Outline of Today’s Lecture

• I: Overview of course objectives and handout.
• II: Brief overview of the course
  – human/computer interaction
  – Origins of speech processing
  – Speech synthesis
  – Speech and auditory perception
  – Automatic speech recognition
  – Applications in industry
Human/Computer Interaction (HCI)

- Human-Human contact
- Human-Machine contact (WIMPy interface)
  - Windows, Icons, Mouse, Pointer
- Other examples
  - Writing pad, scanner, cell-phone, camera, wearable computer,
- Overall Goal: efficient, easy, and ergonomic human-computer interaction.

Speech as Human/Computer Interaction

- Speech and Hearing, most natural mode of human-human communication
- Speech can be used for human-computer interface as well.
- This class is an introduction to how such systems work.
- For a speech HCI system, need two things:
  - Speech synthesis (at most 1 lecture)
  - Speech recognition (rest of course)
Computer Speech Processing

Speech recognition and synthesis require learning about all of the following topics

- Human Anatomy & Physiology
- Phonetics
- Speech Production & Vocal Mechanism
- Physics
- Acoustics
- Human (Auditory) Perception
- Computational Auditory Scene Analysis
- Psychoacoustics
- Signal processing
- Information Theory
- Communications Theory
- Statistical Pattern recognition
- Probability & Statistics
- Multiple Languages and Language Universals
- Machine Learning
- Linguistics
- Text Processing
- Word Pronunciations & Dictionaries
- Human Language Learning
- Natural Language Processing
- Spoken Dialog Understanding
- Cognitive Science
- Artificial Intelligence
- Computer Science & Algorithm Design
- Human Computer Interface Design
- Software Engineering & Computer Architecture
- High-performance & Parallel computing
- Fixed-point DSP micro-processors

Origin of Speech Processing

- Key original motivation is simple bandwidth reduction – how to describe the intelligibility information on speech as cheaply as possible.
- 0-4kHz => 8kHz sps, 8 bits => 64k bps vs. 1.2k bps for modern speech coders
- Cell phones – goal: reduce bandwidth
  - LPC – linear prediction coefficients
  - CELP – code excited linear prediction
  - SOLA (synchronous overlap & add)
    - Reduce time-span w/o reducing spectral quality or pitch.
Origins of Speech Processing

• Desire to build a speaking machine:
  – Wolfgang Ritter von Kempelen (late 1700s)

  *The speaking machine of Wolfgang von Kempelen*, Dudley & Tarnoczy, JASA 22, 1950

Figure shows Wheatstone’s version, 1835.

Origins of Speech Processing

• “If I could determine what there is in the very rapidly changing complex speech wave that corresponds to the simple motion of the lips and tongue, if I could then analyze speech for these quantities, I would have a set of speech defining signals that could be handled as low frequency telegraph currents with resulting advantages of secrecy, and more telephone channels in the same frequency space as well as a basic understanding of the carrier nature of speech by which the lip reader interprets speech from simple motions”

  » -- Homer Dudley, 1935
Homer Dudley’s Voder

- One of the first to understand the information bearing element of speech (all done at Bell Labs)
- Originally demonstrated 1939 Worlds Fair

• Finger keyboard
• Voiced/unvoiced excitation
• Resonator filters (BPF)
• 10-key keyboard
• Foot controls pitch
Branches

• **Speech Synthesis**
  – Goal: from simple description (e.g., text) produce a natural sounding voice that someone will both understand and not mind hearing.

• **Speech Coding**
  – From a human’s voice, “code” and compress it down to its fundamental information bearing element so that it can be transmitted efficiently (e.g., by radio, where bandwidth limitations are severe)

• **Speech Recognition**
  – From a human’s voice, determine the string of words and other sounds that were said, and represent that in a form amenable to action by computer (e.g., ASCII, or some computer-based representation)

• **Dialog systems**
  – Build a computer system that is able to have a “conversation” with a human in as natural a way as possible.

• These all draw from one or more of the list of subjects.

A History of Speech Synthesis

Lots of examples at: [klatt_examples\klatt.html](klatt_examples\klatt.html)
Speech Synthesis Today

- A real industry (some say it is a solved problem)
- Goal: Overcoming consumer resistance
- Humanizing the machine’s voice for the human for the speaker
  - Augmentative Communication
  - Deaf or disabled Speech
  - Example: Steven Hawking
- Humanizing the machine’s voice for the listener

How do Today’s Synthesizers Work?

