AUTOMATIC SPEECH RECOGNITION

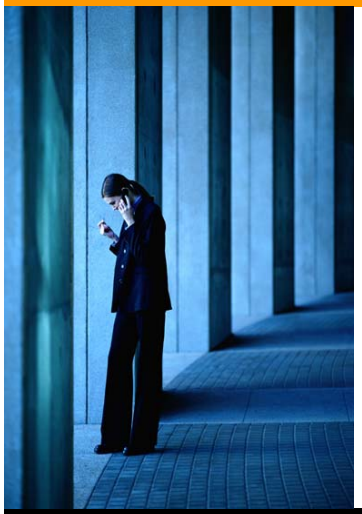
Lecture 1

Overview of Speech Technology and ASR

General Information

- Lecturer: Atiwong Suchato
- email: atiwong.s@chula.ac.th
- TA: Isarun Chamveha (Klui)
- Course Webpage:
  http://www.cp.eng.chula.ac.th/~atiwong/2110432

Speech Communication



- Human use speech as the major mean to communication to other people.

Speech Communication

Speech Signal in Different Representations

- The brain controls various articulators in the vocal tract to create the sound wave in some systematic forms (language).
- Auditory system sends nervous signal according to the received sound wave to the brain.

Speech and Language-related Technologies

- Human-machine interaction
  - Automatic Speech Recognition
  - Speech Synthesis
  - Text-to-Speech (TTS)
  - Natural Language Understanding
  - Natural Language Generation
- Telecommunication
  - Speech Coding
  - Speech Enhancement
- Human speech communication
  - Speech-to-speech Language Translator
  - Text-to-text Language Translator

Department of Computer Engineering, Chulalongkorn University Spoken Language Systems Research Group

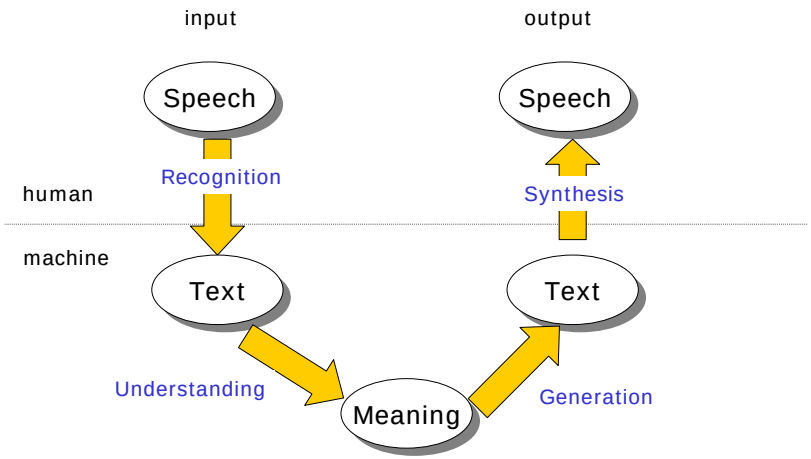
Speech and Language-related Technologies

- Security
  - Speaker ID
  - Speaker Verification
  - Speaker Authentication
- Speech and Hearing Disorder
  - Speech Synthesis
  - Hearing Aid
- Speech Manipulation
  - Speaking Rate Adjusting
  - Special Effect

2110432 Auto. Speech. Recog. | First Semester 2008 | Lecture 1
ATIWONG SUCHATO

Department of Computer Engineering, Chulalongkorn University Spoken Language Systems Research Group

Human-machine comm. via spoken language



MIT Open Course Ware

2110432 Auto. Speech. Recog. | First Semester 2008 | Lecture 1
ATIWONG SUCHATO

Department of Computer Engineering, Chulalongkorn University Spoken Language Systems Research Group

Speech-based Interface Applications

- Mostly input (recognition only)
  - Simple Command and Control
  - Simple Data Entry (Over the phone)
  - Dictation
- Interactive Conversation (understanding needed)
  - Information kiosks
  - Transactional processing
  - Intelligent agents

