Abstract

The development of Indonesia Large Vocabulary Continuous Speech Recognition requires a phonetically balanced sentence set that accommodate as many triphones and punctuations as possible. However, the existing Indonesian phonetically balanced sentence set does not accommodate triphone optimally since it is extracted from only 500 thousand sentences. In addition, the sentence set does not pay attention to punctuation so it can not be used in the construction of a speech corpus that accommodates variations of intonation. This final project focuses on building phonetically balanced sentence set that accommodates triphone and punctuation. The sentence set is extracted from 10 million unique sentences using a proposed Semi LTM Greedy algorithm. Experiments show that Semi LTM Greedy is more flexible to balance the number of triphones and their frequencies than the previous Modified LTM Greedy algorithm.

Keywords: *Triphone*, *Greedy*, Corpus Text.