Global TIMIT Mandarin Chinese-Guanzhong Dialect

Authors: Yue Jiang, Juhong Zhan, Hongjian Han, Zuohao Xu, Haiyan Zhou, Jiahong Yuan, Mark Liberman

Introduction

Global TIMIT Mandarin Chinese-Guanzhong Dialect is a TIMIT-like corpus in Guanzhong dialect of Mandarin (in Shannxi province), 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 Global TIMIT Mandarin Chinese-Guanzhong Dialect were selected from the corpus of Chinese Gigaword Fifth Edition (LDC2011T13). Twenty "Calibration" sentences were selected to cover the maximum number of syllable types in the language. "Shared" sentences were selected to cover the maximum number of tones and tonal combinations. Finally, "Unique" sentences were randomly selected. 50 high school students, 25 females and 25 males, were recruited to read the sentences. The speakers were born in Weinan, Shannxi, and speak the Guanzhong dialect of Mandarin. The recording was made in a quiet room.

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

Data

The corpus contains 5999 utterances (one utterance was missing). Each utterance has file data files as listed below:

*.flac: flac files
*.phones: phone segmentation files
*.words: word segmentation files
*.tones: syllable/tone 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" sentences, and 061-120 are "Unique" sentences.