Speech Generation& Recognition

Reid Simmons
Illah Nourbakhsh

Speech Generation

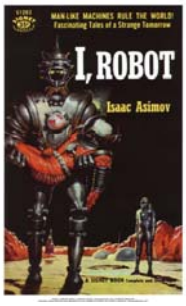
Desirable Speech Characteristics

- Naturalness
  - Sounds human-like

- Intelligibility
  - Easily understandable

Speech Synthesis (Text-to-Speech)

- Why is this difficult, technically?



Speech Synthesis

- Normalization
  - Pre-process text to contain only words
- Text Analysis
  - Syntactic parsing
  - Semantic parsing
- Text-to-Phoneme
  - Pronunciation
- Prosody
  - Pitch, loudness, duration => stress
- **Affect**

Issues with Normalization

- Numbers
- Abbreviations
- Acronyms

Issues with Pronunciation

- Ambiguous words
  - Often depends on part-of-speech
  - May depend on semantics
  - May depend on tense!
- Many exceptions to “sounding out” rules
  - though, through, bough, cough, tough
  - comb, tomb, bomb
  - dose, hose, lose

Issues with Prosody

- Punctuation
  - Pause, after comma
  - Rising tone for questions?
- Syllabic stress
- Word stress

Issues with Affect

- Speed of speech
- Emotional content of speech

Exercise

- Get into your teams
- Create a four word sentence
  - noun-verb-noun-adverb or noun-verb-adjective-noun
  - Read and record using four different emotions (happy, angry, sad, disgusted, fearful, surprised)
  - listen and analyze how prosody changes
- Take notes: We will then discuss

Techniques for Speech Generation

- **Formant** (rule-based)
  - Use acoustic models
  - Compact program
  - Tends to be quite intelligible, but limited prosody
- **Concatenation** (unit selection)
  - Use human speech, “sliced and diced”
  - phones, diphones, triphones, …
  - Layer on prosody using signal processing
  - Domain-specific synthesis

SSML

- Semi-standard markup language for specifying pronunciation and prosody
  - <emphasis level=“strong”>
  - <break time=“4500ms”>
  - <prosody rate=“fast”>
  - <prosody pitch=“+25Hz”>
  - <prosody volume=“33%”>
  - _{Dr.}
  - <phoneme ph=‘t ah0 m ey1 t ow0”>tomato</phoneme>
  - <say-as interpret-as=“digits”>123</say-as>
  - <say-as interpret-as=“number:ordinal”>VIII</say-as>

Speech Recognition

Speech Recognition

- Why is this difficult, technically?

Diverse Sources of Ambiguity

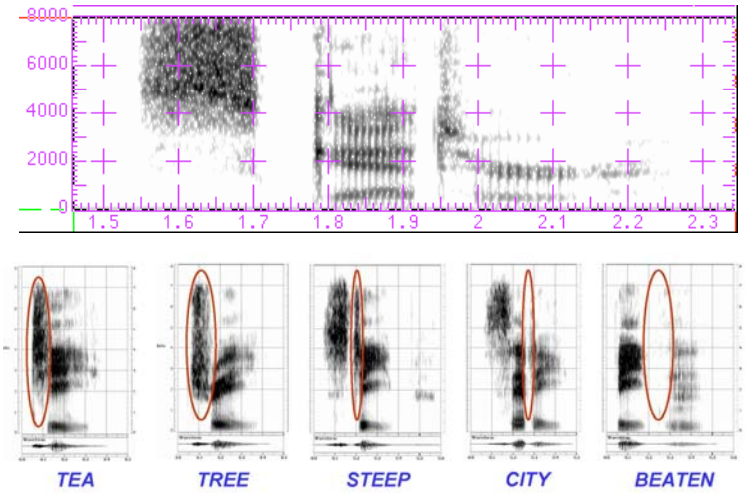
- Acoustic/Phonetic
  - Let us pray
  - Lettuce spray
- Syntactic
  - Meet her at the end of Main Street
  - Meter at the end of Main Street
- Semantic
  - Is the baby crying
  - Is the bay bee crying
- Discourse Context
  - It is hard to recognize speech
  - It is hard to wreck a nice beach

Phonemes

- ~40-45 phonemes in English
  - Variance depends mostly on dialect
- Voiced vs. unvoiced
  - vowels vs. consonants
- Phonetic sounds may differ based on preceding and succeeding phonemes

The Acoustic Signal

Sad:



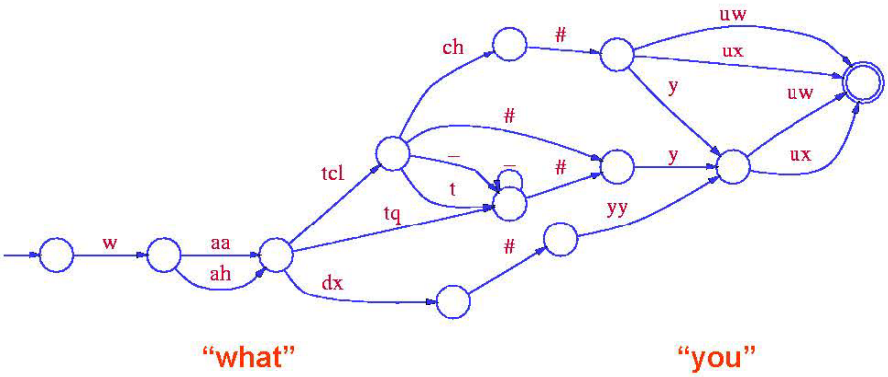
Techniques for Speech Recognition

- Almost all current approaches use statistical modeling and massive amounts of data
- Maximize probability of word sequence
  - P(W* | A) ~= P(A | W*)P(W*)
- Typically, language model uses ***trigrams***
  - Probable sequence of phonemes
  - $P(W_n \mid W_{n-1}, W_{n-2})$
- Other constraints
  - Syntax
  - Semantics
  - Domain / context

Representing Language Model

- Hierarchical ***Hidden Markov Model*** (HMM)
  - Atomic units (sub-phoneme)
    ~20ms slices, characterized by power in bands of frequencies
  - HMM of phonemes
  - HMM of diphones or triphones
  - HMM of words
  - HMM of phrases (trigrams)
  - Put together into single HMM
- Use Viterbi algorithm to find best path
  - May use backwards search to refine path

Simple Word-Level HMM



Grammatical Issues

- Incomplete Sentences
- Non-Grammatical Sentences
- Fillers
  - er, um, …
- Disfluencies
  - cutting off mid-word
  - corrections
  - hesitations

Further Issues

- Recognizing Prosody
  - Stress is important in interpreting pragmatics
- Recognizing Emotion/Affect
- Current Status
  - Siri and Google Voice have been able to use millions of training examples to create fairly good continuous, speaker-independent speech recognition