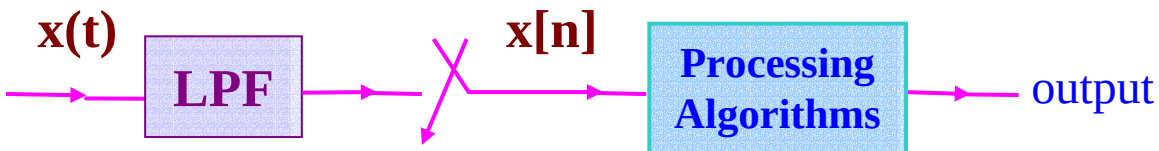
Digital Speech Processing

數位語音處理

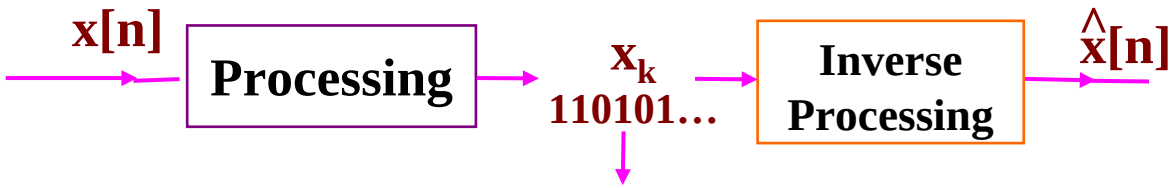
李琳山

Speech Signal Processing



- **Major Application Areas**

1. Speech Coding:Digitization and Compression

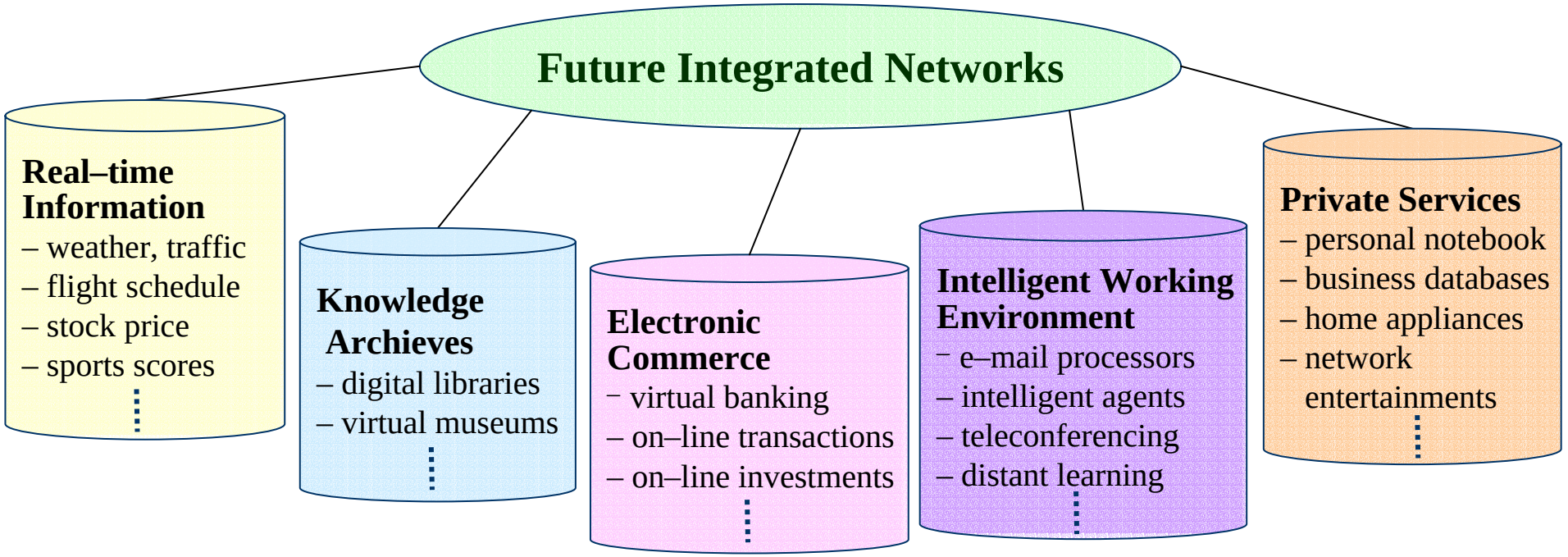


Considerations : 1) bit rate (bps)
2) recovered quality
3) computation complexity/feasibility

2. Voice Interface for Human-Network Interaction

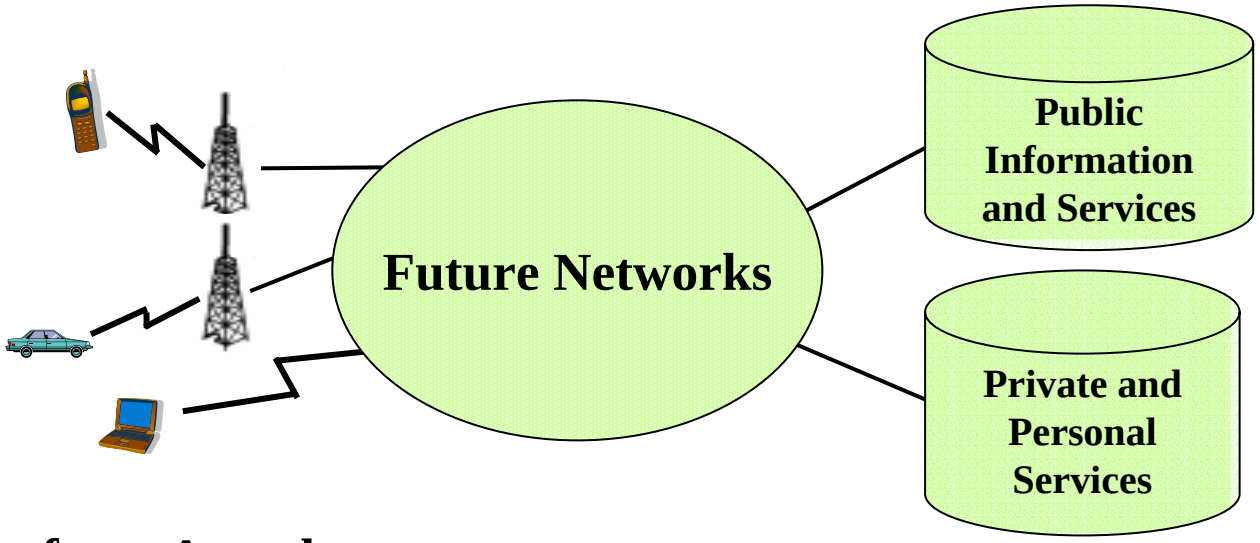
- **Speech Signals**
  - Carrying Linguistic Knowledge and Human Information: Characters, Words, Phrases, Sentences, Concepts, etc.
  - Double Levels of Information: Acoustic Signal Level/Symbolic or Linguistic Level
  - Processing and Interaction of the Double-level Information

