Abstract

This paper aims to see the impact of different sample rates toward voice cloning results on text-to-speech (TTS) system. TTS used in this paper based on Tacotron to handle voice coning and Soundex for swapping syllables. Pronunciation of English syllables and Indonesian syllables is different, so we need to swap them in order for the system that built based on the English language to produce speech in the Indonesian language. Swapping syllables based on phonetic similarity of both languages. The audio sample that will be cloned may be from a recorded voice, video camera, or even from other sources. Each audio sample has a different sample rate depending on their format and source of the audio, e.g. voice recorder with Audio CD format using 44100 Hz sample rate, camrecorder using 32000 Hz in general, and other sources with different sample rates. This paper found that the higher sample rate used, the clone result will have a high similarity voice with the sample, but the listeners are harder to recognize each word that they heard from the audio result of TTS.

Keywords: TTS, sample rate, phonetic, voice cloning, soundex