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.

Acknowledgements

My sincerest thanks to all those who helped make this thesis possible. This Ph.D. degree is a big milestone of my life.

From the bottom of my heart, I would like to thank my adviser, Professor Hideyuki SAWADA, for his endless support and encouragement. Many thanks to him for getting me interested in this research by providing tips and advice. He introduced me to research on human speech synthesis and mechanical vocalization systems. I learned a lot from him about how to write a good technical paper, and how to make a good presentation when participating in seminars and international conferences. After more than 3 years working on my Ph.D. under his supervision, I have been trained to be a researcher and obtained a basic knowledge as a prerequisite to carry out my own research in my own country.

I gratefully acknowledge Prof. Ishii for giving me constructive comments and helpful advice to improve the content of the thesis. I would also like to thank Professor Hirata and Professor Guo for their valuable comments, suggestion, and advice in the preliminary defense.

I could summarize that the purposes of the Ph.D. study are not only to obtain a good achievement in research but also on how to report the research outcomes to the world by writing technical papers and joining international meetings. The most important things are that we should share the research outcomes with others, and together discuss the direction of research in the future. Even though we have a good result from researching, if we do not report it to the world, the situation is the same as if we had done nothing.

I would like to thank the Japanese Government for awarding me the MEXT scholarship to pursue the doctoral degree. I really appreciate the kind support of Ms. Ai Sakamoto, Ms. Asano, and all of the staff in the International Office during my stay at Kagawa University. I would also like to thank the administration staff at the Faculty of Engineering and the Department of Intelligent Mechanical Systems, Kagawa University for giving me full co-operation. I am also thankful to my Vietnamese, Japanese, and foreign friends in Japan for their friendship and kindness. I would also like to thank many people outside the laboratory. I am also grateful to my friends, all Vietnamese students in Kagawa University for their contributions to the excitement of my research journey.

I would not be successful without my family and my dear friends, who have always stood beside me and given me their love no matter what. My parents and siblings have been my inspiration for all of my life. I wish to express my affectionate gratitude to the woman of my life, Phan Nguyen Linh Thao, who has always stood by my side. Last but not least, I would like to extend my warmest thanks to our dear son, Vo Nhu Khoi, the center of our love and the motivation for me to complete the research work.

Vo Nhu Thanh
Kagawa, Japan, 2017

TABLE OF CONTENTS

Chapter **Page**

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Humans use their voice as a primary communication method in daily activity [1]. Although the animal has a voice, only humans can use their voices to communicate with each other effectively. It doesn’t matter how advanced the technology, the importance of speech communication remains unchanged in human society. The reason is that speech is the easiest way to exchange information without using any other devices. The human sound is produced by complicated movements of the vocal organs. In speech production, the vocal cords are vibrated at a certain frequency to generate a sound source. Besides, the air turbulence caused by narrowing or instantaneous changes in the oral cavity generates the sound source for some consonants. After that, the sound source is led through the human vocal track. Various resonances occur according to the shape of the vocal tract to determine the output sound characteristic which is perceived by the human ear as a voice. Human speech production is a complicated mechanism which has drawn the attention of many researchers for a long time [1].

The earliest speech synthesis system was the Brazen Head created by Roger Bacon [2] in the 13th century which was a self-operating machine that is able to answer several simple questions or to speak a simple sentence as shown in Figure 1.1. Since then, the mechanism of speech production began to attract the attentions of researchers. The mathematical models of speech production are also investigated which provided a good knowledge for making a human-like vocalization system. The static synthesis systems such as the static vocal tracts were created to mimic a human-like sound. As the modern technology advanced, the software synthesis to produce a speech through a speaker was established. Finally, with the fast growing field of robotics, many advanced mechanical speech synthesis systems were developed.

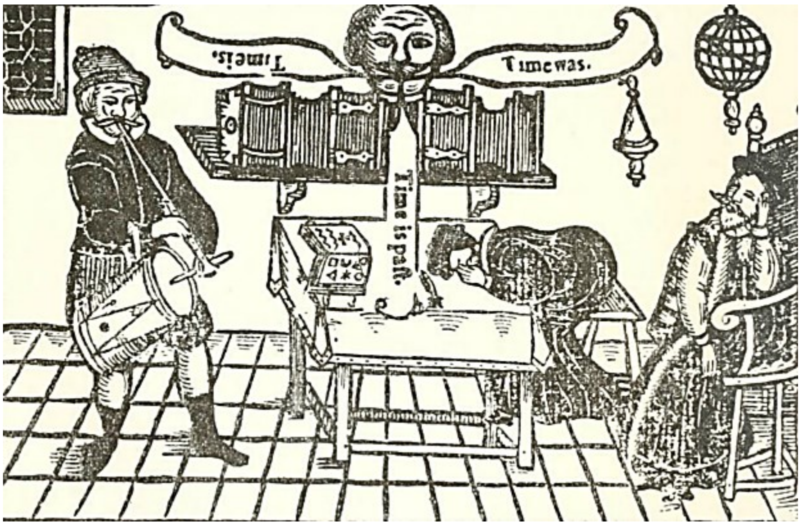


Figure 1.1: Brazen Head speaks: "Time is. Time was. Time is past."

In the Sawada laboratory of Kagawa University, an engineering team develops an automatic mechanical vocalization system, which is an anthropomorphic robot, to produce a sound in a similar way to the way the human vocalization system works [3]-[7]. The prosody of speech, which is determined by the fundamental frequency, duration, and stress, is an important factor of human voice [8]. However, almost none of the mechanical vocalization systems include these prosodic features in their speech production up to this moment. These features are sometimes called suprasegmentally features which are considered as the melody, rhythm, and emphasis of the speech at the perceptual level. In addition, generating a sound with precise intonation, stress, and duration is possibly one of the toughest challenges for hardware speech synthesis system today. The intonation is defined as how the pitch pattern or fundamental frequency changes during the speech. The prosody of continuous speech depends on different characteristics, such as the speaker individualities and feelings, and the purpose of the sentence (question, normal, or impression). The timing of each phoneme at sentence level is also important in speech generation. If there are no pauses when talking or the pauses are put in the wrong places, the output sound will be very strange or the meaning of the sentence is entirely different. For example, the sentence "Tom says Jerry is a bad guy" can be spoken as "Tom says*, [pause]* Jerry is a bad guy" which means Jerry is a bad guy or "Tom, *[pause]* says Jerry*, [pause]* is a bad guy" which means Tom is a bad guy. Similarly, for the Japanese language, “Hashi” has two different meanings (“bridge” or “chopstick”) if it is spoken differently [9]. Some significant features of prosody are pitch, intonation, tempo and stress, and they are subjected to the vocalization characteristics of the talking robot in this study. The prosodic features are shown in Figure 1.2 [10].

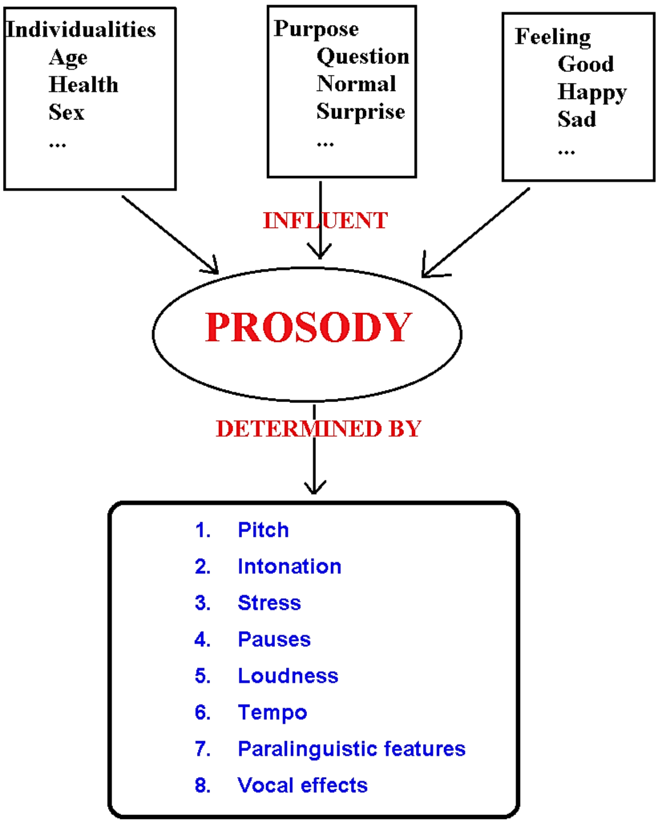


Figure 1.2: The prosody of speech

In this study, the author provides for the talking robot the ability to generate a speech that has prosodic features by two approaches which are the modification and upgrade of the current mechanical system and the development and implementation of new control algorithms to it. In specific, for the refined talking robot version, the author designs and adds a voiceless system that allows the robot to speak fricative sounds. The voiceless system consists of a motor controlled valve and a buffer chamber to control the amount of airflow to the narrow cross-sectional area of vocal tract which generates a fricative sound. The intonation characteristic of output sound is determined by a newly redesigned artificial vocal cord. The fundamental frequency of the output sound varies from 50 Hz to 250 Hz depends on the air pressure and the tension of the vocal cord. A newly redesigned vocal cord, which its tension is easily controlled by a motor, provides the widest pitch range for a mechanical vocalization system so far in comparison with other available mechanical vocalization systems (Waseda Talker, Burby Robot …). The speaking capability of the talking robot is greatly increased with this newly redesigned vocal cord.

Secondly, the control software of the talking robot is also upgraded using the Matlab software which includes many built-in functions for sound processing and motor control [10]. Most of hardware-based speech synthesis systems are developed to generate a specific language; thus, these systems have many difficulties in producing a new language. For the Japanese language, nearly all the syllables are in a consonant-vowel form which makes the synthesis easier than with other languages. Numerous languages include special features that make the development of the speech synthesis system either easier or harder. For instance, some languages, such as Italian and Spanish, have very regular pronunciation with almost one-to-one correspondence with a letter to sound and it is a little easier to develop a speech synthesis system. While some other languages, such as French and German, have many irregular pronunciations, so the development process is a bit harder. For some languages that have a tonal effect, such as Vietnamese [11], the intonation is significant because it can change the meaning of the spoken word totally. The prosodic characteristic is important when a speech synthesis system learns to speak a foreign language. Thus, a real-time interactive modification system, which allows the user to visualize and manually adjust the motions of the artificial vocal system in real-time to get a more precise sound output especially when it learns to vocalize new language, is developed. An original equation about the formant frequency change due to vocal tract motor movements is derived from acoustic resonance theory. Based on this equation, the strategy to interactively modify the speech is established. This interactive modification system greatly increases the speaking capability of the talking robot in term of foreign language generation. [3]

The prosodic features of speech are important when speaking a sentence or a phrase. Thus, the capability to mimic the human sound and reproduce a sentence is another important aspect of mechanical speech synthesis system. However, this capability has not been introduced in any mechanical vocalization system so far. 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. Then, several phonemes detection methods including the direct cross-correlation analysis, the linear predictive coding (LPC) association, the partial correlation (PARCOR) coefficients analysis, and the formant frequencies comparison are applied to each sound element to give the corrected command for the talking robot to repeat the sound sequentially. [12]

A new text-to-speech (TTS) system is also developed for the talking robot based on the association of the input texts and the motor vector parameters for more effectively training the auditory impaired hearing patient to vocalize. The intonation feature is also added to this TTS system so the talking robot can be able to produce an input text with different tonal effect. This system is very helpful when letting the robot generate a tonal language like Vietnamese or Chinese. The author notices that this is the only hardware-based TTS system that has tonal effect at the moment.

Timing property is very important in the prosody of speech. The timing function for the talking robot can be straightforwardly calculated by using mathematical analysis and be implemented to the talking robot. However, the author has a different approach, and that is using a human-like spiking neural network to control a human-like mechanical system. The spiking neural network is a third generation of neural network. It is different than the traditional neural network because of its capability of encoding the timing characteristic in its neural signal pattern. Also, the cerebellum has been widely known for its role in precision, coordination and accurate timing of motor control [13], [14]. Therefore, in this study, the author designs and builds a simple cerebellum-like neural network using bio-realistic spiking neuron, and uses it to control the timing function of the talking robot. In detail, a cerebellum-like neural network is implemented to FPGA board for timing signal processing. Then, the timing data is combined with motor position control function to fully control the talking robot's speech output that has prosodic features.

1.2 Human Articulatory System

Speech is a complex process that must be learned. It takes roughly about 3 to 5 years for a child to properly learn how to speak. The speech learning process is done via many trials of speaking and hearing. The information is stored and organized in the central nervous system to control the speech function. Impairment of any part of the vocalization system or auditory control area within the brain will degrade the performance of speech generation [1].

1.2.1 Vocal Organ

People use vocal organs to generate speech. Figure 1.3 shows the structure of human vocal organ (adapted from http://roble.pntic.mec.es). Vocal organs include lungs, vocal cords, nasal cavities, vocal tract, tongue, and lips. These vocal organs form a single continuous tube as a whole. The air discharged from the lungs by the abdominal muscles pushing up the diaphragm passes through the trachea and then passes through the glottis. While in normal breathing, this glottis is wide open, but when speaking the vocal cords physically come close to each other. As the airflow from the lung attempts to pass through this intersection, the interaction between the air flow and the vocal cords periodically opens and closes the glottis, resulting in a harmonic sound wave. This is the sound source of the sound. It is known that this sound source wave can be approximated by an asymmetrical triangular wave.

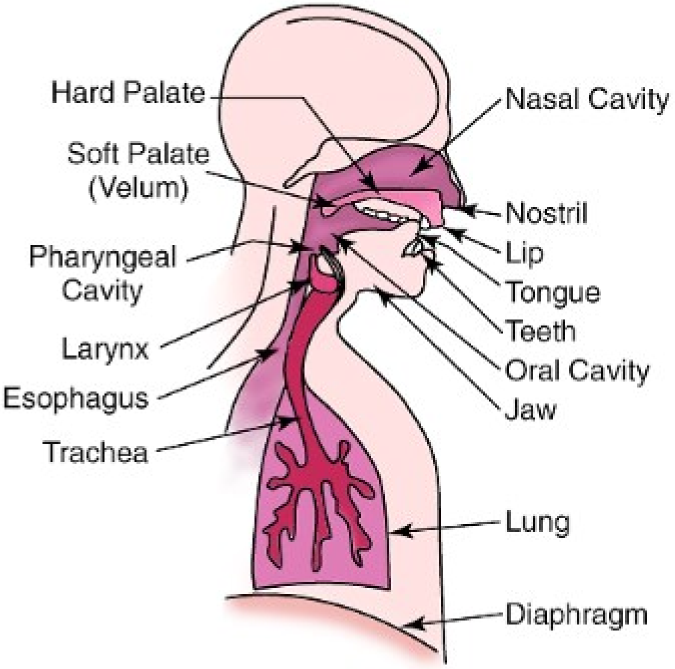


Figure 1.3: Structure of human vocal organ

The part above the larynx is called a vocal tract, and it has a length of about 15 ~ 17 cm in adults, and it can form into various shapes by the movement of the jaw, tongue, and lips. As a result, acoustic features are added to the sound source waves, and they are emitted as sounds from the lips. Also, the nasal cavity can be opened or closed by the movement of the soft palate and posterior tongue. When generating nasal consonants, the soft palate and posterior tongue come close to the vocal tract. Thus, the sound wave comes toward the nostrils and released. This nasal air flow resonates within the nasal cavity. As a result, the nasal consonant is generated, and it is added to acoustic features of the output sound.

1.2.2 Vocalization Mechanism

Vocalization is performed by firstly making a sound source. The sound source is the fundamental element of articulation. A voiced sound source caused by the vibration of a vocal cord which is commonly known as phonation. This provides the periodic sound source for all voiced speech sounds. The unvoiced sound source, which is known as the turbulent sound source, is the source for noise-type sound in speech. The unvoiced sound source is necessary for generating fricative, whisper, and aspiration sound. It is caused by a turbulent flow through a narrow cross-sectional area. Among the turbulent sound sources, the ones caused by the narrow cross-sectional area of the larynx are called breathing noise sources, and those caused by the narrow cross-sectional in the vocal tract are called frictional noise sources [1]. Figure 1.4 (adapted from http://studylib.net/doc/5604423/lecture-6) shows a schematic diagram of vocal cords vibration. The vocal cord vibration mechanism is described below.

- Vocal cords close due to muscle contraction during vocalization.
- The pressure of the glottis is increased due to the closure of the glottis.
- The air from the lung passes through the glottis between the left and right of the vocal cords.
- Vocal cords are pushed up by exhalation pressure, and the upper part of the vocal cords is opened.
- The air flows between the vocal cords.
- Exhalation flows out from the glottis gap which temporarily lowers the subglottic pressure.
- The air blows out to the vocal tract as an acoustic shock.
- Glottis closes due to Bernoulli effect of the vocal cords elasticity and exhalation flow.
- Repeat a series of actions.

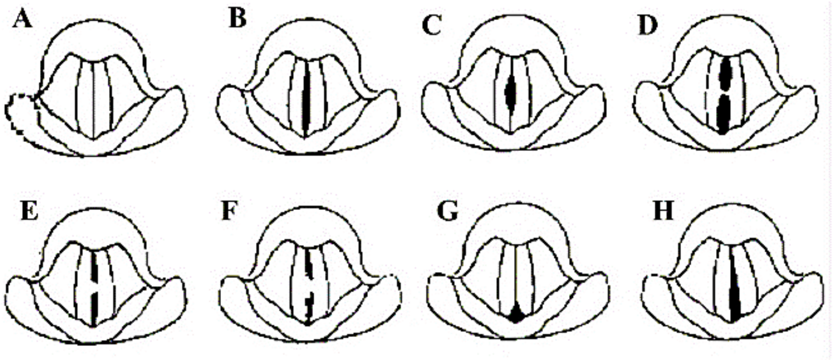


Figure 1.4 Vocal cords vibration: (A) Complete closure, (B) Spindle-shaped gap along entire edge, (C) Spindle-shaped gap at middle, (D) Hourglass-shape gap, (E) Gap by unilateral oval mass, (F) Gap with irregular shape, (G) Gap at post glottis, (H) Gap along entire length

The sound source generated by the vibration of the vocal cords is emitted from the lips with various acoustic features controlled by the movements of the articulatory organs. The vocal tract is the non-uniform tube from the glottis to the lips that has a significant contribution to the generation of the speech. The vocal tract acts as an acoustic filter and is deformed variously by the movement of the mandible, tongue, lips and palate sail, and the resonance feature is added to the sound source. The average length of the vocal tract is about 180 mm for an adult male, 150 mm for an adult female, and 122 mm for a child. The pharynx, the oral cavity, and the nasal cavity also affect the articulation, and they add resonance characteristic of the output sound. The ratio of the pharynx and oral cavity length, which determines the shape of the vocal tract, are different between men, women, and children. Assuming the pharynx length has the value of 1, the proportion of the oral cavity to the total length of the vocal tract is large for women and children. The proportion of the oral cavity to the total length of the vocal tract is 0.49 for men, 0.60 for women, and 0.66 for children.

The joint of the mandible is in front of the ear canal. Movement of the mandible can be regarded as rotational motion about the temporomandibular joint. A group of muscles connecting the lower jaw and hyoid bone activates when opening the lower jaw. Also, the muscles connecting the mandible and cranium are active when closing the mandible. Movement of the lower jaw occurs synchronously with the movement of the tongue. When the tongue moves upwards, the lower jaw closes. When the tongue moves downward, the lower jaw opens.

Movement of the tongue is manipulated by the external tongue muscles and the internal tongue muscles which are called extrinsic and intrinsic, respectively. There are five extrinsic muscles, which are the genioglossus, the hyoglossus, the chondroglossus, the styloglossus, and the palatoglossus, that extend from bone to the tongue. The extrinsic muscles control the protrusion, retraction, and side-to-side movement of the tongue. Four paired intrinsic muscles, which are the superior longitudinal muscle, the inferior longitudinal muscle, the vertical muscle, and the transverse muscle, are located within the tongue and attached along its length. These muscles manipulate the shape of the tongue by expansion and shortening it, curling and uncurling its apex and edges, and flattening and rounding its surface. This provides various shapes for vocal tract area and helps facilitate speech.

The lips are controlled by facial muscles and are involved in the articulation of vowels and consonants by the protrusion, closing, rolling, lateral opening, and other movements. The movement of the levator veli palatini muscle manipulates the movement of the palate sail, which affects nasal cavity resonance. When generating sounds other than nasal consonants, the palate sail raises backward and upward, and the pharynx together with nasal passages are closed. When generating nasal sounds, the muscles lower and the palatal sail descends, and that allows the air pass to the nasal cavity. Since the motion of the palate sail is slower than the movement of other articulatory organs, vowels sounded before and after the nasal consonant are also nasalized.

1.2.3 Sound Characteristic

Speech waveform:

A waveform is defined as a two-dimensional representation of a sound. The horizontal and vertical dimensions in a waveform respectively display time and amplitude. Thus, the waveform is also known as a time domain representation of sound as it shows the changes of sound amplitude over time. The sound amplitude is actually the measurement of the vibration of the air pressure**.** The greater the amplitude, the louder the sound a person perceives. The physical characteristics of speech waveform include wavelength and frequency. Wavelength is defined as a period of time when the waveform starts to repeat itself. Frequency is defined as the number of times the pattern repeats over a second and is measured in Hertz (Hz). A human can hear a sound within 20 Hz to 20,000 Hz range.

The sound people hear every day is a compound sound. The pure tone sound can be represented by a sine wave, but the complex tone is the combination of many pure tones and forms a complicated waveform. Thus, it is necessary to analyze the sound waveform for speech synthesis purpose.

Vowel sounds:

The vowel sounds are produced by the vibration of vocal cords together with the resonation of the vocal tract. During vowel generation, the vocal tract is opened in a stable configuration. There is no build-up of air pressure at any point within the vocal tract. In Japanese language, there are five main vowels of */a/, /i/, /u/, /e/, /o/.* In speech synthesis, vowel sound is much easier to synthesize in comparison with a consonant sound. The resonance frequency band of the vocal tract is called a formant, and it characterizes each vowel. The number of formant frequency is different for each sound. However, the first two formant frequencies are most important because the relation between the first formant and the second formant is important in discriminating vowels. Figure 1.5 shows the relationship between the first formant and the second formant of Japanese 5 vowels.

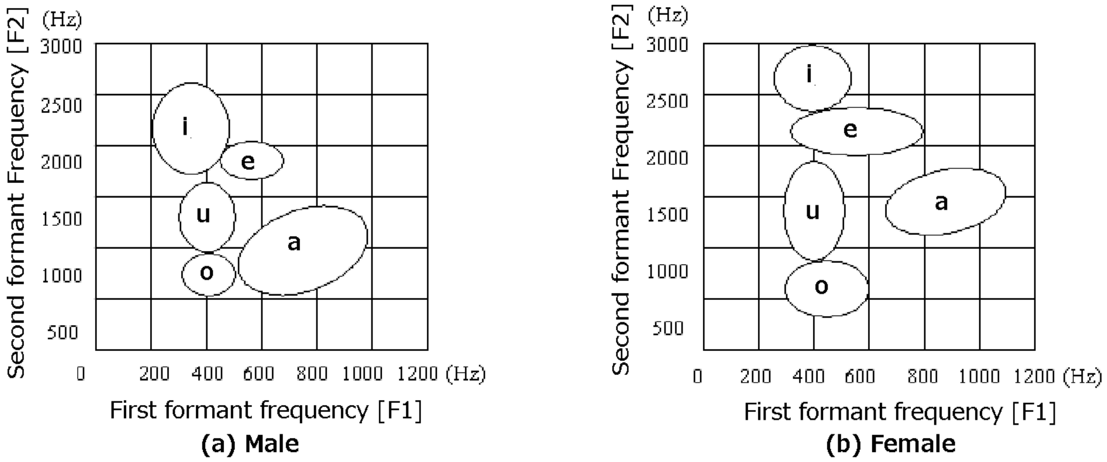


Figure 1.5 Nature of vowel

Consonant sounds:

By definition, a consonant is a speech sound that is articulated with a relatively unstable vocal configuration, especially the vocal tract shape. Depending on a consonant, the vocal tract shape can be either completely or partially closed for a very short time. For example, */p/* pronounced with the lips; */t/* pronounced with the front of the tongue; */k/* pronounced with the back of the tongue; */h/* pronounced in the throat; */f/* and */s/* (fricative sounds) pronounced by forcing air through a narrow channel; and */m/* and */n/* (nasal sounds) have air flowing through the nose. Different languages have unique consonant sounds such as */ch/, /tr/* consonant in the Vietnamese language which is a combination of fricative and plosive sound. The consonant sound usually combines with the vowel sound to form a syllable sound. In the Japanese language, most of the sounds are in CV form (consonant followed by a vowel). The vocal tract cross-sectional area can be very narrow or completely closed by movement of the tongue, soft palate, and lips, and these movements are combined to generate a consonant sound. When a strong air current is generated at the stenotic site, turbulence can occur, and the fricative sound is generated. When the air from the lung is completely blocked by the tongue, lips, etc., then suddenly opened, the air current flows intensely, and a plosive sound is generated. The location of stenosis or complete closure is called the articulation point. Table 1.1 shows the classification of consonants, and Figure 1.6 shows the articulation points of each consonant.