Future Network Era



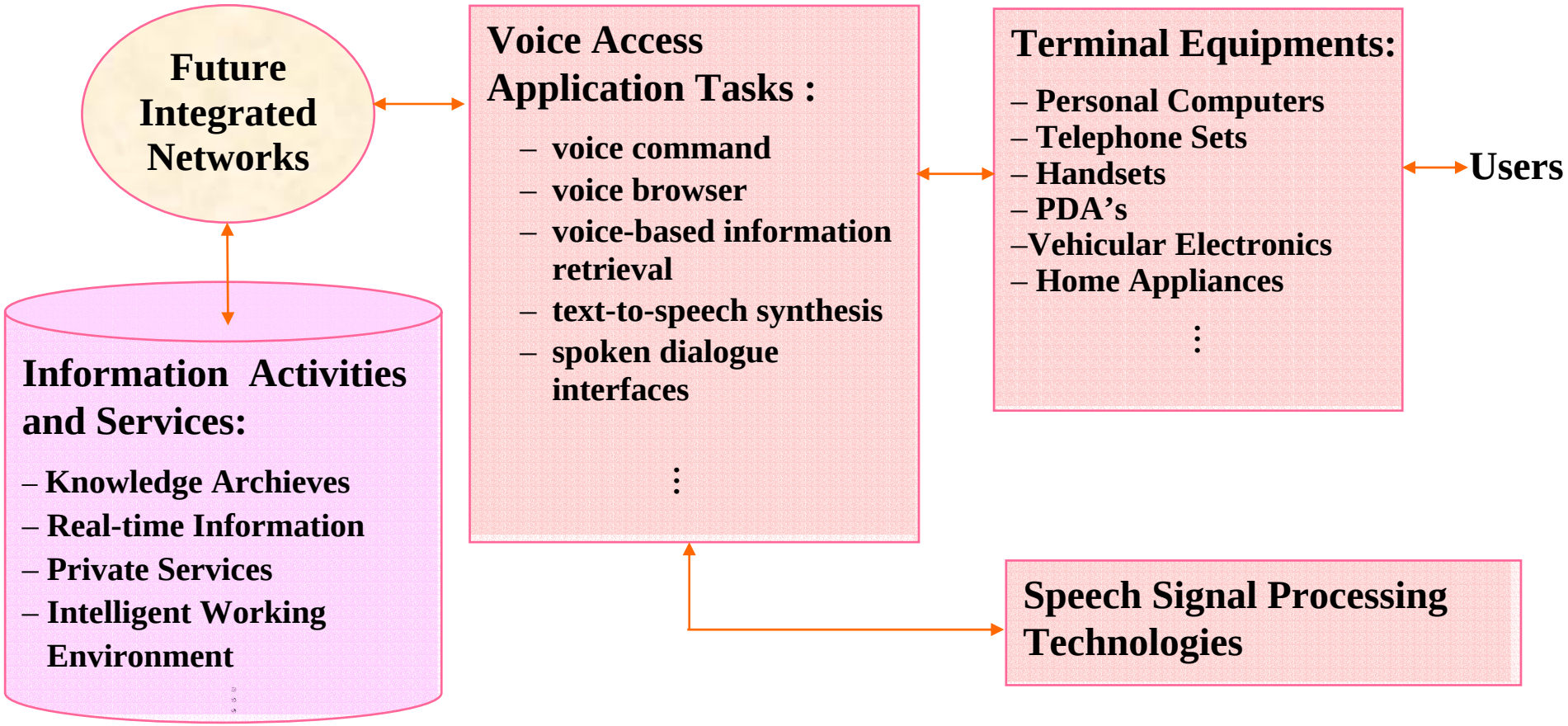
- **Multi–media, Multi–lingual, Multi–functionalities**
- **Cross–cultures, Cross–domains, Cross–regions**
- **Integrating All Knowledge Systems and Information–related Activities and Services Globally**

Wireless Communications Technologies are Creating a Whole Variety of User Terminals



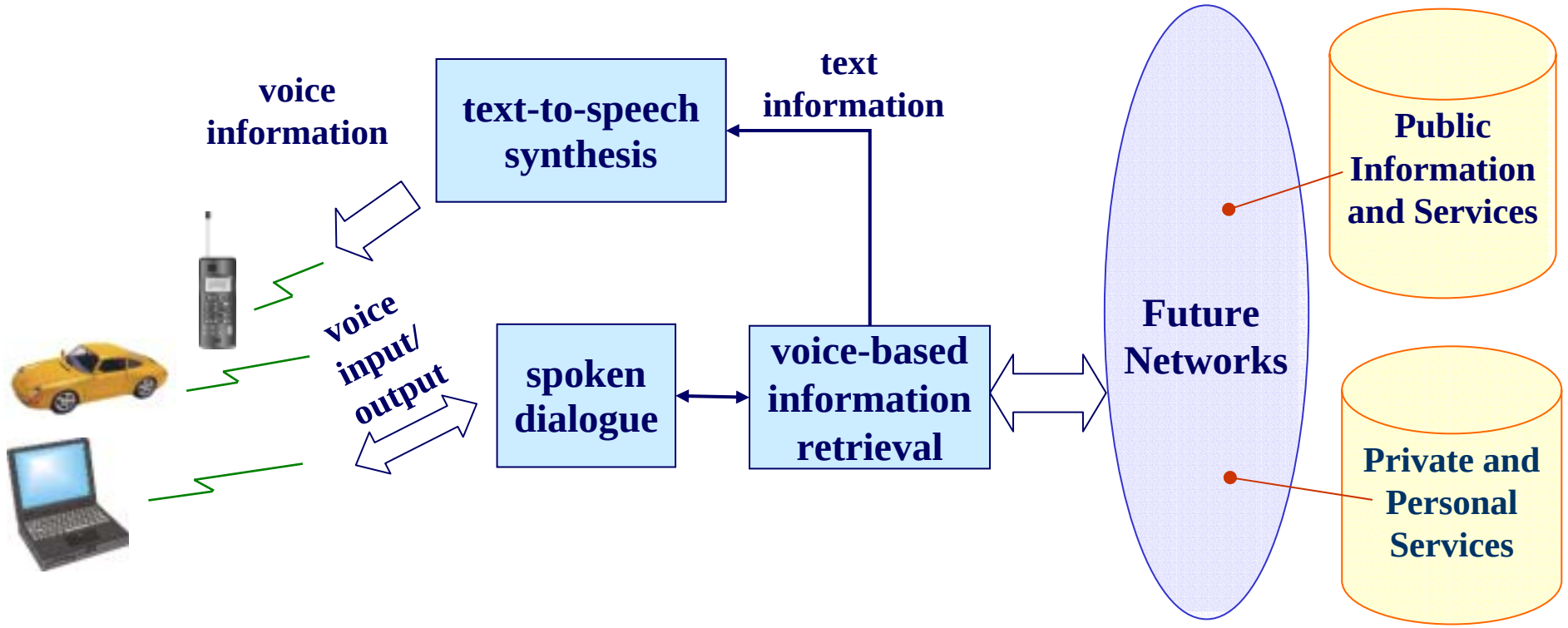
- **at Any Time, from Anywhere**
- **Personal Notebooks, Handsets, Hand-held Devices, PDA's, Vehicular Electronics, Hands-free Interfaces, Home Appliances, Wearable Devices…**
- **Small in Size, Light in Weight, Ubiquitous, Invisible…**
- **Popularity of Personal Computers Continuously Diminished — Evolving towards a "Post-PC Era"**
- **Keyboard/Mouse Most Convenient for PC's not Convenient any longer**
  - human fingers never shrink, and application environment is changed
- **Voice is the Only Interface Convenient for ALL User Terminals at Any Time, from Anywhere**

Evolution of Speech Processing Technologies towards a Future Network Era



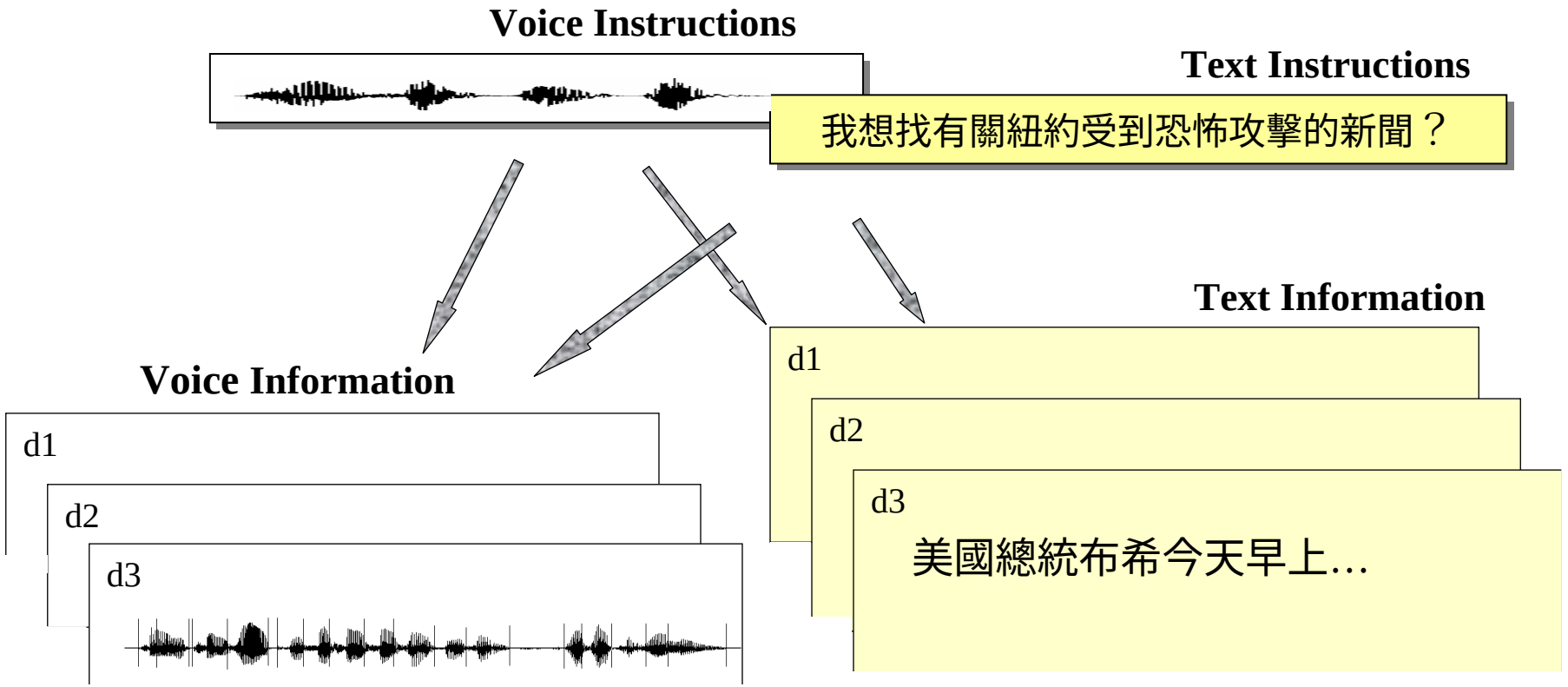
- From " PC-Era " to " Post-PC Era "
- From PC-based Document-Processing Applications to Network-based Information/Service-Access Applications

