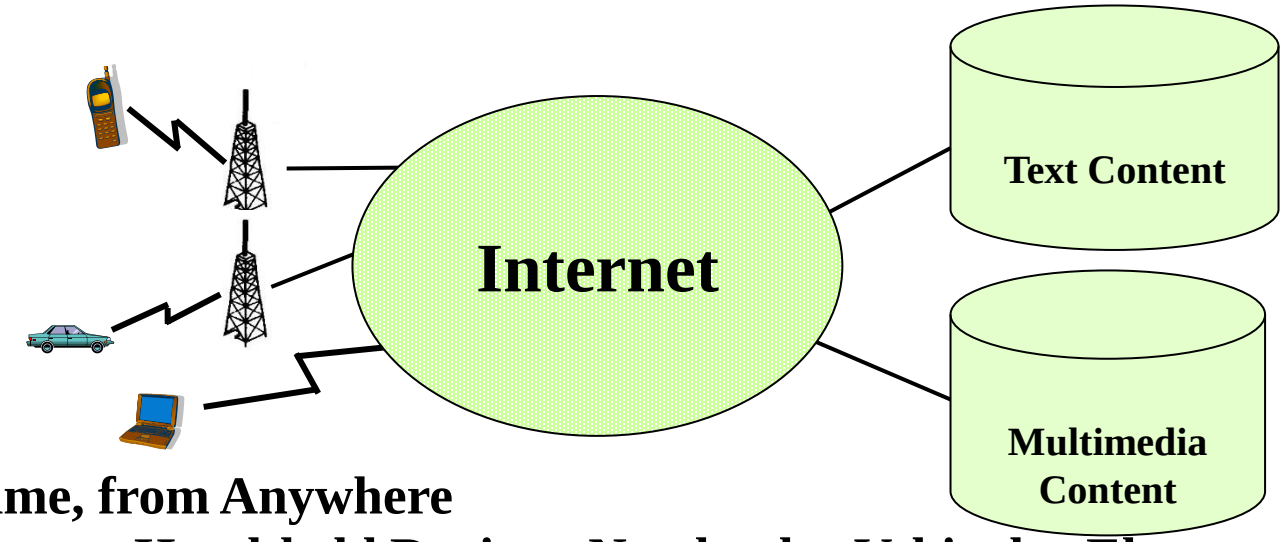
- Content Analysis

  — help in browsing/retrieval of multimedia content

- User-Content Interaction

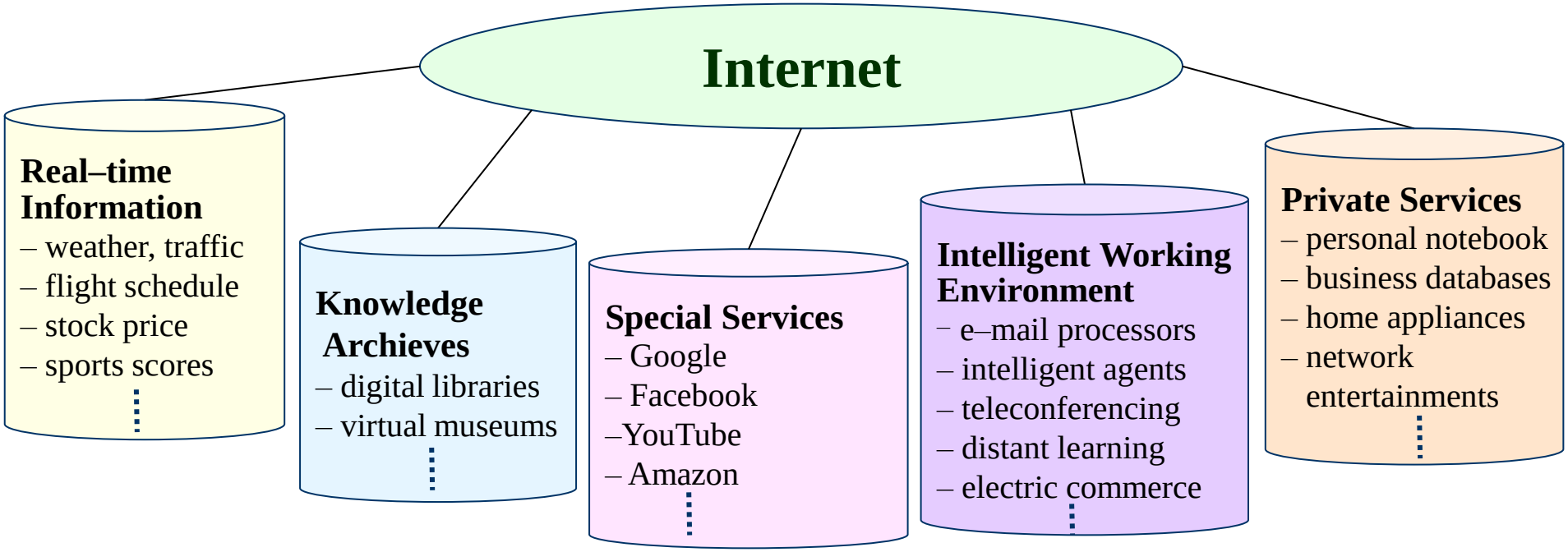
  —all text-based interaction can be accomplished by spoken language

User Interface —Wireless Communications Technologies have Created a Whole Variety of User Terminals



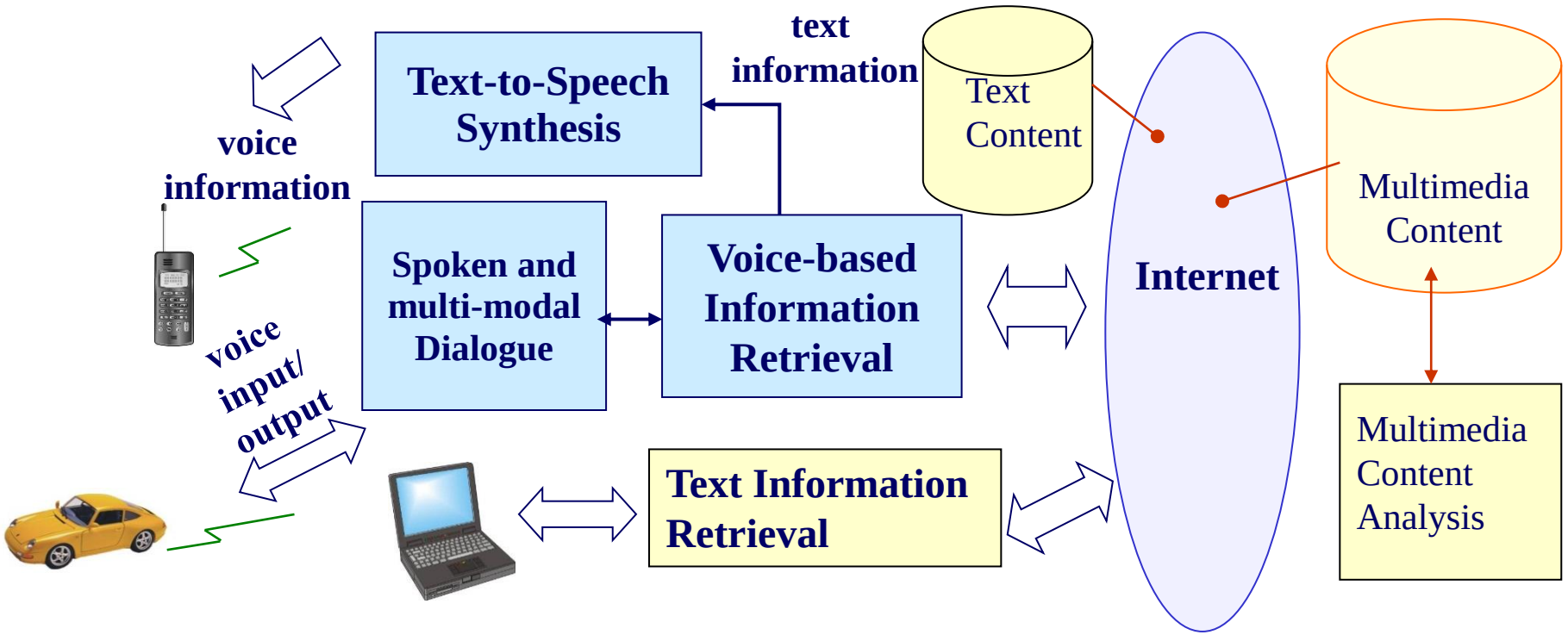
- **at Any Time, from Anywhere**
- **Smart phones, Hand-held Devices, Notebooks, Vehicular Electronics, Hands-free Interfaces, Home Appliances, Wearable Devices…**
- **Small in Size, Light in Weight, Ubiquitous, Invisible…**
- **Post-PC Era**
- **Keyboard/Mouse Most Convenient for PC's not Convenient any longer**
  - —— human fingers never shrink, and application environment is changed
- **Service Requirements Growing Exponentially**
- **Voice is the Only Interface Convenient for ALL User Terminals at Any Time, from Anywhere, and to the point in one utterance**
- **Speech Processing is the only less mature part in the Technology Chain**