My office was on St. Mary’s St. one block from the coffee shop.

My office was on Saint Mary’s Street, one block from the coffee shop.

*My office | was on Saint *Mary’s Street ||
*one block from the *coffee shop.
Text Analysis and Generation

- Unrestricted text-to-speech synthesis requires:
  - Text normalization: interpreting abbreviations and non-standard words (Example from Sproat et al. 99)
  - Linguistic analysis: parsing, word sense disambiguation (bass: fish or instrument?), ...
- Human-computer dialogs use text generators that can potentially provide this information
- Inherent trade off in synthesis, flexibility vs. naturalness

Speech/Auditory Perception

- Important for both speech recognition and speech coding
- Human auditory system
  - We have a working system, so should study it.
- The auditory periphery and critical band analysis
  - Auditory system does quite a bit before sending signal to brain.
- Analytical results:
  - perceptually inspired spectral analysis
  - autocorrelogram processing
  - speech intelligibility lies within modulation domain
- Moving up the auditory nerve towards brain, knowledge becomes much more sketchy.
Machine Learning & Statistical Pattern Recognition

• Fundamental problem is too difficult to program explicitly (e.g., rule-based system)
• Solution: Write a program that can learn how to solve the problem
• Many possible learning algorithms (beyond this course)
• Need lots of “training data”, so the program can learn.
• This is what statisticians have been doing for years (but not necessarily on speech)

Speech Recognition Overview

Outline

• I: Machine Speech Recognition: a Solved Problem?
• II: How do Speech Recognition Systems work?
• III: Current and future speech applications
Year 2005: Speech Recognition

• Can computers actually recognize speech at this point? 🎧
• HAL9000’s legacy, it should be solved by now.
• “it” = Seamless interactive intelligent conversations with a computer using naturally spoken conversational speech.

Year 2005:
Isn’t speech recognition by machine a solved problem by now?

Yes and No
**DARPA Broadcast News Benchmark Test Results**

More than **10% error** at best (2001)

**Human vs. Machine Performance**

DARPA Switchboard & Call-Home Benchmark Test Results (2001)

Automatic Recognition Performance:
- Switchboard: about 28% error at best
- Call-Home: about 40% error at best
- Human Performance: about 3% error

Recent LVCSR experiments on Switchboard (RT’2004)

Figure 2: English speech-to-text results for primary systems on the Conversational Telephone Speech data.
Why is the problem difficult?

• Background noise, “cocktail party” effect.
• Channel differences between training and testing
  – Head-mounted vs. desktop mic: 10% ➔ 70% WER for a speaker-trained commercial system
• Read versus spontaneous speech
  – yeah yeah I’ve noticed that that’s one of the first things I do when I go home is I either turn on the t v or the radio it’s really weird

Technology Development

We Are About Here

Research & Development needed to move to the right.
Overview on the basics on how modern speech recognition systems work.

Automatic Speech Recognition

- Audio speech signal
- Feature Extraction
- Probability Calculation
- Select Maximum
- Priors (Language Model) $p(M)$
- Likelihoods $p(X|M)$

Model Database
- $M_1$ = “How to wreck a nice beach”
- $M_2$ = “How to recognize speech”
- $M_3$ = ...

“How to wreck a nice beach”
Inspiration: Communication Theory

Semantic Intent $S$ → Linguistic Representation $L$ → Word Sequence $W$ → Sub-Word Sequence $U$

- glotal & articulatory control sequence $A$
- Markov state sequence $Q$
- acoustic waveform $X$
- acoustic waveform at ear $Y$

Fundamental Equation of Speech Recognition

$$W^* = \arg\max_W \Pr(W | X)$$

$$= \arg\max_W \Pr(X | W)P(W)$$

• Actually, this is the “fundamental” equation of all Bayes-optimal pattern classification, or Bayes Decision Theory (see Duda & Hart 73). This gives minimal error.
**Five Stages of Speech Recognition Systems**

- Stage 1: Signal Processing/Feature Extraction
- Stage 2: Acoustic Modeling
- Stage 3: Pronunciation Modeling
- Stage 4: Language Modeling
- Stage 5: Spoken Dialog Systems

**Stage 1: Signal Processing/Feature Extraction**

- Spectral Processing, Downsampling, and orthogonalization: Mel-frequency cepstral coefficients (MFCCs)
- Speech signals are “framed” (short-time analysis)
Stage 2: Acoustic Modeling

• Mixtures of Gaussian Densities

\[
Pr(q) = \sum_j c_j N(q | \mu_j, \Sigma_j)
\]

• Other “densities” are also used such as neural networks.