Voice Access of Network Information — Voice may Take the Place of Texts in the Future



- Network Access is Primarily Text-based today, but almost all Roles of Texts can be Replaced by Voice in the Future
- Human-Network Interactions can be Accomplished by Spoken Dialogues
- Voice-based Information Retrieval/Spoken Dialogues

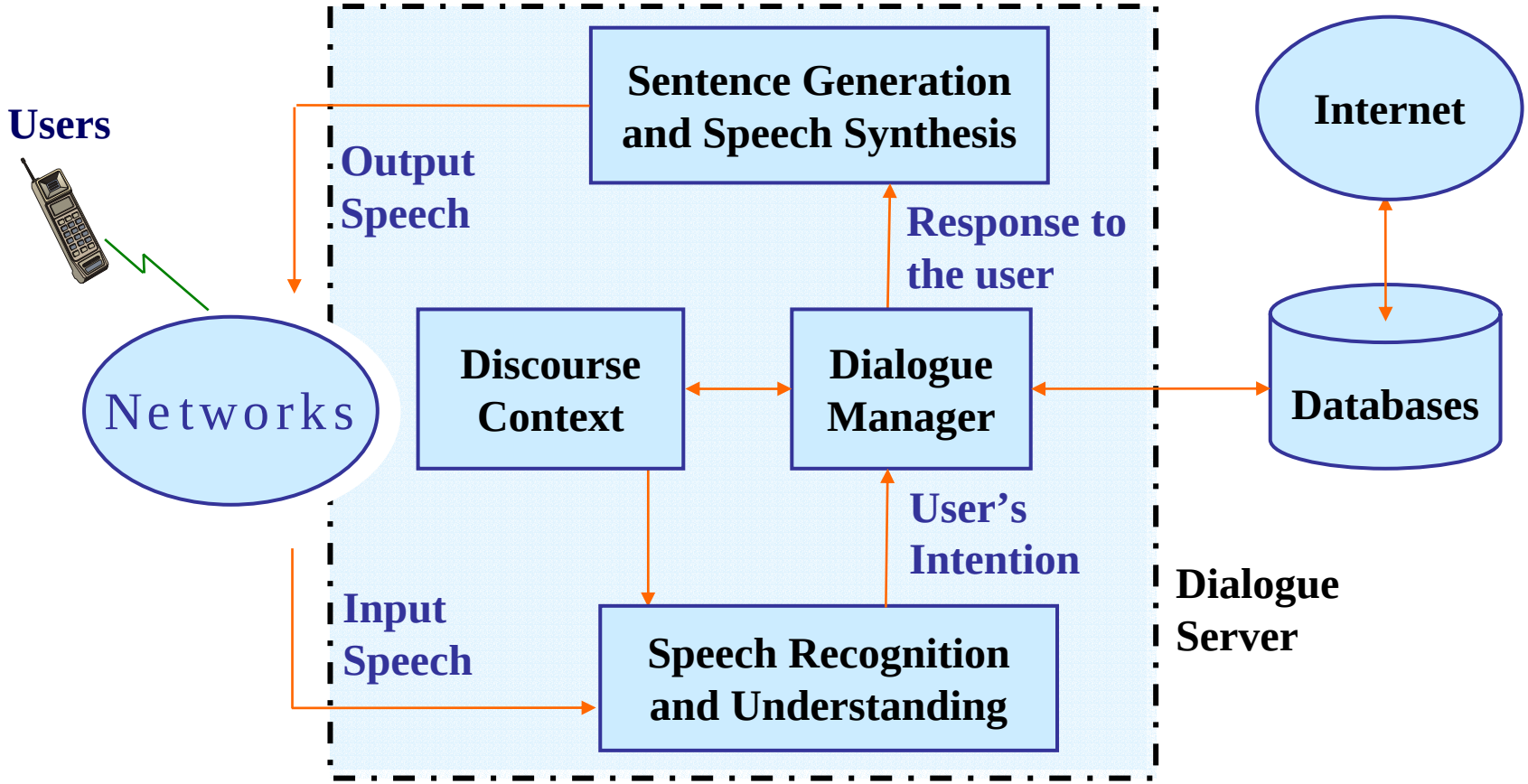
Voice-based Information Retrieval



- **Speech may become a New Data Type, if the Difficulties in Browsing and Retrieval can be Overcome**
- **Application Examples: Personal Memo、 Meeting Minutes、 Personal Phone Records、 Voice Mail Databases、 Course Lectures Databases、 Broadcast Programs …**

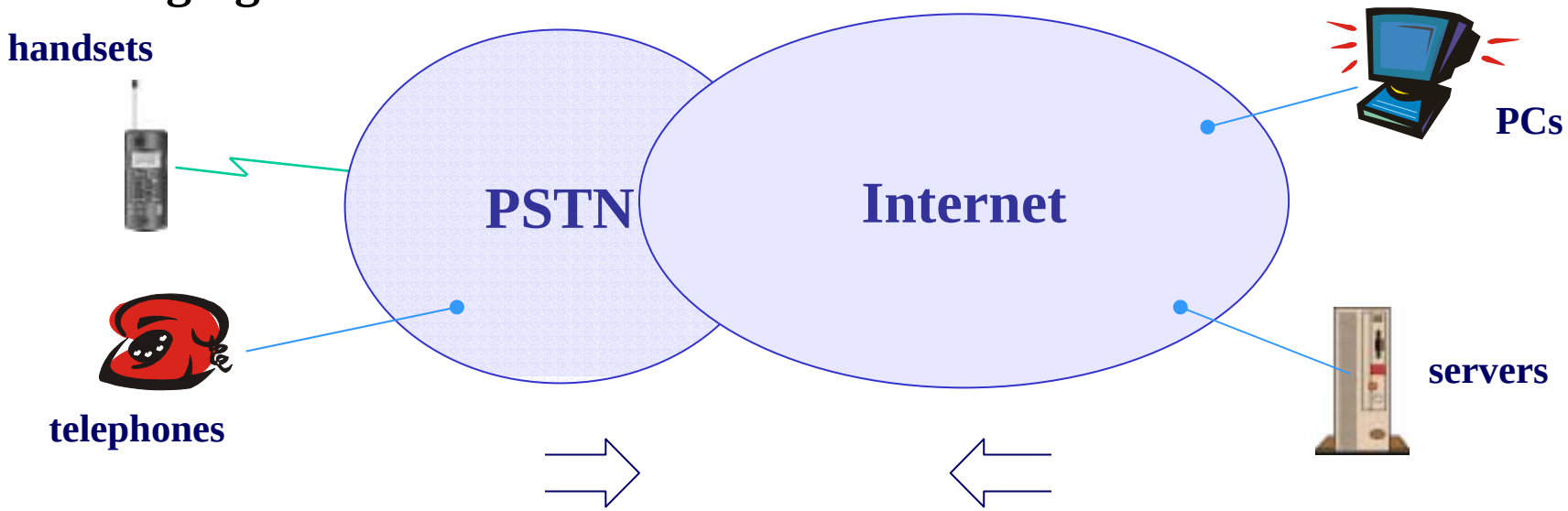
Spoken Dialogue

- **Almost All Human-network Interactions can be Accomplished by Spoken Dialogues**
- **An Example of Client-Server Computing Environment**