Content Analysis—Multimedia Technologies have Created a World of Multimedia Content



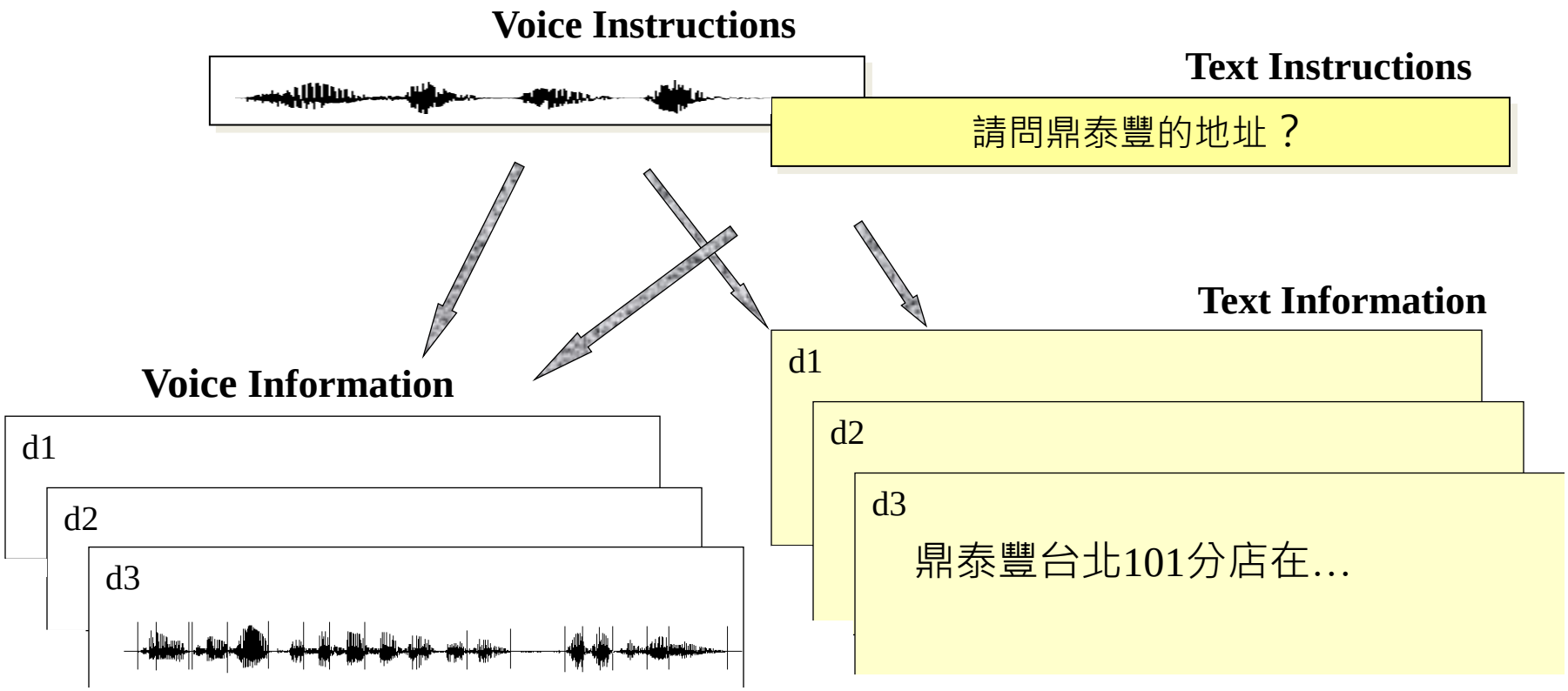
- **Most Attractive Form of the Network Content is Multimedia, which usually Includes Speech Information (but Probably not Text)**
- **Multimedia Content Difficult to be Summarized and Shown on the Screen, thus Difficult to Browse**
- **The Speech Information, if Included, usually Tells the Subjects, Topics and Concepts of the Multimedia Content, thus Becomes the Key for Browsing and Retrieval**
- **Multimedia Content Analysis based on Speech Information**

User-Content Interaction — Wireless and Multimedia Technologies are Creating An Era of Network Access by Spoken Language Processing



- **Hand-held Devices with Multimedia Functionalities Commonly used Today**
- **Network Access is Primarily Text-based today, but almost all Roles of Texts can be Accomplished by Speech**
- **User-Content Interaction can be Accomplished by Spoken and Multi-modal Dialogues**
- **Using Speech Instructions to Access Multimedia Content whose Key Concepts Specified by Speech Information**

