Introduction to Spoken Language Processing/Systems

Hsin-min Wang

References:
1 X. Huang et al., Spoken Language Processing, Chapter 1
Motivations

- Speech communication has been and will be the dominant mode of human social bonding and information exchange -> the human preference for spoken language communication finds a reflection in human-machine interaction as well.
- A spoken language system needs to have both speech recognition and speech synthesis capabilities. In addition, an understanding and dialogue component is required to manage interactions with the user.
- The goal of building commercially viable spoken language systems has long attracted the attention of scientists and engineers all over the world.
Motivations (cont.)

- **spoken language interface ⇔ graphical user interface**
  
  **Speak and listen**
  **Understand and learn**
  **GUI:** windows, icons, menus, pointers
  **Keyboards strokes and mouse clicks**
  **Display monitor for feedback**
  **Touch-screen, joystick**

  *How to raise the voice output volume?*
  *Wearable computers?*
  *Driving?*
  *Speech-only-device: telephone?*

- **Speech-to-speech Translation**
  - Need multilingual speech recognition, synthesis, and understanding

- **Multimodal interactions**
  - Speech, visual orientation, gesture, facial expression…
Spoken Language System Architecture

- **Spoken language processing** refers to technologies related *speech recognition, text-to-speech, and spoken language understanding*.

- **Speech recognition** converts speech into words.
- **Text-to-speech synthesis** conveys spoken information.
- **Spoken language understanding** maps words into actions and plans system-initiated actions.
Automatic Speech Recognition

A source-channel model for a speech recognition system
Acoustic models include the representation of knowledge about acoustics, phonetics, microphone and environment variability, gender and dialect differences among speakers, etc.

Language models refers to a system’s knowledge of what constitutes a possible words, What words are likely to co-occur, and in what sequence
Text-to-speech

Raw text or tagged text

Text Analysis
- Document Structure Detection
- Text Normalization
- Linguistic Analysis

Phonetic Analysis
- Grapheme-to-Phoneme Conversion

Prosodic Analysis
- Pitch and Duration Attachment

Speech Synthesis
- Voice Rendering

TTS Engine
Basic Architecture of a Spoken Language Understanding System
Statistical Modeling Paradigm in Speech Recognition

Training Data → ANALYSIS → Feature Sequence → TRAINING ALGORITHM → STATISTICAL MODEL → RECOGNITION SEARCH

Ground Truth

TRAINING

RECOGNITION

Input Data → ANALYSIS → Feature Sequence → Recognized Sequence
Speech Recognition - Problem Definition

\[ W^* = \arg \max_W P(W \mid X) \]

\[ = \arg \max_W \frac{P(X \mid W)P(W)}{P(X)} \]

\[ = \arg \max_W P(X \mid W)P(W) \]
Speech Recognition - Acoustic Processing

Speech Waveform

Framing

Signal Processing

Feature vector sequence

\[ o_1, o_2, o_3, o_4, \ldots, o_t, \ldots, o_N \]

\[ s_1, s_2, s_3, s_4, \ldots, s_t, \ldots, s_M \]

Hidden Markov Model

\[ a_{ij} = P(s_t = j | s_{t-1} = i) \]

\[ b_i(o_t) = P(o_t | s_t = i) \]

\[ S^* = \arg \max_S P(O | S) \]

\[ W^* = \arg \max_W P(O | W) \]

\[ S^* = \arg \max \sum_{k=1}^{M} c_{ik} N(o_t; \mu_{ik}, \Sigma_{ik}) \]
Speech Recognition - Linguistic Processing

\[
p(W) = p(w_1, w_2, ..., w_i, ..., w_K)
\]

\[
= p(w_1) \cdot \left[ \prod_{2 \leq i \leq K} p(w_i \mid w_1, w_2, ..., w_{i-1}) \right]
\]

Assume \( p(w_i \mid w_1, w_2, ..., w_{i-1}) \approx p(w_i \mid w_{i-1}) \)

then \( p(W) = p(w_1) \cdot \left[ \prod_{2 \leq i \leq K} p(w_i \mid w_{i-1}) \right] \)

Assume \( p(w_i \mid w_1, w_2, ..., w_{i-1}) \approx p(w_i \mid w_{i-2}, w_{i-1}) \)

then \( p(W) = p(w_1) \cdot p(w_2 \mid w_1) \cdot \left[ \prod_{2 \leq i \leq K} p(w_i \mid w_{i-2}, w_{i-1}) \right] \)
Large Vocabulary Continuous Speech Recognition

Input Speech → Front-end Signal Processing → Feature Vectors → Linguistic Decoding and Search Algorithm → Output Sentence

Speech Corpora → Acoustic Model Training → Acoustic Models

Lexicon → Language Model → Language Model Construction → Text Corpora

Lexical Knowledge-base → Grammar
Historical Review

1952, isolated digits
(a single speaker)
Davis et al, Bell Labs.