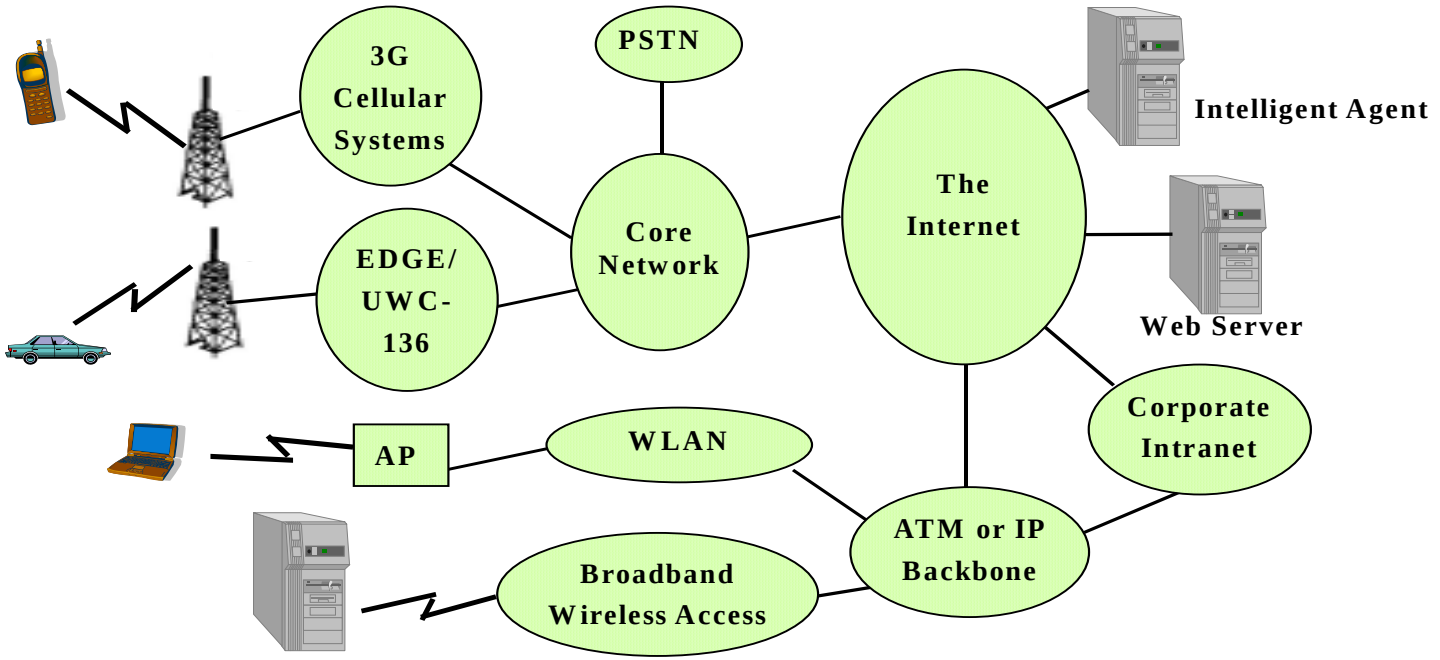
Convergence of PSTN and Internet

- **PSTN (for Voice) and Internet (for Data and Multi-media Contents) are Converging**



- **Driving Force for the Convergence**
  - "anywhere, any time" of wireless services
  - voice provides the most convenient and natural interaction interface
  - attractive contents over the Internet
  - contents (human information) are why the Internet is attractive, while voice directly carries human information
  - Speech-enabled Access of Web-based Applications

Wireless Access of Global Information



- **Global Information**
- **At Any Time, from Anywhere**
- **As Handset Size Shrinks While Required Functionalities Grows and the User Environment Changes, Voice Interface will be Useful for all Different User Terminals**
- **Integration of Many Different Technologies**
  - information processing, networking, transmission, internet, wireless, speech processing

Voice Interface for Human-network Interaction

– huge volumes of data disseminated across the globe by optical fiber networks

– any time, from anywhere by wireless terminals

– vehicular electronics, PDA, handset, home appliance, etc.
  new platforms accessing the global network information/services

– traditional keyboard/mouse not adequate any longer
  size shrinkage, different user environment, etc.
  desired functionalities/human–network interactions increasing

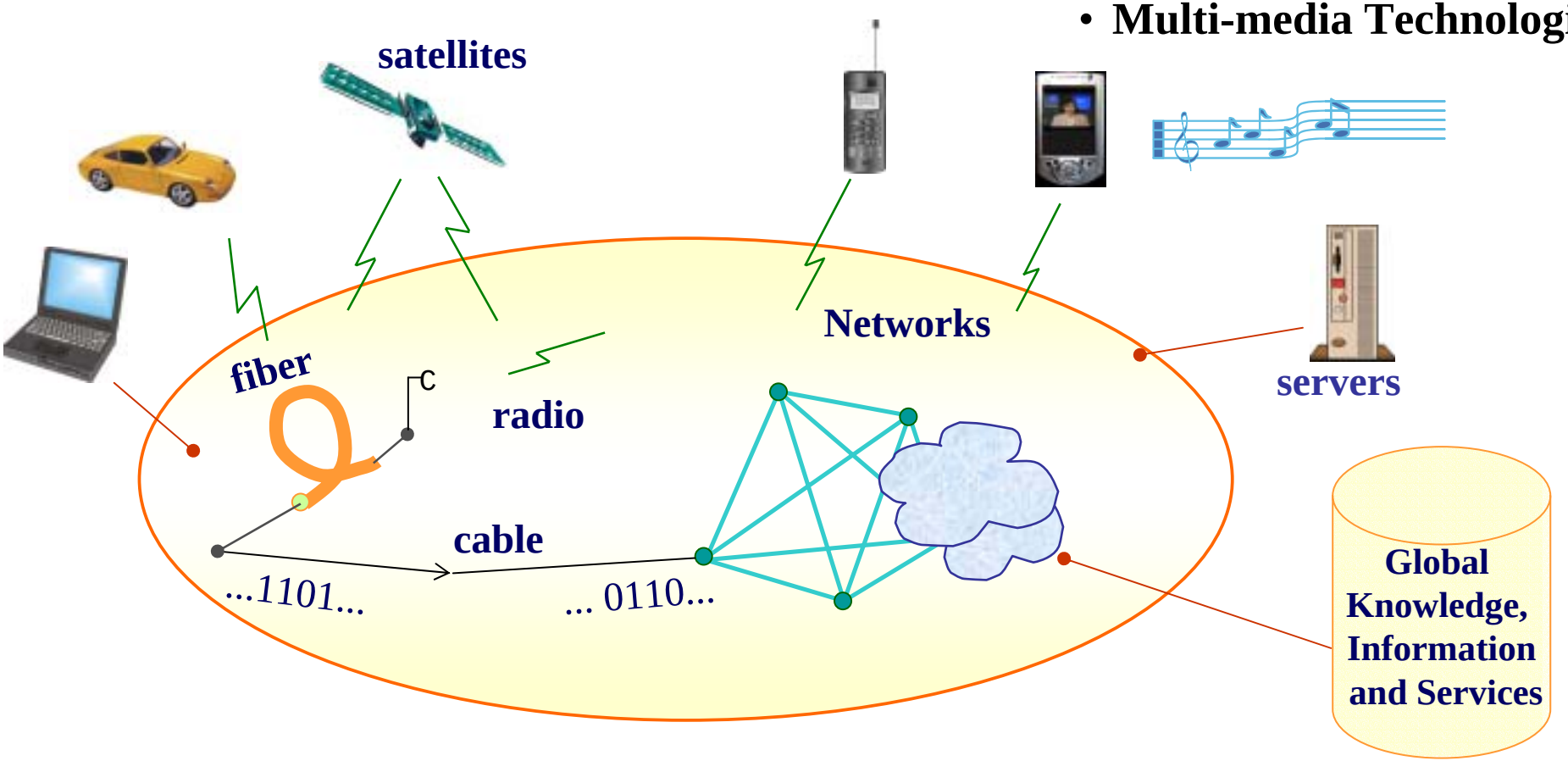
– voice interface will be one out of the few most important, natural, user friendly, attractive interface

– examples: voice retrieval, voice browser, voice portal, voice web
  speech–enabled access of Web-based applications
  voice–based web tools/application interfaces, etc.

