Global TIMIT Mandarin Chinese

Authors: Hongwei Ding, Sishi Liao, Yuqing Zhan, Jiahong Yuan, Mark Liberman

Introduction

Global TIMIT Mandarin Chinese is a TIMIT-like corpus in Mandarin Chinese, 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 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 college students at Shanghai Jiao Tong University, 25 females and 25 males, were recruited to read the sentences. The speakers all achieved Class 2 Level 1 or better on *Putonghua Shuiping Ceshi* (which is the national standard Mandarin proficiency test). The recording was made in a sound-treated recording booth.

HMM/GMM-based forced alignment, with employment of explicit phone boundary models, was applied to obtain phonetic segmentation. To evaluate the accuracy of automatic segmentation, 50 randomly selected sentences were manually corrected. Excluding the boundaries between silence and a stop or an affricate, where the boundary cannot be determined because of the stop closure, there are 1431 boundaries in the 50 sentences. 93.2% of the boundaries have an agreement of within 20 ms between forced alignment and manual segmentation, which is on par with state-of-the-art results in terms of accuracy of automatic phonetic segmentation.

Data

The corpus contains 6000 utterances. Each utterance has four data files as listed below:

*.flac: wave 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" sentences, and 061-120 are "Unique" sentences.