ABSTRACT

Syllable based Indonesian continuous speech recognition system needs a balanced corpus containing as many syllables and punctuations as possible. There's no corpus accommodating enough syllables due to its being built from a mother set composing of only around 500 thousand sentences. This research aims on tackling that very problem by using a mother set consisting of around 10 million sentences. Results show that several possible child set are feasible to be used as train sets for a continuous speech recognition system.

Keywords:*text corpus*, *greedy algorithm*, *syllable*, *punctuation*