Stage 3: Pronunciation Modeling

• Markov Chains determine possible sequences of phones or syllables.

• Ex: “and”
Stage 3: Pronunciation Modeling

• Many possible units may be used to specify a pronunciation
  – most common: phonemes and their realizations, phones, ex using ARPAbet
    • chocolate pudding => CaKxlIt pUdG
    • phones can often be characterized acoustically (using formants, and their characteristic frequencies)
  – syllables
  – individual articulatory gestures
    • semi-asynchronous and independent parallel controls in the human vocal tract. Promising research area for pronunciation modeling.

Stage 4: Language Modeling

• Goal: describe the probabilities of sequences of words, \( p(w|h) \)
• Most common language models: Trigrams \( p(w_n|w_{n-1},w_{n-2}) \)
• Difficult parameter estimation problem (e.g., 60k words, 2.16e14 entries)
• Estimation is difficult (smoothing, backoff)
Stage 2+3+4: Hidden Markov Models

- Different utterances will be different length
- E.g., stop consonants (‘k’, ‘g’, ‘p’) are always short, but vowels are typically long(er).
- Need a way of comparing variable length features
- Dynamic Time Warping (earlier solution)
- Modern systems use statistical approach: Hidden Markov Models (HMMs)
Speech Recognition Requires multi-level Classification

\[ W \rightarrow A \rightarrow Q \rightarrow X \]

\[
p(x \mid w) = \sum_a \sum_q p(x, q, a \mid w)
\]

\[
= \sum_a \sum_q p(x \mid q) p(q \mid a) p(a \mid w)
\]

Speech Recognition Requires large search space

\[
p(x \mid w) = \max_{a, q} p(x, q, a \mid w)
\]

\[
w^* = \arg\max_w \Pr(x \mid w)p(w)
\]

Need special search algorithms (Viterbi decoding, stack decoding)

\[
(a, q, w)^* = \arg\max_{a, q, w} p(x, q, a, w)
\]
Stage 5: Spoken Dialog Systems

• Language Understanding (artificial intelligence)
• Discourse modeling
• Language Generation
  – speech synthesis, prosody is important:
  – Synthetic: Natural:

Quality is critical for user acceptance.
Which example sounds more natural?

Computer Speech Processing
Speech recognition and synthesis require all of the following topics

• Human Anatomy & Physiology
• Phonetics
• Speech Production & Vocal Mechanism
• Physics
• Acoustics
• Human (Auditory) Perception
• Computational Auditory Scene Analysis
• Psychoacoustics
• Signal processing
• Information Theory
• Communications Theory
• Statistical Pattern recognition
• Probability & Statistics
• Multiple Languages and Language Universals
• Machine Learning
• Linguistics
• Text Processing
• Word Pronunciations & Dictionaries
• Human Language Learning
• Natural Language Processing
• Spoken Dialog Understanding
• Cognitive Science
• Artificial Intelligence
• Computer Science & Algorithm Design
• Human Computer Interface Design
• Software Engineering & Computer Architecture
• High-performance & Parallel computing
• Fixed-point DSP micro-processors
What’s wrong???
With all this technology, why are speech recognition systems still doing so poorly?