Table 1.1 Classification of Consonants

	Stop	**Fricative**	**Affricate**	**Nasal**	**Liquid**	**Glide**
Bilabial	p,**b**			m		w
Labio-dental		f,**v**				
Dental		th				
Alveolar	t,**d**	s,**z**		n	l,r	
Palatal		sh,**s**	ch,**j,g**			j
Velar	k,**g**			**ng**		
Glottal		h				
Voiced sound are **bold**						

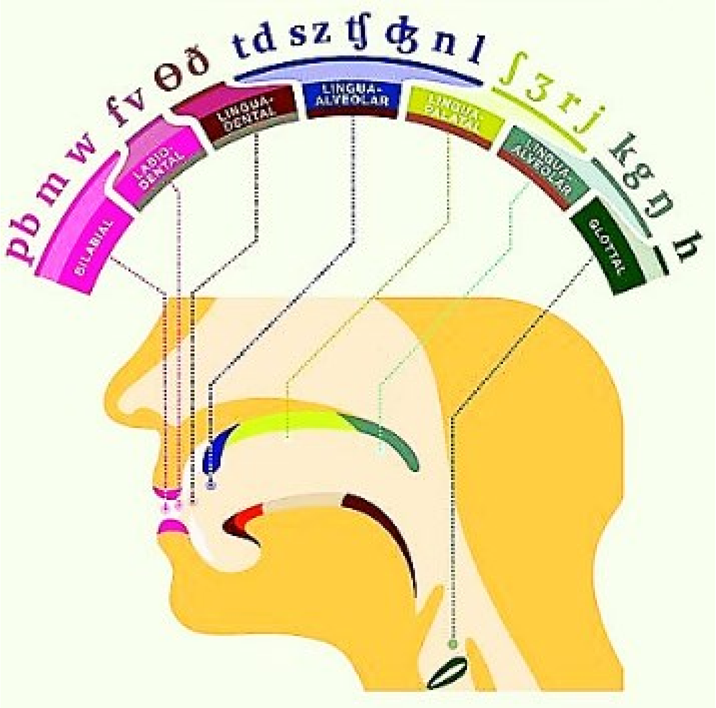


Figure 1.6 Point of a consonant in vocal organ

Plosive sounds:

Plosives or stop sounds are consonant sounds that are formed by completely stopping the sudden release of airflow. Plosive sounds such as */p/, /t/,* and */k/* are voiceless sounds, while */b/, /d/,* and */g/* are voiced sounds.

The total duration of articulatory closure for plosive sound is about 50 to 100 milliseconds then sudden opening of the the vocal tract causes the air pressure to instantly increase. Because the vocal tract is closed, little or no sound is generated. However, when releasing, an immense source of energy is formed as blocked air flows out. The duration of opening phase in plosive sound articulation is about 5 to 40 milliseconds. In order to identify a plosive sound, the analysis tool must have a very short time resolution, and typically it is about 5-10 milliseconds.

Fricative sounds:

Fricatives are consonants produced by forcing air through a narrow cross-sectional area of the vocal tract. The narrow cross-sectional area may be formed by the lower lip against the upper teeth; the back of the tongue against the soft palate; or the side of the tongue against the molars. When the air passes through the stenosis at an appropriate flow rate, turbulence is produced as a result. Turbulence airflow becomes significantly complicated in the region right behind the narrowed section of the vocal tract.This turbulent airflow is called frication. Sibilants are the subset of fricative sound. The mechanism to form sibilant sound is similar to fricatives together with the tongue is curled lengthwise to direct the air to the edge of the teeth. Some sibilant consonants are */s/, /z/, /sh/,* and */j/.*

Turbulent flow aerodynamic conditions are related to turbulence noise generation in acoustic signals. As a result, fricatives can be classified by the following characteristics. They are the position where the turbulence noise is formed in the vocal tract and the development of the turbulence (voiced or unvoiced). If the vocal cords operate in conjunction with the noise source, the sound is classified as voiced fricative. In contrast, if only the noise source is used without the operation of vocal cords, then it is classified as unvoiced fricative

Nasal consonants:

Nasal consonants, such as */n/* and */m/*, are formed by a closure in the oral cavity and emitting sound from the nasal cavity. The inhibited oral cavity acts as a branch or side branch resonator. That means the oral cavity is closed at a certain point, but still contributes to the resonance characteristics of the nasal consonant sound.

Unlike vowels that only have formants, nasal sounds have formants and anti-formants. Anti-formants are frequency regions in which the amplitudes are weakened due to the absorption of sound wave energy by the nasal cavities. Nasal consonants also have the following three main features. First, if there is a consonant, the first formant always exists around 300 Hz. Second, the formant of consonants tends to attenuate strongly. Thirdly, the formant density of nasal consonants is high, and there is an anti-formant.

Semivowel and glide sounds:

Semivowels and glides are two small groups of consonants that are the sounds with vowel properties. The semivowels, such as */r/* and */l/,* and the glides, such as */w/* and */j/*, are characterized by the continued, gliding motion of the articulators into the following vowel. For semivowel consonants, the oral chamber is narrower than vowels, and the tongue tip is not down.

Affricative sounds:

An affricate is a consonant characterized as having both a fricative and stop manner of production. Affricatives are vocalized by firstly forming a stop sound and immediately followed by a fricative. Consonant */ch/* in “chapter” and “chop”, or */g/* in “cage” are the examples of affricatives sound. The affricative sounds are common in English but not very popular in Japanese.

Diphthongs:

Diphthongs are the combination sounds that are formed by two vowels produced consecutively in the same syllable by moving the articulators smoothly from the position of one to the other. A diphthong is very similar to vowels; however, due to the unstable vocal configuration during vocalization, it is classified as a consonant sound.

Voiced consonants and unvoiced consonants:

There are other types of consonant classification such as voiced consonant and unvoiced consonant in addition to the difference in articulation style such as fricatives, plosives, nasals, etc. as mentioned above. This is because the articulation style and the presence or absence of vibration of the vocal cords are independent of articulation. Thus, those with vocal cord vibrations are called voiced consonants, and those without vocal cords are called unvoiced consonants. The unvoiced sound source can be approximated by white noise, whereas the voiced sound source can be approximated by pulse or triangular wave. Besides, since the vocal cord vibration is accompanied, the voiced sound tends to have a larger amplitude of the speech waveform.

1.2.4 Prosody of Speech

Prosody is the study of the melody and rhythm and the influence of these features on the meaning of speech. Prosodic features are typically considered in a sentence level or group of words. These features are above the level of the phoneme (or "segment") and referred to as suprasegmental [8].

At the phonetic level, prosodic features include:

- Stress
- Vocal pitch
- Intonation
- Pause
- Loudness
- Tempo

- Vocal effects
- Paralinguistic features

Stress is the emphasis that is placed on a syllable or words when a person speaks. The same word can have the stress in different places, which changes the meaning of the spoken word. For example, “present” will be pronounced differently in the two sentences below.

“I give him present.”

“I present my research.”

The stress of the word “present” of the first sentence is placed at the first part of the word, while it is placed at the second part of the word in the second sentence. The stress in speech is very important when asking a question or giving an impressive. Pitch is the variation of an individual’s vocal range, from high to low, used in speech to convey the individual’s relation to a current topic of conversation. Usually, the pitch is different when speaking normally or asking a question.

The intonation is the difference in the pattern of the pitch which an individual uses in speech. Intonation is significant in tone languages like Vietnamese or Chinese because the different intonation can result in a different meaning of a word. For example, the word “ba” has different meanings in Vietnamese when pronouncing with different intonation [9].

Pauses are the period of silence in speech. Depend on the context, the pauses could give different meanings. For example, when announcing a winner, the speaker usually takes a long pause to create a dramatic situation.

The loudness or volume is often used to enhance the meaning of speech. People change their volume in a different context or when they speak with a different person. For example, the loudness of a person’s voice is different when he speaks to his children from when he speaks to his parents.

The tempo is the speed at which an individual makes speech. This also can vary depending on what the situation is and who the individual is talking to. For example, a person talks faster with his friends than he talks with his teachers.

Vocal effects such as laughing, crying, yawning, coughing, etc. can give additional information in an individual’s speech.

Paralinguistic features such as body language, gestures and facial expressions can also give extra information about the spoken language. This feature can support the spoken message or contradict the message and change the meaning of the message completely [11].

1.3 Dissertation Organization

This dissertation addresses a study of mechanical vocalization system and a cerebellum-like neural network and its application to control the prosodic features of the output speech. The study includes two parts which are the improvement of the current talking robot and the development of new algorithms and control systems for the mechanical vocalization system. In detail, the talking robot is added a new unvoiced sound system which allows the talking robot to generate fricative sound. A new design of vocal cords is employed to the talking robot. The new vocal cords mechanism is controlled by a command type servomotor. The new vocal cords with a wider pitch range from 50 Hz up to 250 Hz greatly increase the speaking capability of the talking robot, especially its singing performance. The new algorithms include interactive modification system, the sequence of vowels generation system, text-to-speech system, and the cerebellum-like neural network are developed for the talking robot. The prosodic features of the speech are also employed in the new algorithms. The remainder of this thesis is organized as follows:

In Chapter 2, the author surveys previous works on speech synthesis systems and their related research. The authors also address the limitation of these systems in term of prosodic features of the generated sound from these systems. Likewise, these limitations are the objectives of the study in the current mechanical vocalization system. In addition, the author reviews about the spiking neural network, which is more biologically plausible compared to its non-spiking predecessors and its potential application in the robotics field. In particular, since the author has been engaged in the study of timing function of the cerebellar neural network, a literature review is done focusing on this field.

Chapter 3 presents the construction of the talking robot. The talking robot consists of an air compressor, artificial vocal cords, artificial vocal tract, a nasal chamber, a silicone tongue, and a microphone-amplifier system which respectively represent the lung, vocal cords, vocal tract, nasal cavity, tongue, and auditory feedback of a human. The new unvoiced system with a separated airflow input not going to the artificial vocal cords is introduced in this chapter. The new design of artificial vocal cords, which provides the widest pitch range in comparison with other available systems, is also described. Experiments to verify the working behavior of the new voiceless system and the new vocal cords design is also reported in this chapter. The author has two publications for new unvoiced system and new vocal cords design.

In Chapter 4, the interactive modification system for the talking robot is developed and introduced. The interactive interface is built using Matlab Graphic User Interface (GUI) to take the advantage of built-in function of this software. The formula about formant frequencies changes due to the variation of vocal tract cross-sectional area is derived. Then, an adjustment strategy for the vocal tract area based on the formant frequency difference is described. The interactive modification system is used to quickly adjust the robot articulation to achieve the target formant frequency. The experiment verified the working performance of the interactive modification system. The author has two publications on the interactive modification system.

Chapter 5 describes the proposed algorithm of sequential vocal generation for the talking robot. In this chapter, 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 human speech and translate into a sequence of sound elements for the sequence of vowels reproduction. Then, several phonemes detection methods including the direct cross-correlation analysis, the linear predictive coding (LPC) association, the partial correlation (PARCOR) coefficients analysis, and the formant frequencies comparison are applied to each sound element to give the corrected command for the talking robot to repeat the sound sequentially. A survey experiment is conducted with eight people including Japanese and foreigners, and the result of the experiment verifies the ability of the talking robot to mimic a sentence. The author has three publications on this sentence repeating system.

In Chapter 6, a new text-to-speech (TTS) for the talking robot is introduced. This TTS system is developed based on the association of the input texts and the motor vector parameters. The main purpose of this system is to provide a more effective training tool for the auditory impaired hearing patient to vocalize. The intonation feature is added to this TTS system so the talking robot is able to generate an input text with different tonal effect. This system is very helpful when letting the robot generate a tonal language like Vietnamese or Chinese. The author acknowledges that this is the only hardware-based TTS system that has tonal effect at the moment. The author has one publication currently under review for this TTS system.

Chapter 7 discusses the timing property in the prosody of speech. The author's approach is using a human-like spiking neural network to control a human-like mechanical system. The spiking neural network with its capability of encoding the timing characteristic in its neural signal pattern is applied as the short-range timing function for the talking robot. Therefore, in this chapter, the author describes a cerebellum-like neural network model assembled by spiking neuron, and uses it to control the timing function. In particular, a cerebellum-like neural network is implemented to an FPGA board for timing signal processing. Then, the timing data is combined with motor position control function to fully control talking robot’s speech output that has prosodic features. The author has one publication on this timing function for the talking robot. Finally, the dissertation is settled with conclusions and future research in **Chapter 8**.

Chapter 2

Literature Review of Speech Synthesis and Biological Spiking Neural Network

In this chapter, the author gives a literature review on speech synthesis and an introduction on biological spiking neural network. The first part of this chapter describes the historical development together with some major speech synthesis systems. Then, an introduction and the significance of the spiking neural network is followed in the second part. The last part of this chapter discusses the limitations and challenges in speech synthesis, and proposes the objectives of this study.

As the human technology developed and expanded, the scientist had curiosity and began to explore more seriously into the nature of things. Physiological functions of the human were a reasonable target of research. Hence, the physiological mechanism of speech belonged in this domain. The comparatively complex vocal organs movement was often considered as a more tractable model. Therefore, many clever designs of mechanical synthesis models have been developed. As mentioned in section 1.1, the earliest speech synthesizer system was the “Brazen Heads" which was developed by Pope around 1003, Albertus Magnus at the beginning of the 12th century, and Roger Bacon in around 1249. In the late 17th and early 18th century, some original mechanical speech synthesis systems were proposed, such as the Kratzenstein speech synthesizer in 1779, the Kempelen machine in 1791, and the Wheatstone speaking machine in 1837. At the same time, some other different speech synthesizer approaches were proposed by Helmholtz as turning forks vibrator, Miller and Stumpf as organ pipes, and Koenig as Spectra siren. [15]- [20]

Later on, with electrical technology evolution, speech synthesis approach using electronic circuit was considered. Some of the circuit speech synthesizers were proposed by Stewart as two coupled resonant circuits, Wagner as four resonators, and most famously Dudley, Riesz, and Watkins as the Voder. After that, with the development of computer technology, PC system began to be used for speech synthesis purposes. This system was called computer speech synthesizer which was usually applied to a software product as a Text-To-Speech (TTS) system. Professor Stephen Hawking is a famous person using TTS to communicate with people. TTS systems translate the input text into speech via a speaker. TTS is very useful to assist visually impaired people on screen readout, to assist handicapped patients to speak or to help the foreigners to understand other languages. However, for studying the nature of the speech mechanism, a TTS system is not a good choice. Therefore, human-like mechanical vocalization system with robotics technology was developed to fulfill this task. Some of the famous robotics speech synthesis systems include Anton, Burpy, Lingua, Sawada Talking Robot and Waseda Talker. Among them, the anthropomorphic Waseda Talker was the most advanced robotics speech synthesis system so far.

Artificial neural networks (ANN) were often used to train robotic systems the mechanical speech synthesis systems to vocalize. However, the old generation of ANN, which its neural signals were either using digital on-off signal or activation function, are relatively old technology. Thus, a new generation of ANN, in which its neural signals mimic the behavior of biological neurons, was proposed and considered as the third generation of ANN. At the moment, the third generation of ANN, which is known as the spiking neural network, is mostly at theory building and modeling level and the engineering applications level is not very common yet.

2.1 Overview of Speech Synthesis

2.1.1 Static Synthesis System

Helmholtz Tuning Forks:

This brilliant device was among some of the very first sound synthesizers in human history. This device was developed by Herman von Helmholtz in around 1860. These tuning forks were used for resembling the musical tones and the human sound as shown in Figure 1.6 [15].

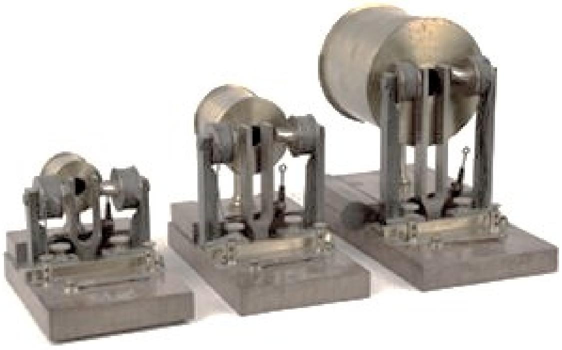


Figure 2.1: Helmholtz tuning forks

A tuning fork is a sound resonator in the shape of a two-pronged fork made of flexible metal. It generates a sound with a specific constant pitch when set vibrating by striking it against a surface. The sound pitch that a tuning fork generates depends on the length and weight of the prongs. Hence, it is frequently used as a standard of pitch measurement device. Helmholtz used a set of tuning forks to generate fundamental frequencies by combining the pitch of those tuning forks in varying proportions. The tuning forks were controlled by electromagnets. The sound amplitude of each fork was adjustable by controlling a shutter using a keyboard. Thus, Helmholtz was able to control various resonances and recreated some vowel sounds of a human. An online simulator of Helmholtz is available at http://www.sites.hps.cam.ac.uk. Figure 1.7 shows the picture of Helmholtz tuning forks online simulation software.

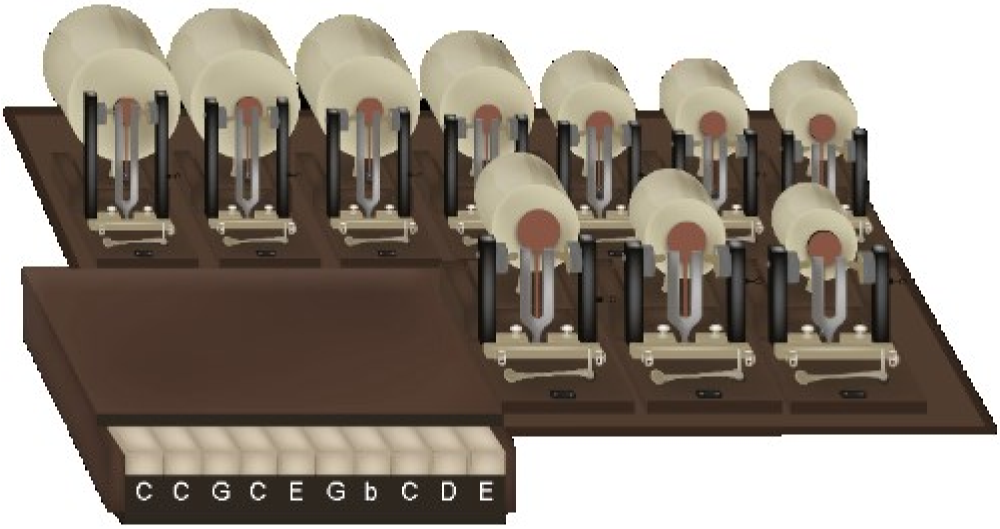


Figure 2.2: Helmholtz tuning forks online simulation software.

Miller Organ Pipes:

In 1916, Prof. Miller of the Case School of Applied Science, USA, introduced the vowel formation by combining the operation of several organ pipes simultaneously. Miller was successfully able to reproduce the vowel sound, and his work confirmed Helmholtz's idea about the vowel sounds were the resonance of several base tones together. The number of separate organ pipes, which required to build up the effect of a single vowel sound, varies for each vowel [1]. The typical organ pipes set are shown in Figure 1.8 below.

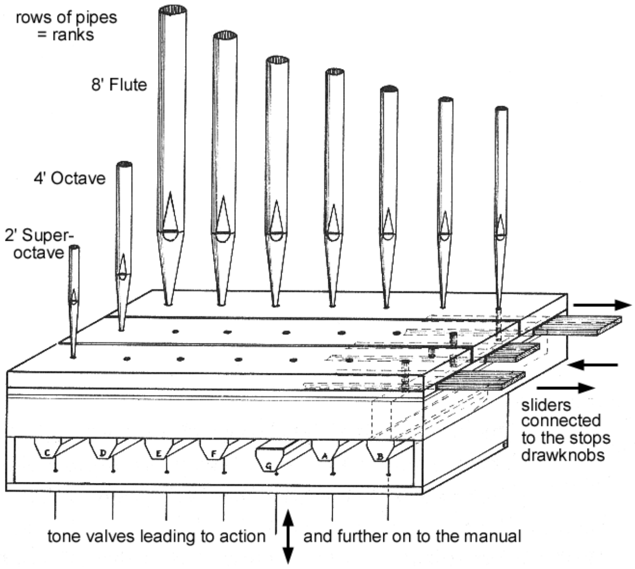


Figure 2.3: Organ pipes set for vowel generation.

Christian · Kratzenstein's Resonance Tube:

In 1779, Kratzenstein, a Russian scientist, was interested in vowel utterances of humans and investigated how the vocal tract deforms when people utter. From the deformation shape of the vocal tract according to each vowel, Kratzenstein created a speech synthesis system which was able to perform */ a /, / e /, / i /, / o /, / u /* using an acoustic tube imitating the shape of the human vocal tract [1]. Figure 2.4 shows a schematic view of the manufactured resonance tube. A reed was installed in the lower part of the resonance tube. When a person blew air into the tube, the reed vibrated, and a sound source was generated. The generated sound source was a delay and echo inside the resonance tube, and sound like a human vowel was generated. For the */ i /* resonance tube, no reed was installed; and only a turbulent sound source was used.

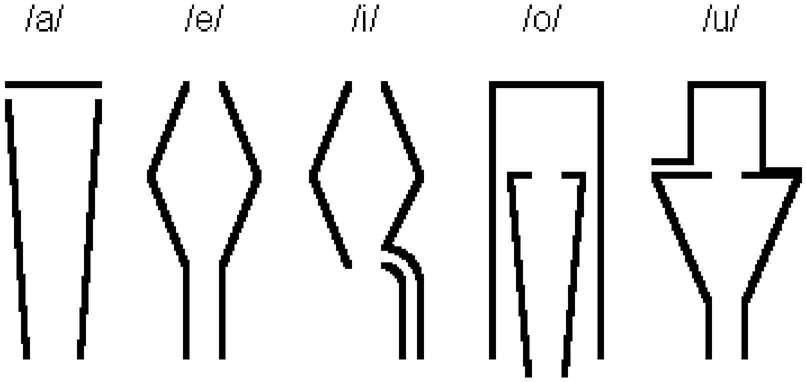
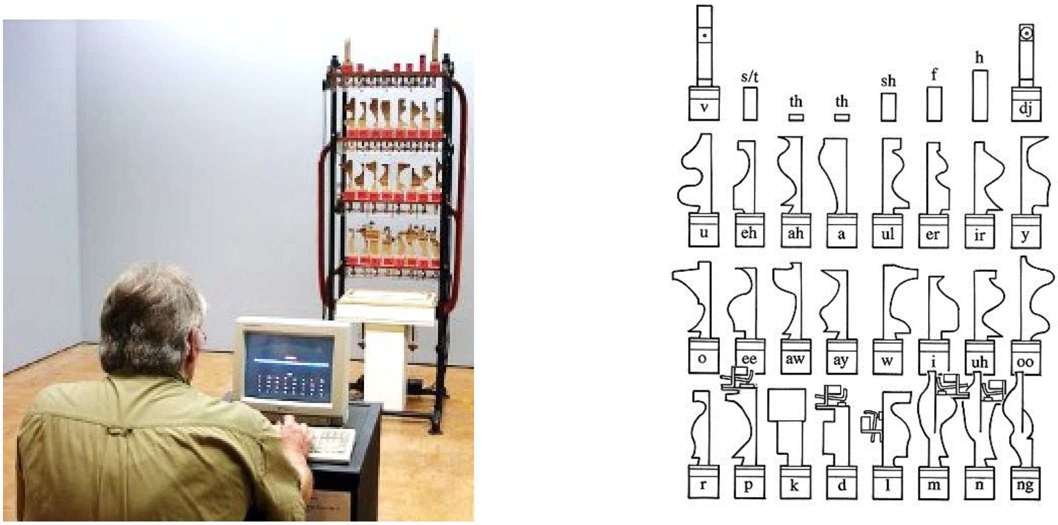


Figure 2.4: Kratzenstein's resonance tube

Martin Richard's Talking Machine:

In 1990, Martin Richards, a British scientist, produced a speech synthesizer using 32 resonator tubes which resemble the human vocal tracts (http://martinriches.de/talkmore.html#top). This talking machine reproduced the vocal tract cross-sectional area obtained by human X-ray photograph with a resonance tube, and it was possible to generate various voices by sending air from the end of the resonance tube controlled by a computer. At this time, it was reported to be able to speak about 400 words in English and 100 simple words of simple Japanese words. Figure 2.5 shows the appearance of this and the structure of the resonance tube.



(A) Picture (b) Structure

Figure 2.5: Martin Richard's Talking Machine

Vocal Vision II:

With the development of 3D-printing technology. Artists and scientists began to use 3D-printed vocal tracts to generate sound. In 2013, David M Howard, a British engineer and musician, developed the vocal tract organs and used them as a new musical instrument (https://www.york.ac.uk/50/impact/synthesised-speech/). Vocal Vision II system consisted of five 3D-printed vocal tracts, which were attached on loudspeakers, to produce static vowel sounds. This system was controlled by a keyboard. When a key was pressed, a typical waveform was generated accordingly. Figure 2.6 shows the picture of Vocal Vision II system.

