Automatic Speech Recognition
Lecture Note 1: Overviews

1. Historical notes

   - 1920s: Radio Rex
   - 1950s: IBM digit recognition system: using formant estimates based on the number of zero crossings of signal in two subbands.
   - 1960s: FFT (Fast Fourier Transform), cepstral analysis and LPC (Linear Predictive Coefficient) were developed as new approaches for short-term spectral estimation. DTW (Dynamic Time Warping) and HMM (Hidden Markov Model) were developed as new methods for pattern matching of sequences. HMM-based ASR systems were being developed.
   - 1970s: The start of task-oriented projects with definite objectives funded by Advance Research Projects Agency (ARPA). First project was 1000-word ASR with less than 10% semantic error.
   - 1980s: Large corpora collection (TIMIT, Resource Management (RM), Wall Street Journal); new front ends (speech signal processing modules) such as MFCC; HMM became dominant. Second ARPA ASR program, with RM and Wall Street Journal tasks, was launched.
   - 1990s: More engineering efforts towards improving performance of current mainstream framework of ASR. These efforts include vocal tract normalization, speaker adaptations, new model training criteria, gender-dependent models, etc.
   - Recent work: Broadcast News task; Switchboard & Call Home task; speaker identification; language identification; noise-robustness; pronunciation models; language models.

2. Components of ASR

   - Front end: input acquisition, preprocessing and feature extraction.
   - Back end: implementation of acoustic model and language model.

3. Challenges in ASR

- uncertainty in speech (dialects, disfluencies, styles, rates, etc)
- mismatch problem (noises)
- efficient training and decoding algorithms
- limited computational resources
- limited knowledge about human speech perception

4. Applications of ASR

   - Telephone applications: ASR can replace the currently common touch-tone systems and possibly cut through the menu hierarchy and get to the point. For example, in a customer-service line, this can reduce the cost of the service provider (since they often pay for the phone) and the time of the caller.
   - Hands-free operations: For example, to initiate a phone call while driving.
   - Helper for the physically handicapped: For example, alternative interface to computers for those with limited eyesight or mobility in their arms or hands.
   - Dictation (auto transcribing): This has achieved 99% of accuracy in quite rooms for English users.
   - Translation: A communication between two persons using two different languages can be implemented with ASR and TTS in between them.
   - Information systems: stock quoting system.

5. Task Parameters of ASR

   - speaker-dependent (SD) vs. speaker-independent (SI)
   - lexicon size and perplexity
   - isolated speech vs. continuous speech
   - read speech vs. spontaneous speech vs. conversational speech
   - clean speech vs. noisy speech

6. An instance of ASR

- **Robot** I am Stanford's handyman robot. Tell me a task, and I will do it for $5 per hour. This money will be applied to further research in artificial intelligence.
- **Human** $5 dollar an hour? Sounds great! Can you paint?
- **Robot** My painting is of the highest quality.
- **Human** OK. See that paint brush and bucket of paint? Take them out and paint the porch.
- **Robot** Your request will be fulfilled, courtesy of Stanford.

  (The robot trundles off to do his job, and return in an hour.)
- **Robot** The task is complete. Please deposit $5 to aid in further research.
- **Human** (Handing over the cash) This was a great deal! Come back again!
- **Robot** (While leaving) Oh, by the way, it wasn't a Porsche. It was a BMW.