1959, 4 vowels & 9 consonants
(statistical syntax for phoneme sequence)
Fry&Denes, Univ. College in England
10 vowels in /b/-vowel-/t/
(speaker-independent)
Forgie&Forgie, MIT Lincoln Labs.

1956, 10 monosyllabic words
(a single speaker)
Olson&Belar, RCA

1960s, Vintsyuk proposed Dynamic Time Warping

1970s Voice-activated typewriter
(dictation, speaker-dependent)
Jelinek, IBM

Telecommunication – voice dialing, command&control
(keyword-spotting, speaker-independent)
Bell Labs.

1980s Rapid development of statistical methods, HMMs

Nontelecommunication applications: dictation systems (IBM, Dragon Systems, L&H, Philips, etc.)

Telecommunication applications: AT&T’s VRCP applies a 5-word keyword spotter to automate billions of calls every year, resulting in saving operating cost in hundreds of millions of dollars
Progress of Speech Recognition

Fig. 7. Progress of spoken language technology along the dimensions of vocabulary size and speaking styles. Furui & Juang Proceedings of IEEE, 2000
Progress of Speech Recognition (cont.)

Benchmarks of ASR performance in word error rates in DARPA-sponsored tasks

Courtesy NIST 1999 DARPA
HUB-4 Report, Pallett et al.
Progress of Speech Recognition (cont.)

- Benchmarks of ASR performance: Broadcast News Speech
Uncertainties in Human Speech that Degrade Recognition Accuracy

- Pronunciation Variation
- Speech style
  - Speaking rate
  - Emotion
  - Accent
- Intra-speaker variability
- Inter-speaker variability
- Linguistic variability
- Variability caused by the environment
- Variability caused by the context

- Speaker-independency
  - Speaker-adaptation
  - Speaker-dependency
- Non-native speaker
  - Gender
  - Dialect
- Robustness Enhancement
  - Microphone
  - Background sound
- Context-Dependent Acoustic Modeling
  - Isolated word v.s. Continuous speech
  - Read speech v.s. Spontaneous speech
Golden Mandarin Dictation System

Golden Mandarin (I)
- Isolated syllable input mode
- Speaker dependent
- Transputer with 10 CPU’s
- Completed in March 1991

Golden Mandarin (II)
- Isolated syllable input mode
- Speaker adaptive
- PC with an extra DSP card
- Completed in Sept. 1993
Golden Mandarin Dictation System (cont.)

- Golden Mandarin (III) - Version 3.0
  - Complete sentence/Continuous speech input mode
  - Speaker adaptive
  - Sparc 20 workstation
  - Completed in March 1995
Golden Mandarin Dictation System (cont.)

- Golden Mandarin (III) - Version 3.1
  - Prosodic segment/Continuous speech input mode
  - Speaker adaptive
  - PC with an extra DSP card
  - Completed in March 1995
Golden Mandarin Dictation System (cont.)

- Golden Mandarin (III) - Version 3.2
  - Complete sentence/Continuous speech input mode
  - Speaker adaptive
  - Pentium PC with a Sound Blaster card
  - Completed in 1996

Training user interface

Dictation user interface

Testing user interface
Speech-based Information Retrieval

Input speech query
Record speech information

Spoken Dialogue

TTS

speech

text, video, audio, speech

text

Voice portal

Internet resources:
- broadcast radio
- stock price
- TV programs
- train info.

Personal information service:
- voice memo
- voice mail

Search engine

Video/audio search engine
Retrieval of Very Large Chinese Text Databases Using Continuous Mandarin Speech (A)
Retrieval of Very Large Chinese Text Databases Using Continuous Mandarin Speech (A) (cont.)

- Example: Internet News
  - example input query: 柯林頓的醜聞 (in voice)
Retrieval of Very Large Chinese Text Databases Using Continuous Mandarin Speech (B)

- Example: electronic Chinese dictionary (Ministry of Education)
  - 160,000 word entries
  - 20,000,000 words of information (explanations, example sentences, etc)

- Problems
  - the first character of the word may not be known to the user
  - the desired word may not be an entry, but similar words should be retrieved
  - retrieving all relevant words using some “general concept”

- Syllable-based architecture
- Database-specific language models provide higher accuracy
Retrieval of Very Large Chinese Text Databases Using Continuous Mandarin Speech (B) (cont.)

- Example: Electronic Chinese Dictionary
  - Example input query: 形容女子容貌美麗的詞彙 (in voice)
Telephone Directory Services

- Example: telephone directory services for financing/banking organization in Taipei
- Syllable-based architecture, corpus-statistics-based approach
- Speaker Act Type (SAT) determined by key phrase spotting and understanding, System Response Type (SRT) determined with a N-gram-based dialogue model
Telephone Directory Services (cont.)

example input queries:

喂! 好! 請查台灣銀行的電話

是的

我要的是中小企業金融部的電話

是的
TV/Radio News Retrieval System (A)

柏林陳，現於NTNU
TV/Radio News Retrieval System (B)

http://sovvideo.iis.sinica.edu.tw