5.3 An Example of Corpus-based Text-to-speech Synthesis

- **Basic Problems for Text-to-speech Synthesis**
  - text analysis
  - prosodic modeling
  - speech synthesis

- **Corpus Design/Construction**
  - phonetically balanced/phonetically rich

    to include as many different phonetic combinations as possible in a not-too-large corpus

    intra-syllabic/inter-syllabic/intra-word/inter-word structures, tonal combination patterns
  - prosodically rich

    to include as many different prosodic behavior as possible in a not-too-large corpus

    modalities of utterances : interrogative, exclamatory, declarative-word patterns for the modalities

    combination patterns of parts-of-speech/ punctuation marks

    maximizing bigram/trigram coverage for parts-of-speech/punctuation marks in the corpus
  - automatically selected from a large corpus
  - produced and recorded by selected speakers

- **Automatic Phonetic Labeling of Corpus**
  - alignment of the speech data with the HMMs for the known words in the corpus
  - automatic correction algorithm based on some phonetic combination rules (e.g. : silence + affricate) and some parameters (e.g. : short-time energy, sub-band energy from FFT)
- **Automatic Prosodic Labeling of Corpus**
  - 6 categories of breaks (break indices)

    B0 : reduced syllable boundary
    B1 : normal syllable boundary
    B2 : minor phrase boundary
    B3 : major phrase boundary
    B4 : breath group boundary
    B5 : prosodic group boundary
  - 4 levels of emphasis

    E0, E1, E2, E3
  - a parameter vector P_j is generated at the end of each syllable, and a break index BK (K = 0, 1, 2, …5) is labeled accordingly

    $(P_1, P_2, \ldots P_n) \rightarrow (BK_1, BK_2, \ldots BK_n)$
  - multiple-pass hierarchical break labeling : first locating B5, then B4 between B5's, then B3 between B4's, etc.

    one level of break located at a time
  - tree-structured decision algorithm trained by human-labeled corpus, then tested with unknown corpus

- **Hierarchical Prosodic Analysis for Arbitrary Sentences**
  - all words in training corpus tagged with parts-of-speech (POS's) automatically
  - "minor phrase table" for POS's of minor phrases between two B2 constructed including frequency counts based on training corpus
  - input arbitrary sentences segmented into words, tagged with POS's automatically
  - a lattice of possible "minor phrases" constructed by looking up the "minor phrase table"
  - best paths in the lattice found based on the frequency counts in the "minor phrase table", some extra rules, and a dynamic programming procedure
  - B3 then located similarly, B4 afterwards
- **Voice Units Selection, Modification and Speech Synthesis**
  - desired acoustic features found based on break indices
  - desired voice units (longer or shorter) found in the corpus
  - modification/concatenation performed when needed
  - final concatenation and smoothing gives the synthesized speech