A Study of Cerebellum-Like Spiking Neural Networks for the Prosody Generation of Robotic Speech

A Doctoral Thesis submitted to

Graduate School of Engineering, Kagawa University

Vo Nhu Thanh

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Engineering

September 2017

Abstract

Speech synthesis has been an interesting subject for researchers in many years. There are two main approaches for speech synthesis, which are software-based systems and hardware-based systems. Between them, hardware-based synthesis system is a much appropriate tool to study and validate human vocalization mechanism. In this study, the author introduces a new version of Sawada talking robot with new design vocal cords, additional unvoiced mechanism, and new intelligent control algorithms.

Firstly, the previous version of the talking robot did not have voiceless speech system, so it has difficulty when generating fricative sounds. Thus, a voiceless speech system which provides a separated airflow input is added to the current system in order to let the robot generate fricative sounds. The voiceless system consists of a motor controlled valve and a buffer chamber. The experimental results indicate that the robot with this system is able to generate fricative sound at a certain level. This is significant in hardware-based speech synthesis, especially when synthesizing a foreign language that contains many fricative sounds.

The intonation and pitch are two important prosodic features which are determined by the artificial vocal cords. A newly redesigned vocal cords, which its mechanism is controlled by a servomotor, is developed. The new vocal cords provide the fundamental frequency from 50 Hz to 250 Hz, depending on the air pressure and the tension of the vocal cord. The significant contribution of this vocal cords to speech synthesis field is that it provides the widest pitch range for a hardware-based speech synthesis systems so far. Thus, it greatly increase the synthesizing capability of the system

Most of the existing hardware speech synthesis systems are developed to generate a specific language; accordingly, these systems have many difficulties in generating a new language. A real-time interactive modification system, which allows a user to visualize and manually adjust the articulation of the artificial vocal system in real-time to get much precise sound output, is also developed. Novel formula about the formant frequency change due to vocal tract motor movements are derived from acoustic resonance theory. Based on this formula, a strategy to interactively modify the speech is established. The experimental result of synthesizing German speech using this system give an improvement of more than 50% in the similarity between human sound and robot sound. The contribution of this system provides a useful tool for speech synthesis system to generate new language sounds with higher precision.

The ability to mimic human vocal sounds and reproduce a sentence is also an important feature for speech synthesis system. In this study, a new algorithm, which allows the talking robot to repeat a sequence of human sounds, is introduced. A novel method based on short-time energy analysis is used to extract a human speech and translate into a sequence of sound elements for the sequence of vowels reproduction. Several features include linear predictive coding (LPC), partial correlation coefficients

(PARCOR) and formant frequencies are applied for phoneme recognition. The average percentage of properly generated sounds are 53, 64.5, 73, and 75 for cross-correlation, LPC, PARCOR, and formant method, respectively. The results indicate that PARCOR and formant method achieve high accuracy, and it is suitable for applying in speech synthesis system for generating phrases and sentence.

A new text-to-speech (TTS) is also developed for the talking robot based on the association of the input texts and the motor vector parameters for effectively training the auditory impaired hearing patient to vocalize. The intonation feature is also employed in this system. The TTS system delivers clear sound for Japanese language but needs some improvement for synthesizing foreign language.

For prosody generation, the author pays attention to the employment of cerebellum-like neural network to control the speech-timing characteristic in vocalization. Using bio-realistic neural network as robot controller is the tendency robotics field. Thus, for the timing function of the talking robot, a cerebellum-like neural network is implemented to FPGA board for timing signal processing. This neural network provides short-range learning ability for the talking robot. For the experimental result, the robot can learn to produce the sound with duration less than 1.2 seconds. The significant contribution of this section is that it proposes the fundamental finding to construct and apply the bio-realistic neural network to control the human-like vocalization system. Confirming the timing encodes within the cerebellum-like neural network is another contribution of this section.

This study focused on the prosody of a speech generated by a mechanical vocalization system. This dissertation is summarized as follow: (1) the mechanical system is upgraded with newly redesigned vocal cords for intonation and an additional voiceless sound system for fricative sound generation, (2) new algorithms are developed for sentence regeneration, text to speech, and real-time interactive modification, (3) the introduction of a timing function using cerebellum-like mechanism installed in an FPGA board is employed in the talking robot.