Abstract

Large Vocabulary Continous Speech Recognition System or LVCSR is state of art of speech recognition system. This system is capable at recognizing various speech and words uttered by a person. The recognizing capability comes from training the system with reading speech corpus and spontaneous speech corpus.

Speech corpus is an integral element in order to train the system, especially spontaneous speech corpus. This corpus is a pronounciation reference for the system mentioned. For several languages such as English, building such recognition system is relatively easy as there are many speech corpus available. However, for several languages such as Bahasa Indonesia, the speech corpus is scarce in number.

Combining the design that exists in Eyra, Woefzela and Datahound, we develop a similar application for building speech corpus to ease the corpus development, especially in speech data acquisition.

Keywords: Under-resourced, spontaneous speech corpus, triphone coverage, question generation, balanced sentence set.