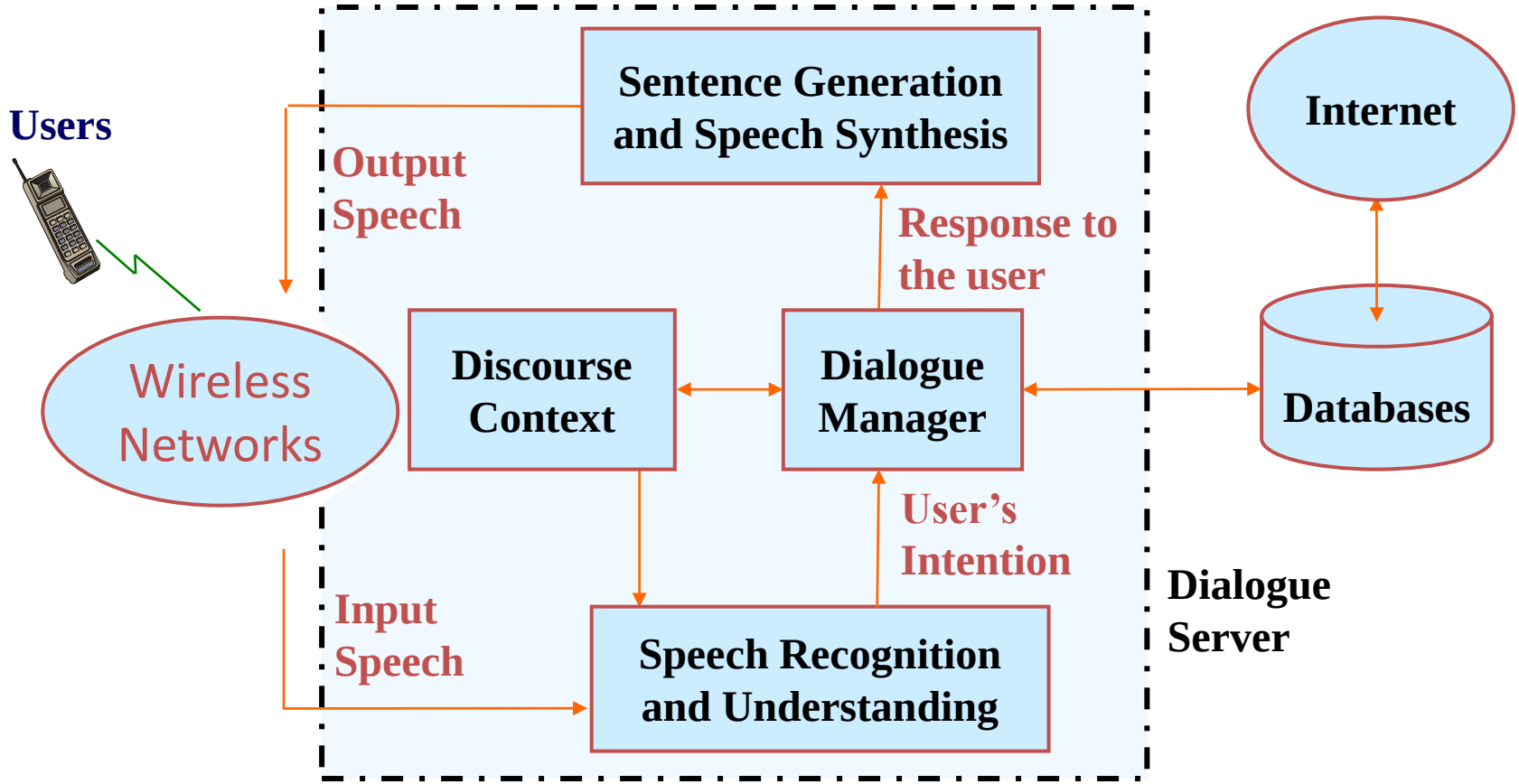
Voice-based Information Retrieval



- **Both the User Instructions and Network Content Can be in form of Speech**

Spoken and Multi-modal Dialogues

- **Almost All User-Content Interaction can be Accomplished by Spoken or Multi-modal Dialogues**



Outline

- **Both Theoretical Issues and Practical Problems will be Discussed**
- **Starting with Fundamentals, but Entering Research Topics in the Second Half**
- **Part I: Fundamental Topics**
  - 1.0 Introduction to Digital Speech Processing
  - 2.0 Fundamentals of Speech Recognition
  - 3.0 Map of Subject Areas
  - 4.0 More about Hidden Markov Models
  - 5.0 Acoustic Modeling
  - 6.0 Language Modeling
  - 7.0 Speech Signals and Front-end Processing
  - 8.0 Search Algorithms for Speech Recognition
- **Part II: Advanced Topics**
  - 9.0 Speech Recognition Updates
  - 10.0 Speech-based Information Retrieval
  - 11.0 Spoken Document Understanding and Organization for User-content Interaction
  - 12.0 Computer-assisted Language Learning(Call)
  - 13.0 Speaker Variabilities: Adaption and Recognition
  - 14.0 Latent Topic Analysis
  - 15.0 Robustness for Acoustic Environment
  - 16.0 Some Fundamental Problem-solving Approaches
  - 17.0 Spoken Dialogues
  - 18.0 Conclusion

References

- **教科書：無**
- **主要參考書：**

1. X. Huang, A. Acero, H. Hon, “Spoken Language Processing”, Prentice Hall, 2001,松瑞
2. F. Jelinek, “Statistical Methods for Speech Recognition”, MIT Press, 1999
3. L. Rabiner, B.H. Juang, “Fundamentals of Speech Recognition”, Prentice Hall, 1993, 民全
4. C. Becchetti, L. Prina Ricotti, “Speech Recognition- Theory and C++ implementation”, Johy Wiley and Sons, 1999, 民全
5. D. Jurafsky, J. Martin, “Speech and Language Processing- An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics, 2nd edition”, Prentice-Hall, 2009 (3rd edition draft parts on-line)
6. G. Tur, R. De Mori, “Spoken Language Understanding- Systems for Extracting Semantic Information from Speech”, John Wiley & Sons, 2011
7. D. Yu, L. Deng, “Automatic Speech Recognition - A Deep Learning Approach”, Springer, 2015.
8. 其他參考文獻課堂上提供

Other Information

- **教材：**

  available on web before the day of class (http://speech.ee.ntu.edu.tw)

- **適合年級：三、四 (電機系、資工系)**

- **成績評量方式**

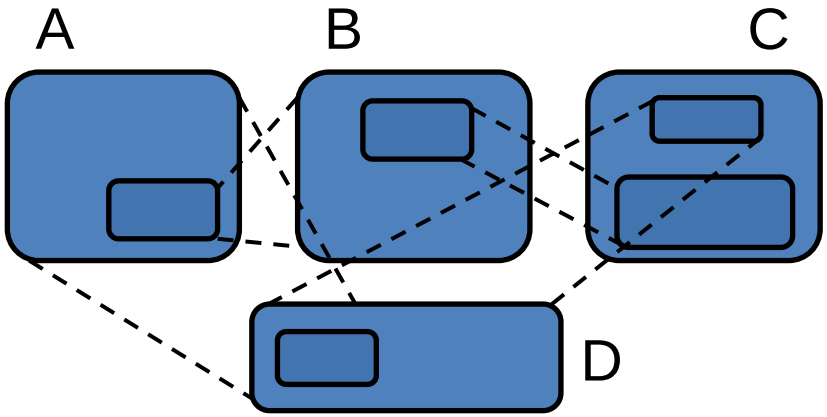
Midterm Exam	25%
Homeworks (I) (II) (Ⅲ)	15%、5%、15%
Final Exam	10%
Term Project	30%

