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

- Speech synthesis takes text as input and output speech signal for the text. It is also known as text-to-speech (TTS).
- We will focus on *concatenative synthesis* here, but note the alternative *formant synthesis* using source-filter model, as well as *articulatory synthesis* using physical model for all articulators.
- One can also categorize synthesis systems as *rule-based* or *data-driven*.

Concatenative Synthesis

- limited-domain vs. unrestricted synthesis
  - A limited-domain synthesis offers high-quality synthesized speech with only a small number of recorded segments.
  - An unrestricted system is much more difficult.
- without waveform modification vs. with waveform modification
  - Although waveform modification provides flexibility in concatenation, it may also harm the naturalness of speech.

Attributes

- delay
- memory resource
- CPU resource
- variable speech rate
- pitch control
- voice characteristics

Design Issues

- What type of speech segments to use? diphone? phoneme? word?
- How to design the acoustic inventory from a set of recordings?
- How to select the best string of speech segments given a phonetic string and possibly its prosody?
- How to alter the segments to best fit the desired output prosody?

Choice of Units

- One must balance quality and quantity.
  - the longer the units, the better the quality
  - the longer the units, the larger the number of unit types
- design objectives
  - low distortion
  - generalizability (for unrestricted synthesis)
  - covered by training data

Concatenative Units

- CI phonemes
- diphones
- CD phonemes
- subphonetic units (acoustic states)
- syllable
- word and phrase

Decoding

- Choose the optimal string of unit instances for a given phonetic string with desired prosody.
- Synthesis quality is dominated by discontinuities at unit boundaries, due to
    - difference in phonetic context
    - incorrect segmentation
    - acoustic variability
    - difference in pitch

Objective Function

- ought to approximate synthesis quality
- ought to facilitate fast search
- We define *transition cost* and *unit cost* and use them in the dynamic-programming search of optimal unit string.

$$d(\Theta, T) = \sum_{j=1}^{N} d_u(\theta_j, T) + \sum_{j=1}^{N-1} d_t(\theta_j, \theta_{j+1})$$

$$\hat{\Theta} = \arg\max_{\Theta}\ d(\Theta, T)$$

Unit Inventory Design

- The minimal requirement is recording a number of utterances covering all units in the inventory.

Prosodic Modification

- goal: to change the amplitude, duration, or pitch of a speech segment
- methods
  - OLA
  - SOLA
  - PSOLA

Intelligibility Tests

- whether the synthesized speech is clean to a human listener
- diagnostic rhyme test (DRT): intelligibility of 96 pairs of initial consonants
- modified rhyme test (MRT) is a variant that include 50 six-word lists, each differing in initial consonants.
- phonetically balanced word lists are used to consider context effect
- semantically unpredictable sentences
- Harvard psychoacoustic sentences

Other Tests

- overall quality: mean opinion score (MOS)
- preference test: for comparing two systems directly, an ITU recommendation is comparison category rating (CCR)
- functional test
- automated test