氏　　　名（本籍）	Vo Nhu Thanh（ベトナム）
専　　　　　攻	知能機械システム工学専攻
学位の種類	博士（工学）
学位記番号	博甲第127号
学位授与の要件	学位規則第4条第1項該当者
学位授与の年月日	平成29年9月29日
学位論文題目	A Study of Cerebellum-Like Spiking Neural Networks for the Prosody Generation of Robotic Speech
論文審査委員	（主査）　石井　明 （副査）　郭　書祥 （副査）　平田　英之 （副査）　澤田　秀之

論文内容の要旨

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.

審査結果の要旨

The defense committee examined the doctoral dissertation entitled "A Study of Cerebellum-like Spiking Neural Networks for the Prosody Generation of Robotic Speech", submitted by the above applicant for a doctoral degree. The aim of this study is to design a new version of a talking robot and its intelligent control algorithms based on the cerebellum-like spiking neural networks for the prosody generation. Speech synthesis has been an attractive subject for researchers, and two different approaches are primarily being conducted recently, the software-based method and the hardware-based method. A hardware-based synthesis system would be an appropriate tool to study and validate the human vocalization mechanism using robotic technologies. In this study, the author develops a new version of a talking robot by installing new vocal cords and the control methods for the different vocalization of voiced and unvoiced sounds with various pitches, and installs a cerebellum-like spiking neural network, which is the reproduction of the human brain activities in learning, for prosodic speech generation.

The dissertation consists of 8 chapters as follows.

Chapter 1 is concerned with the objectives and background of this study, and gives the research targets and outline of the dissertation.

Chapter 2 reviews the previous and existing works on speech synthesis systems and biological neural networks. This chapter presents literature on the algorithms that have been used in human-like speech generation from the viewpoint of computation and robotic technologies. Another issue presented in this chapter is artificial neural networks to be compared with the spiking neural networks that have been introduced and studied by the author. In addition, it overviews and discusses the advantages and disadvantages of all the algorithms used, for clarifying the objective of the study.

Chapter 3 gives an overview of the developed talking robot, together with learning algorithms of speech articulation based on the auditory feedback control using the self-organizing neural network. The mechanical ability of the intonation and pitch control, which is the greatest among the previous and existing mechanical speaking systems, is described in this chapter.

In Chapter 4, the design and the establishment of an interaction between the talking robot and a human is described for the interactive robotic speech modification. Novel formulas about the formant frequency change due to vocal tract motor movements are derived from acoustic resonance theory, and a strategy to interactively modify the speech articulation is established for the learning of difficult articulations such as foreign language and uncertain speech sounds.

Chapter 5 describes a sentence reproduction system for the natural robotic speech. The ability to mimic human vocal sounds and reproduce smooth sentences is also an important feature for a 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 it into a sequence of sound elements for the sequence of vowel reproduction. In this study, several acoustic features including the linear predictive coding (LPC), the partial correlation coefficients (PARCOR) and the formant frequencies are examined and applied for the phoneme recognition in human speech.

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 auditorily hearing impaired patients to vocalize. The TTS system delivers clear sounds for Japanese language, however it requires some improvements for synthesizing foreign language, which is described in Chapter 6.

Chapter 7 presents the novel cerebellum-like neural network, which is a bio-realistic network of the human brain activities in learning, to control the speech-timing characteristics in robotic vocalization. The cerebellum-like neural network is implemented to a FPGA board for timing signal processing, and provides short-range learning ability for the talking robot. From the experimental results, the robot proved to learn the reproduction of human speech with different sound durations. The significant contribution of this chapter is that it proposes the fundamental finding to construct and apply the bio-realistic neural network to control the human-like vocalization system.

Chapter 8 concludes the dissertation with a summary of the work that has been accomplished, together with the prospects for the application to the new robotic technologies and the bio-realistic neural networks.

As presented above, the dissertation describes that (1) the mechanical system is upgraded with newly redesigned vocal cords for intonation and additional vocal sound generation, (2) new algorithms are developed for the sentence regeneration, the text-to-speech, and the real-time interactive modification, and (3) the introduction of a timing function using cerebellum-like mechanism installed in an FPGA board is employed in the talking robot for the speech timing control.

The achievements presented in the dissertation were published in the following two journal papers and one international conference paper as the first author. All the publications were made during his doctoral period.