– voice interface is the only major “missing link” in the “semi–mature” technology chain

Future World of Communications and Computing

- Wireless Technologies
- Speech Processing Technologies
- Multi-media Technologies



- Communications and Networking Technologies
- Information Processing Technologies

Outline

- **Both Theoretical Issues and Practical Problems will be Discussed**
- **Starting with Fundamentals, but Entering Research Topics Gradually**
- **Part I: Fundamental Topics**
    1. Introduction
    2. Basic Concepts in Speech Recognition
    3. Hidden Markov Models (HMM's)
    4. Acoustic Modeling
    5. Speech Signal Representation and Feature Parameters
    6. Language Modeling
    7. Linguistic Decoding and Search Algorithm
    8. Keyword Spotting
- **Part II: Advanced Topics**
    1. Speaker Adaptation/Recognition/Verification
    2. Robustness with respect to Noise & Channel Distortion
    3. Pronunciation Modeling
    4. Advanced Topics in Linguistic Processing
    5. Speech and Language Understanding
    6. Text-to-speech Synthesis
    7. Voice-based Information Retrieval
    8. Spoken Dialogues
    9. Distributed Speech Recognition and Wireless Environment

Outline

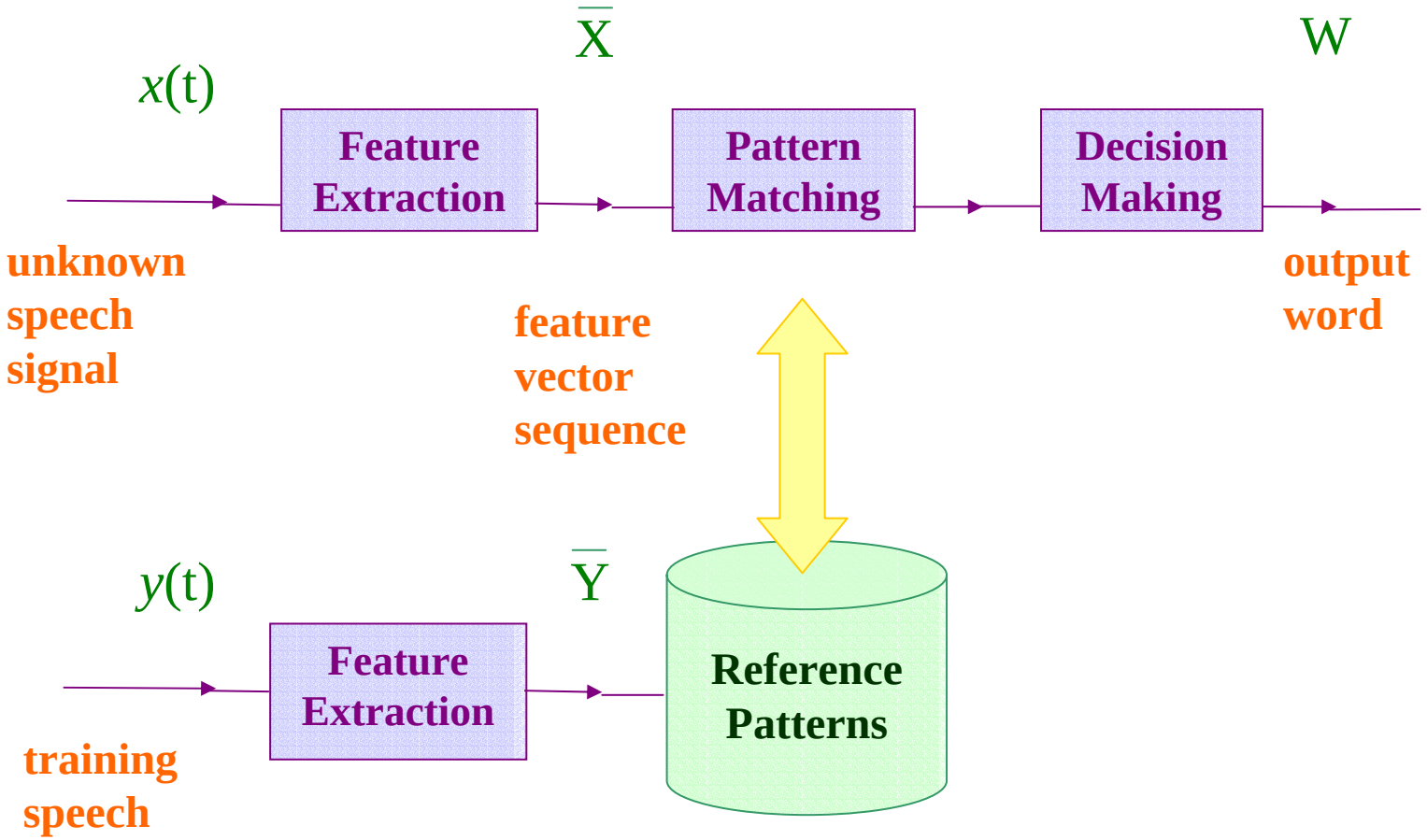
- **教科書：無**
- **主要參考書：**
  1. X. Huang, A. Acero, H. Hon, “Spoken Language Processing”, Prentice Hall, 2001,松瑞
  2. F. Jelinek, “Statistical Methods for Speech Recognition”, MIT Press, 1999
  3. L. Rabiner, B.H. Juang, “Fundamentals of Speech Recognition”, Prentice Hall, 1993, 民全
  4. C. Becchetti, L. Prina Ricotti, “Speech Recognition- Theory and C++ implementation”, Johy Wiley and Sons, 1999, 民全
  5. 其他參考文獻課堂上提供
- **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）、研（電信所、資研所、電機所）**
- **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**
- **成績評量方式**

Midterm	35%
Homeworks	15%
Final Report	50%

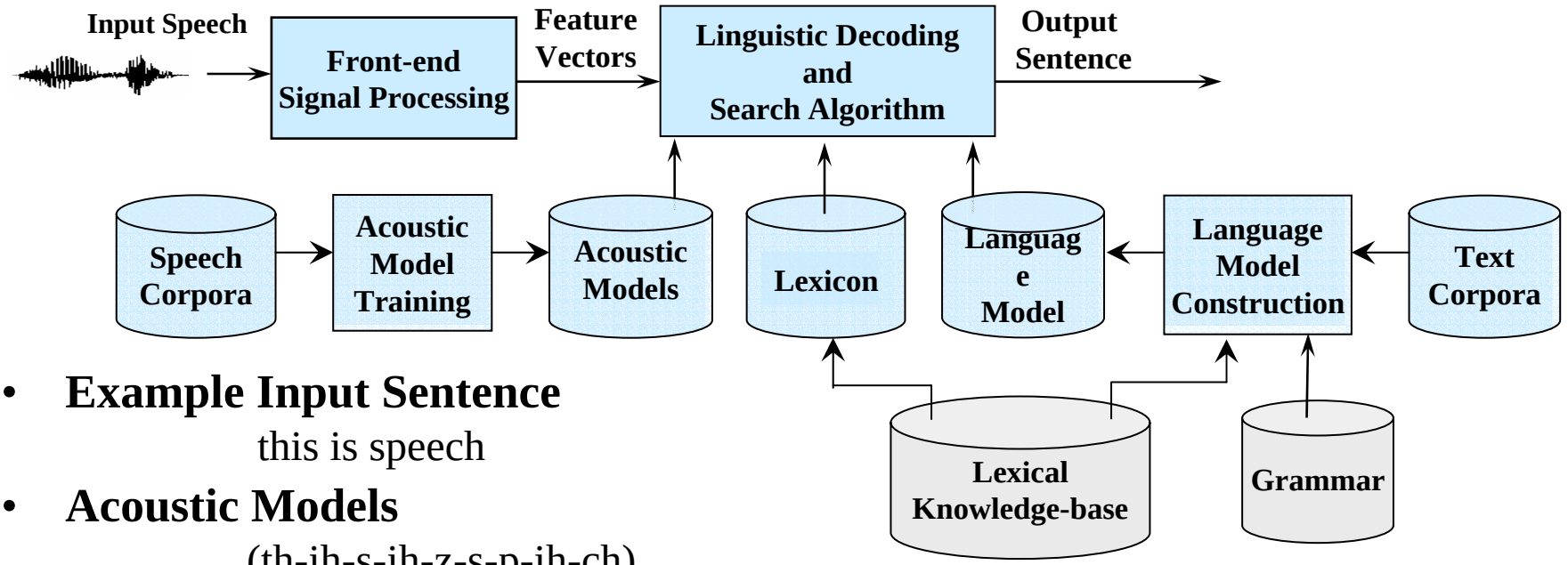
1.0 Introduction — A Brief Summary of Core Technologies and Current Status