MIT Open Course Ware

2110432 Auto. Speech. Recog. | First Semester 2008 | Lecture 1
ATIWONG SUCHATO

Benefits of Speech Interface

- requires no special training
- leaves hands free
- leaves eyes free
- spoken language has high data rate

Limitation of Speech Interface

- cannot be used in noisy environment
- unnatural for some tasks (coding?)
- create distraction or annoyance in some situations
- confidential information

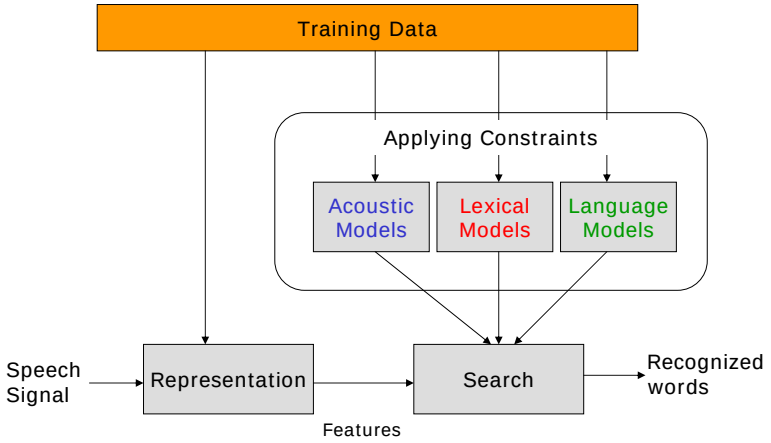
ASR vs. Touchtone

- Benefits of using an automated system, such as touchtone telephone and ASR, in stead of human customer representatives?
- Problems of touchtone system? Can ASR overcome each of the problem?
- Problem caused by ASR that will not happen if the touchtone system is used?
- Ideal ways for a call center to handle incoming calls?

Automatic Speech Recognition

- Uncover words from acoustic signal

Major Components of an ASR System

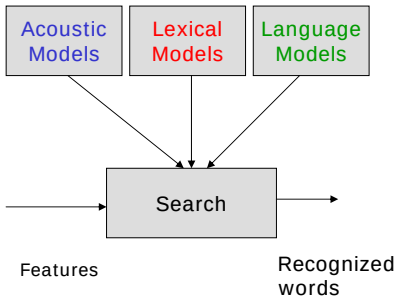


MIT Open Course Ware

Representation

- must select the right features for the representation used for the desired task
- tasks:
  - recognition of fruits in a basket containing mangos, guavas, durians, and coconuts
    - weight?
    - skin color?
    - meat color?
    - shape?
  - recognition of different sound classes?

Constraints



- Acoustic
  - signal generated by human vocal apparatus
- Lexical
  - Words in a language are limited.
- Language
  - A sentence must be syntactically and semantically well formed.

Domain Constraint

- Domain example
  - Movie listing
  - Phone directory
  - Weather information
- Specifying domain makes lexical and language constraint more specific.

Characterizing ASR System Capability

Parameters	Range
Speaking Mode	Isolated Word – Continuous Speech
Speaking Style	Read Speech – Spontaneous Speech
Enrollment	Speaker Dependent – Speaker Independent
Vocabulary	Small (<20words) – Large (>50,000 words)
Language Model	Finite-state – Context-sensitive
Perplexity	Small (<10) – Large (>200)
SNR	High (>30dB) – Low (<10dB)
Transducer	Noise-canceling mic. – cell phone

ASR Trends

	before mid 70's	mid 70's–mid 80's	mid 80's - now
Recognition Units	whole-word and sub-word units	sub-word units	sub-word units
Modeling Approaches	heuristic and ad hoc (rule-based)	template matching, deterministic and data-driven	probabilistic and data-driven
Knowledge Representation	heterogeneous and complex	homogeneous and simple	homogeneous and heterogeneous
Knowledge Acquisition	intense knowledge engineering	simple structure	automatic learning

ASR performance (English)

- High performance, speaker-independent speech recognition is now possible
  - Large vocabulary for cooperative speakers in benign environments
  - Moderate vocabulary for spontaneous speech over the phone
- Commercial recognition systems are now available
  - Dictation
  - Telephone transaction
- When well-matched to applications, technology is able to help perform real work.

