1.1 Human Computer Interaction

The main goal of most computer speech research is to create an improved method for human-computer interaction.

- Human-Human Contact:
  The study of how humans communicate can give us insight on how to improve human-computer interaction. Speech is the most natural method for human-human communication, and the study of this particular method of human-human interaction is of great interest to people who study HCI.

- Human-Machine Contact:
  The most common method for human-machine interaction is the Windows, Icons, Mouse, and Pointer (WIMPy) interface.
  Other HCI methods that are becoming widespread are the writing pad, scanner, digital camera, and cell-phone. Another emerging technology is a wearable computer where the screen is embedded on a pair of glasses.
  Haptic interfaces are another topic of HCI research. A haptic interface is one which provides mechanical feedback to the user. Examples are glove or pen-type devices that allow the user to "touch" and manipulate 3-dimensional virtual objects, or to simulate the operation of machinery.

- Overall Goal of HCI:
  The overall Goal of HCI is to be efficient, easy, and ergonomic. Ideally the interface should be so natural that the user does not know that they are using a computer, it is just second nature.

Speech is the standard communication mode of humans, and seems to be an obvious choice for a human-computer interface as well.

For a speech HCI system, we need two things:

- Speech Synthesis
- Speech Recognition

1.2 Further Information on HCI

- ACM SIGCHI (HCI professional organization)
  http://www.acm.org/sigchi/
• Human Computer Interaction Index
  http://degraaff.org/hci/

• Haptics community web page
  http://haptic.mech.nwu.edu/

• Speech Generation for Human-Computer Interaction project at UW SSLI lab:

• Human Interface Technology Lab at UW:
  http://www.hitl.washington.edu/

1.3 Computer Speech Processing

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

• Human Anatomy and Physiology
  The study of the way humans produce voice is important for two reasons:
  1. So we can understand what speech is, and what the information bearing element in speech is
  2. So we can understand how to build speech synthesizers.

• Phonetics
  Phonetics refers to the study of speech sounds and their classification and transcription. Phonetics does this in a less anatomical way.

• Speech Production and Vocal Mechanism
  What is it that someone else is doing with their voice such that you hear what it is they are saying.

• Physics
  Understanding of the nature of sound propagation through the air is also important. One can study what is needed for something to be transferred from someone’s mouth to someone else’s ears in an efficient manner.

• Acoustics
  How do we deal with any potential acoustic environment? Humans can easily tune out ambient background noise, but this is quite difficult with a computer.

• Human (Auditory) Perception
  It is useful to study human auditory perception because humans have already solved the problems of speech recognition and synthesis, and perception is the first step a human takes in this process. This class will study some of the peripheral things which take place in the human perception system. Unfortunately, what goes on after these peripheral things are not well understood.

• Computational Auditory Scene Analysis
  This field attempts to try to parse or understand auditory scenes. Imagine walking down a busy city block with the noise of cars, barking dogs, and screaming people. Humans can easily parse and pick out the different sounds occurring in this scene. Another example is listening to a particular instrument which is part of a larger ensemble. Computational auditory scene analysis is an attempt to get computers to perform these same types of tasks. Examples are to separate out multiple sources in a recording, or to reproduce sounds recorded using old noisy technology.
Psychoacoustics
Psychoacoustics is study of acoustics from a psychological perspective. This field attempts to study why humans hear in a certain way, but not in another. For example, why are speech and other sounds still intelligible even after after undergoing certain dramatic distortions.

Signal Processing
Methods for speech feature extraction are typically based on spectral analysis methods.

Information Theory
Information theory is the study of how we quantify information. It is the root behind the study of communications.

Communications Theory
One can apply many of the same tools people have developed to study radio communications to any source of information. When we are speaking and hearing we are communicating. One could argue that speech recognition systems are based entirely on this communications system point of view. Speech recognition uses the term “decoding” for determining words from speech signals, meaning it is determining the word which has been “encoded” on some noisy channel.

Statistical Pattern Recognition
This is a large field itself and provides the basis for most speech recognition methods.

Probability and Statistics
This field is needed in order to study information theory and pattern recognition techniques.

Multiple Languages and Language Universals
Many people have attempted to learn what universals exist across all languages in reference to both language and speech. One might attempt to determine if information from one language can be used in the automated recognition of another.

Machine Learning
The science of trying to build computers that can learn. In the 40’s and 50’s, and even into the 70’s it was thought that computers would be able to do anything though an explicit program. Machine learning instead tries to create program which can learn new behaviors from data. Humans might not be able to write a program which is intelligent, but instead they might be able to write a program that can learn to be intelligent.