Speech Recognition as a pattern recognition problem



Basic Approach for Large Vocabulary Speech Recognition

- **A Simplified Block Diagram**



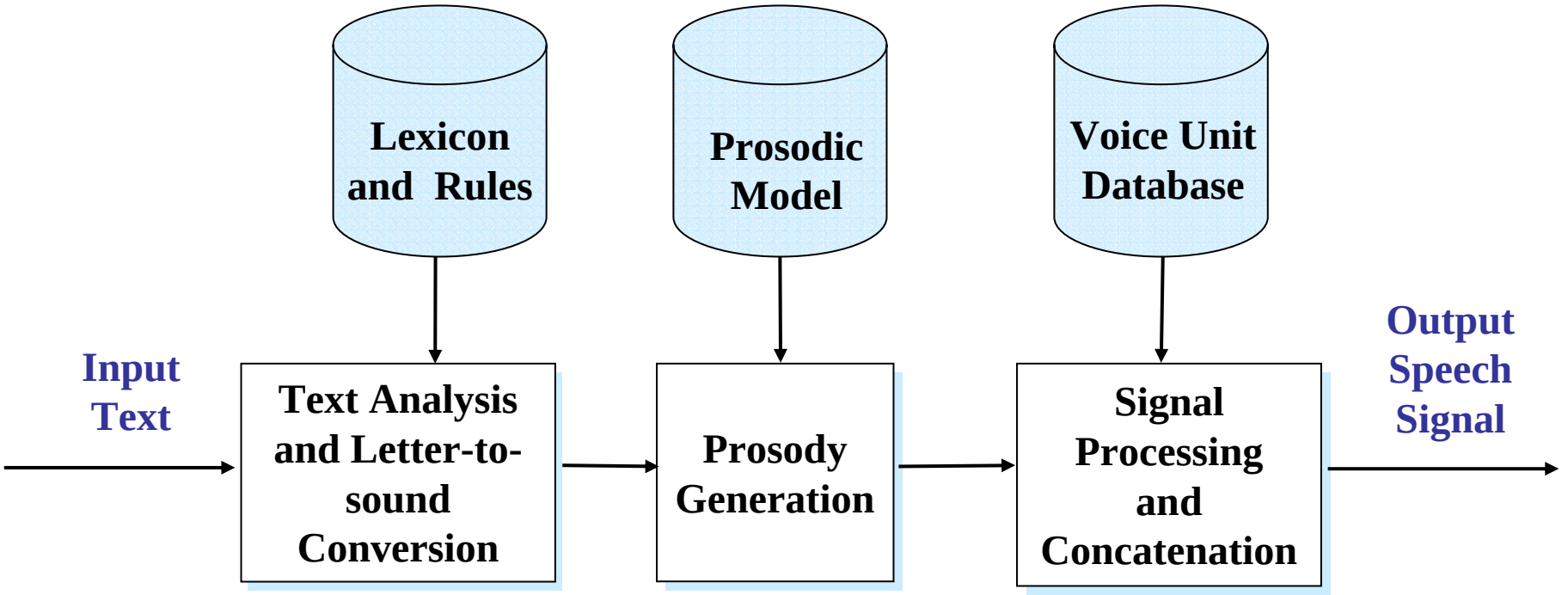
- **Example Input Sentence**
  this is speech
- **Acoustic Models**
  (th-ih-s-ih-z-s-p-ih-ch)
- **Lexicon** (th-ih-s) → this
  (ih-z) → is
  (s-p-iy-ch) → speech
- **Language Model** (this) – (is) – (speech)
  P(this) P(is | this) P(speech | this is)
  $P(w_i|w_{i-1})$ bi-gram language model
  $P(w_i|w_{i-1},w_{i-2})$ tri-gram language model,etc

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**

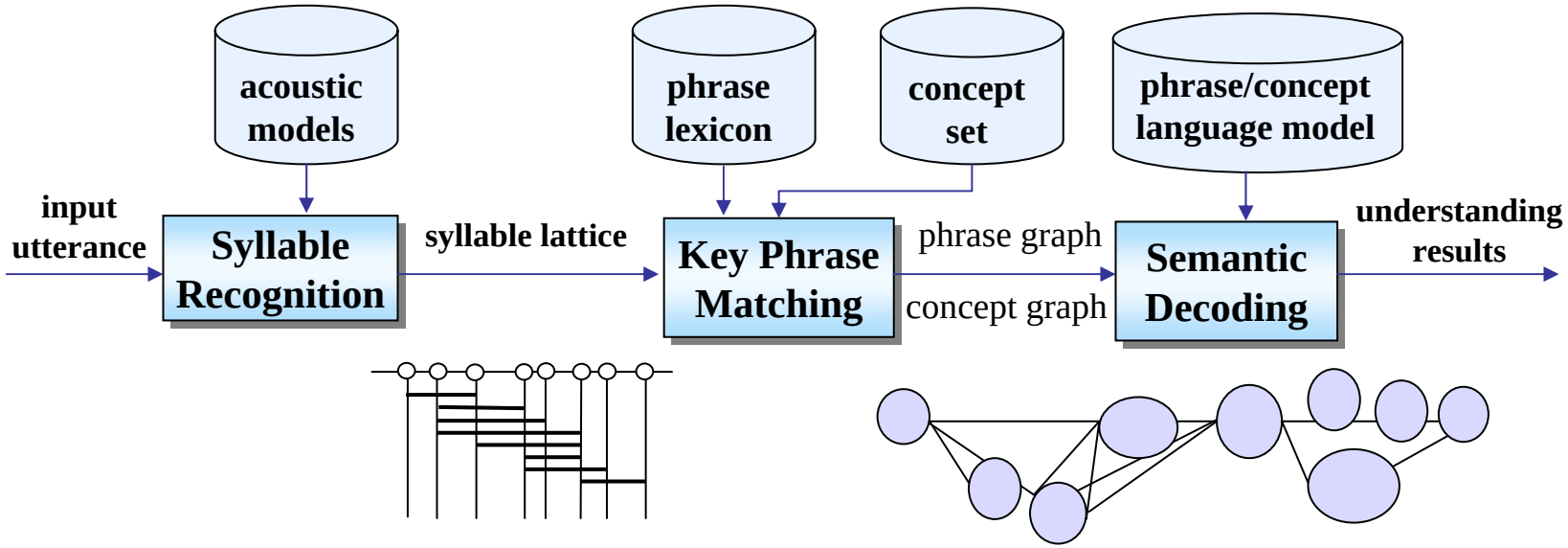
Text-to-speech Synthesis

- 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



Speech Understanding

- **Understanding Speaker's Intention rather than Transcribing into Word Strings**
- **Limited Domains/Finite Tasks**
- **Grammatical Approaches (e.g. partial parsing)/Statistical Approaches (e.g. corpus-based by training)**
- **Semantic Concepts/Key Phrases**



Prob (C_i | C_{i-1}, C_{i-2})

Prob (ph_j | C_i)

- **An Example**

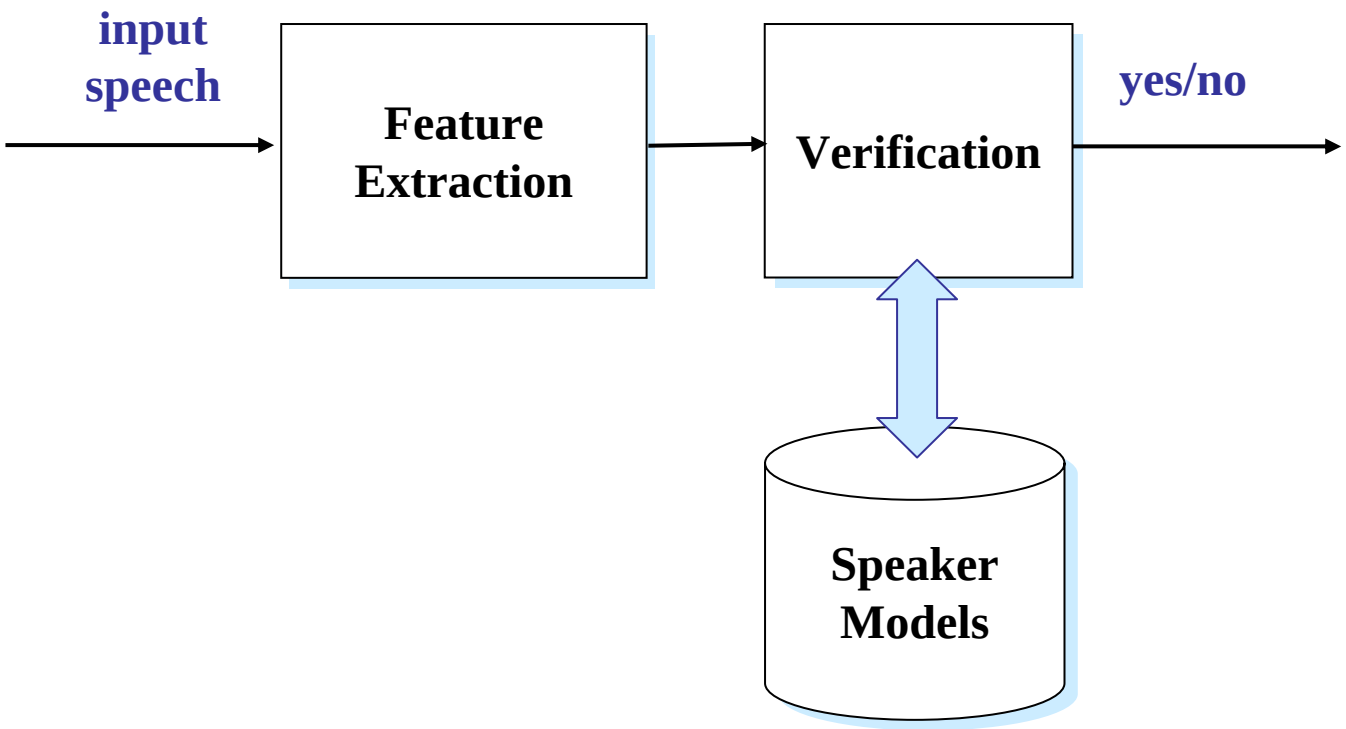
utterance:請幫我查一下 台灣銀行 的 電話號碼 是幾號?

