Speech Synthesis

Notes on Speech and Audio Processing

Chia-Ping Chen

Department of Computer Science and Engineering
National Sun Yat-Sen University
Kaohsiung, Taiwan ROC

Introduction

- The output of a speech synthesis system is speech.
- The input is a specification of the speech to be synthesized. This is often the text, but can also be a meaning representation.
- The old-fashion methods for speech synthesis are often rule-based, using principles such as linear prediction and formant synthesis.
- We will focus instead on the concatenation-based approaches.

Concatenation Synthesis

- Acoustic units are stringed together for an intended speech.
- As units are concatenated, discontinuity leads to bad quality.
- Using larger units can deal with the issue of discontinuity but may raise the issue of prohibitive number of units.
- Apparently there is a trade-off between using larger and smaller units.

Concatenation Units

- Words
- Syllables
- Demi-syllables
- Diphones
- Phones
- Sub-phone units (states)

From Text to Units

- Text normalization
  - Editing: “hte” → “the”
  - Acronyms: “a.k.a. TTS” → “also known as text to speech”
  - Abbreviations: “St.” can be “street” or “saint”
  - Numbers: “10” can be “one-zero” or “ten”
  - Symbols: $ (dollar), % (percent), @ (at)
  - Dates and times
- Word pronunciation
  - letter-to-phoneme rules
  - pronunciation error rate

Prosody

- Pitch, stress, rhythm; f_0, energy, duration.
- Need to implement modules to decide
  - intonation phrase (IP) boundaries
  - segmental durations
  - pitch (f_0) contour