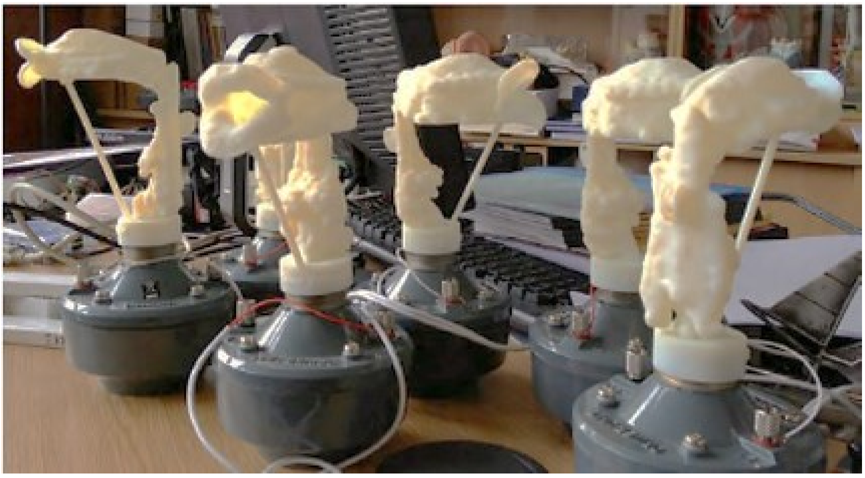
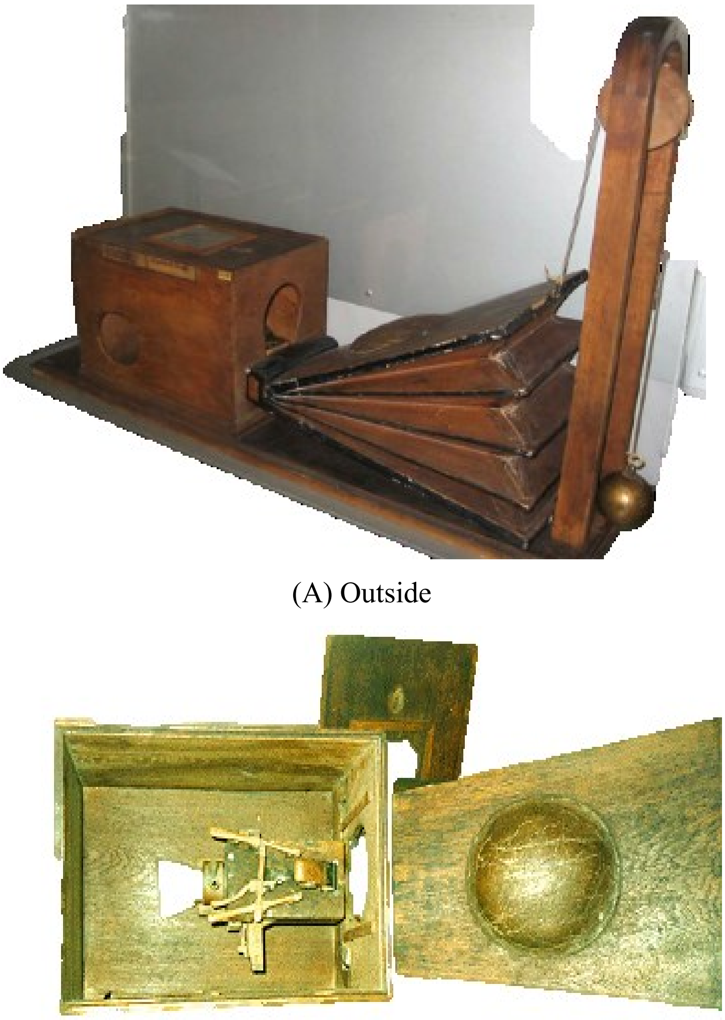


Figure 2.6: Vocal Vision II system

2.1.2 Dynamic Synthesis System

Wolfgang Von Kempelen's Speaking Machine

One of the most popular speaking machines in history was developed by Wolfgang von Kempelen in 1791. Kempelen, a Hungarian inventor, produced a mechanical speech synthesizer consisting of a bellow, reed, and a leather resonator [16]. Figure 2.7 (A) and (B) respectively shows the outside and inside view of Kempelen's Speaking Machine. Figure 2.7 illustrates the configuration of the machine. This system was the first real speaking machine that mimicked the functional behavior of a human vocalization system. This speech synthesizer substituted the bellows for the air from the lungs. The reed, which provided the sound source for the system, is installed in front of the rubber tube. The reed generated a sound source according to its vibration manipulated by the input airflow. Also, the vibration of the reed could be stopped by adjusting the amount of air with the lever.



(A) Outside

(B) Inside

Figure 2.7: Kempelen's speaking machine

Von Kempelen's speaking machine was the first mechanical system which was able to produce not only some vowels sounds, but also whole words and short sentences. It was reportedly able to generate the Latin, French, or Italian language. The resonance box attached to the exit of the bellows played the role of the the human vocal tract. Inside the box, there was a rubber tube whose shape could be manipulated by the human hand. Resonance characteristics were added to the sound source, and various sounds were generated by adjusting the shape of the resonance tube.

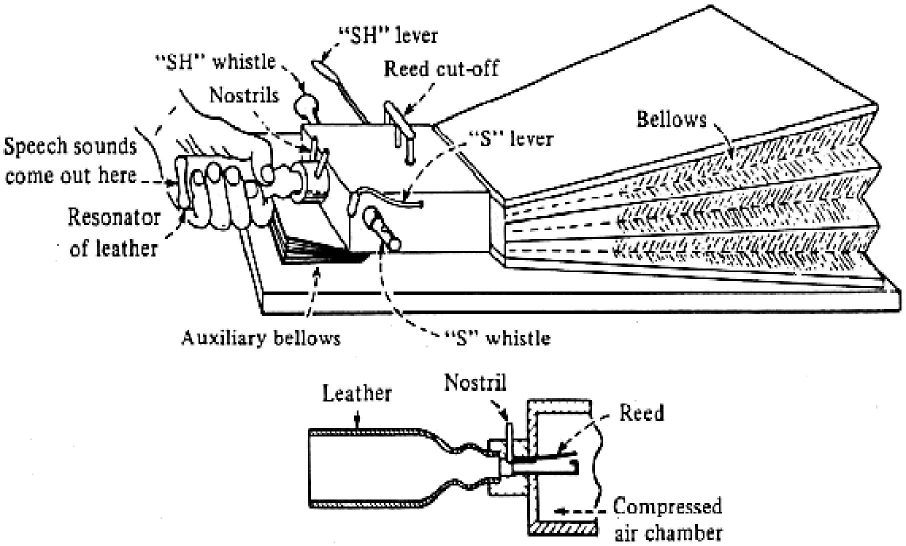


Figure 2.8: Kempelen's speaking machine outline

When synthesizing speech, a person operated a bellows with one hand to generate air flow. When the bellows moved, it created the airflow to the reed. The airflow was compressed in a chamber, and this compressed air vibrated the reed at a certain frequency to create the sound source for the system.Then, he used the other hand to manipulate the leather acoustic tubes to make a specific sound. It was possible to utter vowels and simple words by adjusting the shape of the leather acoustic tubes.

Riesz's Talking Mechanism:

In 1937, Riesz proposed the mechanical speech synthesizer as shown in Figure 2.9 (http://www.haskins.yale.edu/featured/heads/simulacra/riesz.html). The device had the shape similar to the human vocal tract. This device was constructed using mainly rubber and metal. There were 10 control keys (or valves) to operate this machine. This mechanical talking device was reported to produce pretty good sound [17].

For the operation, the air was firstly stored in the tank on the right side, and the airflow amount was adjusted by V1 and V2 valves. Valve V1 passed air through the cavity L1 where the reed was installed. The reed acted as a vocal cord mechanism. The cylinder mounted on the top of the reed adjusted the vibration of the reed; thus, the fundamental frequency was controllable. Unvoiced sounds were generated by letting the airflow passing through valve V2. The shape of the vocal tract was formed by deforming nine parts corresponding to lips (1,2), teeth (3, 4), tongues (5, 6, 7), pharynx (8) and soft palate (9).

Later, Riesz modified the system in a way that its vocal tract shape could be adjusted by playing keys. Figure 2.10 shows the configuration of the new version of Riesz mechanical vocal tract. Playing keys # 4 and # 5, corresponded to valve V 1 and V 2, were used to control the airflow. The other keys controlled the pitch and vocal tract shape of the device. By operating these adjustment keys, a person could produce sound continuously.

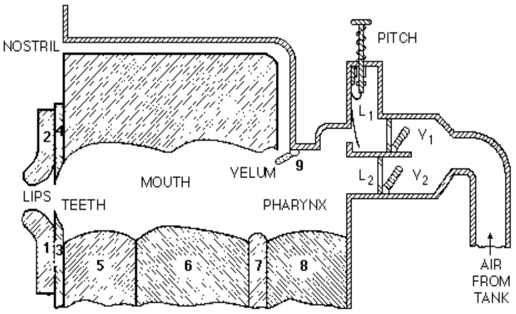


Figure 2.9: Riesz mechanical vocal tract (old version)

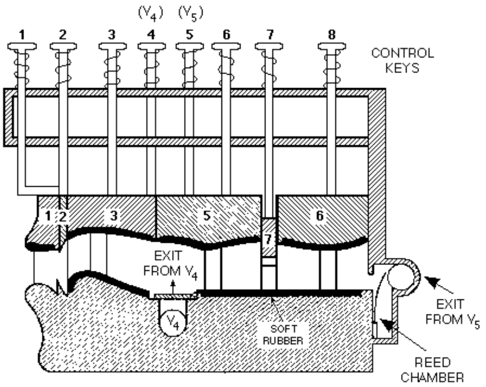
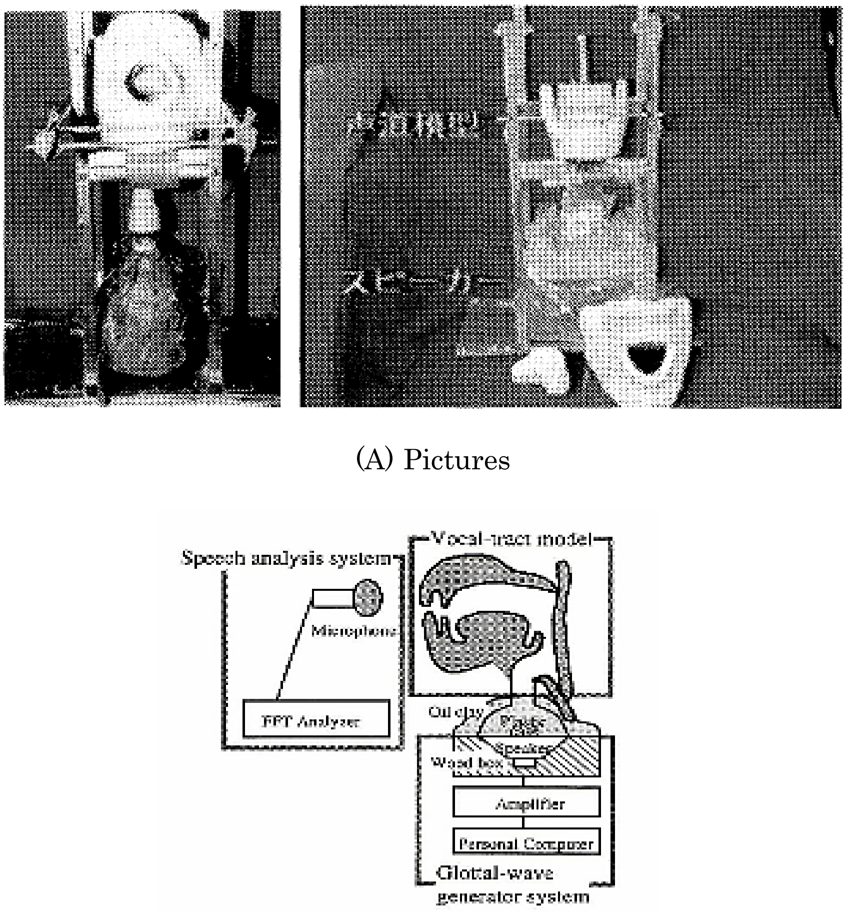


Figure 2.10: Riesz mechanical vocal tract (new version)

Mechanical Speech Synthesizer by Koichi Osuka:

Koichi Osuka and colleagues developed a mechanical speech synthesizer with the aim to apply to a humanoid robot. In their study, they constructed a 3D physical model of the articulatory organ to generate 5 Japanese vowels [18].

In their research, they used a speaker to produce the sound source. The sound source generated a triangle wave with the period of 100 Hz. The vocal tract was made with clay model; the tongue, cheeks, and lips were made with plastic; and the skin is made with silicone. The synthesizer was reportedly able to produce a vowel sound whose characteristics were close to human, and revealed that it was possible to generate speech close to a person using a machine model. Figure 2.11 shows the picture and outline of this synthesizer.



(A) Pictures

(B) Osuka system outline

Figure 2.11: Mechanical speech synthesizer by Koichi Osuka

2.1.3 Software Speech Synthesis

With the invention of the sound speaker and the development of computer systems, the researchers began to shift to the direction of using software and speaker for speech synthesis systems. Firstly, software synthesis speech could be created by associating the output sounds with the pieces of prerecorded speech stored in a database[19]-[21]. This was a very old fashioned technique which was referred to as recording editing method or unit selection synthesis method. Later on, to save the database memory, only important features of sound were stored in the database. This technique was known as parameter selection method or domain-specific synthesis method. However, the database was always limited, but the human sounds were unlimited. Thus, a rule-based synthesis system was proposed to not only reduce the database size but also can synthesize any sound. The challenge for this technique was finding a suitable rule for the synthesizer. The neural network was one of the best options for solving the problem of rule-based synthesis systems. Then, a full model of speech synthesis was developed by referring to the biological neuronal mechanism of speech production and perception as occurring in the human nervous system. This kind of model was often known as neurocomputational speech synthesis. The DIVA (Directions into Velocities of Articulators) model developed by Guenther et al. and the ACT (vocal tract ACTions) model developed by Bernd et al. were the two famous models of neurocomputational speech synthesis [22]-[26].

Recording Editing Method (or Unit Selection Synthesis):

An artificial sound that came close to a natural human voice could be generated by recording the voice of a person and synthesizing sound by editing it. Figure 2.12 shows the process of the recording and editing method. As preprocessing, the human voice was recorded beforehand, and a speech database was constructed. Then, the speech in the database was separated at short intervals to make speech segments. The user related the target speech to the sound unit to create the synthetic speech. By using a human voice, it was possible to synthesize high-quality sounds. However, the database size was limited due to the significant memory size of each recorded sound; thus, a simple phrase or sentence was frequently synthesized this way. For example, train station or airport announcements were generated using this method [9].

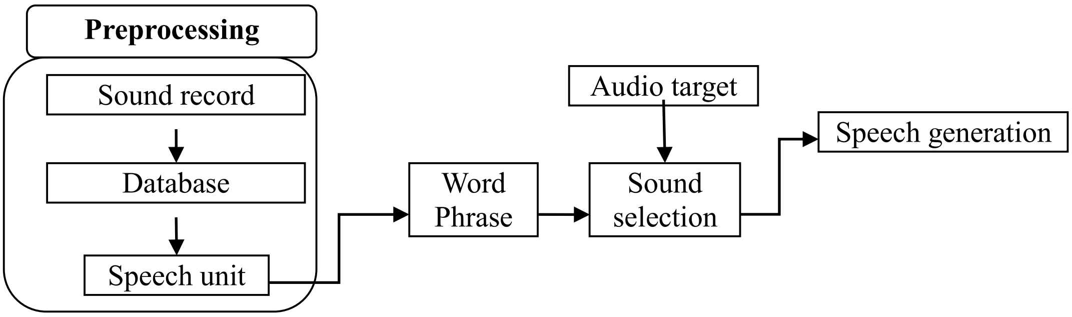


Figure 2.12: Recording editing method process

Parameter Editing Method (or Domain-Specific Synthesis Method):

The parameter editing method was similar to the recording and editing method in term of words and phrases being used as synthesis units. However, in this synthesis method, the database was formed by extracting important acoustic features. Thus, compared with the recording editing method, the naturalness of synthesized speech was lower because the original human speech was not directly used. However, since only some features were used to build up the database, it was possible to perform substantial information compression. Also, it was possible to extend or reduce the speech duration and smooth the pitch and spectrum of the conjunction sound. The most common method was based on stochastic process theory. Parameters of the frequency spectrum were statistically analyzed from the voice signal and encoded. Then, the speech was synthesized from these parameters. For example, a vocoder was a speech synthesizer based on a linear prediction analysis. At first, the speech signal was subjected to frequency analysis by a large number of band pass filters and a rectification smoothing circuit, and then the sound spectrum was extracted. At this time, the pitch was extracted according to the periodicity of the signal. The speech synthesizer controlled the pulse generator and the noise generator to generate sound sources. Then, the sound source was modulated according to the frequency spectrum, and the speech was finally performed [20].

Rule-based Synthesis Method:

In the rule-based synthesis method, the unit of the speech unit was further reduced as compared with the recording editing method or the parameter editing method to construct a database. Rule speech synthesis method was performed by connecting small sound units such as syllables, phonemes, waveforms, etc. with a special rule so that it was expected that all kinds of speech could be synthesized. However, it was difficult to establish the rules for connecting speech units, the pitch, and also the control rules for prosodic information. Generally, a neural network model, such as hidden Markov (HMM), was often applied to establish the rule for a rule-based synthesizer. HMM was a reinforcement learning neural network that established a rule for the relations of the sound features input and the synthesized speech output. A new type of speech synthesizer, which translated mouth movements directly into speech, had just been released by Bocquelet and his group in France. This speech synthesizer used nine sensors to capture the movements of human lips, tongue, jaw and soft palate. Then a neural network was applied to interpret the sensor data into vowels and consonants. The output speech was reported to sound like a robot, but the syllables were well distinguishable [21].

Table 2.1 shows the features of the software speech synthesis method. The sound comprehensibility and naturalness are the highest in the recording editing method, and then the parameter editing method. Regarding the rule synthesis method, the comprehensibility and naturalness of synthesized speech are lower than those of the other two methods. Regarding the number of vocabularies that can be synthesized, the recording editing method, in which the unit of the speech unit is a word, phrase or sentence, is the smallest. For the rule synthesis method, the unit of the speech unit is small in comparison with other systems, so the number of vocabularies is large, but the synthesis is the most complicated. It is not easy to decide which method is superior or inferior completely, and depending on the specific purpose, a suitable synthesis method should be used.

Table 2.1 Comparison of speech synthesis

	Recording editing	**Parameter editing**	**Rule synthesis method**
Comprehensibility	High	High	Medium
Naturalness	High	Medium	Low
Vocabulary	A little	Large	Infinite
Speech unit	Word, phrase, sentence	Word, sentence, phrase	Phoneme, syllable
Difficulties	Easy	Normal	Complex

Neurocomputational Speech Synthesis:

Recently, with the development of neuroscience, neurocomputational speech synthesis systems which focused on the human nervous system and its influence on speech production and perception functions were developed. Neurocomputational speech synthesis models were usually comprised of three neural maps which were linguistic, motor function and sensory map. The linguistic map contained the neural generations of phonemic for speech production. The motor function map compromised motor plans to articulate particular speech item. The sensory map was an auditory representation of input sound signal and associated the phonemic representations for that sound. DIVA model [25], as shown in Figure 2.13, and ACT model [26] were two leading approaches in the neurocomputational modeling of speech production. The result of these model are association of three maps as shown in Figure 2.14

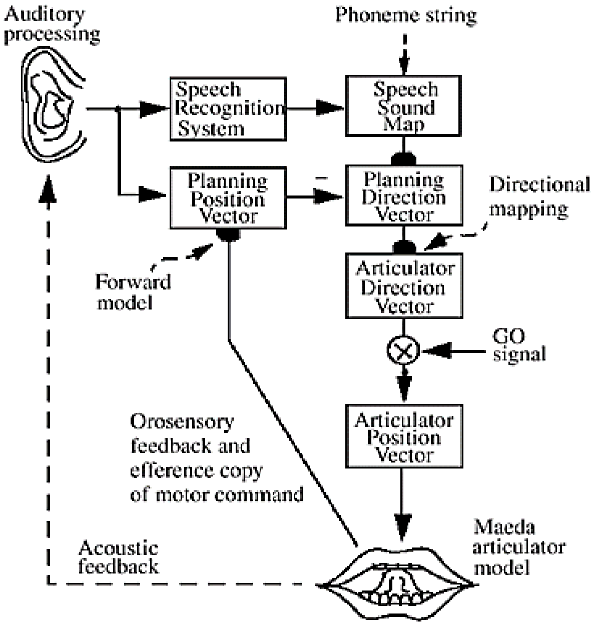

Auditory processing
Phoneme string
Speech Recognition System
Speech Sound Map
Planning Position Vector
Planning Direction Vector
Directional mapping
Forward model
Articulator Direction Vector
GO signal
Articulator Position Vector
Orosensory feedback and efference copy of motor command
Acoustic feedback
Maeda articulator model

Figure 2.13: The DIVA Model

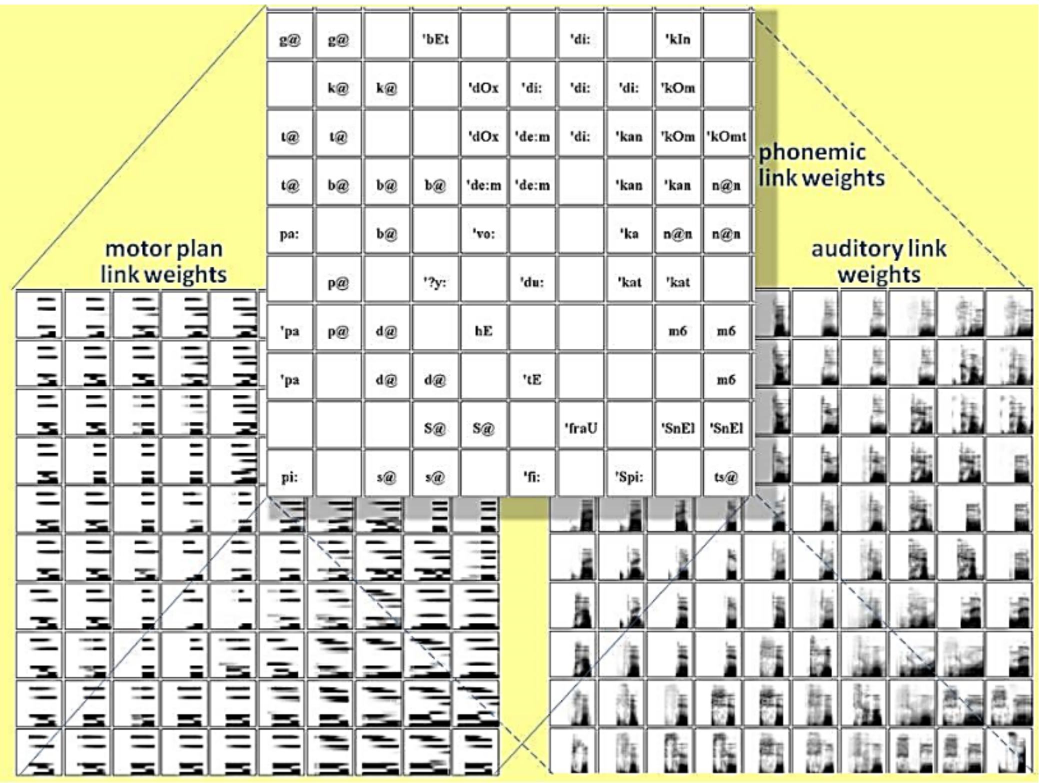


Figure 2.14: The mapping of 200 German syllables in ACT model

Nevertheless, no matter how advanced a software-based speech synthesis is, its output sounds are always produced by a sound speaker. Thus, for studying the naturalness of human vocalization mechanism, the hardware-based system is obviously the better option.

2.1.4 Speech Synthesis by Robotic Technology

Since the development of the computer, software-based speech synthesis occupied the major part of the speech synthesis method used in recent years. However, compared with software-based speech synthesis, hardware-based speech synthesis was performed by moving mechanical speech organs, and it was possible to generate more natural speech. The mechanical structure of artificial articulation organs and its control method were extremely difficult to establish in comparison with the case of a software-based system. Since there were many challenges and complications in the development of a hardware-based speech synthesis system, not so many new robotic speaking systems were built in the past few years.

Speaking Robot Burpy:

Yoshikawa and his group made a speaking robot, Burpy, and trained the robot to generate vowel sounds by an infant-caregiver interaction approach [27]. During operation, the compressor sent air to the vocal band to generate a sound source. Then, the robot spoke out the sound by resonating this sound source with an artificial silicone vocal tract. Figure 2.15 shows the configuration of Burpy. This robotic speaking system used four motors to control the vocal tract shape and two motors to control the movement of the lips. This robot learned to vocalize by imitating the caregiver's articulatory movements.