key phrases: (查一下) - (台灣銀行) - (電話號碼)

concept: (inquiry) - (target) - (phone number)

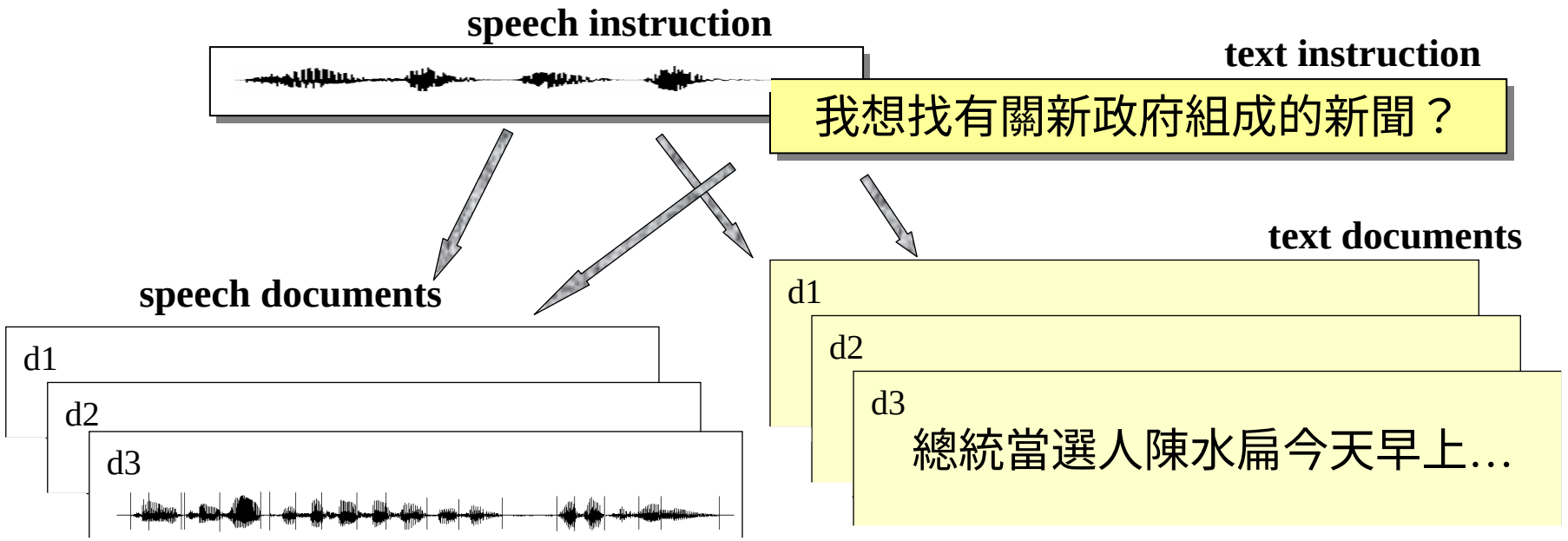
Speaker Verification

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



Speech-based Information Retrieval

- **Speech Instructions**
- **Speech Documents (or Multi-media Documents including Speech Information)**
- **Voice Personal Memo, Private Database , Meeting Minutes, Personal Phone Records.....**
- **Indexing Features/Relevance Evaluation**
- **Recall/Precision Rates**

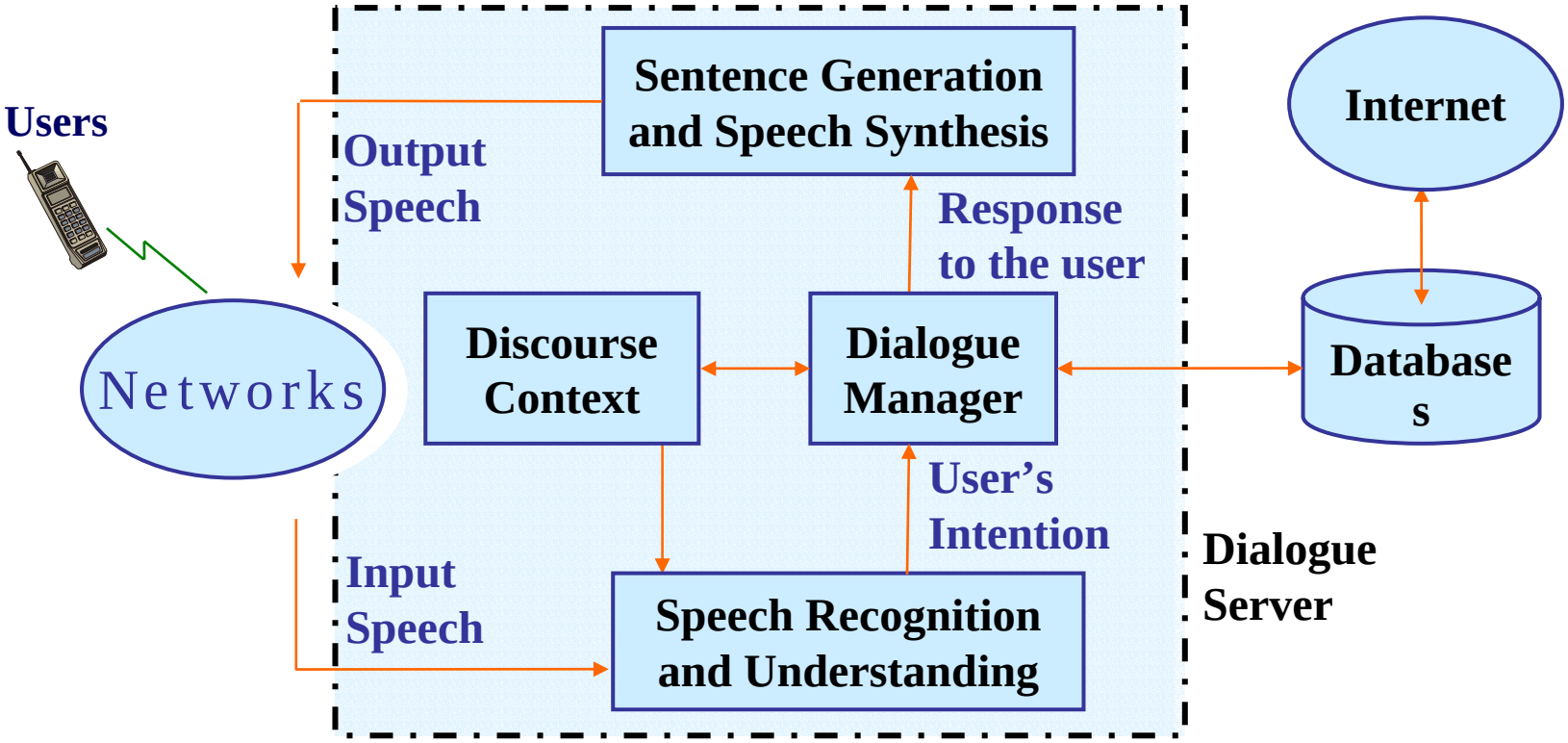


Multi-lingual Functionalities

- **Code-Switching Problem**
  - English words/phrases inserted in spoken Chinese sentences as an example
    人人都用Computers，家家都上Internet
  - the whole sentence switched from Chinese to English as an example
    準備好了嗎？Let's go!
- **Cross-language Network 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**
- **Shared Acoustic Units/Integrated Linguistic Structures**

