Outline of lecture 2:

- Overview of Speech Recognition
- Speech Processing:
  - Motivation, speech communication
  - Engineers’ view of speech production

There are many possible end-goals for computer speech processing: speech modification; speech coding; speech enhancement; speech recognition; speaker recognition/speaker ID; language recognition/ID; speech synthesis; spoken dialog management; spoken language processing; spoken language analysis, etc. In this course we will focus mainly on automatic speech recognition.

### 2.1 Overview of Automatic Speech Recognition

#### 2.1.1 Automatic speech recognition system configuration

A diagram of the automatic speech recognition (ASR) process is shown in the following figure:

![Diagram of ASR process](image)

Figure 2.1: Automatic speech recognition process.

In the first step, the spectral features are extracted from the input speech waveform. Let \( X = x_1, x_2, \cdots, x_N \) denote the acoustic observation (feature vector) sequence and \( W = w_1, w_2, \cdots, w_M \) be the corresponding word sequence. The decoder chooses the word sequence with the maximum \textit{a posteriori} probability:

\[ P(W | X) = \max_{w} P(w | X) \]

These notes are mostly based on the in-class lecture slides by Prof. Bilmes and Chapter 1 and 2 in the textbook [TEXT] by Xuedong Huang et al.
\[ \hat{W} = \arg\max_W P(W|X) = \arg\max_W P(X|W)P(W) \]

where \( P(X|W) \) is the \textit{acoustic model} and \( P(W) \) is the \textit{language model}. The equation 2.1.1 is the fundamental equation of all Bayes-optimal pattern classification problems. It is also called by some the “fundamental equation of speech recognition”.

### 2.1.2 Five stages of speech recognition systems

There are following five stages in speech recognition systems: 1) Signal Processing/Feature Extraction; 2) Acoustic Modeling; 3) Pronunciation Modeling; 4) Language Modeling; 5) Spoken Language Understanding/Dialog Systems.

#### 2.1.2.1 Stage 1: Signal Processing/Feature Extraction

In stage 1, we apply signal processing techniques to extract acoustic features from the speech waveform. The most widely used acoustic feature sets for ASR are Mel-frequency cepstral coefficients (MFCCs) and Perceptual Linear Prediction (PLP) coefficients. A diagram of the MFCC feature extraction procedure is shown in Figure 2.2. First, the input waveform signal goes through a Mel-scaled filter bank. Then it is followed by low pass filtering and downsampling. Finally discrete cosine transform (DCT) is performed on the log-energy of the filter outputs.

![Figure 2.2: MFCC feature extraction procedure](image)

A diagram of PLP feature extraction is shown in Figure 2.3, which is from the reference [PLP]. PLP technique uses three concepts from the psychophysics of hearing to derive the estimate of the auditory spectrum: (1) the critical-band spectral resolution, (2) the equal-loudness curve, and (3) the intensity-loudness power law. More detailed description about MFCC and PLP is in the textbook [TEXT] Section 6.5 and the reference [PLP].

![Figure 2.3: PLP feature extraction procedure](image)

#### 2.1.2.2 Stage 2: Acoustic Modeling

After feature extraction, how do we model the distribution of the feature vectors \( x_{1:N} \)?
- Mixture of Gaussians:
  \[ Pr(x_i) = \sum_j c_j N(x_i | \mu_j, \Sigma_j) \]

- Hidden Markov Models:
  \[ Pr(x_{0:T}) = \sum_{q_0:T} \pi_{q_0} \prod_{t=0}^{T-1} a_{q_t,q_{t+1}} \prod_{t=0}^{T} p(x_t | q_t, \eta) \]

- Neural networks, etc.

### 2.1.2.3 Stage 3: Pronunciation Modeling

There are multiple possible pronunciations for a word in continuous speech. For example, the word “and” has more than one possible pronunciations, as shown in Figure 2.4.

![Figure 2.4: Pronunciations of word “and”](image)

Meanwhile, many possible pronunciation units may be used to specify the pronunciation of a word:

- Most common: phonemes and their realizations (phones)
- Syllables: intermediate unit between the phones and word level
- Individual articulatory gestures

### 2.1.2.4 Stage 4: Language Modeling

The goal of language modeling (LM) is to describe the probabilities of sequence of words, or \( p(w|h) \). There are many ways to describe the language models. The most common language models used in ASR are trigrams, described by \( p(w_n|w_{n-1}, w_{n-2}) \). The trigrams can be generalized to N-gram: \( p(w_n|w_{n-1}, w_{n-2}, \ldots, w_{n-N+1}) \). However, it is difficult to estimate the parameters of N-grams. For example, for trigram with 60k words, there will be 2.16e14 trigram probability entries. Some techniques have been adopted to solve the data sparsity problem, such as smoothing, backoff (e.g., trigram backoff to bigram or unigram).

Combining stage 2+3+4, Hidden Markov Models (HMMs) have been used widely in ASR:

- Different utterances will be different length
  e.g., stop consonants (’k’, ’g’, ’p’) are always short, but vowels are typically longer.
- Ways of comparing variable length features
  - Earlier solution: Dynamic Time Warping (DTW)
  - Modern systems: Hidden Markov Models
The Hidden Markov Model is illustrated in Figure 2.5, which is from the HTK Book [HTKBOOK]. The parameters $a_{ij}$'s are the transition probabilities from state $i$ to state $j$. The observation probability $b_j(o_t)$ represent the output probability of observation $o_t$ given the state $j$. Since the HMM states $j$ are not observed, they are called “hidden” states.