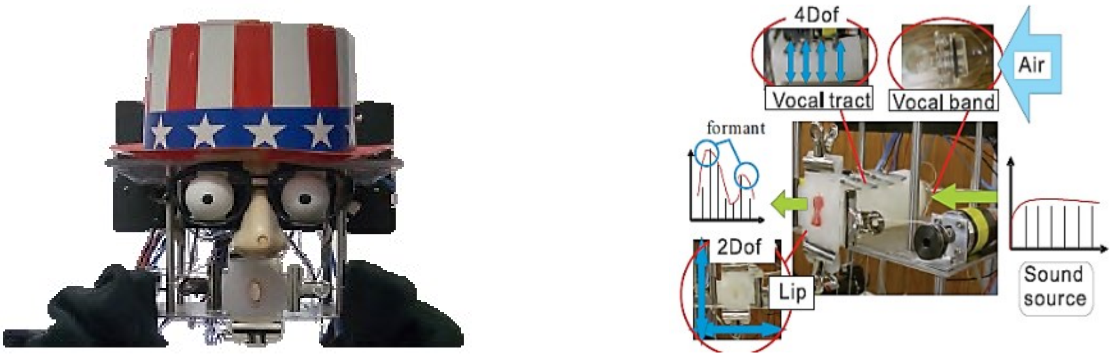
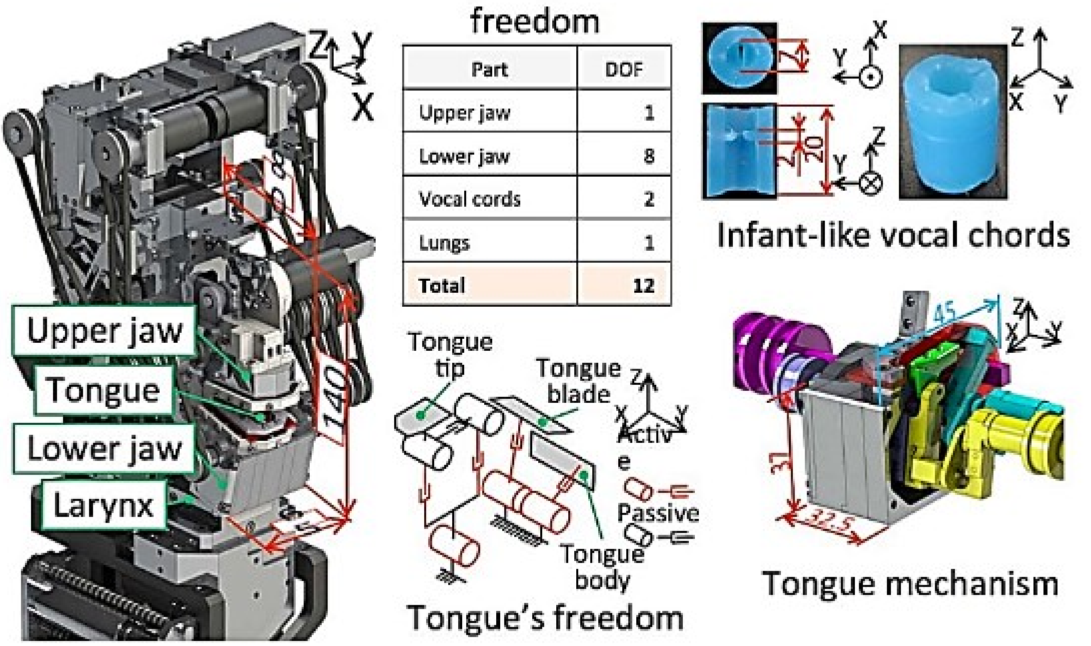


Figure 2.15: “Burpy” the speaking robot

Infant-like Vocal Robot “Lingua”:

Prof. Asada and his team at Osaka University also developed an infant-like vocal robot called “Lingua” as shown in Figure 2.16. Lingua has seven DOFs of tongue articulation which were realized using a complex design consisting of linkage mechanisms inside a miniaturized vocal tract [27]. This enabled the achievement of a high articulation performance. The shapes of its vocal cords and vocal tract were similar to those of a six-month-old infant as determined from anatomical data. Lingua robot could produce an infant-like voice and had a good articulation capability. This robot was a research platform for modeling early articulatory development through mother-infant interaction.

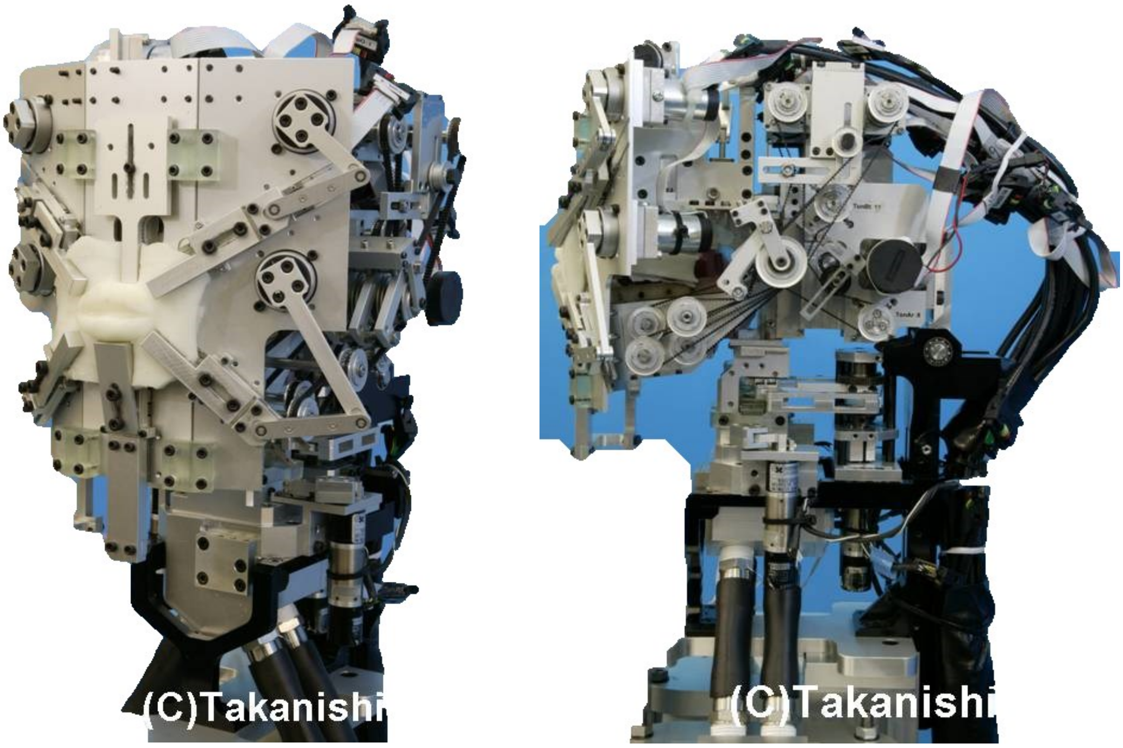


Part	DOF
Upper jaw	1
Lower jaw	8
Vocal cords	2
Lungs	1
Total	12

Figure 2.16: “Lingua” infant-like speaking robot

Waseda Talker No. 7 Refined II (WT-7RII) (Takanishi Laboratory):

Waseda Talker was one of the legendary systems in hardware-based speech synthesis, which was developed by the Takanishi laboratory. WT-7RII consisted of vocal cords (5 degrees of freedom), lungs (1 degree of freedom), lips (4 degrees of freedom), jaws (1 degree of freedom), and tongues (7 degrees of freedom). The vocal tract length of was about 180 [mm] and had almost the same length as the adult male. WT-7RII had a total of 20 degrees of freedom including the soft palate (1 degree of freedom) [29]. WT-7RII also had a moving jaw with an independent driving mechanism. Also, the tongue and palate, which were made of thermoplastic elastomer, were able to reproduce three-dimensional shape like a human. The picture of WT-7RII is shown in Figure 2.17. The vocal cords could provide the sound source attenuating in the acceptable frequency range. In the tongue, the movable range was up to 7 mm due to the high density of the linkage mechanism integrated into the tongue that allowed a more accurate reproduction of the vocal tract shape. In addition, the inner space of the tongue composed of an elastic body was filled with ethylene glycol liquid to improve vocal tract resonance characteristics. For this, a clear utterance of five vowels was realized. Moreover, using the acoustic feature quantity extracted from the generated speech optimized the controlled variable of the robot. Therefore, this WT-7RII is considered the most advanced hardware-based speech synthesis system at the current time.



(Front View)

(Side View)

Figure 2.17: WT-7RII robot

2.2 Biological Spiking Neural Network [30]-[37]

2.2.1 Overview of Artificial Neural Network

Humans continuously receive information about the surrounding environment via their sensory organs. Then, they process this information and take appropriate actions. All of the procedures occur in their nervous system. Billions of neurons are interconnected and cooperate with each other to process information from sensory input and then to transmit the action commands to the motor system [30]. On average, a single neuron transmits its signals out to over 10.000 other neurons [34]. Thus, the signal flow is a very complicated process and hard to analyze. Thus, understanding and then constructing a neural network has been a very interesting topic for scientists since James Chadwick discovered the neuron in 1932 [30]. Through many years of development, the neural network has grown to the 3rd generation.

The first generation of artificial neural networks [35] was a very simple model. A neuron gave a digital 'high' signal if the sum of incoming signals was greater than a threshold value. Despite their simple structure and its digital output, the first generation neural networks have been successfully applied in many complex systems. For example, in digital computations, any function with Boolean output could be processed by a multilayer perceptron with a single hidden layer. Artificial neural networks, which were over fifty years old, were becoming a pretty old technology.

Neurons of the second generation of the artificial neural network used a continuous activation function to present their signals. Sigmoid and hyperbolic tangent were the two commonly used activation functions in the second generation. Typical models of the second-generation neural networks were feed-forward and recurrent neural networks. These models were more powerful than their first generation predecessors because they could work well with any analog system, making them suitable for analog computations. If these models were equipped with a threshold function as the first generation, they would also work very well for digital computations and even with fewer neurons within the network [36].

Biological neurons transmit information by using short and rapid increases in voltage. The biological neural signals are generally known as action potentials or spikes. The neurons encode information not only in their average firing frequency but also in the timing of single spikes. Spiking neural networks, which are recognized as the third generation of neural network, are more powerful and biologically plausible than their non-spiking predecessors because they encode temporal information within the neuron signals. However, they are more complex to model and analyze than the other artificial neural networks.

Humans continuously receive information about their surrounding environment via their sensory organs. Then, they process this information and take appropriate actions. All of the procedures have occurred in their nervous system. Billions of neurons are interconnected and cooperate with each other to process information from sensory input and then to transmit the action commands to the motor system. On average, a single neuron transmits its signals out to over 10.000 other neurons. Thus, the signal flow is a very complicated process and hard to analyze.

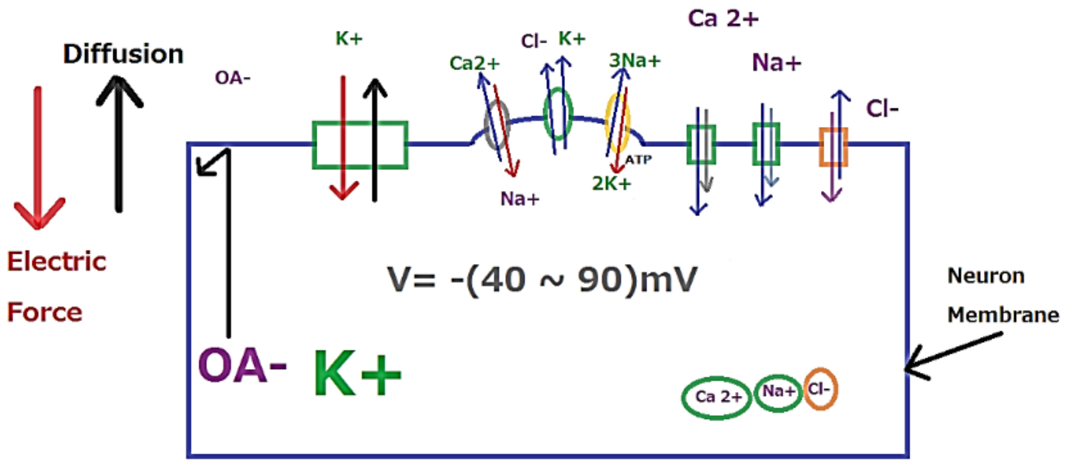


Figure 2.18: Biologically neuron model

All cells in a physical body are electrically polarized. This means a voltage difference is maintained across the cell's membrane and it is commonly known as membrane potential. Figure 2.18 shows a simple model of a biological neuron. The complex interaction between protein structures, which are ion pumps and ion channels, on the membrane creates this electrical polarization. Ion pumps or ion transporter are the membrane proteins that continuously moves ions across a plasma membrane against their concentration gradient. Thus, the ion pumps are responsible for the voltage difference across the cell's membrane in their stable stage. The ion channels are responsible for an immense change in a voltage difference across the membrane within a short period. There are different types of ion channels vary across the membrane of the cell, which gives the dendrites, axon, and cell body different properties. As a result, some parts of the membrane can be excitable, and some others are not. Voltage-gate sodium and voltage-gate potassium are the two main ion channels that are responsible for action potential or neuron spikes. At their normal state, the neuron cells have a low concentration of sodium ion and high concentration of potassium inside the cell's membrane. The ion channels are closed during this stage. When the cell's membrane potential rises above the threshold potential, the ion channels open. This allows sodium ion together with potassium ion rush through the membrane due to concentration difference, and an action potential is formed [37].

In detail, each membrane has two important levels of potential which are the resting potential and a higher value called the threshold potential. Resting potential is the value of the membrane potential in their stable state. The resting potential is around –70 millivolts (mV) and the threshold potential is around –55 mV. Neuron signal inputs or synaptic inputs to a neuron make the membrane potential to rise or fall, which are called depolarize or hyperpolarize, respectively. Action potentials are triggered when the sum of depolarization raises the membrane potential above the threshold level. When an action potential occurs, the membrane potential quickly increases to a certain value, typically around 40mv and then falls down. The membrane's potential falls below the resting level, and then slowly recovers to the resting potential. The phase, which the potential of the membrane has a value smaller than resting potential, is called refractory period. Thus, a typical action potential process contains three phases of depolarization, hyperpolarization, and refraction. In most neurons, the whole process takes roughly 4-5 milliseconds. The entire process is shown in Figure 2.19 [37].

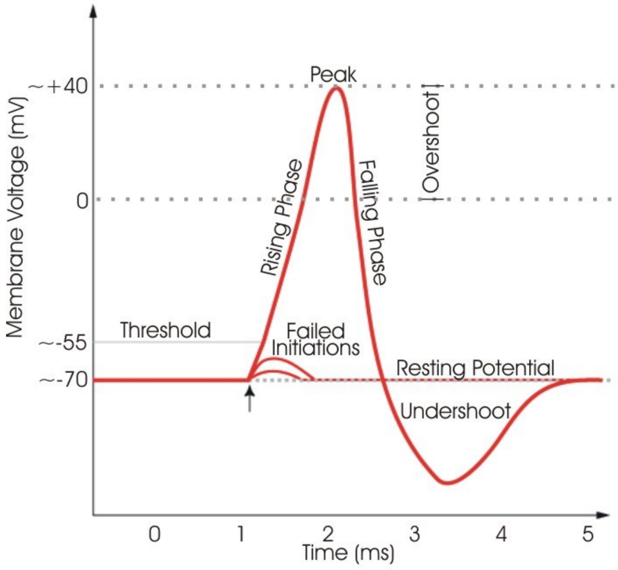


Figure 2.19: Biological neural signal

When an action potential arrives at the axonal or the input side of the neuron, the cell membrane releases its neurotransmitter content into the extra-cellular fluid and fills the synaptic gap. Neurotransmitter molecules react with a matching receptor on the other cell’s membrane to open its ion-channel. The incoming potential or postsynaptic signal can either be positive or negative which are called excitatory (EPSP) or inhibitory (IPSP), respectively. The excitatory signal increases the membrane potential, while inhibitory signal decreases the membrane potential of the receiving neuron.

2.2.2 Spiking Neuron Model

The earliest model of spiking neural network integrated and fire model which was proposed by Louis Fapicque as shown in equation (3.1). When an input current was applied, the membrane voltage increased with time until it reached a constant threshold V_{th}. Then, a refectory period t_{ref} was followed (equation 3.2) [30],[31].

$$I_c = C_m \frac{dV_m(t)}{dt} \tag{3.1}$$

$$f(I) = \frac{I}{C_m V_{th} + I t_{ref}} \tag{3.2}$$

In 1952, after studies about squid neural mechanism, Hodgkin and Huxley proposed the first scientific model of a spiking neuron that described how biological neural signals or spikes were initiated and propagated. Hodgkin and Huxley modeled excitable cells as an electrical element. Their model is shown in equation (3.3) to (3.5).

The membrane potential is "V_m." The lipid bilayer of the membrane is represented as a capacitance "C." Voltage-gated ion channel of i-type is considered as electrical conductance "g_i" whose value varies depending on voltage and time. "g_L" is the electrical conductance of leakage channel. Voltage sources "E_n" are the electrochemical gradients, and its value is determined by concentrations of the ionic species. "Ip" is the current source of the ion. "Vi" is the reversal potential (equilibrium potential) of the i-th ion channel.

The current flowing through cell's membrane is:

$$\sum_i I_i(t, V) = -C_m \frac{dV_m}{dt} \tag{3.3}$$

Current through a given ion channel is:

$$I_i = g(V_m - V_i) \tag{3.4}$$

The total current through the membrane is given by:

$$I = C_m \frac{dV_m}{dt} + g_k(V_m - V_k) + g_{Na}(V_m - V_{Na}) + g_L(V_m - V_L) \tag{3.5}$$

Following Hodgkin and Huxley various neuron models, such as the FitzHugh–Nagumo model (1961–1962), Hindmarsh–Rose model (1984), etc. were developed to describe the other features of the biological neural signal.

2.2.3 Learning Mechanism of SNN

The most important characteristic of a neural network is its learning mechanism. And similar to their previous generation, the learning in SNN also based on synaptic weights adjustment over time. By adjusting the synaptic weights, the neural signals flow within a neural network is altered; thus, the signals transmitted to a neuron are changed. As a result, the output signals of the neural network are also modified. This is a basic mechanism of learning, which is referred as synaptic plasticity in neuroscience [37].

There are two main types of plasticity, which either weaken or strengthen the input signal and are respectively known as depression or potentiation. In regard to duration, short-term synaptic plasticity occurs in about tens of milliseconds to a few minutes, while long-term plasticity, which lasts from several minutes to few hours. The NMDA and AMPA glutamate receptors are mainly involved in molecular mechanisms of synaptic plasticity (Eric Kandel laboratories). The opening of NMDA channels result in a rise in postsynaptic transmitting capability, which is linked to long-term potentiation (LTP); while NMDA ion channels lower the transmitting capability of post-synaptic, which linked to long-term depression (LTD). There are many types of plasticity have been discovered such as short-term depression (STD), long-term depression, short-term potentiation (STP), long-term potentiation (LTP), Excitatory postsynaptic potential, Inhibitory postsynaptic potential, Activity-dependent plasticity Spike-timing-dependent plasticity (STDP), Synaptic augmentation (Short-term plasticity), Neural facilitation (Short-term plasticity), etc. These synaptic plasticities are important in learning and memory of the nervous system.

2.2.4 Potential Applications

The spiking neural network can be used for information processing as traditional artificial neural networks. However, with higher bio-realistic properties, they often are used as the research tool for studying the mechanism of the nervous system. For example, a model of a spiking neural network built based on a hypothesis about the anatomical structure of a biological neuronal circuit and its function. The spiking neural network model is then simulated with a computer system to get the output signals. The output signals of this model are then compared with the electrophysiological recordings of the real biological neural network circuit for validation and determining the plausibility of the hypothesis model.

In practice, the modeling spiking network has shown its usefulness in neuroscience, but not yet in engineering and application. Some large-scale neural network models have been developed and confirmed the theoretical advantage of the pulse coding of spiking neural networks. However, these models heavily rely on the development of computer systems. In fact, the application of large-scale spiking neural networks has been limited because of the increased computational costs associated with simulating bio-realistic neural models. As a result, there have been very few applications of large-scale spiking neural networks. In addition, it is difficult to employ large-scale spiking neural network models into a hardware system to obtain the output signals in real time [32].

In this study, the author attempts to design a cerebellum-like spiking neural network model and then employ this network to an FPGA board with the aim of obtaining the timing signal in real-time. The timing signal is then encoded in the talking robot command to generate output speech with prosodic features. The detailed of the cerebellum-like spiking neural network and its achievement is reported in chapter 7.

2.3 Problems Overviews and Objective of Study

In the previous section, the author presented the literature review on the static synthesis system, dynamic synthesis system, software-based system, and robotics system. This section discusses the problems of each system and the objective of the study related to the challenges of the other systems.

Firstly, as mentioned in section 2.1.3, all the software-based systems used speakers to synthesize speech; thus, it was easier to generate and control sounds in comparison with mechanical systems whose sound was limited by their physical behaviors. Also, the capability of speech generation of the software-based system is related to the computational capacity of the computer system; for example, recording and editing method requires a powerful computer system to be able to perform efficiently in real-time. In addition, the more speech data is needed, the higher resolution is required. As a result, this requires massive memory size and high-performance processors to achieve fast random access to the database. When testing, the same data was often used over and over again, and the developer tuned the system based on these data. This could result in a much better performance in comparison when using randomness test data. However, the mechanism that a speaker generates a sound in software-based synthesis system is very different to the human vocalization system. Hence, the naturalness of the sounds generated from the software-based system would never be similar to the real human voice. Therefore, for the purpose of studying and clarifying human vocalization mechanism, it is impossible to use a software-based synthesis.

As for a static synthesis system, these systems were very old and mostly based on musical instruments to generate very few human simple sounds. Thus, the number of sounds is very limited, and the level of control of these systems to generate a properly recognizable sound for a human to recognize is somewhat complicated and required some training. Similarly, the dynamic synthesis systems, whose mechanical structure is very complex, are very difficult to control and operate by a new user who is not familiar with the system. Also, due to the complexity of its construction, the employment of automatic control systems is nearly impossible. In both of these systems, the mechanism to generate a sound is similar to the human vocalization system; however, the quality and number of the sounds are very limited.

Robotics approach in speech synthesis is the most suitable tool for studying and verifying the mechanism of the human vocalization system. The robotics approaches can be considered as the revolution of a dynamic synthesis system. The automatic control mechanism in a robotic speaking system enhances the capability of vocalization. However, there are limitations in each specific automatic talking machine. For the Anton system, the developers built this robot with the purpose of investigating the energy consumption in speech in relation to the environmental noise level. Although it was constructed by biologically referring to the human vocalization system, it could not mimic the sound. It is difficult to consider Anton as a real robotics speech synthesis system. The Burpy robot designed by Yoshikawa is very simple in construction. It also used an artificial electro-larynx to generate a sound source which results in very unnatural output sound. This system was reported to be able to generate only 4 out of 5 basic Japanese vowel sounds. Also, the first and formant frequencies of the generated sound are very close to each other in the four sounds; thus, it was difficult to distinguish these sounds. The Lingua speaking robot design is very similar to the Burpy robotic system. However, the Lingua system has vocal cords to generate the sound source; thus, its output sound is more natural then Burpy system. Nevertheless, the Lingua robot has only four motors to control the vocal tract shape, and the output speech was not very clear. Both Burpy and Lingua are not able to generate consonant sounds, fricative sounds, and mimic sentences.

Waseda Talker, WT-7RII, is probably the most advanced mechanical speech synthesis available at the moment. However, this robot is very complicated with 20 DOF, and its vocalization movement is also difficult to observe. Its vocal cords can produce pitches of 129–220 Hz, which is a little narrow in comparison with the human voice that has a pitch range of roughly 50Hz to 500Hz. Also, this robot is trained to vocalize based on human articulatory and vocal organs using Electro Magnetic Articulography (EMA) system to collect data. Thus, the number of sounds is limited by the size of collected data. Also, it is difficult to get precise articulatory data using the EMA system. The algorithm to repeat a sentence is not introduced in the WT-7RII system yet. Some prosodic features, such as tempo and melody, are ignored in its output speech.

Therefore, in this study, the author focuses on solving the problems described above together with improving the limitation of the current talking robot version. First of all, the author upgrades the mechanical system of the talking robot with voiceless sound for fricative sound generation. It is important in speech synthesis systems because there are many languages that have many fricative sounds. Without the capability of generating fricative sounds, it is impossible to synthesize those languages.

From the literature review that has been presented, all robotic speaking systems are developed and trained to perform the sound of one specific language. The capability of generating a language has not been investigated and introduced so far. Therefore, in this study, the author explores the capability of foreign language generation by the talking robot. In specific, the author designs and builds a real-time interactive modification system for easily modifying Japanese sounds to foreign sounds. The pitch range of all the current system is still narrow in comparison with human sound. Thus, the author also designs new vocal cords with simple operation mechanism to increase the pitch range of the current talking robot system. The newly designed vocal cords provide a pitch range of roughly from 50 Hz up to more than 250 Hz for the talking robot. The other limitation of these above systems are the ability or algorithm to repeat a sentence is not introduced yet. Thus another aim of this study is developing an algorithm for the talking robot to be able to generate a sentence or a series of sounds.