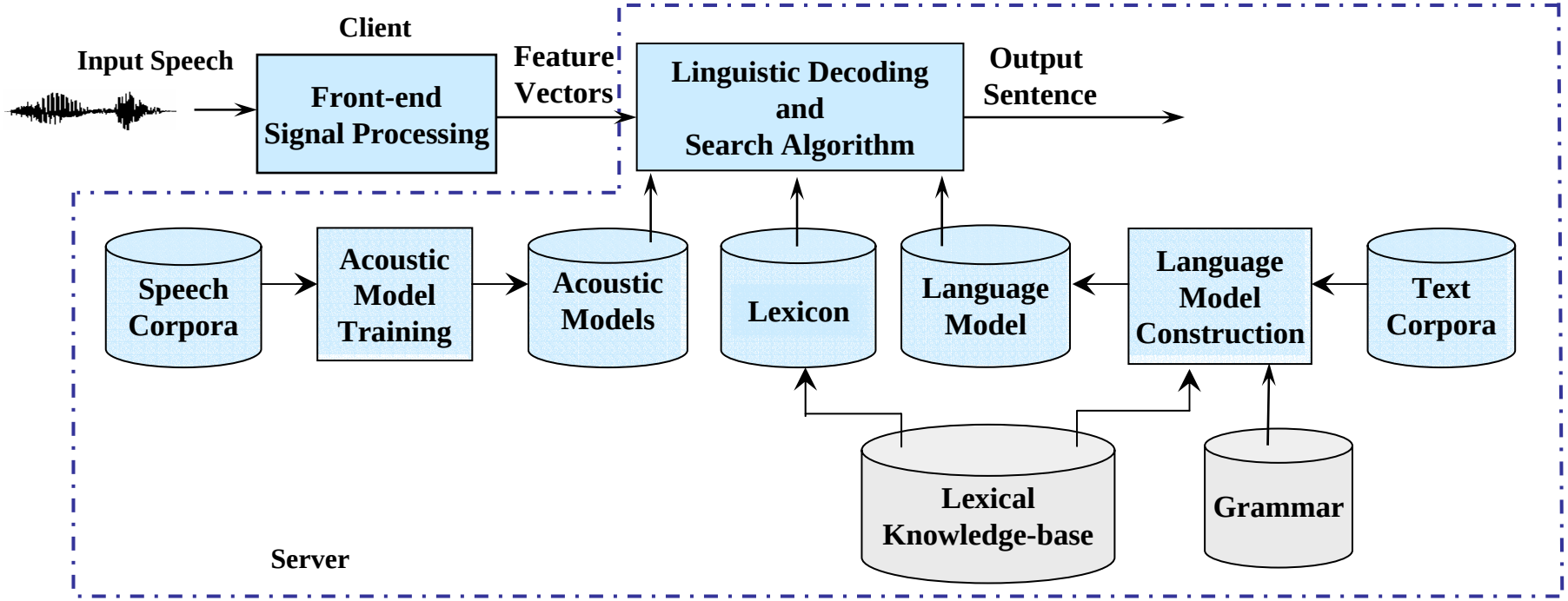
Spoken Dialogue Systems

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



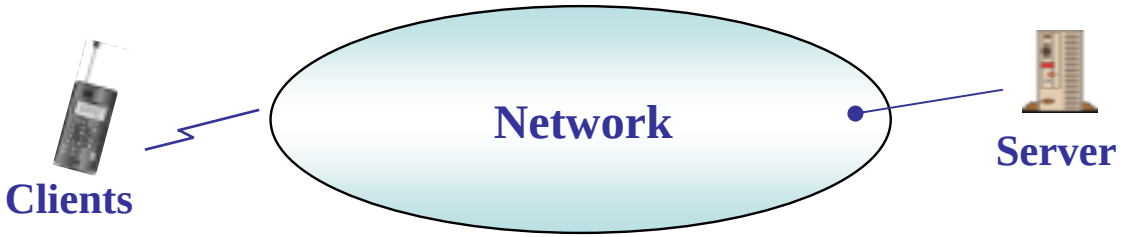
Distributed Speech Recognition (DSR) and Wireless Environment

- **An Example Partition of Speech Recognition Processes into Client/Sever**

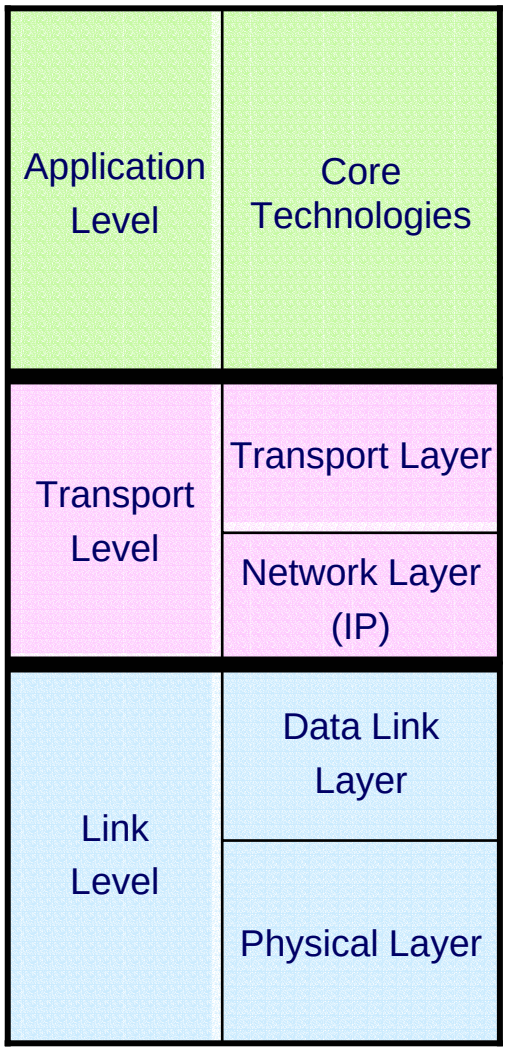


– encoded feature parameters transmitted in packets

- **Client/Server Structure**

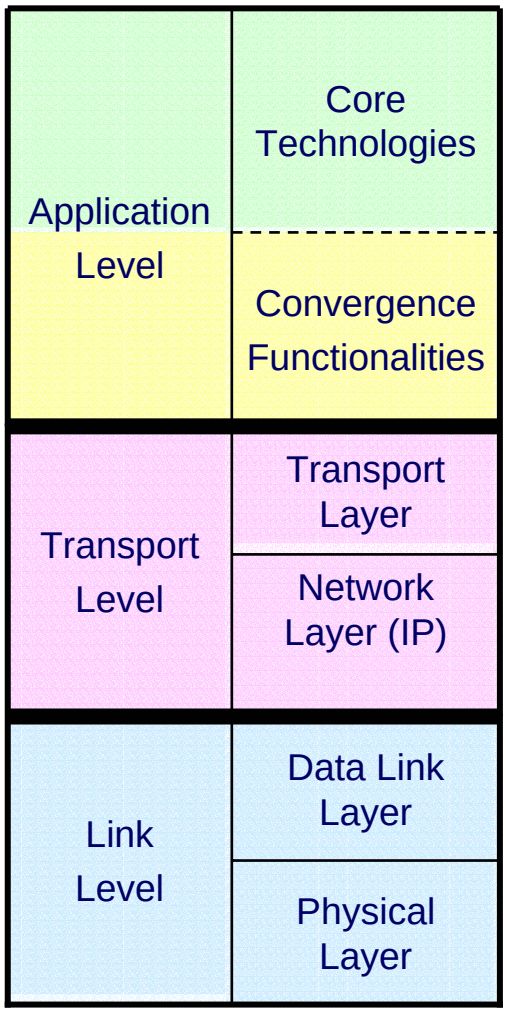


Distributed Speech Recognition (DSR) and Wireless Environment



- **Wireless Environment**
  - examples: Personal Area Networks (Bluetooth, etc.), Wireless LAN (IEEE 802.11), Cellular (GSM, GPRS, 3G), etc.
- **Link Level**
  - time-varying fading and noise characteristics
  - time-varying signal level and signal-to-noise ratios
  - bursty errors with much higher error rates
  - much smaller and dynamic bandwidth, much lower and changing bit rates
- **Transport Level**
  - TCP/IP: errors ⇨ retransmission ⇨ delay
  - UDP/IP: errors ⇨ real-time/no delay ⇨ packet loss
  - packets out of sequence

Distributed Speech Recognition (DSR) and Wireless Environment



- **Developing "Convergence Functionalities" in the Application Level**
  - wireless link level problems and transport level variations become transparent to "core technologies"
  - "core technologies" shielded from and equally applicable to all different link/transport environments
  - examples: packet re-sequencing, error concealment,etc.
- **Robust "Core Technologies"**
  - Combatting with residual errors
  - Examples: signal verification,error resilience, etc.