Modeling with HMMs is a multi-level classification. From the high level to low level, it can be described as:

$$W(\text{word}) \rightarrow A(\text{acoustic unit}) \rightarrow Q(\text{HMM state}) \rightarrow X(\text{acoustic observation})$$

$$p(x|w) = \sum_a \sum_q p(x, q, a|w) = \sum_a \sum_q p(x|q)p(q|a)p(a|w)$$

where $a$ is the phone, $q$ is the HMM state, $w$ is the word. Speech recognition requires large search space to search for the best sequences $(a, q, w)$, as in:

$$(a, q, w)^* = \operatorname{argmax}_{a, q, w} p(x, q, a, w)$$

Therefore we want to maximize the joint probability using Viterbi decoding or stack decoding techniques, as described in later part of the textbook [TEXT].

2.1.2.5 Stage 5: Spoken Language Understanding/Dialog Systems

- The operational definition of “language understanding” is: to react in a way that the user of the speech system is satisfied and has achieved a desired goal, using only a speech human-computer interface.
- When the feedback to users is speech, the spoken language dialog system needs to respond in natural ways, i.e., it needs to have discourse modeling:
  - Need to generate appropriate textual responses
  - Need to generate natural and pleasant sounding speech synthesis
2.1.2.6 Why isn’t ASR as good as humans?

- Best result in the latest English conversational telephone speech recognition has a word error rate (WER) of around 15%. Humans do much better than this. Especially in noisy environment, the ASR performance degrades significantly while the humans are only slightly affected.

- Problems exist in all five stages of the ASR process.

- The gain from optimization in one stage may not translate into the gain in the whole system. Therefore, people working in different stages should work together to optimize the whole system.

2.2 Speech Processing

2.2.1 Speech communication

The goal of speech communication is to communicate information over a distance:

- How far does speech travel? Speech gets absorbed by objects in environment. Ancient ways to increase the travel distance include: centurians, voice towers. Centurians are the people with loud voices, so their voice can travel farther in the air. Voice towers are used to transfer speech from one point to another by humans.

- Definition of speech: organized fluctuation of acoustic pressure waves, which disperse spatially from a source consisting of the human vocal apparatus.

- Properties of speech waves: outward propagation, Doppler, reflection, absorption, refraction, diffraction, superposition, etc.

- Information in speech:
  - Shannon’s formula: \( C = BW \cdot \log(1 + (S/N)) \)
    For speech, \( BW = 4kHz \), \( SNR = 30dB \), around 40kbps (ISDN 8bit, 8kHz, 64kbps)
  - 42 symbols are sufficient for describing the phonemic alphabet. Entropy under uniform distribution is \( \log(42) = 5.4 \) bits. Under “true” distribution the entropy is about 4.9 bits. At 10 phones/second, the speech information is around 50 bits/second.
  - For sine-wave speech (modulation < 10Hz, 8 bits, first 3 formants), the information is \( 3 \times 8 \times 20 = 480 \text{bits/sec} \). At this bit rate, the speech quality is still very low.
  - Best modern speech coders are at approximately 1.2kbits/sec. The quality loss is noticeable but not awful.
  - Reading: 150-1500wpm (2.5-25 wps) with 1.75 bits/char, 4.5 chars/word \( \Rightarrow \) 19 bits/second - 196 bits/second
  - Conclusions: “Information” in speech is less than one would expect.

2.2.2 Engineers’ view of speech production

Speech organs may have multiple roles: speaking, eating, breathing, smelling and facial expressions; ears are just for hearing. The human speech production system majorly consists:

- Air expansion from lungs forces air through larynx and vocal tract

- “articulators” determine shape of vocal tract and form the sound
- Vocal tract: single acoustic tube after larynx, between lips and vocal chords, non-uniform area
- Nasal tract: velum to nostrils

Figure 2.6 shows the human speech production apparatus.

Figure 2.7 shows the glottal cycle. In (a) the vocal folds are closed with sub-glottal pressure buildup; in (b) trans-glottal pressure differential causes folds to blow apart; in (c) pressure equalization and tissue elasticity force temporary reclosure of vocal folds, then it is ready to begin next cycle. The number of such cycles per second is called fundamental frequency \(F_0\).

There are several points we can find from vocal tract and acoustic analysis:

- Vowels are the sounds with no major constriction, while consonants have some constriction
- Speech is not a string of discrete well-formed sounds
  - It is continually varying with periods of “relative” stationarity
  - Speech has coarticulation effect: effects of the surrounding context of a speech segment will greatly influence that speech segment

Since the glottal wave is periodic, speech consists of harmonics with fundamental frequency \(F_0\) and integral multiples of \(F_0\). The resonances of the vocal tract are excited with the glottal air pressure. When the shape of the vocal tract changes, the resonance frequencies change as well. The harmonics near the resonances are emphasized. The resonance frequencies caused by the vocal tract are called formants. When we talk about formants, we generally mean \(F_0\), \(F_1\), \(F_2\), etc:
• F0 = pitch, fundamental frequency of periodic excitatory source (vibrating folds)
  – Males: F0 = 50 - 250Hz
  – Females: F0 = 120 - 500Hz
• F1 = first formant, F2 = second formant
• F1, F2, F3 … are resonance frequencies of vocal tract
  – Resonance frequencies (formant locations) determine vowels
  – Shape of vocal tract $A(x, t)$ determines frequencies

In spectrogram, we can see the formants of the speech. In Figure 2.8, the three lines denote the formants F1, F2 and F3 respectively. The fundamental frequency $F0$ is the frequency distance between the harmonics.

(a) Spectogram of a speech utterance
(b) The first 3 formants are labeled in black line

Figure 2.1: Illustration of formants of speech.

References