Goals

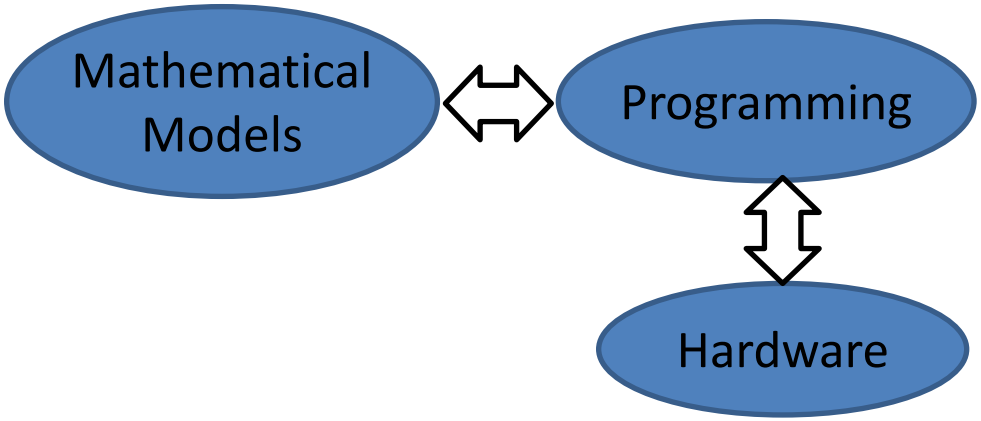
- 課程目的：

  提供同學進入此一充滿機會與挑戰的新領域所需的基本知識，體驗數學模型與軟體程式如何相輔相成，學習進入一個新領域由基礎進入研究的歷程，體會吸收非結構性知識(Unstructured Knowledge)的經驗

- **Unstructured Knowledge**



- **Math & Programming**

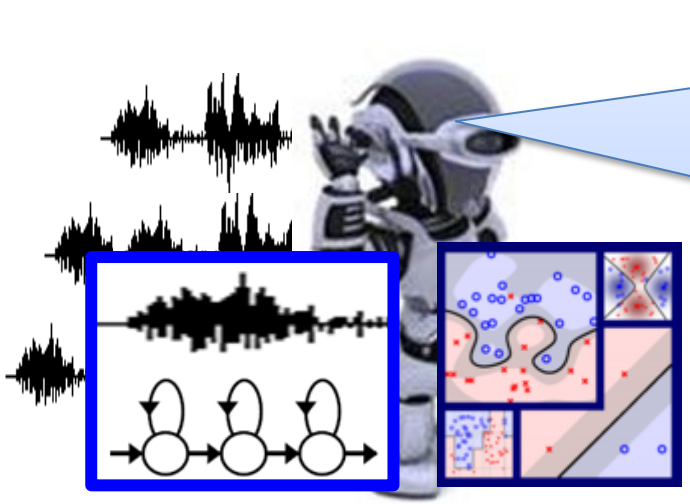


1.0 Introduction — A Brief Summary of Core Technologies and Example Application Seenarios

References for 1.0

1.“Speech and Language Processing over the Web”, IEEE Signal Processing Magazine, May 2008

Speech Recognition



Signal Modeling

Machine Learning

Deep Learning

Machine mimics human brains

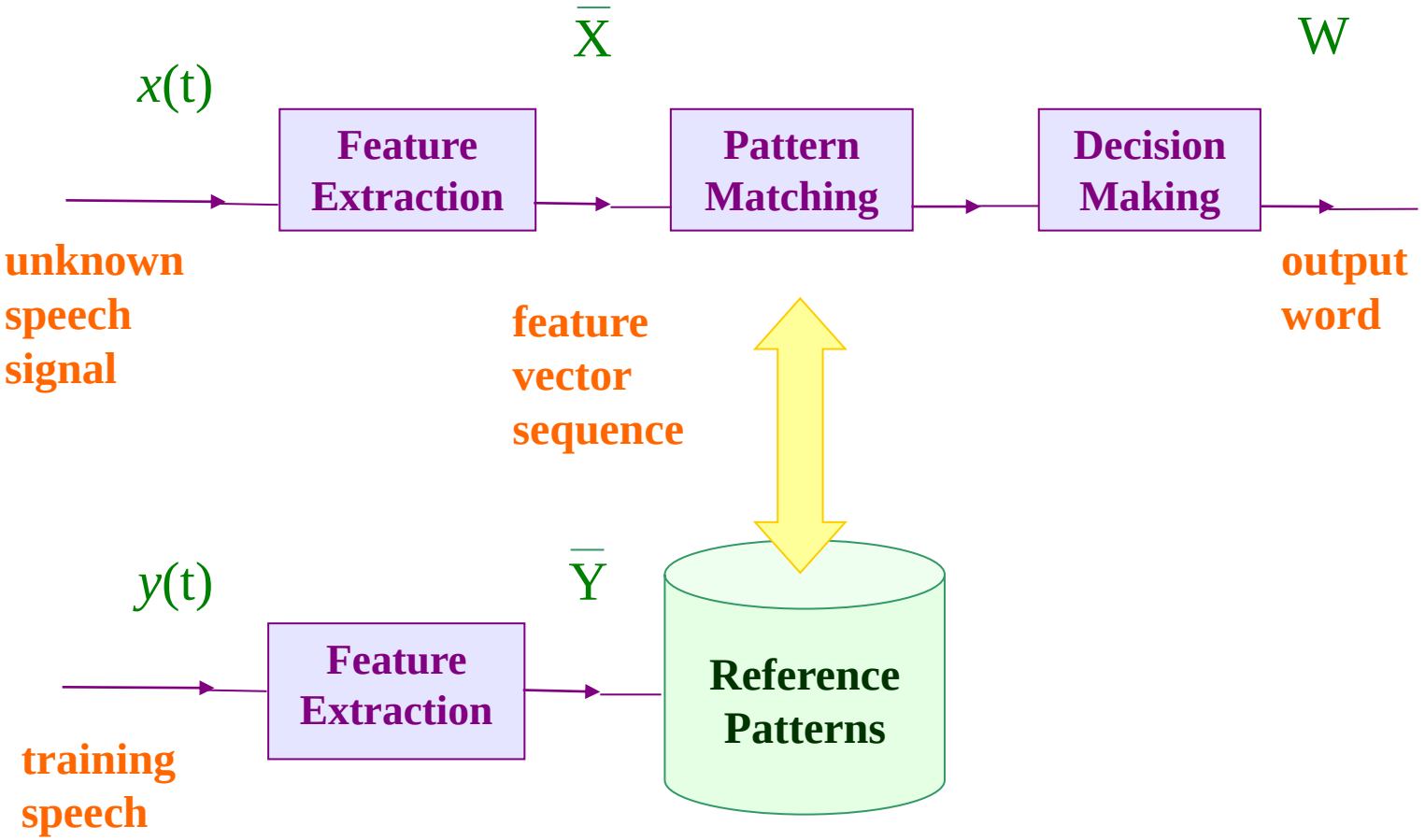
Not only used in speech, but many other fields

Speech Signal

Text

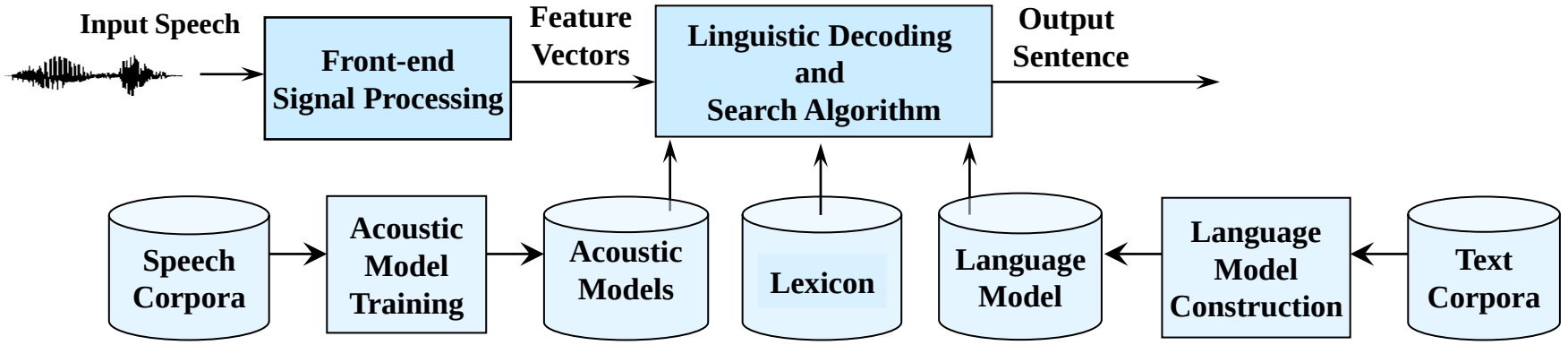
今天天氣很好

Speech Recognition as a pattern recognition problem



Basic Approach for Large Vocabulary Speech Recognition

- **A Simplified Block Diagram**



- **Example Input Sentence**

  this is speech

- **Acoustic Models (聲學模型)**

  (th-ih-s-ih-z-s-p-ih-ch)