Speech prosody is not clearly mentioned in any of these above systems, especially the speech rhythm and tempo which is related to timing characteristics in vocalization. Thus, employing timing function with a cerebellum-like spiking neuron network model for the talking robot is proposed in this study.

Chapter 3

Robot Structure and SONN Learning

The author constructed a talking robot which is a mechanical vocalization system to reproduce speech in the same way that the human vocalization system works. In detail, the robot generates a sound source by artificial vocal cords and the shape of the resonance tube, which plays the role of the vocal tract, characterizes the output sound. By performing speech synthesis using a robot, it is possible to create a more realistic sound output which is difficult to deliver by software-based speech synthesis. In addition, it is possible to study how a robot has learned to acquire its vocalization through numerous repeat and error trials similarly to the way an infant learns to speak. In this section, the author mainly describes the configuration of the speech robot and the autonomous learning of vowels generation using the self-organizing neural network (SONN). Important upgrades such as the new vocal cord design and the voiceless system for fricative sound generation are also introduced.

3.1 Main Structure

The human voice is generated by a combination of vocalization organs such as lung, trachea, vocal cords, vocal tract, nasal cavity, tongue, and muscle. Based on this, the robot was constructed using a pressure valve, an artificial vocal cord, a resonance tube, a nasal cavity, and a speech analyzer which respectively represents the function of human lungs, vocal cords, vocal tracts, nasal cavities, and ears. Figures 3.1 and 3.2 show the configuration and appearance of the talking robot, and Table 3.1 shows the equipment of the robotic system. When speaking, firstly, airflow is sent out from the air compressor. The air passes through a pressure reducing valve and a flow control valve to vibrate the artificial vocal cord to generate a sound source. It is possible to change the volume of the generated voice by the flow control valve. The sound source is led to a resonance tube that plays a role as a vocal tract. By changing the shape of the vocal tract, it is possible to add various acoustic characteristics to the sound source. The material to make the vocal tract is silicone rubber. This material has an acoustic impedance close to human's vocal organs; thus, the sound characteristic resonated by this vocal tract is very close to human speech. In detail, the vocal tract length is about 180 mm, which is similar to the average vocal tract length of an adult male. Eight servo motors and stainless steel bars are installed in the lower part of the vocal tract, and by moving the motors, the vocal tract shape is changed accordingly. The vocal tract shape adds various resonance characteristics to the sound source. The microphone is connected to the speech analyzer, which is an integrated filter/amplifier, and is used for auditory feedback learning [3]-[7]. The detailed structure and function of each part are described below.

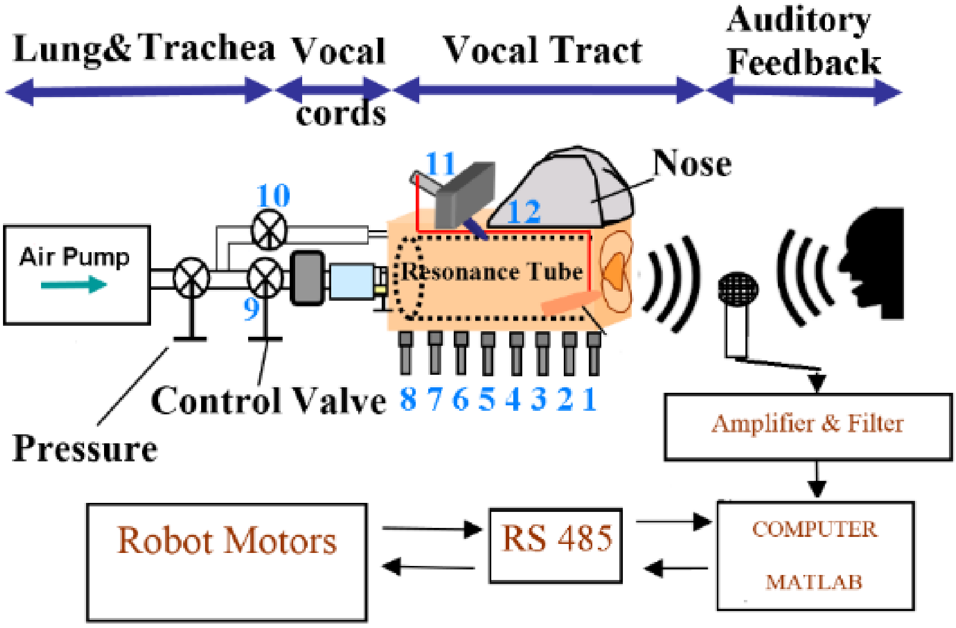
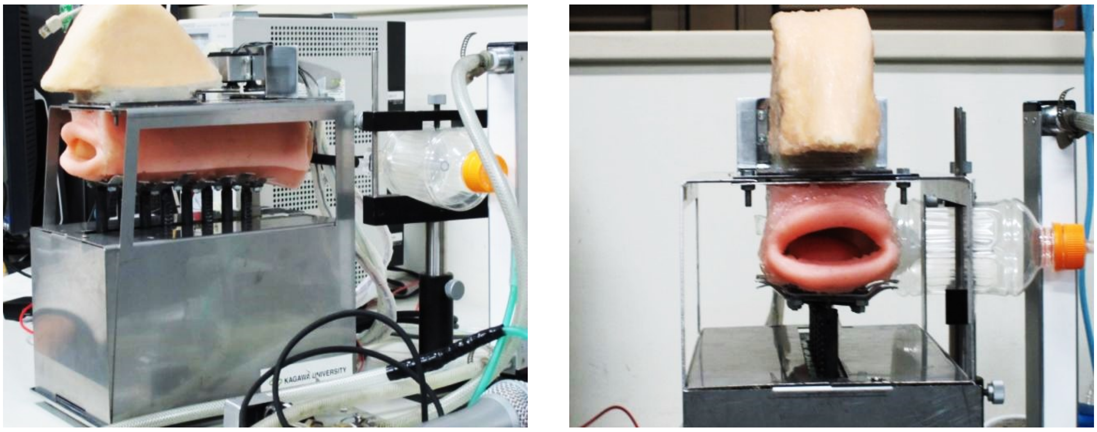


Figure 3.1: Structure of the talking robot



(a) Side View

(b) Front View

Figure 3.2: Appearance of the talking robot

Table 3.1 Equipment of the system

Computer	Mouse Computer i7-3820 3.6 GHz ; 16GB ram
OS	Microsoft Windows 7 Professional
Development Software	Matlab 2012,2016
Air Compressor	Hitachi Bebicon 40P-7s 400W
Low-pass filter	Entropy Software Laboratory FLT-02
Motor	Futaba RS-301CR
Power supply	COSEL PBA1000F 9V-10.5A
Silicone rubber	Asahi Kasei Wacker Silicone Co., Ltd. RTV-2 VP 7550 Shin-Etsu Silicone X-32-2428-4
Silicone rubber catalyst	Asahi Kasei Wacker Silicone Co., Ltd. CATALYST T 40 Shin-Etsu Silicone CX-32-2428-4
Microphone	SONY F-V820
FPGA	Xilinx Spartan 6 SP605 Evaluation Kit
Pressure valve	NKS 0-15 Kg/cm^2

3.1.1 Vocal Tract

The structure of the robot's artificial vocal tract is shown in Figure 3.3. The length of the artificial vocal tract is 180 mm which is close to the average length of an adult male’s vocal tract. The inner diameter of the vocal tract is 48 mm. The material of the vocal tract is silicone rubber whose properties are very similar to human tissue. Eight stainless steel rods are attached to the lower part of the vocal tract of the robot at even intervals of 20 mm. Eight vocal tract motors connect to these stainless steel rods using a rack and pinion mechanism. The cross-sectional area of the vocal tract is alternated by the movement of the vocal tract motors. The deformation of the cross-sectional area varies from 0 to 18.2 cm^2 which are nearly the same as the human one.

The requirements of silicone rubber material to make the vocal tract are high durability, flexibility, and reasonable tensile strength. Therefore, the author mixed two types of silicone, one with high tensile strength (Shin-Etsu Silicone X-32-2428-4) and the other one with softness and flexibility (Asahi Kasei Wacker Silicone Co., Ltd. RTV-2 VP 7550). The tensile strength, hardness and resonance characteristics of the artificial vocal tract are changed depending on the mixing ratio of the two silicone types. In this study, the ratio of mixing to make one batch of the vocal tract, lips, and tongue is shown in Table 3.2.

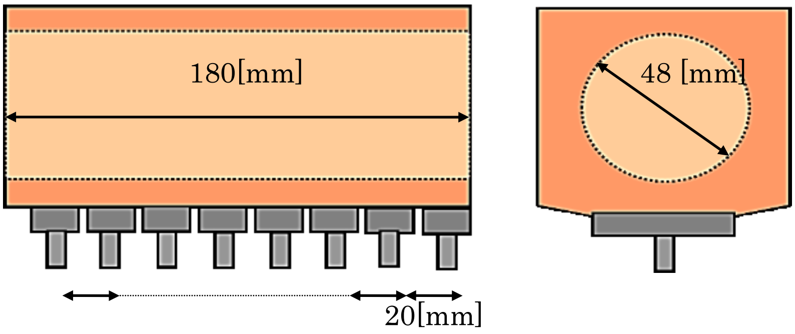


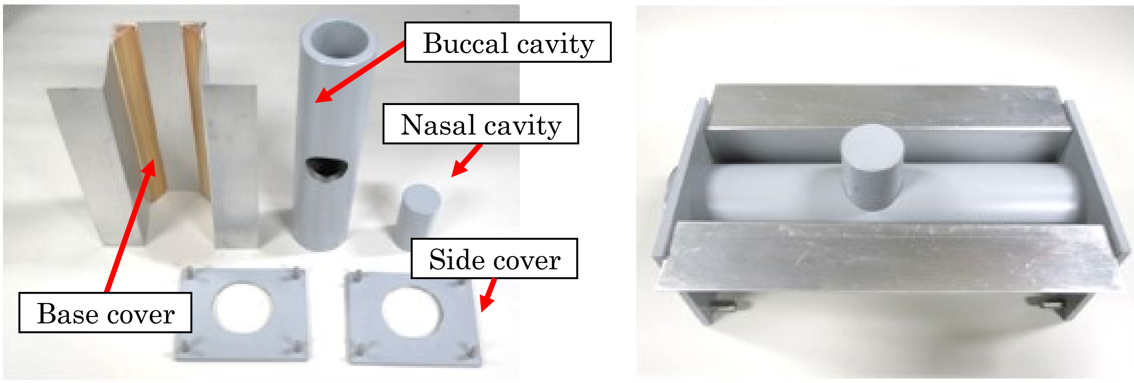
Figure 3.3 Vocal tract shape

Table 3.2 Mixing material for vocal tract, lips, and tongue

Material	Amount (ml)	%
Asahi Kasei Wacker Silicone Co., Ltd. RTV-2 VP 7550	450	87.46
Shin-Etsu Silicone X-32-2428-4	50	9.72
Asahi Kasei Wacker Silicone Co., Ltd. catalyst T 40	13.5	2.62
Shin-Etsu Silicone CX-32-2428-4 catalyst	1	0.19

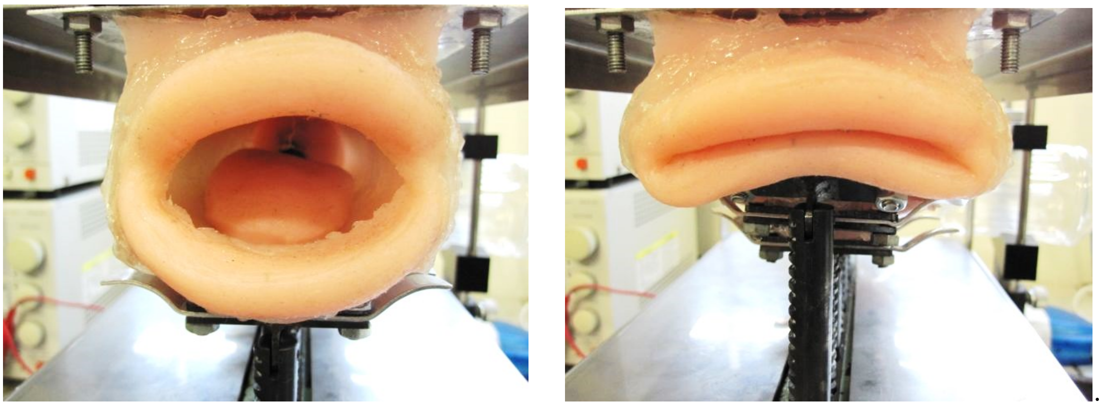
After all materials are well mixed together, the author pours them into a mold as shown in Figure 3.4 to make a vocal tract. The mold consists of a base cover, a buccal cavity, a nasal cavity rod, and two side covers as shown in Figure 3.4 (A), and the mold after assembled is shown in Figure 3.4(B). Chrome plating treatment is applied to a buccal cavity, a nasal cavity rod, and two side covers to enhance reliability. After the mold is assembled, the edges of the mold are applied with silicone adhesive to prevent leakage. It is necessary to wait for about a few hours for the silicone adhesive to dry before pouring the silicone mixture inside the mold; otherwise, leakage will occur. The silicone lips and tongue are also made by the same mixture. It takes about 50 to 60 hours for the silicone rubber to completely dry and be ready to use.

Artificial lips are glued to one end of the vocal tract of the robot with a silicone adhesive. The lips are used to close the vocal tract completely. Figure 3.5 shows the closing and opening states of the vocal tract. The robot can generate plosive sound by quickly closing and opening the lips.



(A) Parts for mold (B) Mold assembled

Figure 3.4 Molding for making artificial vocal tract



(A) Open mouth (B) Close mouth

Figure 3.5 Lip closure and opening

3.1.2 Artificial Tongue

In the talking robot, a tongue made of the same silicone rubber is attached inside the vocal tract. Figure 3.6 shows the shape and dimensions of the tongue. The base of the tongue is glued on a position that right above the third motor of the vocal tract. A long transparent string is attached to the middle of the tongue and connected to a motor installed at the end of the vocal tract. With this mechanism, the tongue can be raised and lowered, and the robot can generate the sound that required tongue movement such as */l/* and */r/* sound.

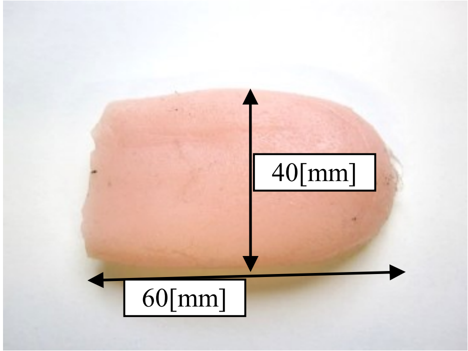
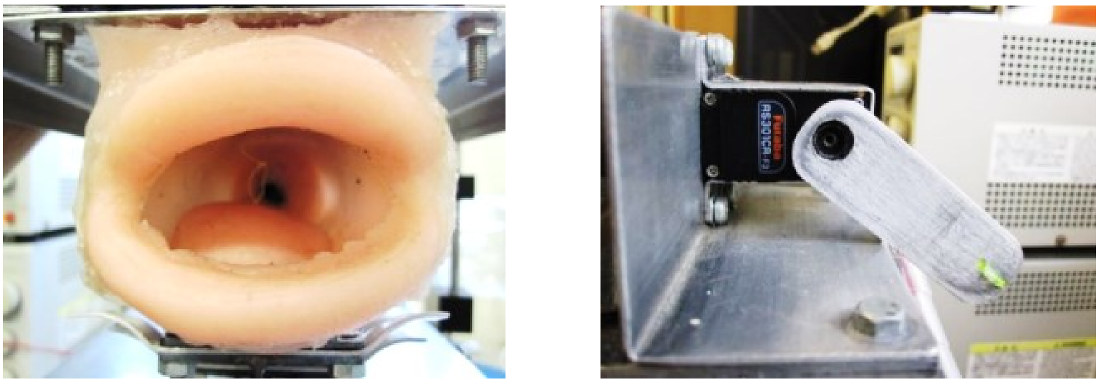


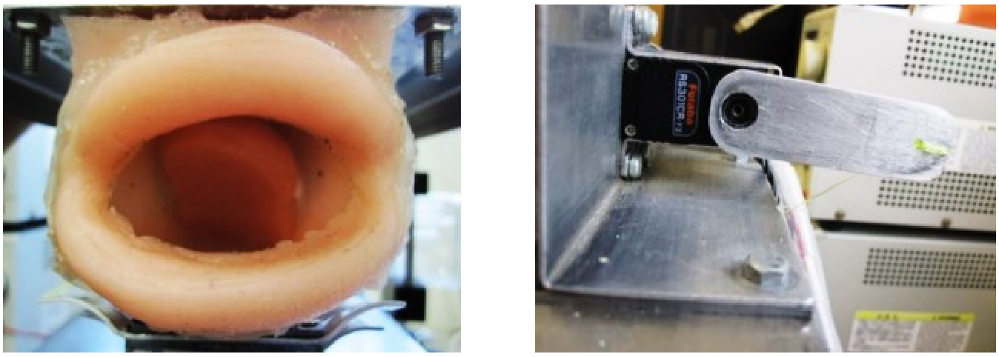
Figure 3.6: Tongue shape and dimensions

The author compares the waveform of the produced sounds when the talking robot moves the tongue up and down. The state of the tongue-down and tongue-up is shown in Figure 3.7 and 3.8. Furthermore, the speech waveform when raising and lowering the tongue is shown in Figure 3.9 and 3.10. Also, the author performs spectrum analysis on each sound for comparison. When the robot moves the tongue up and down, a difference is observed in the speech waveform and the spectrum. The waveform and spectrum of the sound with tongue up are more fluctuations than the other. However, this is not enough to generate liquid sound yet. The tongue has to move up and down to generate liquid */l/* sound. The author programmed the robot's tongue to move up and down in a short duration to generate */la/* sound. The result of */la/* sound is indicated in Figure 3.11. By using this feature, the talking robot performed the Spanish language generation on a national Spanish TV program in 2016. In this show, the robot performed the sound */ola/* which means "hello" in Spanish.



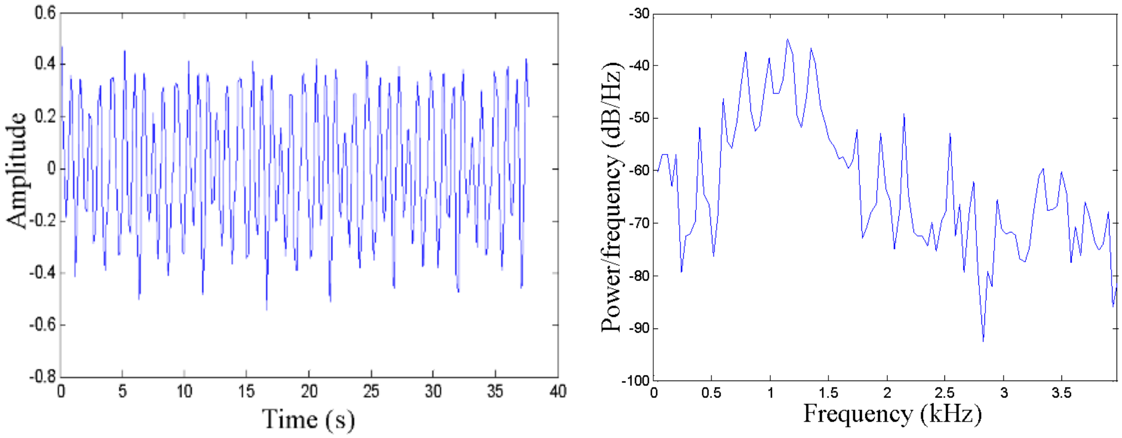
(A-1) Front view of the robot (A-2) Motor # 11 Down

Figure 3.7: Vertical motion of the tongue (tongue down)



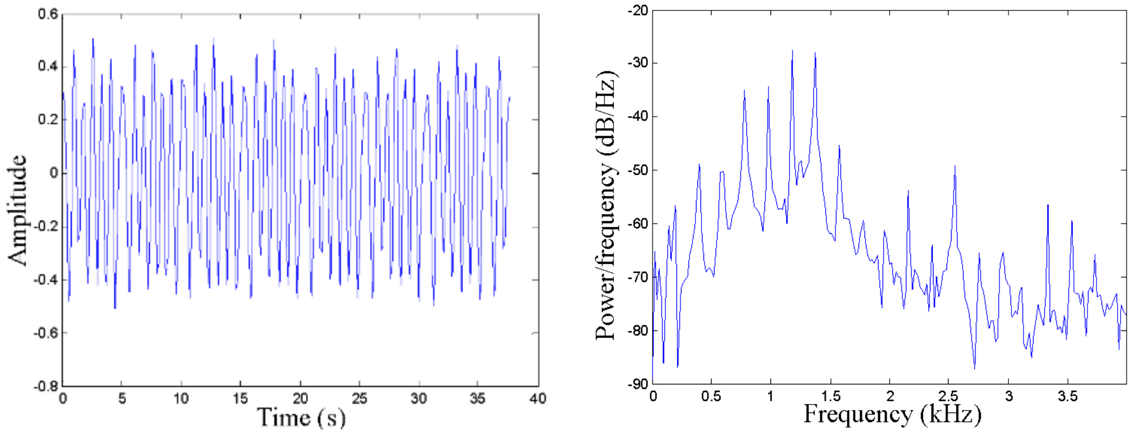
(B-1) Front view of the robot (B-2) Motor # 11 Down

Figure 3.8: Vertical motion of the tongue (tongue up)



(A) Waveform (B) Spectrum

Figure 3.9 /a/ sound with tongue-down



(A) Waveform (B) Spectrum

Figure 3.10 */a/* sound with tongue-up

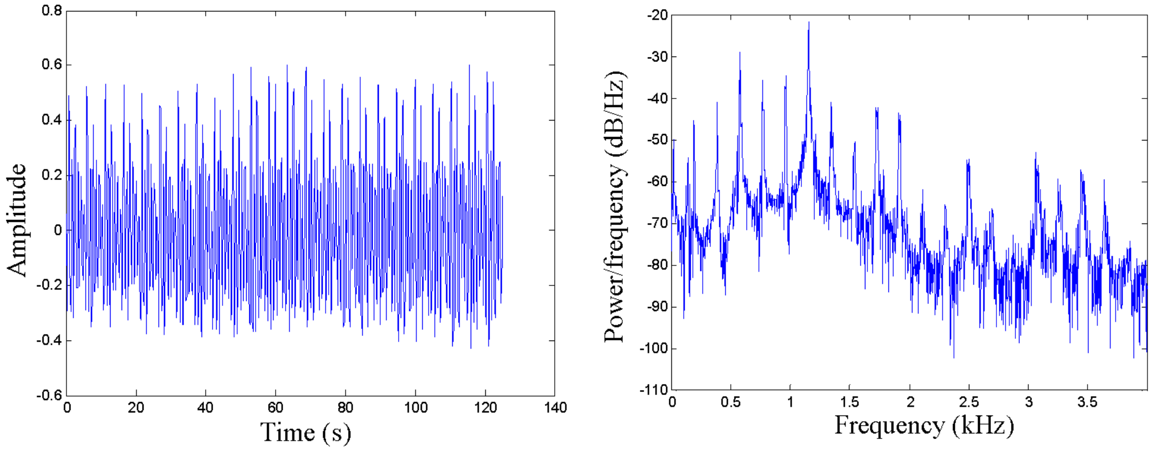


Figure 3.11: The talking robot's liquids */la/* sound with up and down tongue movement

3.1.3 Artificial Nasal Cavity

A person can block the oral cavity and nasal cavity with the soft palate to generate a nasal sound. Similarly, the talking robot has a rotary valve to open and close the nasal cavity chamber connecting to the vocal tract. By manipulating the presence or absence of nasal cavity resonance, the robot is able to generate nasal sounds, such as */m/* and */n/* sounds. Figure 3.12 and 3.13 shows the shape of the nasal cavity and the external view of the rotary valve, respectively. The nasal cavity is constructed by referring to the size and shape of the human nasal cavity chamber. By closing the lips on the vocal tract and opening the rotary valve to let the air flow to the nasal cavity chamber, the air flow is repeatedly echoed or delayed to produce nasal sounds. With this mechanism, the talking robot is able to generate nasal sounds such as */n/* and */m/*.

