Global TIMIT L1 Simple English

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

Introduction

Global TIMIT L1 Simple English is a TIMIT-like corpus in L1 English, part of the "Global TIMIT" series. The sentences of the corpus were selected from the original TIMIT as "simple" to read by Chinese learners of English (which are the same sentences used in Global TIMIT L2 English Simple).

Each dataset has 50 speakers and 120 sentences per speaker. Among the 120 sentences, 20 are read by all speakers ("Calibration"); 40 are read by 10 speakers ("Shared-1"); and 60 are read by 5 speakers ("Shared-2"). The total number of sentence types is, therefore, 20 + 40*(50/10) + 60*(50/5) = 820. Global TIMIT L1 Simple English 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 5983 utterances (17 were missing). Each utterance has four data files as listed below:

> *.flac: flac files
> *.phones: phone segmentation files
> *.words: word segmentation files
> *.TextGrid: Praat TextGrid files

Base filenames have the form SP##_###, where the first digit string is a 0-padded subject number and the second is a 0-padded sentence number. Odd subject numbers (SP01, SP03, …) represent male speakers; and even subject numbers (SP02, SP04, …) represent female speakers. Sentences numbered from 001 to 020 are "Calibration" sentences, 021-060 are "Shared-1" sentences, and 061-120 are "Shared-2" sentences.