- **Lexicon** (th-ih-s) → this

  (ih-z) → is

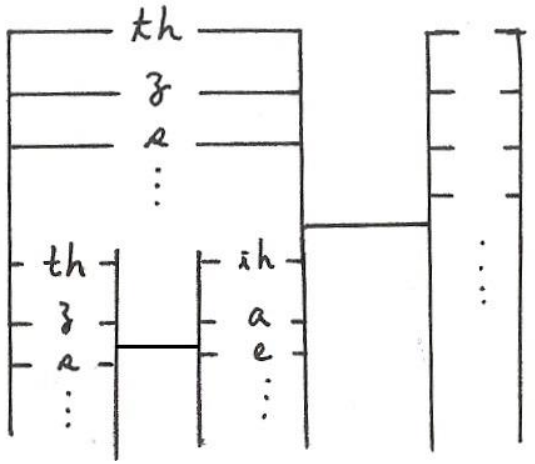
  (s-p-iy-ch) → speech

- **Language Model (語言模型)** (this) – (is) – (speech)

  P(this) P(is | this) P(speech | this is)

  $P(w_i|w_{i-1})$ bi-gram language model

  $P(w_i|w_{i-1},w_{i-2})$ tri-gram language model,etc



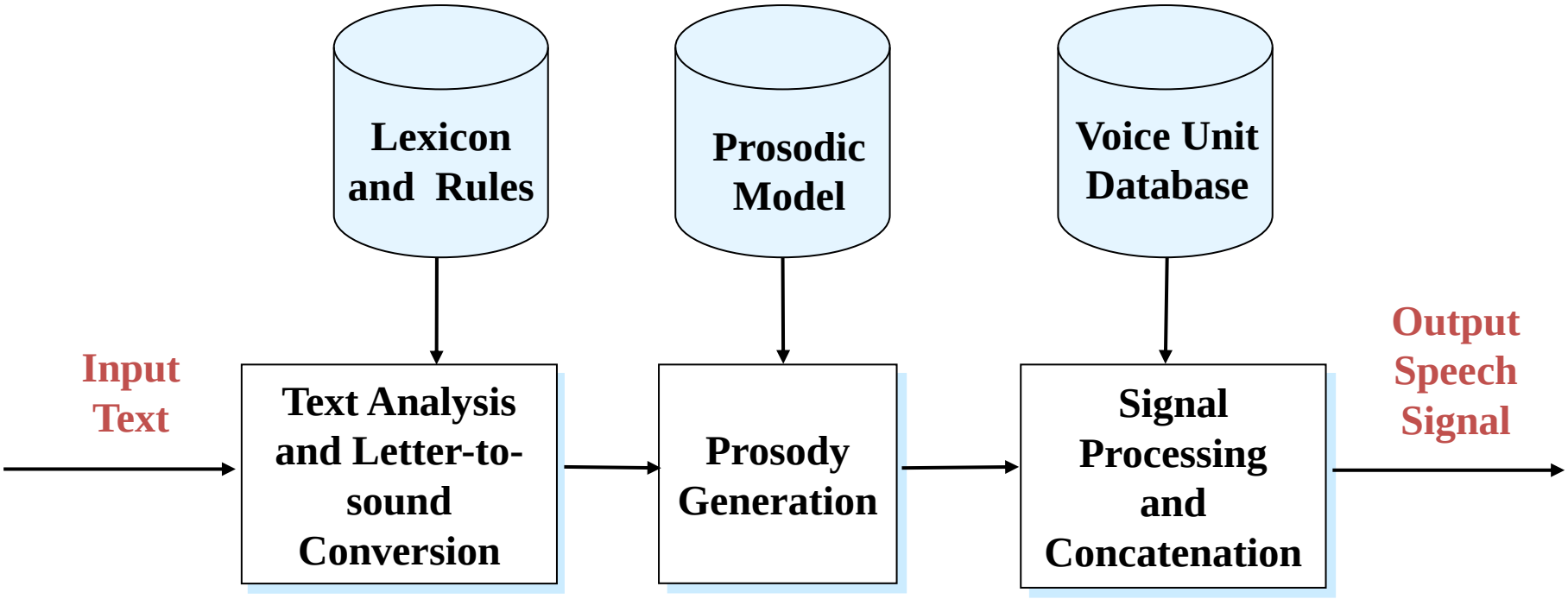
- **Deep Learning Approaches**

Speech Recognition Technologies, Applications and Problems

- **Word Recognition**
  - voice command/instructions
- **Keyword Spotting**
  - identifying the keywords out of a pre-defined keyword set from input voice utterances
- **Large Vocabulary Continuous Speech Recognition**
  - entering longer texts
  - remote dictation/automatic transcription
- **Speaker Dependent/Independent/Adaptive**
- **Acoustic Reception/Background Noise/Channel Distortion**
- **Read/Spontaneous/Conversational Speech**
- **Deep Learning Approaches**

Text-to-speech Synthesis

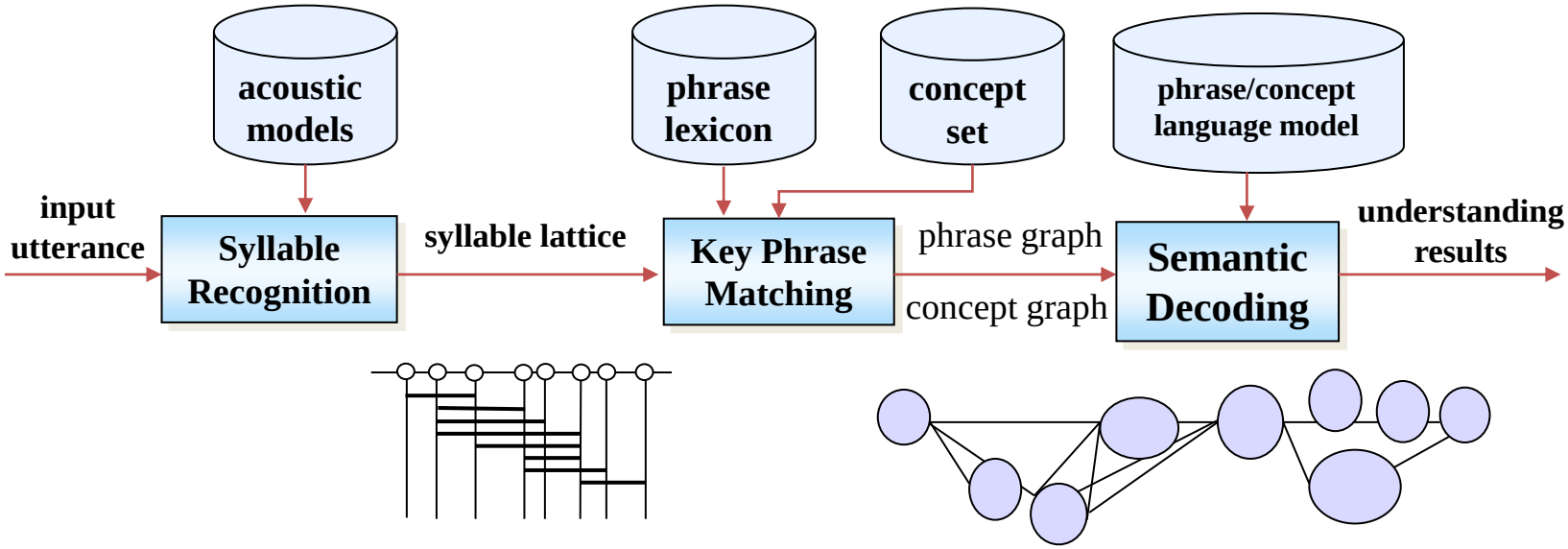
- Transforming any input text into corresponding speech signals
- E-mail/Web page reading
- Prosodic modeling
- Basic voice units/rule-based, non-uniform units/corpus-based, model-based