Linguistics
The study of the science of language.

Text Processing
Text processing is also important to speech research. This is the process of transforming text from something humans are familiar with to something a computer can process. When humans read text it can include all kinds of abbreviations and mistakes, but we can still interpret it. Computers are easily confused by such issues. Text processing was important long before the web became popular.

Word Pronunciations and Dictionaries
What are the underlying components that words should be divided into. Words can be decomposed in to phones, but this may or may not be the best way to decompose words for a speech recognition system.

Human Language Learning
Study ways that humans learn language and apply this to the problem of making a computer learn language.
• Natural Language Processing
  This is the study of how a computer can act upon a string of words as though it understood it. The computer only needs to act like it understands since there really is no good way to define "understanding."

• Spoken Dialog Understanding
  This is a branch of natural language processing which specifically studies dialogs between two agents. An example of this is an automated telephone system which can respond to questions.

• Artificial Intelligence
  This is the study of making computers do things which are normally done through the use of human intelligence.

• Cognitive Science
  This is similar to artificial intelligence, but it implies a broader study of the human mind linking artificial intelligence to other fields such as psychology, neuroscience, and linguistics.

• Computer Science and Algorithm Design
  Many fields are now forms of computer science to some extent. This makes it important to be able to design and evaluate algorithms.

• Human Computer Interface Design
  See previous section.

• Software Engineering and Computer Architecture
  Speech recognition and synthesis programs are software systems which can be quite large. The ability to design and manage such a system can become an issue.
  Considering a computer’s architecture and understanding how the computer works can help one create a design which will perform well.

• High-Performance and Parallel Computing
  If you really want something to run fast, you need to understand how to get it to run in parallel.

• Fixed-Point DSP Micro-Processors
  This is important for portable devices.

1.4 Origin of Speech Processing

Some of the people at Bell Labs foresaw that the telephone was going to be an important invention. The key original motivation for speech processing was simple bandwidth reduction. Their goal was to transmit the intelligibility information in speech as inexpensively as possible.

Speech is remarkably compressible. The inherent information in a signal is the number of bits needed to represent the signal using the best possible compression scheme. For example, if a 10 megabyte file can be compressed to just two bits, the file only has two bits of information.

If we only consider speech from a signal processing point of view:

Most of the energy falls off after 4 kHz ⇒ sample at 8 kHz using 8 bits per sample ⇒ 64 kilobits per second

64 kb per second was long thought to be needed for voice transmission. Long distance service across the world has been digital for some time and was built using 64 kps PCM channels (although some of this is used for signaling).
ISDN service allocates a full 64 kbps channel for voice. [J96] Modern speech coders can transmit speech which is both intelligible and still allows you to determine the identity of the speaker using only 1.2 kb per second. Digital cell phone systems have greatly benefited from speech compression.

This course will cover the basics of speech compression when discussing:

- LPC - Linear Prediction Coefficients
- CELP - Code Excited Linear Prediction
- SOLA - Synchronous OverLap and Add

SOLA is not typically used for speech compression, but is interesting from a signal processing point of view. This method can change the temporal duration which speech lasts without changing the spectral content. For example, some voice mail services will allow you to speed up or slow down messages without changing the pitch of the voice. If this was done by just changing the sampling rate the voice would sound different.

Speech research has a rich history reaching back hundreds of years. Wolfgang Ritter von Kempelen attempted to build a speaking machine in the (late 1700s). It used a bellows blowing through a reed to create vibrations. The sound then passed through a leather tube which was shaped with your hand in order create the various phonetic sounds. It also included a switch which changed the reed from creating a tone to creating noise for emulating a whisper.

![Figure 1.1: Wheatstone’s version of the speaking machine, 1835][D50]

Homer Dudley from Bell Labs was ahead of his time in the realization of the information bearing element of speech.