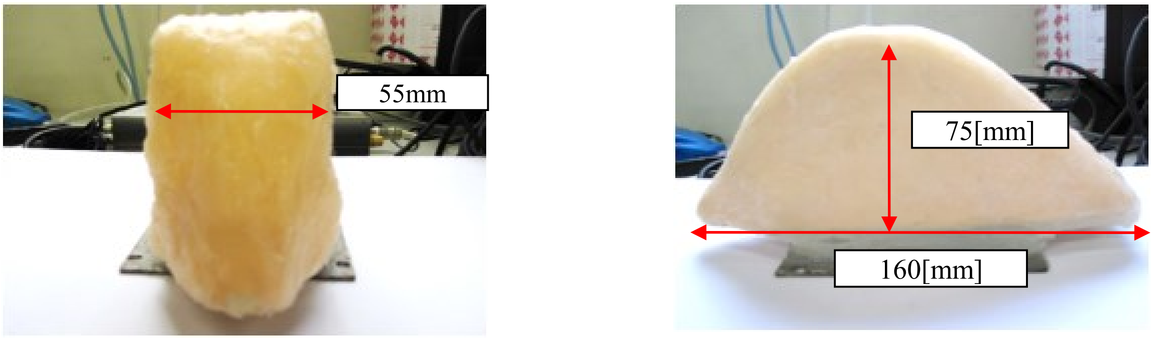
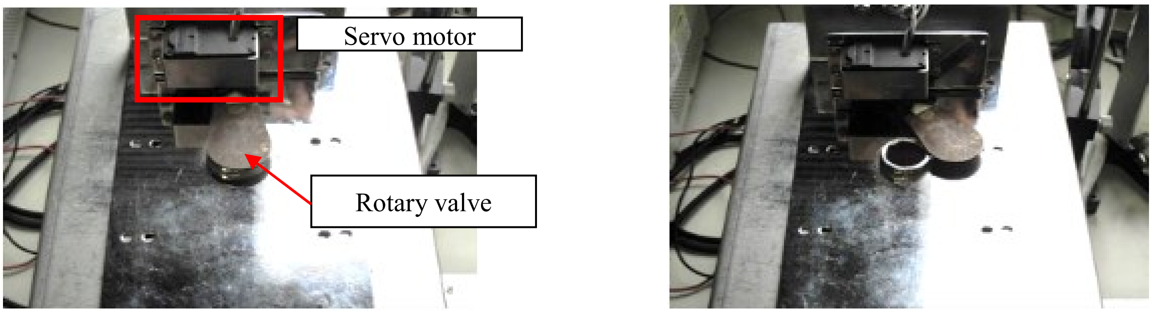


Figure 3.12: Nasal cavity chamber



(a-1) Valve close (a-2) Valve open

Figure: 3.13 Rotary valve

3.1.4 Auditory Feedback System

A sound analyzer system, which includes a microphone (SONY F-V820) and an integrated low-pass filter and amplifier (Entropy Software Laboratory), corresponding to the human auditory feedback system is also employed in the robotic system. Figure 3.14 shows the picture of the acoustic analyzer system. The human voice and the robot voice are recorded by a microphone, and environmental noise is reduced with a low-pass filter. The sounds are stored on the computer system. With many built-in audio processing functions in the Matlab software, it is easier to extract the sound features for further analysis. The auditory feedback system is critical in the autonomous vocalization of the talking robot.

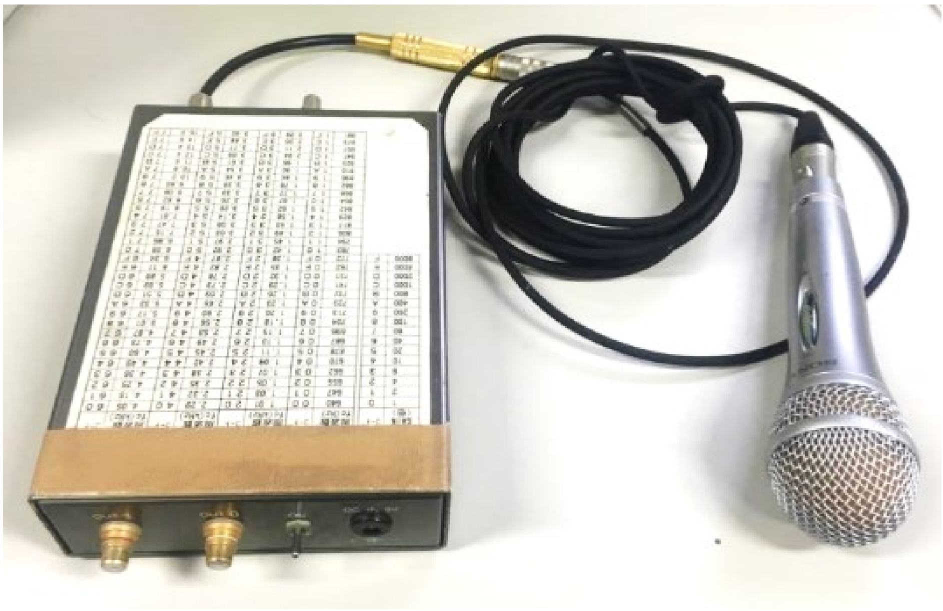


Figure 3.14: Auditory feedback system

3.2 New Vocal Cords Design

As mentioned in section 2.3, the vocal cords of the most advanced mechanical vocalization system (WT-7RII) at the current time is a complex mechanical structure with 4 DOF. However, the pitch range provided by these vocal cords was reported to be about 129-220 Hz, which was pretty narrow in comparison with the human voice. Thus, the target for the author is developing a simpler vocal cord structure with wider pitch range for the talking robot. In this study, a new design of vocal cords is developed for the talking robot using a functional approach. Therefore, the author designs and assembles a new vocal cord system that basically consists of a rubber band attached to a plastic body and is completely sealed in a plastic chamber. The curving shape of the rubber band together with the applied pressure to the rubber band determine the spectrum envelope output of the sound source. The curving shape of the rubber band and the input pressure are controlled by two different command type servo motors. The newly redesigned vocal cords greatly increase the speaking capability of the talking robot by providing a much larger pitch range in comparison with other systems.

The characteristic of a glottal wave, which determines the pitch and volume of human voice, is governed by the complex behavior of the vocal cords. It is the oscillatory mechanism of human organs consisting of the mucous membrane and muscles excited by the airflow from the lung. In the previous study, a simple design of a thin 5mm width rubber band with a fixed length of 8mm is attached to a plastic body creates an artificial vocal sound source [6]. The fundamental frequency can be changed depending on the pressure being applied to the rubber band. This artificial vocal cord design is pretty simple, and its frequency characteristics can be regulated by the amount of airflow to the vocal cord. However, the sound frequency range is very narrow using these vocal cords. In addition, when changing the sound source frequency, the volume also has to be changed. In this study, newly redesigned vocal cords with adjustable length is introduced; thus the frequency range is greatly increased, and the frequency can be changed without changing the sound volume. The length of the rubber band determines the curving shape of the vocal cords. The curving shape affects the fundamental frequency of the talking robot voice.

Figure 3.15 shows a picture of the newly redesigned vocal cords. The vibratory actions of the rubber band excited by the airflow led by the tube, and generate a source sound to be resonated in the vocal tract. Here, assuming the simplified dynamics of the vibration given by a strip of rubber with the length of L, the fundamental frequency f can be roughly estimated by the following equation.

$$f = \frac{1}{2L}\sqrt{\frac{S}{D}} \tag{3.1}$$

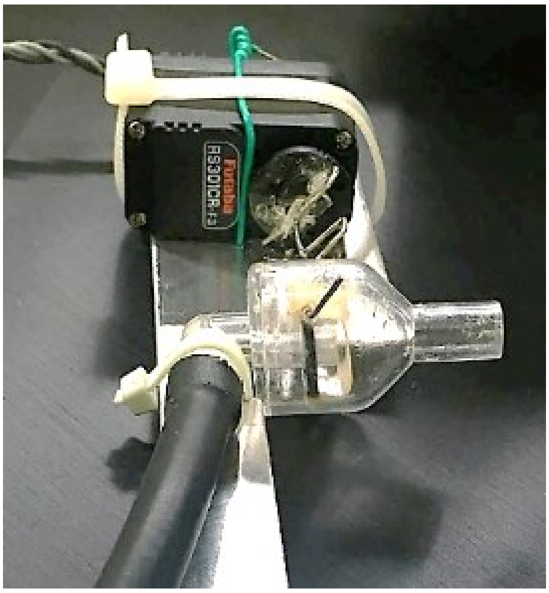


Figure. 3.15: Vocal cords picture

In equation 3.1, D is the density of material, S is the applied tension to the rubber band, and the fundamental frequency f can be changed based on alteration of L, S, and D. The tension of rubber can be adjusted by a rotating mechanism as shown in Figure 3.16.

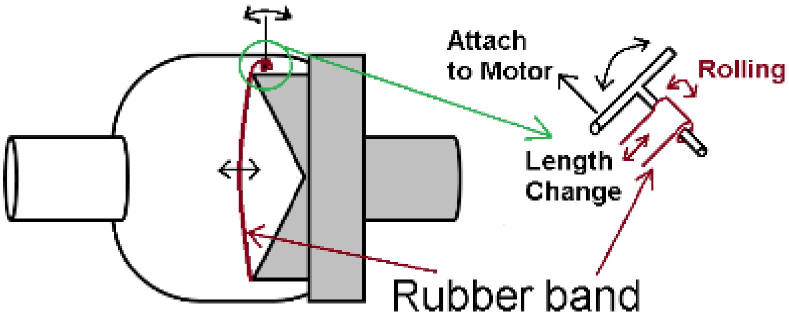


Figure 3.16: Vocal cords adjustment mechanism

For the mechanism of the new vocal cords, there is a small metal rod attached to one end of the rubber band, the other end of the rubber band is fixed on the plastic body. By rotating this metal rod, the length of the rubber band is changed; thus, the curving shape is also varied. The resolution of rotational change is 0.1 degree, and the range of rotational change varies from 150 to -150 degrees. Therefore, there are about 3000 steps to change from the lowest frequency to the highest frequency. Figure 3.17 shows the curving shape of the rubber band to generate sounds with different frequencies. By pulling the cords, the tension increases so that the frequency of the generated sound becomes higher. The structure of the vocal cords proves the easy manipulation for the pitch control since the fundamental frequency can be changed by just giving tensile force for pushing or pulling the rubber band.

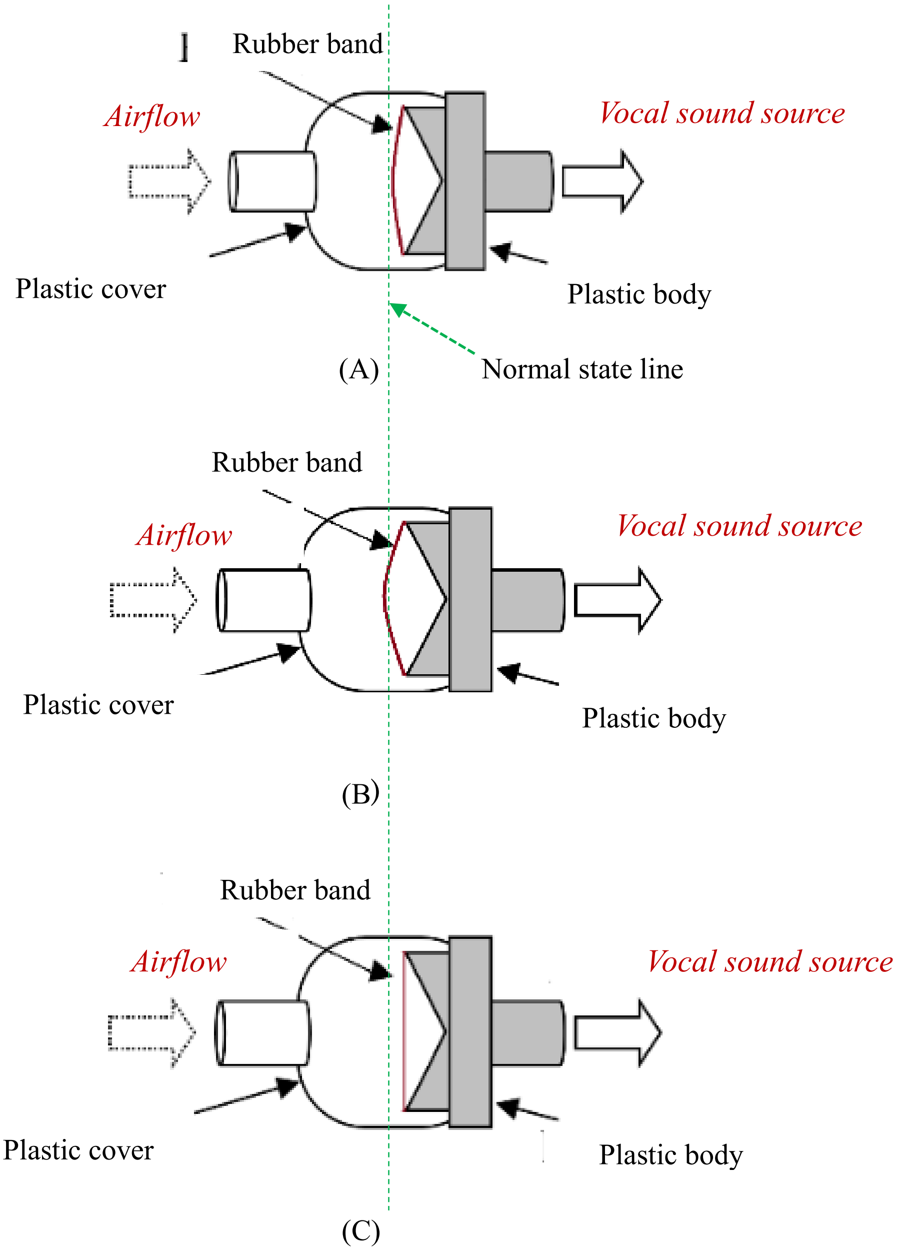


Fig. 3.17 Vocal cords shape for middle frequency (A), low frequency (B), high frequency(C)

3.2.1 Pitch Extraction

The pitch of a sound can be roughly defined as the vibration of the sound wave and measured in Hertz (Hz) unit. It is usually referred as the sensation of a frequency of a sound. The pitch of a human sound varies from 50 to 500 Hz. To determine the quality of the vocal cords, the author extracts and analyzes the pitch of the robot sound. There are several pitch detection techniques such as applying zero crossing, using autocorrelation, applying the adaptive filter, and taking harmonic product spectrum (HPS). In this study, the author applies the HPS technique to detect the pitch of the talking robot sound [12]. This technique is demonstrated by the signal flow of Figure 3.18.

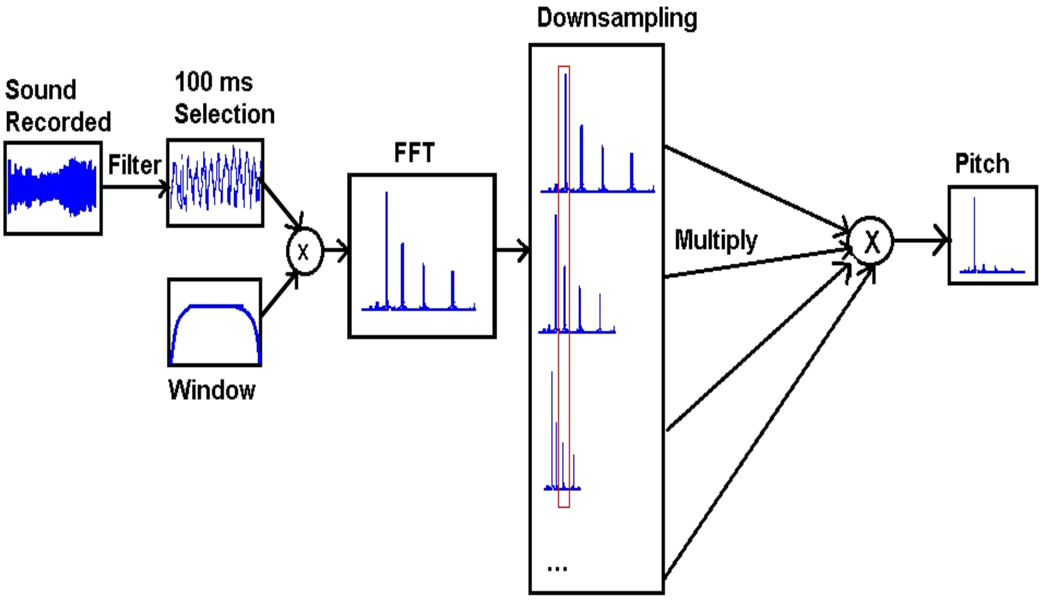


Figure. 3.18 Pitch detection using Harmonic Product Spectrum

HPS method measures the maximum coincidence of harmonics for a spectral frame. The harmonics *(Y)* of a signal *(X)* is calculated using equation (3.2), with *R* is the number of harmonics to consider. The maximum value of possible fundamental frequency *(fᵢ)* specifies the fundamental frequency as shown in equation (3.3).

$$Y(f) = \prod_{r=1}^{R} |X(fr)| \tag{3.2}$$

$$\hat{Y} = max_{f_i} Y(f_i) \tag{3.3}$$

The output sound of the talking robot is recorded using the Matlab software with a single channel, 8000 Hz sampling rate. After filtering with a band pass filter, the author selects about 100ms duration of each sound element to extract the fundamental frequency. The selected sound is windowed using a Hamming window. The signal after windowing is then performed with a Fourier transform. After that, the signal is down-sampled by the factor of 2, 3, 4 … The number of down-sampling depends on the number of harmonic frequencies (peaks) of the FFT signal. Finally, these down-sampling signals are multiplied together to get a histogram. The frequency at the peak of the histogram is the fundamental frequency of the sound signal.

3.2.2 New Vocal Cords Experiment

The author conducts three experiments to validate the improvement of the mechanical vocalization system using these newly redesigned vocal cords. The first experiment is letting the talking robot produce five different Japanese vowels */a/, /i/, /u/, /e/, /o/* with fixed air pressure input and various tensions of the rubber band. The second experiment is letting the robot generate vowel sounds with several combinations of input air pressure and tensions of the rubber band. After that, the author validates the singing performance of the talking robot by letting it sing a short simple song.

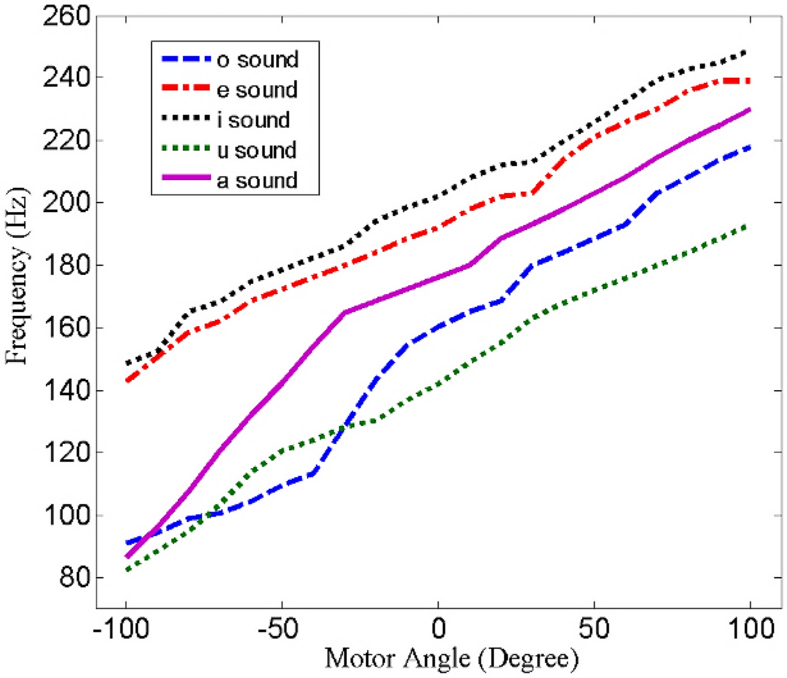


Figure. 3.19: Motor angle vs. fundamental frequency (fixed pressure)

The first experiment is first conducted with a fixed input air pressure of 1.0 Kpa, and the vocal cords shape varies depend on the rotation of the motor. The vocal cords are initially put at middle range frequency (as shown in Figure 5A) corresponding to the motor value of 0 degrees. The motor value varies from -100 to +100 degree with 10-degree stepping. Figure 3.19 shows experimental results of pitch changes versus motor angle changes for five vowel sounds. The fundamental frequency varies from 50 Hz to 250 Hz depending on the tension of the vocal cords and the vocal tract shape. Sounds with narrow cross-sectional area such as */i/* and */e/* have higher fundamental frequency than wide open cross-sectional area such as */u/* and */o/*. The fundamental frequency increases as the motor angle increase with a slope roughly about 4Hz per 10 degrees.

For the second experiment, different combinations of several motor angles and few levels of air pressure input are applied to the talking robot to generate a sound. Figure 3.20 shows the experimental results of robot sound */a/* with 20 levels of angle and five levels of pressure input. The fundamental frequency increased as the input pressure increased. By changing the pressure input and the motor angle, the fundamental frequency of sound */a/* varies from 70Hz to 240 Hz.

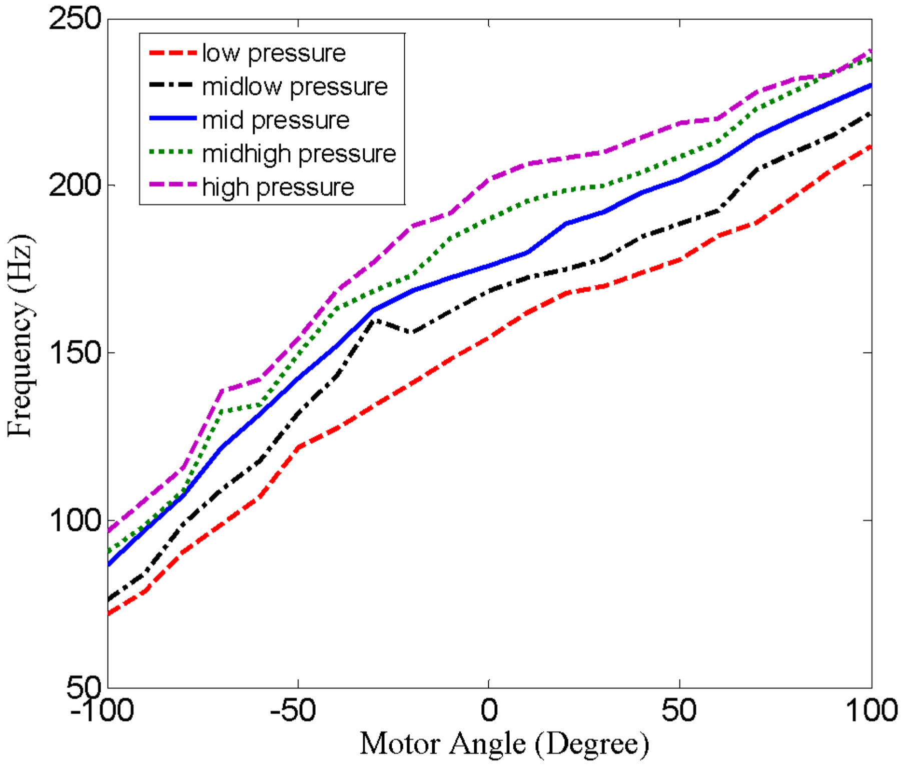


Figure. 3.20: Motor angle vs. fundamental frequency for */a/* sound (varies pressure)

To determine the stability of the fundamental frequency generated by the new vocal cords, the author conducts a new test which lets the talking robot generate each vowel sound in a long period of 2 seconds. In this test, the author only uses a mid-pressure airflow intake. Then, the author analyzes the sound pitch of that sound for every 0.2 seconds. The result of the stability test is shown in Figure 3.21. The maximum difference of the pitch for each vowel in percentage is calculated using equation (3.4). The largest value of the maximum difference of the experiment which belongs to vowel */i/,* is less than 8%, which is an excellent result.

$$Diff_{max} = \frac{|Max(X) - Min(X)|}{Min\ (X)} \tag{3.4}$$

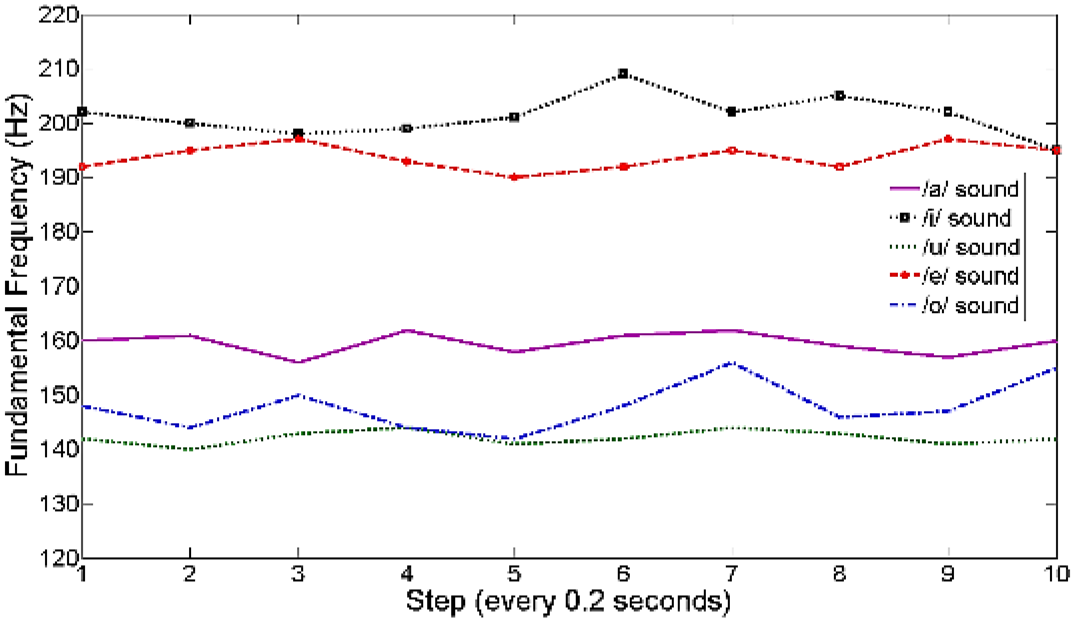


Figure. 3.21: Stability test for five vowels

Next, the author tests the reproducibility of the new vocal cords. The talking robot is programmed to generate five configurations of pressure airflow input, the tension of vocal cords, and the vocal tract shape for ten trials per configuration. The detailed configurations are provided in Table 3.1. Between each trial, the talking robot generates different sound. The result of the reproducibility test is shown in Figure 3.22. The average difference of each trial is calculated using equation (3.5), in which, n is the total of data, X_t is the value of sample t, and Ave(X) is the average value. The maximum value of the average difference is less than 5%, which is a very good result. The result of this experiment confirms the reliability of these newly redesigned vocal cords.