- Deep Learning Approaches

Speech Understanding

- **Understanding Speaker's Intention rather than Transcribing into Word Strings**
- **Limited Domains/Finite Tasks**



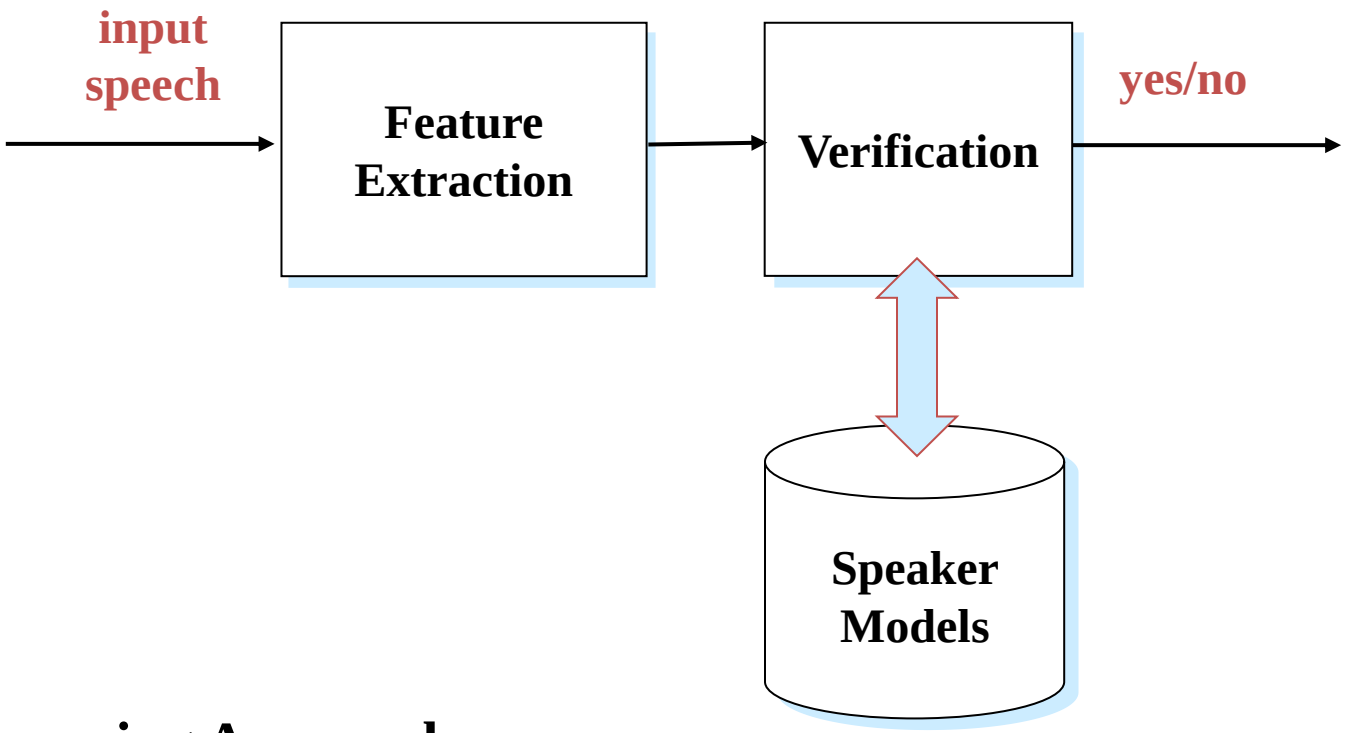
- **An Example**

utterance: 請幫我查一下 台灣銀行 的 電話號碼 是幾號?
key phrases: (查一下) - (台灣銀行) - (電話號碼)
concept: (inquiry) - (target) - (phone number)

- **Deep Learning Approaches**

Speaker Verification

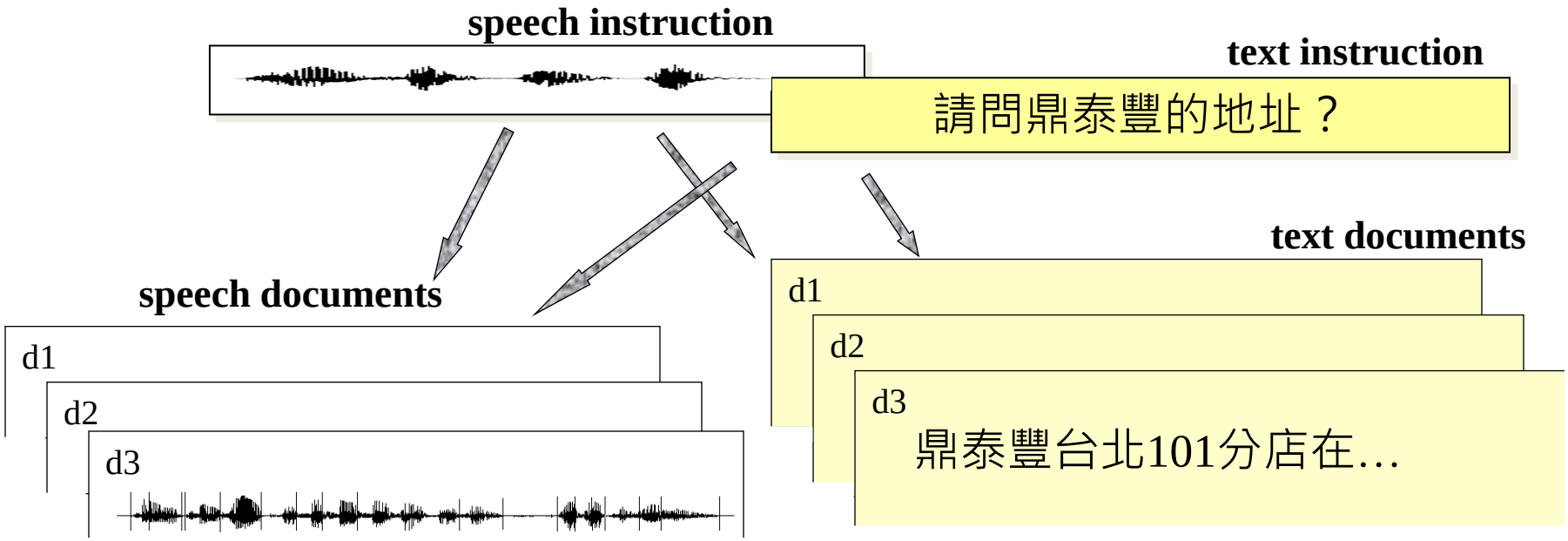
- Verifying the speaker as claimed
- Applications requiring verification
- Text dependent/independent
- Integrated with other verification schemes