"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

The invention of the telephone in the late 19th century, and the subsequent efforts to reduce the bandwidth requirements of transmitting voice, led back to the idea. In the 1930s, the telephone engineers at Bell Labs developed the famous Voder, a speech synthesizer that was unveiled to the public to great fanfare at the 1939 World’s Fair.
Figure 1.2: Pictures of Dudley’s Voder
Figure 1.3: The Voder was demonstrated by trained operators at the World’s Fairs of 1939 (New York) and 1940 (San Francisco). Although the training required was quite long (on the order of a year or more), the operators were able to ‘play’ the machines – literally as though they were organs or pianos – and to produce intelligible speech. (Oizumi and Kubo)

Figure 1.4: The Voder was basically a spectrum-synthesis device operated from a finger keyboard.
The ‘resonance control’ box of the device contains 10 contiguous band-pass filters which span the speech frequency range and are connected in parallel. All the filters receive excitation from either the noise source or the buzz (relaxation) oscillator. This choice of excitation was to duplicate the voiced or unvoiced characteristic of the human vocal system. The wrist bar selects the excitation source, and a foot pedal controls the pitch of the buzz oscillator. The outputs of the band-pass filters pass through potentiometer gain controls and are added. Ten finger keys operate the potentiometers. Three additional keys provide a transient excitation of selected filters to simulate stop-consonant sounds.

Note: H. W. Dudley retired from the Bell Laboratories in October 1961. On the completion of his more than 40 years in speech research, one of the Voder machines was retrieved from storage and refurbished. In addition, one of the original operators was invited to return and perform for the occasion. Amazingly, after an interlude of twenty years, the lady was able to sit down to the console and make the machine speak.) More recently, further research studies based upon the Voder principle have been carried out (Oizumi and Kubo).

1.5 Further Information on the Origin of Speech Processing

- Virtual Voices, Remko Scha. This is an interesting article on the history of speech synthesis. Be sure to check out the anecdote about Alexander Graham Bell and his dog.
  http://cf.hum.uva.nl/computerlinguistiek/scha/IAAA/rs/virtual.html
- Wolfgang von Kempelen’s and the subsequent speaking machines (as well as more information on the Voder)
  http://www.ling.su.se/staff/hartmut/kemplne.htm
- History and Development of Speech Synthesis
  http://www.acoustics.hut.fi/~slemmett/dippa/chap2.html
- History of Speech Recognition, (brief, but a few interesting factoids)
  http://www.stanford.edu/~jmaurer/history.htm

1.6 Branches of Computer Speech Processing

- 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). The recognizer must also be able to discriminate between speech and noise.
- Dialog systems
  Build a computer system that is able to have a ”conversation” with a human in as natural a way as possible.
• Speaker Identification
  Identify someone by their voice from a list of possible speakers.

• Speaker Verification
  Determine if a speaker is who they claim to be, typically for security purposes.

1.7 Speech Synthesis

There is a trade off between flexibility and the naturalness of the speech. A very flexible interface would be text to speech. A less flexible interface would be where the text is known and the output can be hand tuned.

Modern speech synthesis takes recordings of human speech, divides it into smaller chunks, and reconnects the chunks in a new order. This is known as concatenated speech synthesis. There currently are not any systems which generate all of the speech sounds from scratch which sound as good as a concatenated speech system.

1.8 Is Speech Recognition a Solved Problem?

Can computers truly recognize speech at this point? Most companies with speech recognition products would say "yes," but most academics needing speech recognition funding would say "no."

On the DARPA broadcast news speech recognition benchmark, the word error rate is at best 10%. This is even considering that this corpus consists of radio and television news broadcasts which are mostly commentators speaking carefully and clearly.

![Figure 1.5: DARPA Broadcast News Benchmark Test Results, 1999](image-url)
The best case on the Switchboard corpus, which contains conversational speech between strangers, is about 28% error rate.

The best case on the Call-Home corpus, which contains conversational speech between family members and close friends, is about 40% error rate at best.

Humans only have about 3% error on the same tasks. Note that the error rate is inter-person agreement. It is hard to say what the error rate is unless you know truth, but only humans can provide the truth, so at best we can only do the inter-person agreement of a corpus.

Automated speech recognizers also perform poorly in the presence of noise. Humans do not have much difficulty listening to speech under the same conditions.

The current state of speech recognition systems can be compared to the development of the airplane. We have technologies to build systems which can recognize speech. We also have systems which can take you across the country in several hours, but it is only “usable” technology. We can fly, but we do not like to fly. It is uncomfortable, crowded, the food is bad, and it is not as fast or as safe as we would like it to be. What we really want is “great” airplane technology. Speech recognition has recently reached the point of being “usable” technology, but there is a long way to go before it is as widely used as airplane technology, and even further before we consider it “great” technology. (This analogy was created by Hynek Hermansky)
1.9 **Automated telephone systems which use speech recognition:**

- 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
- UW Communicator: 1-877-890-2630

1.10 **References**

[D50] *The speaking machine of Wolfgang von Kempelen*, Dudley & Tarnoczy, JASA 22, 1950