Well-matched Application?

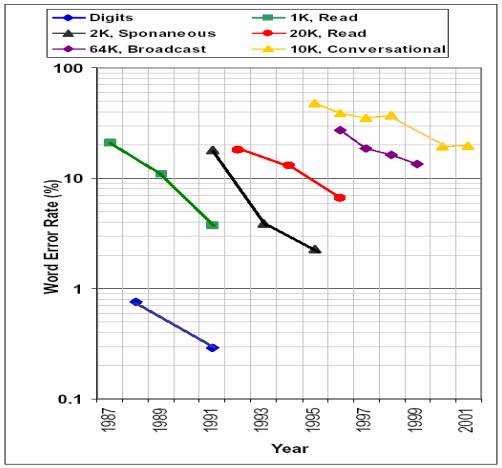
- Realistically, in the foreseeable future, do you expect an ASR system to give a 100% accuracy for an open domain application?
- To which application that ASR system is the least suitable?
  - a) home appliances control using voice command
  - b) putting a patient medical history into a database using conversational speech
  - c) an automatic dictation system
  - d) an ASR-automated call center

ASR performance (English)

- Speaker-independent, continuous speech ASR now possible
- Digit recognition over the telephone with word error rate of 0.3%
- Error rate cut in half every two years for moderate vocabulary tasks
- Error for spontaneous speech more than twice that of read speech
- Conversational speech, involving multiple speakers and poor acoustic environment, remains a challenge
- Tens of hours of training data to port to a different domain
- Statistical modeling using automatic training achieves significant advances

MIT Open Course Ware

ASR performance (English)



MIT Open Course Ware

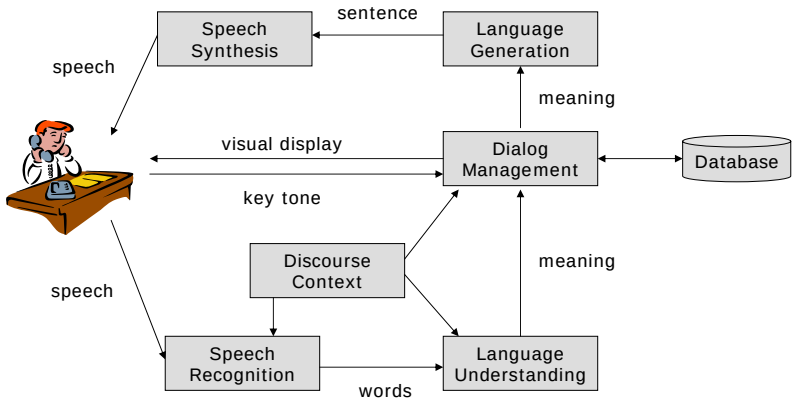
Challenges

- Co-articulation
- Speaker Independent
  - Gender / Children
  - Dialect
  - Non-native
- Spontaneous Speech
  - Disfluencies
- Out-of-Vocabulary Word
  - Fill words
  - Proper names
- Noise / Channel Robustness

A human-machine conversational system

- **System:** *Welcome to the "CU airways" flight information system. Do you know the flight number?*
- **Caller:** Yes.
- **System:** *What is it?*
- **Caller:** Three forty-one.
- **System:** *Thanks. Do you want arrival or departure information?*
- **Caller:** Departure.
- **System:** *Is it Departing Saturday, Sunday, or Monday?*
- **Caller:** Sunday.
- **System:** *Great. Hold on while I check on that flight.*
  - <system checks for the information from the database and report the info.>
- **System:** *Is there anything else?*
- **Caller:** No. Goodbye.
- **System:** *Thanks for calling "CU Airways". Goodbye.*

A human-machine conversational system



Voice Prompt Design

- a challenge by itself
- how to ask? how to convey complex info.?
- Social-Psychological problem
  - male or female voice
  - humorous
  - naturalness