- Deep Learning Approaches

Voice-based Information Retrieval

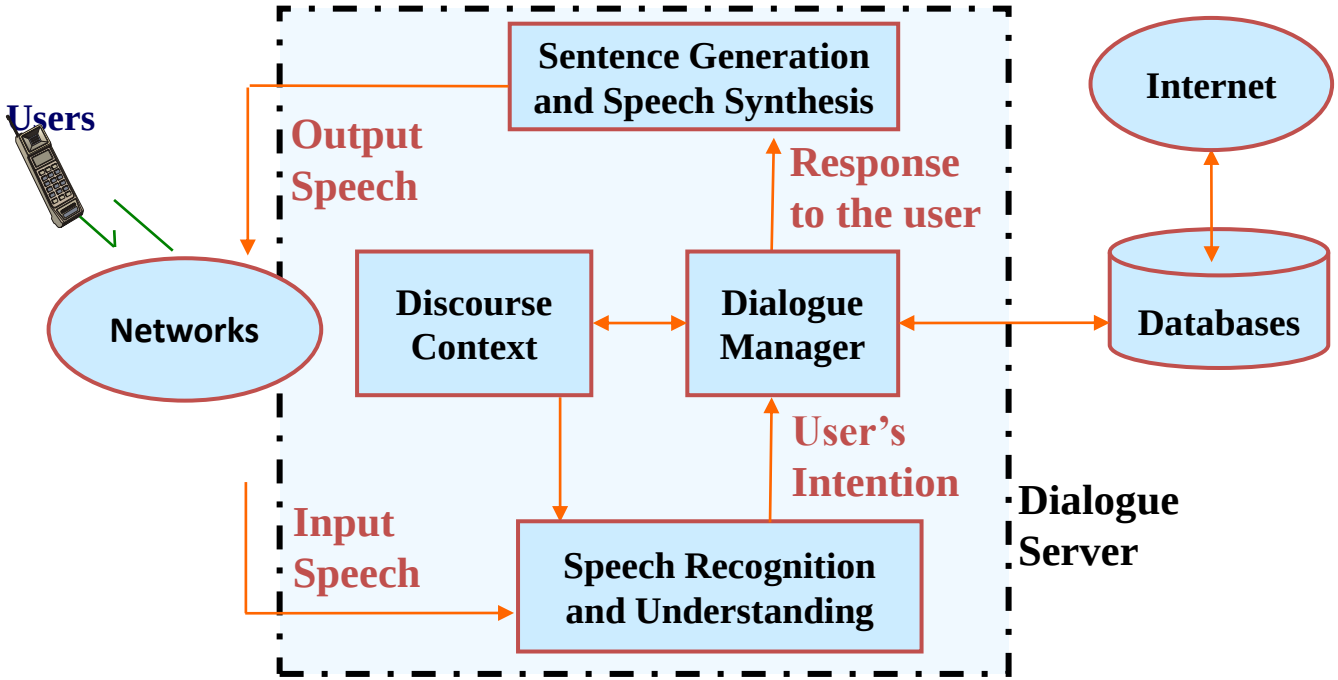
- **Speech Instructions**
- **Speech Documents (or Multi-media Documents including Speech Information)**



- **Locate exactly the desired utterances**
- **Text descriptions not needed for indexing/retrieving purposes**
- **Deep Learning Approaches**

Spoken Dialogue Systems

- Almost all human-network interactions can be accomplished by spoken dialogue
- Speech understanding, speech synthesis, dialogue management
- Mission-oriented/chatbot
- System/user/mixed initiatives
- Reliability/efficiency, dialogue modeling/flow control
- Transaction success rate/average number of dialogue turns



- Deep Learning Approaches

Spoken Document Understanding and Organization

- **Unlike the Written Documents which are easily shown on the screen for user to browse and select, Spoken Documents are just Audio Signals**
  - — the user can't listen each one from the beginning to the end during browsing
  - — better approaches for understanding/organization of spoken documents becomes necessary
- **Spoken Document Segmentation**
  - — automatically segmenting a spoken document into short paragraphs, each with a central topic
- **Spoken Document Summarization**
  - — automatically generating a summary (in text or speech form) for each short paragraph
- **Title Generation for Spoken Documents**
  - — automatically generating a title (in text or speech form) for each short paragraph
- **Key Term Extraction and Key Term Graph Construction for Spoken Documents**
  - — automatically extracting a set of key terms for each spoken document, and constructing key term graphs for a collection of spoken documents
- **Semantic Structuring of Spoken Documents**
  - — construction of semantic structure of spoken documents into graphical hierarchies
- **Deep Learning Approaches**

Multi-lingual Functionalities

- **Code-Switching Problem**
  - English words/phrases inserted in spoken Chinese sentences as an example

    人人都用Computers，家家都上Internet

    OK不OK？OK啦!
  - the whole sentence switched from Chinese to English as an example

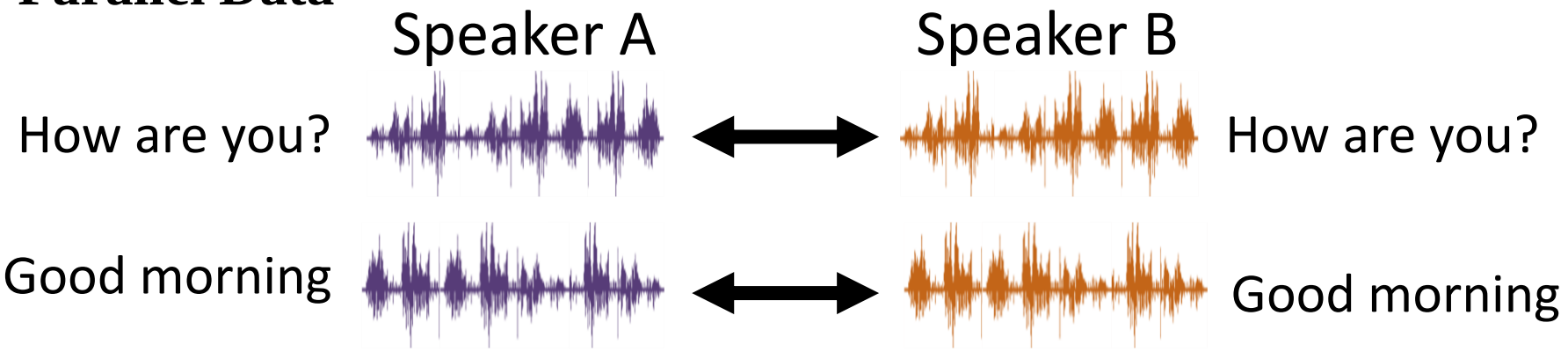
    準備好了嗎？Let's go!
- **Cross-language Information Processing**
  - globalized network with multi-lingual content/users
  - cross-language network information processing with a certain input language
- **Dialects/Accents**
  - hundreds of Chinese dialects as an example
  - code-switching problem─ Chinese dialects mixed with Mandarin (or plus English) as an example
  - Mandarin with a variety of strong accents as an example
- **Global/Local Languages**
- **Language Dependent/Independent Technologies**
- **Code-Switching Speech Processing, Speech-to-speech Translation, Computer-assisted Language Learning**
- **Deep Learning Approaches**

