LING 190 Module 8: Speech Synthesis

Table of Contents

1. Speech Synthesis

Speech synthesis is an applied topic of linguistics, as opposed to the theoretical phonetics and speech development topics that we've covered so far. We'll be learning about speech synthesized by machines.

Speech synthesis is beginning to play an increasingly large role in our lives. We want to understand a bit about **how** speech is synthesized, as opposed to other aspects of machine speech, such as speech recognition.

1.1. Main Ideas

- How is speech synthesized?
  - The history
  - Different approaches
  - Basic pipeline of early speech synthesis approaches
- The goal is to cover base-level knowledge about how speech is synthesized, and what the basic challenges to speech synthesis are.

1.2. The History of Text-to-Speech

1.2.1. Mechanical speech synthesis

In the late 1700s, von Kempelen, the creator of the Mechanical Turk, created a mechanical talking machine. This machine needed lots of practice to operate, and existed before the science of acoustics.

1.2.2. Early Analog Electric Synthesizers

- In the early 20th century, H. Dudley at AT&T Bell Labs created the VODER, or Voice Operation Demonstrator. This machine was based on emerging science around resonance and harmonics. It became the basis of modern-day vocoders.
- In the mid 20th century, Haskins Labs at Yale University created a "Pattern Playback" machine.
  - Developed Post-WWII
  - Based on the science of resonance and harmonics
  - Artificially created formants by painting a pattern

- Towards the end of the 20th century, digital, software-based synthesizers came about. Dennis Klatt of MIT created the first digital synthesizer in 1980, which is still used today.
  - The commercial application become known as DECTalk, which was the technology used by Stephen Hawking.
- Later on, Vocoders, which are software implementations of earlier machines, gained widespread commercial use
  - The origins were in Bell Labs, who wanted to transmit human speech with smaller bandwidth, to save companies costs in material.

1.3. Prior Approaches to Synthesis

1.3.1. Tracing (Pattern Playback)

- Remaking human speech into an electronic voice
  - Take formant information from an existing recording of a human voice
- If you know what the formants are for an utterance, you can create a pattern that reproduces that utterance.

1.3.2. Tracing (Vocoder)

- Takes some acoustic information from an existing recording of a human voice
- Could be formant information, or other acoustic information
- Generates a voice from the source using a "filter."

1.3.3. Parametric Synthesis (Klatt)

- Speech "from scratch," does not use any acoustic information to generate new speech.
- Speech is generated in real time from various parameters, such as:
  - F_0 Frequency and Amplitude
  - F_1 Frequency
  - Burst (plosive) amplitude and duration
  - Fricative noise amplitude and duration

How do we know what the parameters should be?

- F_0, the fundamental frequency, is around 110Hz for men and 200Hz for women.
- Other values can be extrapolated from prior research
  - One example is Peterson & Barney (1952), who measured F1, F2 and F3 for many vowels from men, women, and children.

Consonants also have known patterns for parameters.

- Plosives: Burst of noise, and **voice-onset time** (time for periodicity to start after the burst)
- High-frequency loud noise: /s/
- Low frequency loud noise: /ʃ/
- Quieter Noise: /f/ or /θ/
- Noise and periodicity: /v/ or /z/

Vowel sounds are based heavily on F1 and F2, which correspond to the tongue height and tongue advancement, respectively.

1.3.4. Modified Natural Speech (Siri)

- Take speech that is pre-recorded, tokenize and modify it.
  - This speech tends to sound the most "natural"
  - There are many possibilities for modifications
  - State-of-the-art technologies use artificial intelligance, machine learning and statistical modeling to generate the most natural-sounding speech.
- A simple example is simple concatenation (old Siri)
  - We take individual sounds and re-combine them together
  - Typical results are highly unnatural. Instead, we use more sophisticated techniques.
- A more sophisticated example is pre-recording 2 phonemes, called **diphone** synthesis.
  - We want /bæg/
    - We should record something like /bæd/
    - and /gæg/
  - Diphone synthesis will cut up the phonemes /bæ up to the midpoint of the vowel, and take /æg/ from the midpoint of the vowel, and join the two pieces together.

1.4. Basic steps to early Text-to-Speech (TTS) systems

- Text input & pre-processing
- Linguistic processing
- Phonetic processing
- Synthesis and output