Collecting Speech Data (from SRI)

- Getting users to read a list of complete multi-word sentences.
- Getting a user to read her account number from a pre-distributed sheet
- Getting a user to say natural answers to everyday questions
- Examples of getting people to say their authorization number.
The five major problems with automatic speech recognition:

• Stage 1
• Stage 2
• Stage 3
• Stage 4
• Stage 5

Current and future speech applications. What are some speech-enabled consumer electronic devices?
Four General Uses

- Dictation
- Web/Computer Navigation
- Voice-based Information Retrieval
- Hands-Free operation of devices
- Learning aides (voice training)

Computer Dictation

- What it is:
  - Throw away the typewriter, use speech to do word processing (both text entry and editing)
- The good
  - continuous speech recognition
  - speech to text on desktop
  - avoid RSI
- The bad
  - enrollment needed
  - computationally intensive
- The ugly
  - error prone
  - VSI
  - 3% usage penetration (2003)

Uses:
- email, letters, menu function,
- doc. dictation & transcription,
- school assignments,
- online chats
To compare with acoustical and infrared type of sensing, interdigital dielectrometry is a less invasive, and usually a more cost-effective, especially in manufacturing environment(s).

Great for people with RSI, but also can be faster than typing. Only 3% usage penetration so far.

Speech Enabled Cars

- Driving is becoming complex
- Speech output
  - car state, radio, temperature, speed limit, gas
- Speech Input
  - speech controlled dash board
  - car phones with speech control
- Networked car
  - GPS interface, speech control, bluetooth
- Key Advantage is Safety
Natural Language Interfaces

- Interactive Voice Response (IVR)
- Touch tone is horrible
- Need good natural language interfaces

Mobile Handheld Recorders

Increasing degree of desirability/difficulty
- PDA’s transmit speech to central server
- PDA’s recognize speech locally
- PDA’s do language rec. & translation

Key Concerns:
- Lots of computer power needed
- Power consumption/battery technology
- Good visuals needed on a small visual display
- Need noise robustness
- “Micro-browsers” for networked PDAs
Hands-Free Interactive Radio

Uses Nuance Technology

Other applications

- speech mobile devices for field workers
- computer assisted transcriptions
- voice activated navigation systems
- voice remote control TV
- voice biometrics/security
- VoiceXML
- wearable computers
- speech synthesis
Where Speech Recognition Technology will Ultimately Succeed

• Whenever humans do not have use of their hands, speech recognition will prevail
  – While driving
  – Telephone
  – Assistive devices
  – Other examples?

Hot Research Topics

• Noise Suppression
  – Crucial technology for success
  – noise canceling microphones
  – microphone arrays
  – signal processing for noise reduction
• Large-Vocabulary Continuous Speech (now called speech-to-text)
• Transcribing phone calls
• Speech-to-speech translation
• Alternative Statistical Models
What are companies doing with speech recognition technology?

Speech Recognition Today

Companies

And more ...

Applications

• Dictation
• Voice web browsing
• Hands-free control
• Home shopping
• Call routing
• Unified messaging
• Baggage & package tracking information
• Stock trading
• dictation Viavoice
• control all computer functions
• natural language backends controlled with voice (with One Voice)
• Huge speech research group

---

**Nuance Communications**

• voice web, use web over the phone, voice links
• home shopping over the phone
• American Airlines, Sears, Home Shopping Network, United Parcel Service
• Stock trading by phone/Home banking, etc.
Microsoft

• Speech application-programming interface (SAPI)
• voice controlled PDAs
• MS Speech recognition is now on every desktop running Windows.

Conversa

• browse the web using voice
• voice watch to speech home
• speech & Samsung’s cellular phones
• finding sound files on web using voice
• Company is now defunct!
  – why?
(what used to be)

- continuous speech recognition
- purchased by Learnout & Hauspie which then went out of business
- medical & legal special purpose recognizers
- Now owned by ScanSoft
- Dragon (founded by Jim and Janet Baker) was a pioneer in Speech Recognition. First commercial ASR system on a desktop (followed closely by IBM)

Speech on the phone you can try

- **Homework**: Call one of the following places and come back with your opinion: Does ASR work?
  - Weather information: 1-888-573-8255
  - United flight information: 1-800-824-6200
  - Nuance demo: 1-888-682-6238 (shopping, stock, bank, travel)
  - Tell Me: 1-800-555-TELL
What we will learn in this class.

• Basics of speech production and processing.
• Simple speech synthesis
• Advanced speech recognition techniques at all levels:
  – signal processing
  – acoustic
  – pronunciation
  – language
  – Statistical Modeling

Reading Assignment

• Read Chapters 1 and 2 in the text.