Global TIMIT L1 English Treebank

Authors: Hui Feng, Wenchao He, Xiaoyan Hu, Yu Wu, Jiahong Yuan, Mark Liberman

Introduction

Global TIMIT L1 English Treebank is a TIMIT-like corpus in L1 English, part of the "Global TIMIT" series. The design of "Global TIMIT" adopts a scheme different from that of the original TIMIT. Instead of having 630 speakers and 10 sentences per speaker, the new design has 50 speakers and 120 sentences per speaker. Among the 120 sentences, 20 are "Calibration" sentences, read by all speakers; 40 are "Shared" sentences, read by 10 speakers; and 60 are "Unique" sentences, read by only one speaker. The total number of sentence types is, therefore, 20 + 40*(50/10) + 60*50 = 3220.

The sentences of English Treebank were selected from the corpus of Treebank-3 (LDC99T42). Each sentence contains 10 to 19 words. L1 English Treebank has 50 native American English speakers, 25 female and 25 male, recorded at University of Pennsylvania.

HMM/GMM-based forced alignment, with employment of explicit phone boundary models, was applied to obtain phonetic segmentation.

Data

The corpus contains 5993 utterances (7 were missing). Each utterance has five data files as listed below:

*.flac: flac files
*.phones: phone segmentation files
*words: word segmentation files
*tree: sentence parse trees from Treebank-3
*.TextGrid: Praat TextGrid files

Base filenames have the form S##_####, where the first digit string is a 0-padded subject number and the second is a 0-padded sentence number. Odd subject numbers (S01, S03, …) represent female speakers; and even subject numbers (S02, S04, …) represent male speakers. Sentences numbered from 0001 to 0019 are "Calibration" sentences, 0020-0059 are "Shared" sentences, and 0060-0119 are "Unique" sentences.