[1] Vo Nhu Thanh and Hideyuki Sawada, "Automatic Vowel Sequence Reproduction for a Talking Robot Based on PARCOR Coefficient Template Matching", IEIE Transactions on Smart Processing and Computing, Vol. 5, No.3, pp.215-221, June 2016

[2] Vo Nhu Thanh and Hideyuki Sawada, "A Talking Robot and Its Real-Time Interactive Modification for Speech Clarification", SICE Journal of Control, Measurement, and System Integration, Vol. 9, No.6, pp. 251-256, November 2016

[3] Vo Nhu Thanh and Hideyuki Sawada, "Cerebellum-like Neural Network for Short-range Timing Function of A Robotics Speaking System", The 3rd International Conference on Control, Automation and Robotics (ICCAR 2017), pp. 184-187, April 2017

最終試験結果の要旨

平成 29 年 7 月 28 日に公聴会ならびに最終試験を実施した。公聴会では、申請者が学位論文「機械式音声合成による韻律生成のための、小脳モデルに基づくスパイキングニューラルネットワークに関する研究（和訳）」の内容に関する発表を行った（1 時間）。その後、口述試験として学位論文の内容に関わる審査委員の質疑に的確に答えることを求め、更に、学位論文に関連した分野の専門知識の確認を行って最終試験とした（1 時間）。

最終試験における学位論文に対する質疑応答の概要は以下のとおりであり、申請者はすべて的確に回答した。

1) Cerebellum Neural Network によって、従来の機械式音声生成システムと比較してどれ程良くなったのかを、定量的なデータを示して述べてほしい。

（回答）まず、声帯部の新しい機構と制御手法を開発した結果、音域が 50～280Hz と従

来システムの 3 倍程に広がった。また、Cerebellum Neural Network を音声学習に導入したことにより、人間の発話を十分に再現させるような音韻特徴の生成が可能となった。このような音韻特徴の生成は、従来システムの狭い音域と単純な NN では不可能であった。

2) 人の音声波形から、どのようにエネルギーを算出し、雑音除去、音素の切り出しを行っているのか、更にこれを元に如何に音声解析して、機械的に母音を再生するのか、解りやすく述べてほしい。

（回答）グラフは、横軸が時間で、縦軸が音声波形の振幅になっている。つまり、音声信号のパワーの時間変化を表したものである。このデータを元に、発話時間の切り出しを行ない、音素の認識が実行される。認識した音素に対して音響パラメータを抽出し、Neural Network を使って発話動作との関係を学習している。この学習データを元に、母音音声が機械的に生成されることになる。

3) ビデオデモでは、*/a/*と*/o/*の音の違いが解りにくかった。生成音声の良し悪しをどのようにして評価しているか。

（回答）録音のクオリティが良くなかったのかも知れない。LPC, PARCOR 係数などで音響特徴を解析しているが、数値としては明確に違いが出ている。また、ロボットが獲得した発話動作と口内形状にも違いが現れていることが解る。

4) Spiking NN を導入した理由と導入効果について説明して欲しい。また従来の Spiking NN の課題と、実応用例が少なかったのはなぜか？

（回答）人間の発話の生体機構、脳の学習機構についての biological studies に関して、多くの文献を読んで学習し、Cerebellum Neural Network を着想した。従来研究における Cerebellum Neural Network は、モデルの提案とシミュレーションのレベルに留まっており、ロボットの制御などに使っている例は見られない。学習パラメータが膨大にあり、不安定な機械機構の制御に応用するためには繊細な調整が要求されるため、このことが実応用例がなかった理由と考えられる。実際、本研究では、ロボットの自律学習機構として実装したことが最も苦労した。学習プログラムの構築、実装と自律学習の実行では、パラメータが膨大にあり、これらを調整して脳の機能を上手く再現させることに多くのエフォートをかけた。本研究で、実際のロボットの自律学習と制御に実装してその有用性を示したことは、初の成果であると考えている。

以上、本審査委員会は、学位論文、公聴会及び最終試験における研究内容説明および質疑応答から判断して、申請者が提出した論文は、その新規性と学術的価値から博士（工学）の学位に値するものであり、また、申請者が専門領域に関する十分な学識と研究能力を有すると判断した。よって、本最終試験の評価を合格とする。