$$Diff_{ave} = \frac{\sum_1^n |(X)_t - Ave(X)|}{n * Ave\ (X)} \tag{3.5}$$

Table 3.3 Configuration for reproducibility test

Configuration	**Vocal tract shape**	**Vocal cords Motor angle**	**Pressure Input**
1	*/a/*	0	Mid
2	*/i/*	50	Mid-low
3	*/u/*	-50	Mid-high
4	*/e/*	100	Low
5	*/o/*	-100	High

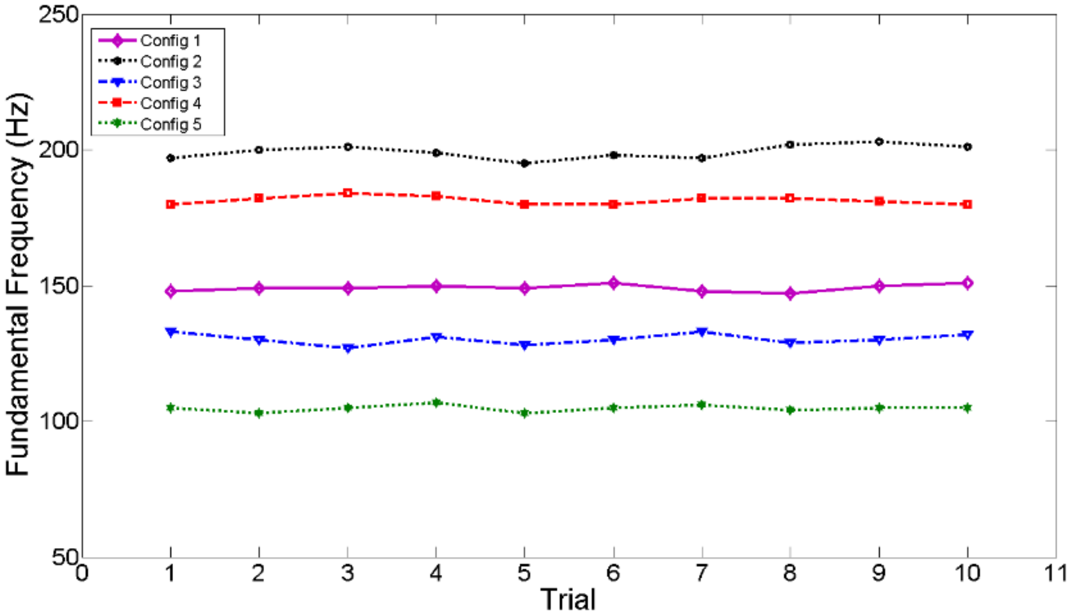


Figure. 3.22: Reproducibility test for 5 configuration

The author applied these results to the singing performance of the talking robot. The talking robot performed the melody of the alphabet song by humming the */a/* sound with the adjusted frequencies based on frequencies of a musical notes chart. The music notes of "the alphabet song" are taken from the Piano-keyboard-guide.com website, and the conversion of music notes to fundamental frequency is processed based on the information of the liutaiomottola.com website. Hence, the target notes and frequencies were established. Then, the talking robot was programmed to generate a sequence of the vowel */a/* sounds that had its frequency changed according to the target frequencies. The output sound of the talking robot was similar to a singer humming "the alphabet song". The result of the singing performance was shown in Figure 3.23. The sound pitch of the singing performance was very good when comparing to the target note. Even though this was a simple experiment, it showed that the singing performance of the talking robot was greatly improved.

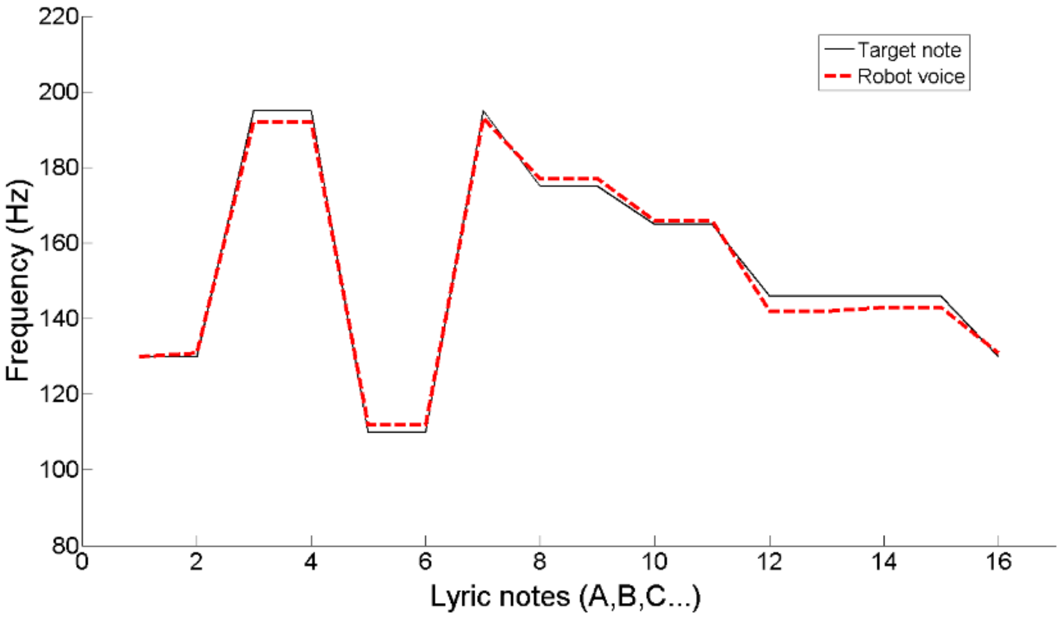


Figure. 3.23: Singing performance of "the alphabet song" with new vocal cords

In this study, the author successfully designed and implemented the new vocal cords to the talking robot which is a human-like mechanical vocalization system. The newly redesigned vocal cords greatly increase the speaking capability of the talking robot, especially its singing performance. The fundamental frequency can be easily adjusted from very low frequency to very high frequency by just changing the rubber band shape. The range of fundamental frequency is from around 50Hz to 250Hz depending on the combination of rubber band shape, input pressure, and vocal tract shape. It is the highest pitch range for a mechanical vocalization system so far.

3.3 New Voiceless System Design

3.3.1 System Configuration

In the previous version, the talking robot was constructed to reproduce simple Japanese sounds including all vowels */a/,/i/,/u/,/e/,/o/* and some consonant sounds such as */k/,/n/,/r/,/g/,/m/*. However, it could not produce unvoiced sounds such as */sa/, /shi/, /su/, /se/, /so/* yet, due to the lack of voiceless input airflow. In this study, a voiceless sound input mechanism is added to the existing system for fricative phonetics generation as shown in Figure 3.24. The airflow supplied from the air pump is led to two valves controlled by motor 9 and 10 for voiced and unvoiced sound inputs, respectively.

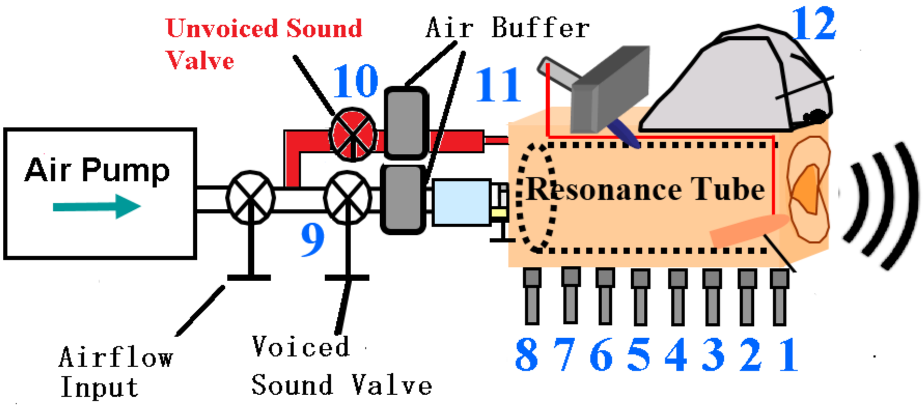


Figure. 3.24: System configuration with voiceless system

To find the appropriate movements of the robot motors for pronouncing */sa/,/shi/,/su/,/se/,/so/,* which are recognized as Japanese fricative phonetics, the authors refer to an online 2D vocal tract animation software from Uiowa university (http://soundsofspeech.uiowa.edu/english/english.html). The talking robot is programmed to speak Japanese fricative phonetics by letting the unvoiced sound go through the narrow cross-sectional area of the resonance tube, then followed by the vowel generation. Firstly, the narrow cross-sectional area is formed at the lips side of the vocal tract. Then motor 10 opens and lets the air pass through it. The air pressure is evenly distributed by the air buffer to prevent high-speed flow air to the vocal tract. At this state, the sound of the robot is similar to whispering sound. Then, motor 9 opens to let the air come to the vocal tract. The vocal tract vibrates at a certain frequency and generates a sound source. The sound source is resonated inside the vocal tract and add phonemic features for the output sound. The fricative sound occurs by slowly changing from the whispering sound to the vowel sound. The mechanism is shown in Figure 3.25.

For experimentation, the talking robot is programmed to produce 5 basic Japanese fricative phonetics, and then these phonetics were compared with those of a human. The sounds of the talking robot and a human for */sa/* phoneme was shown in Figure 3.26 (A) and (B). It is easy to realize an unsounded part, a transition part and a vowel part in both sounds. However, the robotic sound has a shorter transition part by comparison to the human one. Thus, the result of the fricative phonetics of the talking robot is a little less clear than can be recognized. This limitation comes from the motor speed that lets the air rush through the vocal tract too fast.

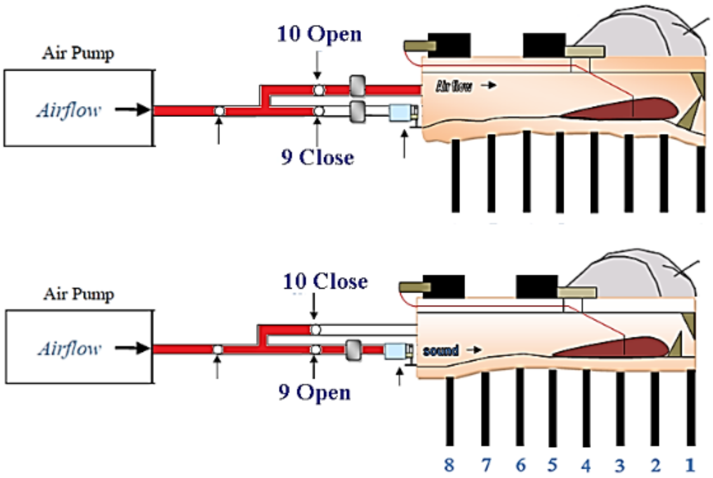


Figure. 3.25: Fricative sound mechanism of the talking robot

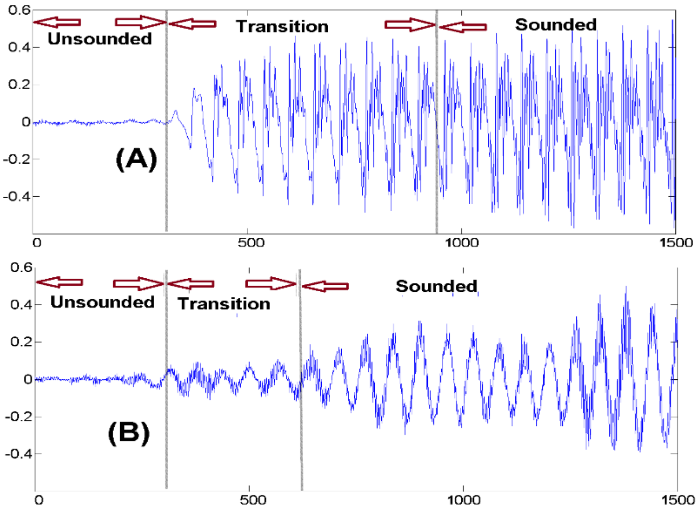


Fig. 3.26: */Sa/* sound of human (A) and robot (B)

3.3.2 Vietnamese language performance

The new vocal cord and voiceless system of the talking robot are applied to reproduce the Vietnamese language, which is a tonal language and contains many unvoiced sounds. The Vietnamese language is different from other languages because of the tonal effect within the language. There are six intonations in the Vietnamese language, which are “normal tone,” “falling tone,” “high rising tone,” “rising then falling tone,” “fast falling tone,” and “mid raising tone.” The tone affects the meaning of a spoken word. For example, */ba/, /bà/, /bá/, /bả/, /bạ/,* and */bã/* mean “father”, “lady”, “governor”, “at random”, “poison”, and “residue”, respectively. To implement the tonal effect in the talking robot, the author divides the airflow to the artificial vocal cords into six levels representing the six tones of the Vietnamese language [41].

The robot is programmed to speak several Vietnamese phrases, and the output speech is evaluated by several Vietnamese students of Kagawa University. There are three phrases, which are */xin chao/* means "hello", */bai-bai/* means "goodbye", and */toi la robot/* means "I am a robot", in the experiment. After listening to the Vietnamese output sounds of the talking robot, the Vietnamese student was to evaluate the clearness, understandable, fricative quality, and intonation quality on the scale of 10. The listeners apprised that the sound from the talking robot had a clear tonal effect, but the fricative sound effect was not very clear. However, the listener stated that he could recognize and understand the Vietnamese sound from the talking robot.

In this study, the performance of the talking robot on Vietnamese language speaking ability was evaluated. The tonal effect performance of the talking robot was pretty good. However, fricative phonetics of the talking robot were a little less clear than could be recognized by a human. Figure 3.27 shows the result of the survey in Vietnamese language performance experiment. On the horizontal are the six Vietnamese students who are subjected to the listening test. There are four columns indicating the four criteria which were questioned for evaluation. The scale of each criterion is 10. From 1-3 is bad, 4-6 is not bad, 7-8 is good, 9 is excellent, and 10 is perfect.

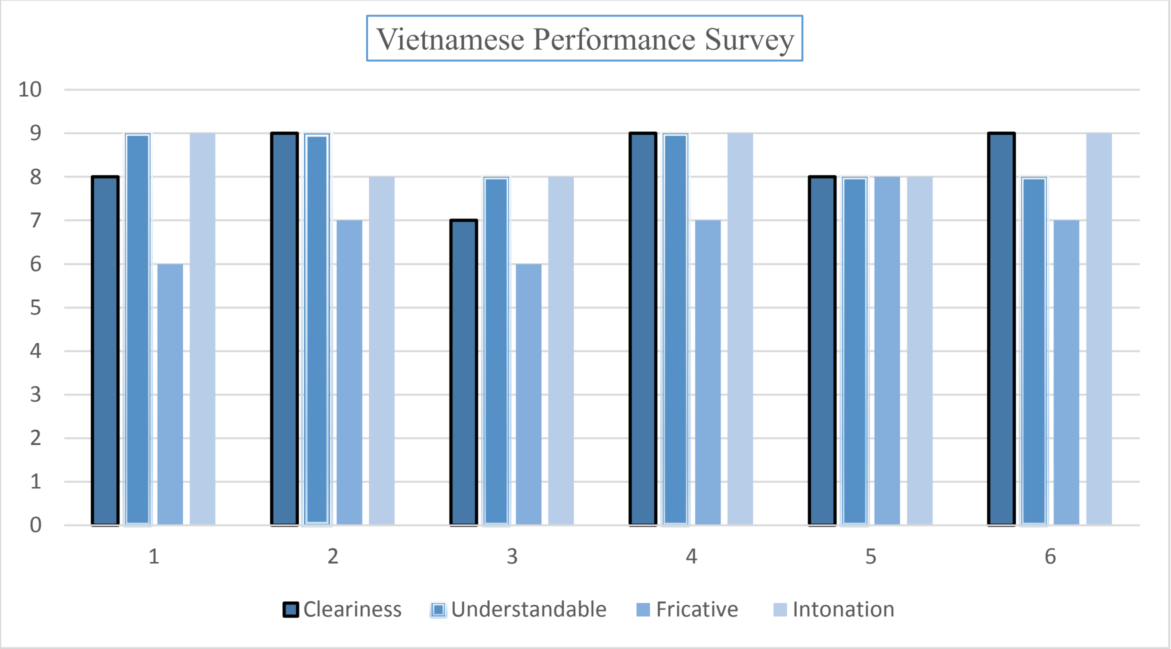


Figure. 3.27: Vietnamese language performance survey result

3.4 SONN Learning

In most of the robotic synthesis models, the mechanical reproductions of the human vocal system were mainly directed by referring to MRI images of the human vocalization system. However, with this method, it was not possible to decide the vocalization of the vocal tract when the input was an unknown sound. To solve this problem, it was necessary to construct an association between the robot's voice and its vocal tract shape. Also, an algorithm that can estimate the shape of the vocal tract for arbitrary voice input is necessary. For the human, an infant learns to speak by repetition of trial and error concerning the hearing and vocalizing of vocal sounds. Therefore, the author uses a neural network to autonomously acquire speech by auditory feedback learning in the sound acquisition period. By doing this, the association between the generated sounds of the talking robot and the motor control parameters was made.

3.4.1 Auditory Feedback

Figure 4.1 and 4.2 respectively show the role of auditory feedback in the learning phase and the vocalization phase. In the learning phase, a neural network is connected in a feedback loop. The relationship between the acoustic features extracted from the sound of the talking robot and the motor control parameters is established.

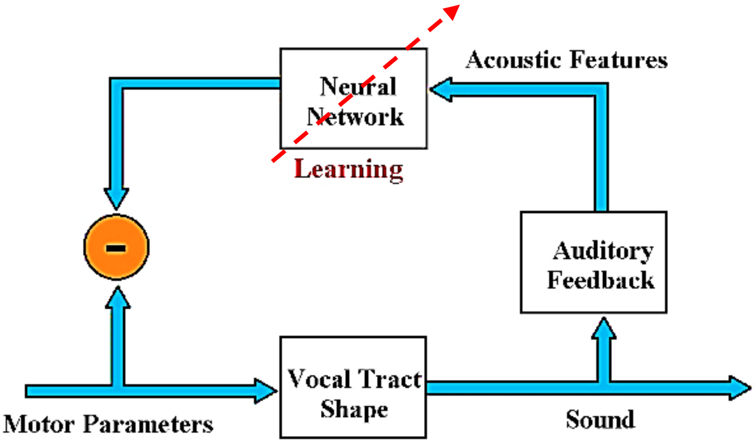


Figure. 3.28: Learning phase

After learning, connect the networks in series. By connecting this to the speech robot, the robot estimates the shape of the vocal tract corresponding to the input target voice and utters the utterance.

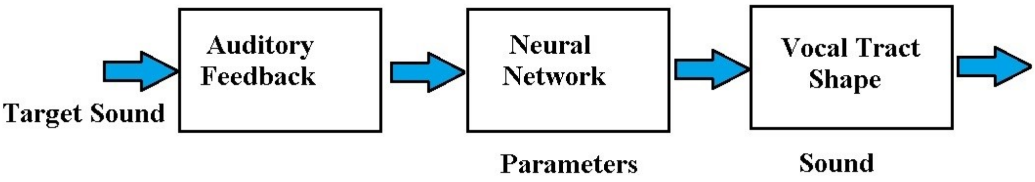


Figure. 3.29: Vocalization phase

Self-organizing mapping (SOM), which is proposed by T. Kohonen of the University of Helsinki, is a type of artificial neural network that is trained using unsupervised learning. SOM is used to present the high-dimensional data on a low-dimensional (usually two-dimensional) map [47]. SOM differs from other types of artificial neural networks as it uses competitive learning rather than error-correction learning, such as backpropagation with gradient descent. SOM constructs a mapping which preserves neighborhood relations by updating weight vectors in the learning algorithm. Figure 4.3 shows the mapping configuration which presents a map from high-dimensional input data R_m onto a two-dimensional feature map $w_i = [w_{i1,}w_{i2,}...,w_{im}]\epsilon\ R_m$. Weight vector w_i at node i on the feature map is connected to the input nodes x_j [$j=1,2,...,m$].

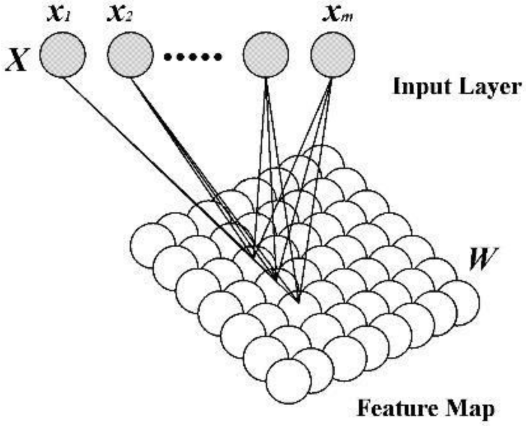


Figure. 3.30: SOM mapping structure

In this study, the inputs for SOM are vectors of nine elements, which are nine coefficients, extracted from vocal sound using Mel-frequency Cepstral coefficients analysis (MFCC), and the weighting vectors, m_i, are initialized with small random values. A Gaussian function, which is initialized to a large value, is employed for the learning of the three-dimensional SOM. Steps for learning are presented below.

The winner c which has the minimum Euclidean distance with x_i is selected by equation (1) on the feature map.

$$c = argmin_t \|m_i - x_i\| \tag{3.1}$$

By using the winner c, the weight vectors w_i-s are updated using equations (2) and (3).

$$m_i(t+1) = m_i(t) + h_{ci}[x_i(t) - m_i(t)] \tag{3.2}$$

$$h_{ci} = \alpha(t)exp\left[-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right] \quad 0 < \alpha(t) < 1 \tag{3.3}$$

Here, $\alpha(t)$ is a learning parameter which indicates the weight of the learning, and its value decreases as the learning proceeds. h_{ci} expresses the range of neighborhood learning. By repeating these steps for a number of times, high-dimensional input sound features can be placed on the low-dimensional feature map. On the feature map, similar feature quantities are embedded in neighborhoods by competitive learning.

3.4.2 Artificial Neural Network

To construct the correspondence between the feature map vector and motor parameters, an artificial neural network (ANN) composed of three perceptron layers with the propagation learning method is used[36]. Figure 4.3 shows the ANN structure.

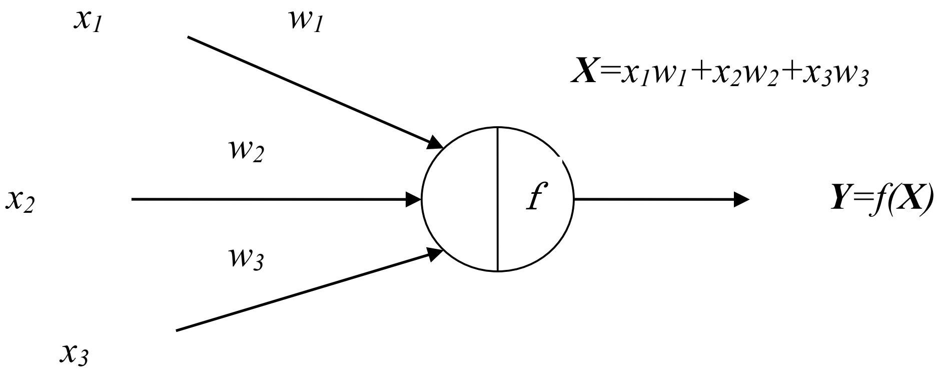


Figure. 3.31: Artificial neural network structure

The input for ANN is a vector *X* that has a certain weight *w,* which can be changed according to learning, for each element. The output value is the transformation of the sum of the weighted input vector using the activation function *f()*. The expression of the output value is shown in equation 3.4 below.

$$X = \sum_i w_i x_i$$
$$y = f(X) \tag{3.4}$$

The author has constructed a correspondence relation between feature map vectors and motor parameters by using back propagation method with three layers perceptron neural network. Multilayer perceptron neural network, which is a feedforward artificial neural network, was proposed by Rosenblatt in the United States for the purpose of pattern classification in 1957. Figure 3.32 shows a simple perceptron model. There is an S-unit (Sensory unit), A-unit (Associative unit) and R-unit (Response unit) in the model. First, when a pattern to be classified is input, the S-unit reacts. Next, the A-unit receives an input from the S unit and outputs a signal. Then, the R-unit responds to input from the A-unit and outputs a signal corresponding to the classification of the input pattern.

