ABSTRACT

Nowadays, the development of Text-to-Speech has been growth very fast. In the beginner generation of Text-to-Speech, the sound that produced from this system is very unnatural. Now, the prosody of Text-to-Speech better than the past. Nowadays, development in Text-to-Speech in the area of it's visualitation.

The video consists of sound and moving picture. Text-to-Video is the integration of two main system, there are Text-to-Speech and Facial Animation. To produce sound in Text-to-Video using Text-to-Speech. While for producing moving picture, we use Facial Animation.

The block synthesizer in Text-to-Speech has three methods to implement, there are formant synthesizer, articulatory synthesizer, and concatenation synthesizer. The synthesizer that we use in this Final Project is diphone concatenation. In the moving picture, we use Facial Animation to produce the video. The main process of Facial Animation is morphing process of picture that has been available in our database. The method of morphing process that we used in this Final Project is Cross Dissolve.

The result of MOS (Mean Opinion Score) from 30 responders, shows that video which produced from Text-to-Video is good enough, moreover the MOS result from Facial Animation shows that animation is good (4.0733).