Computer-Assisted Language Learning

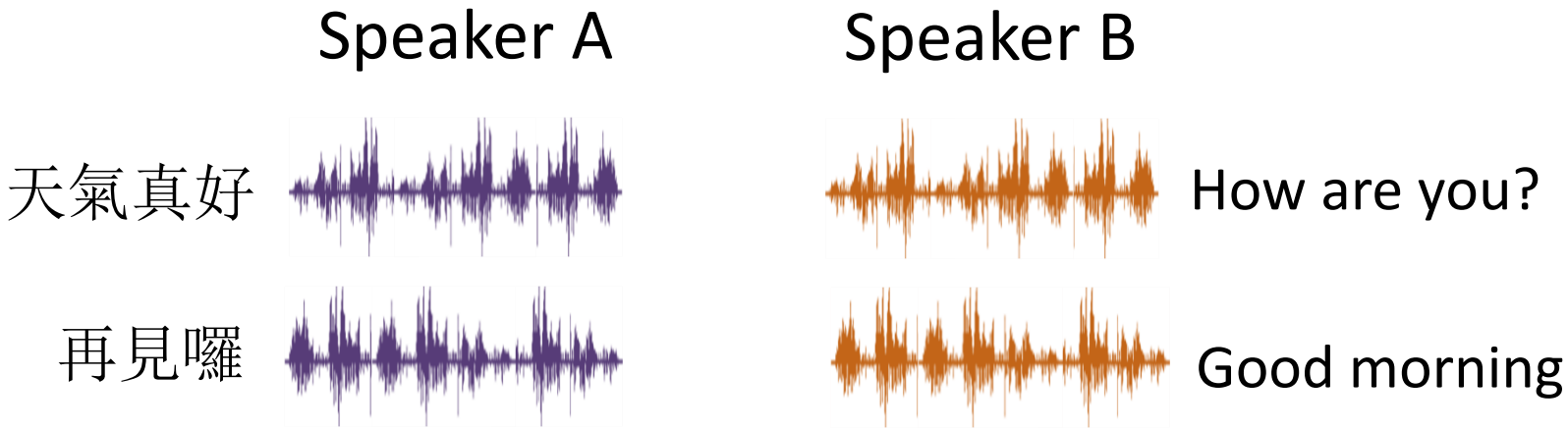
- **Globalized World**
  - every one needs to learn one or more languages in addition to the native language
- **Language Learning**
  - one-to-one tutoring most effective but with high cost
- **Computers not as good as Human Tutors**
  - software reproduced easily
  - used repeatedly any time, anywhere
  - never get tired or bored
- **Learning of**
  - pronunciation, vocabulary, grammar, sentences, dialogues, etc.
  - sometimes in form of games
- **Deep Learning Approaches**

Voice Conversion

- **Parallel Data**



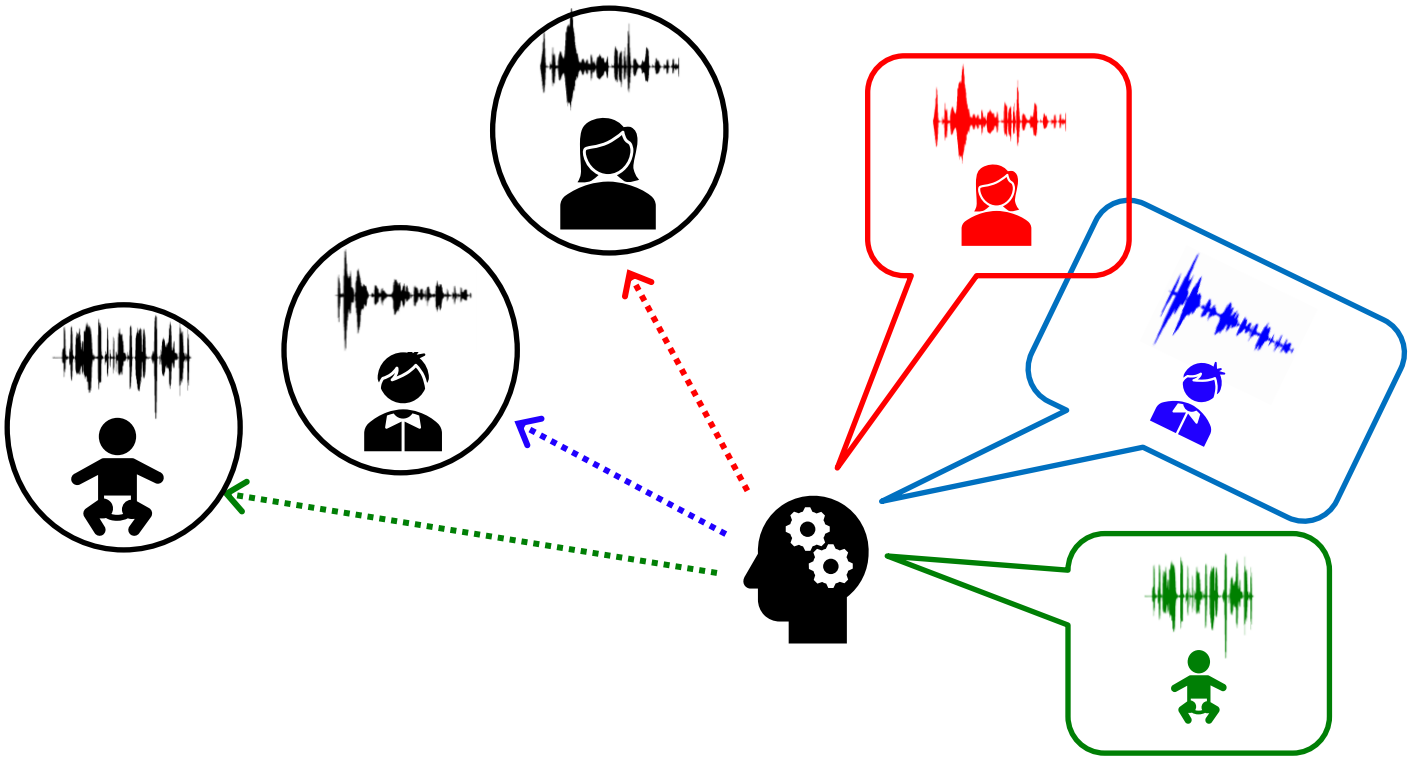
- **Non-Parallel Data**



- **Very Small Data**

Speech Separation

- **Cocktail Party Problem**



Machine Comprehension of Spoken Content

- **TOEFL Listening Comprehension Test by Machine**
  - Audio Story:  (The original story is 5 min long.)
  - Question: “ What is a possible origin of Venus’ clouds? ”
  - Choices:

    (A) gases released as a result of volcanic activity

    (B) chemical reactions caused by high surface temperatures

    (C) bursts of radio energy from the plane's surface

    (D) strong winds that blow dust into the atmosphere