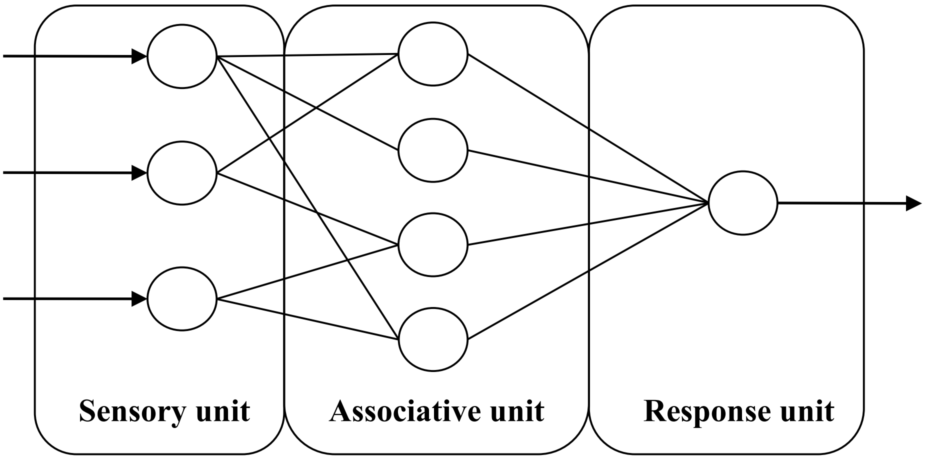


Figure. 3.32: Simple Perceptron Model

When learning a three-layer perceptron, the author updates the coupling weight so that the error becomes smaller while comparing the teacher signal and the signal output by the perception. Figure 3.33 shows the learning procedure of a three-layer perceptron.

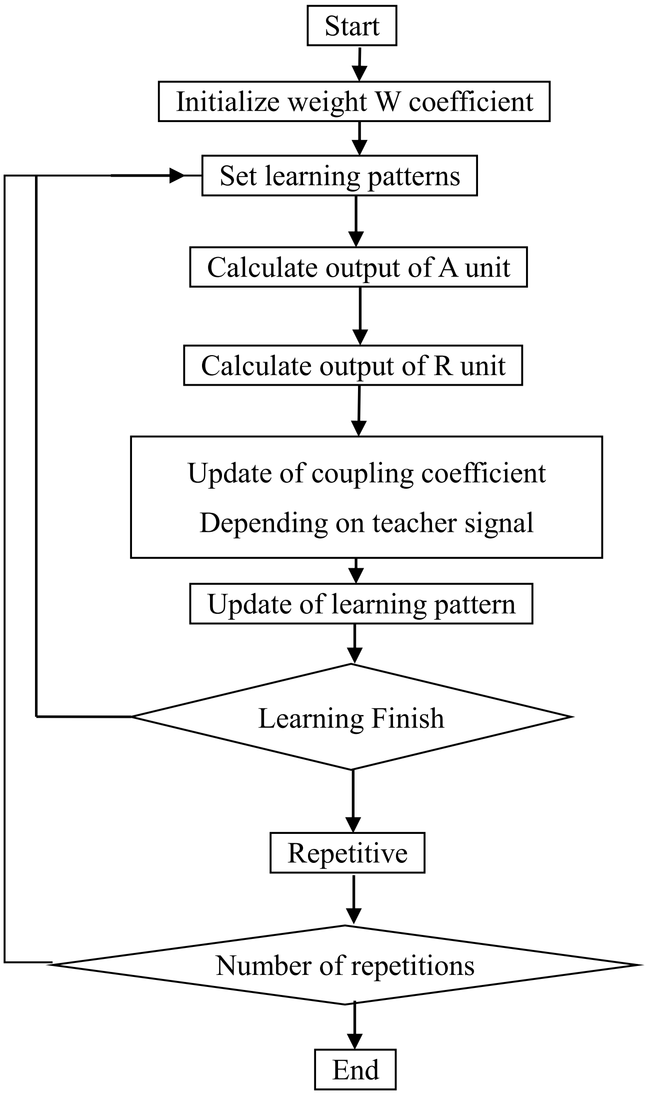

Start
Initialize weight W coefficient
Set learning patterns
Calculate output of A unit
Calculate output of R unit
Update of coupling coefficient
Depending on teacher signal
Update of learning pattern
Learning Finish
Repetitive
Number of repetitions
End

Figure. 3.33: Multilayer perceptron neural network learning procedure

The learning procedure for three layer perceptron neural networks is as follows:

i. Initialize the weight coefficient W from A-unit to R-unit and the threshold value of R-unit with a random number.

ii. Set the first pattern as a learning pattern.

iii. Using the learning pattern as input, obtain the output of the A-unit.

iv. Obtaining the output of the R-unit from the output of A-Unit with the equation $R = f(W \cdot A - \theta)$. At this time, θ represents the threshold of R unit.

v. Update the connection weight W with $W(t+1) = W(t) + C \cdot (Te - R) \cdot A$. *C* is a constant and *Te* is a teacher signal.

vi. Let the next pattern be a learning pattern.

vii. Return to (iii) until the learning pattern ends.

viii. Update the number of iterations of learning.

ix. If the number of iterations of learning is within the limit number, return to (iii).

3.4.3 Self-organizing Neural Network

In this study, a self-organizing neural network (SONN) was constructed by combining a self-organizing map (SOM) and neural network (NN) as the approach of associating acoustic features and motor control parameters. Figure 3.34 shows the configuration of SONN. SONN consists of an input layer, a competitive layer, an intermediate layer, and an output layer. All layers are fully connected by weight vectors v_{ij}, w_{jk}, and w_{kl}. Also, the dimension of the input layer is the 9th order represented for nine MFCC coefficients of the sounds, and the dimension of the output layer is the 8th dimension represented for eight vocal tract motors. The dimensions of the SOM map and hidden layer are arbitrarily determined according to the learning conditions.

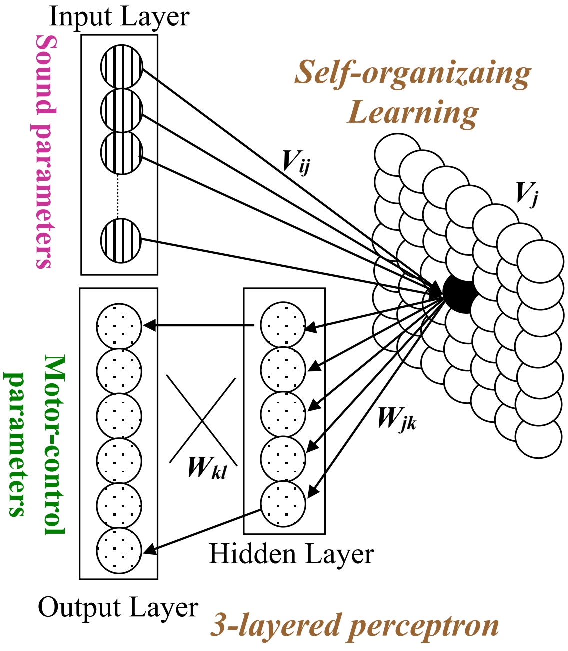
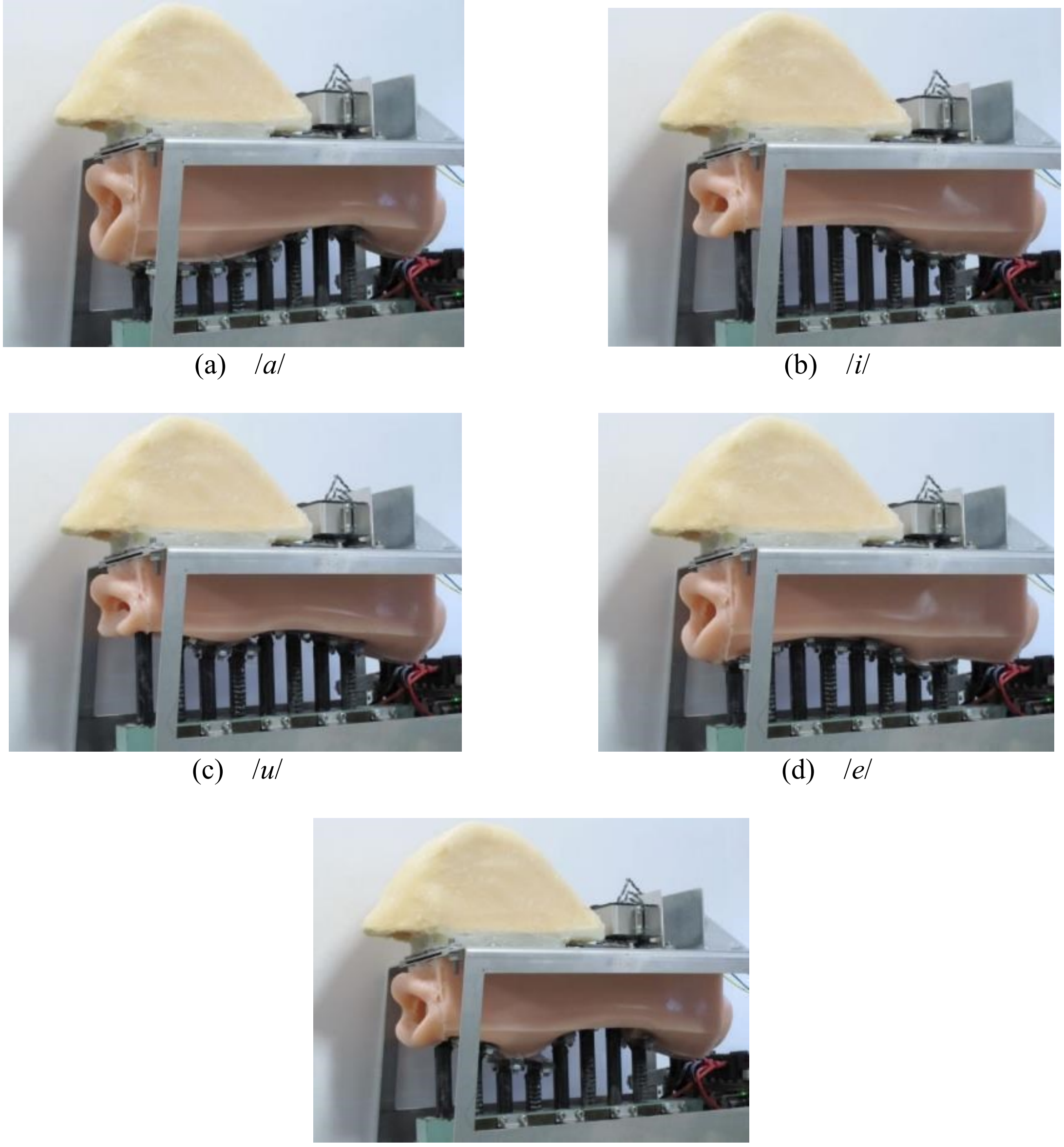


Figure. 3.34: SONN configuration

The acoustic parameters composed of the Cepstrum coefficients obtained from the target speech are input to the input layer of the SONN. The Euclidean distance between the input acoustic parameter and the acoustic parameter arranged on the competitive layer is obtained, and the cell with the smallest distance is selected. The cell pattern on the competitive layer is estimated according to the distance to the selected cell. Estimate motor control parameters from patterns on the competitive layer.

The vocal tract shapes for five basic Japanese vowels after training are shown in Figure 3.35 below.



(a) /*a*/

(b) /*i*/

(c) /*u*/

(d) /*e*/

(e) /*o*/

Figure. 3.35: Vocal tract shape for Japanese vowels

Chapter 4

Interactive Modification System

The talking robot is able to autonomously acquire the control skills of the mechanical system to vocalize stable vocal sounds and mimic a human voice. However, there are many factors that can affect the sound quality of the robot such as the decreased quality of silicone rubber with time, the influence of air temperature and humidity on air pressure and the non-linear characteristics of air flow when the robot is continuously speaking long phrases. Therefore, the authors are developing a real-time control interface, which allows the users to visualize the action of the motors of the talking robot so that they can manually adjust the control parameters to overcome the unpredictable factors described above. Through the manual adjustments, the user can regulate the sounds made by the robot to produce a clearer sound output; especially when setting up the robot to speak a foreign language. Also, novel formulae 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 based on the formant frequency comparison is proposed.

4.1 Human Machine Interface Design

4.1.1 Real-time Interface System [46]

In order to achieve real-time interaction between the robot and a human, the author builds a Graphic User Interface (GUI) using the Matlab software as shown in Figure 4.1.

The interface has buttons programmed to initialize the robot and output utterances for speech demonstration; it also has sliders to visualize the motor actions of the robot, and a graph to display the produced sounds and acoustic features. The amount of air pressure to the robot and the shape of the vocal tract can be adjusted by changing the sliders.

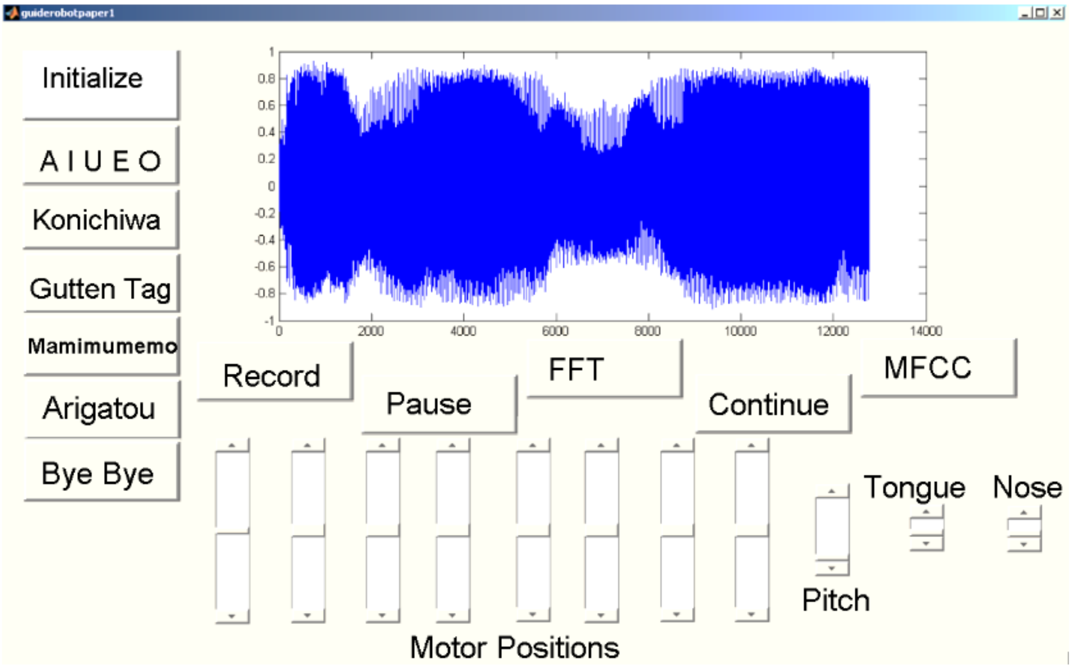


Figure 4.1: Matlab GUI Interface of the robot

Command type servomotors (Futaba) are employed for driving the mechanism of this robot. The advantages of the command type motor are high speed, stability, accuracy, durability, and having built-in feedback signal. In addition, multiple motors can be controlled simultaneously with only one RS 485 serial port. It requires a long command line input to drive the motors through the RS 485 communication protocol. The general structure for sending a packet to control multiple motors is shown below.

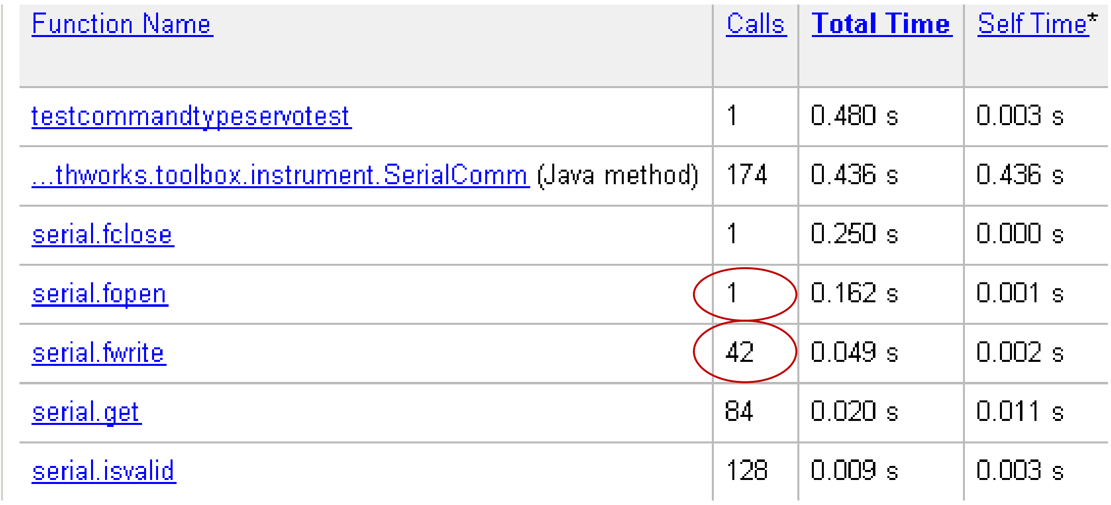
Header (4- bytes) – ID – Flag – Address – Length – Count– Servo ID– Data (4bytes) – Servo ID– Data – ...– Sum

**** Sum is the XOR logic operation of all previous bytes (Header -> last Data)***

Example: For rotating 10 degrees for motors ID1, ID2 and 50 degrees for ID5, the command is:

AFFA – 00 – 00 – 1E – 03 – 03– 01– 64 00 – 02 – 64 00 – 05 – F401 – ED

For the verification of the speed transferring signal, the author uses built-in “run and time” Matlab functions to run a sample program. The result is shown in Figure 4.2 below. The two important values are highlighted with red circles. The time to send 42 commands for 42 articulation movements to the motor is less than 50 milliseconds, and the time to get the feedback signal of all the motor positions is less than 20 milliseconds. The short period of sending and getting signals indicates that the GUI system can be used to observe and modify the movement of the robot via the sliders in real-time.



Function Name	Calls	Total Time	Self Time*
testcommandtypeservotest	1	0.480 s	0.003 s
...thworks.toolbox.instrument.SerialComm (Java method)	174	0.436 s	0.436 s
serial.fclose	1	0.250 s	0.000 s
serial.fopen	1	0.162 s	0.001 s
serial.fwrite	42	0.049 s	0.002 s
serial.get	84	0.020 s	0.011 s
serial.isvalid	128	0.009 s	0.003 s

Figure 4.2: Reaction time testing result

4.1.2 Flow Chart of Real-Time Interaction [3]

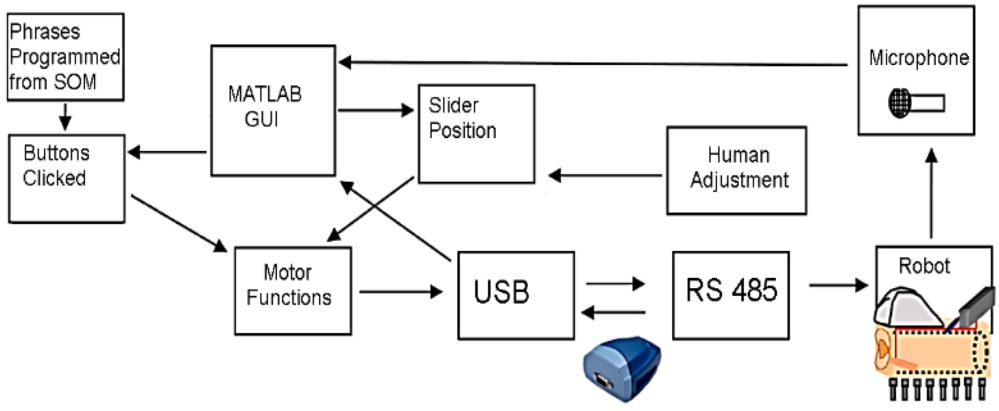


Figure 4.3: Flow chart of real-time interaction system

The flow chart of the system is indicated in Figure 4.3. When the GUI program runs, it establishes the online interconnection between the robot and the computer via the RS-485 USB adapter. Each button arranged in a dialogue is programmed with a motor vector set based on SONN trained data to let the robot speak a phrase. When a button is clicked, the articulatory motors move, and the robot generates the respective sound expressions. The sound of the robot is automatically recorded and displayed on the GUI screen. The recorded sound wave is saved in the Matlab workspace for further analysis. The record button is for recording human sound for the comparison.

4.2 Novel Formula for Formant Frequency Change [3]

4.2.1 Formant Extraction from Speech

To investigate the physical factors that determine human speech to be compared with robot speech, a set of utterances was recorded from one adult male. The phrases such as */Arigatou/* meaning “thank you” in Japanese, and */Gutten Tag/* meaning ”good morning” in German, were recorded at a sampling frequency of 8 kHz using the Matlab recording function. The speech waveform of the */Arigatou/* phrase and its formant frequencies for each sound segment is shown in Figure 4.4.

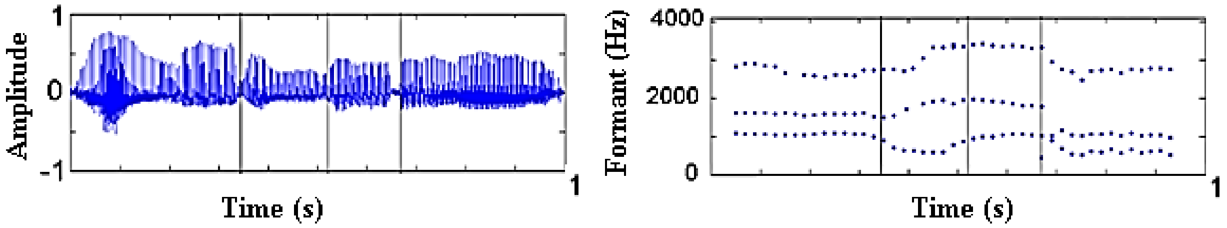


Figure 4.4: Human’s sound wave */Arigatou/* and its formant frequencies

LPC is one of the widely used methods in the sound recognition process that attempts to predict feature values of the input signal based on past signals. There are several techniques to solve the LPC coefficients such as covariance method, autocorrelation method, and lattice method [38], [39]. For this study, the autocorrelation method is applied to solve for LPC coefficients. The waveform for each utterance contains three elements: the sound pressure (amplitude), the phoneme duration and the frequency. The sound features were calculated for each sound segment with Hamming window of frame size 100 milliseconds with frame overlap of 80 milliseconds. The order of linear predictive coding (LCP) coefficients (N$_{lpc}$) for each sound depends on equation (4.1) with a_k is the LPC coefficient of order Kth, and N is the number of coefficients.

$$s(n) = \sum_{i=1}^{N} a_k . s(n-1) \tag{4.1}$$

Let r(k) is the value of the autocorrelation of the signal with k shift samples:

$$r(k) = \sum_{n=-\infty}^{\infty} s(n)s(n+k) \tag{4.2}$$

The LPC coefficients a$_1$, a$_2$, a$_3$ … is the solution of the following equation:

$$\begin{bmatrix} r(0) & r(1) & r(2) & \dots & r(N-1) \\ r(1) & r(0) & r(1) & \dots & r(N-2) \\ r(2) & r(1) & r(0) & \dots & r(N-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r(N-1) & r(N-2) & r(N-3) & \dots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ r(3) \\ \vdots \\ r(N) \end{bmatrix} \tag{4.3}$$

This equation can be solve using Levinson-Durbin recursive algorithm as following.

Let N=1;

The average square difference of 1st order is:

$$E_1 = r(0)(1 - a_1^2(1)) \tag{4.4}$$

with $a_1 = \frac{-r(1)}{r(0)}$

Recursion with n=2,3,4…,N

i. Calculate K_n coefficient:

$$K_n = -\frac{r(n)+r_{n-1}^{bT}a_{n-1}}{E_{n-1}} \tag{4.5}$$

ii. Estimate coefficients of nth order

$$a_n\,(n) = K_n = a_{n-1}(k) + K_p a_{n-1}(n-k) \tag{4.6}$$

iii. The average square difference of nth order is:

$$E_n = E_{n-1}(1 - K_n^2) \tag{4.7}$$

iv. Return to i step, replace n by n+1 if n≤N. In the end, the LPC coefficient is:

$$[(a_N(1) \quad a_N(2) \quad a_N(3) \quad \dots \quad a_N(N)] = -[(a_N(1) \quad a_N(2) \quad a_N(3) \quad \dots \quad a_N(N)] \tag{4.8}$$

The opposite value of the LPC coefficient is referred as the reflection coefficient [39].

Formants are the spectral peaks in the spectrum. Each formant frequency set corresponds to a certain shape of the human vocal tract [3]. The formant frequencies comparison technique has been widely used in vowel synthesis systems for its similarity to the way human auditory systems work.

The formant frequencies for each segment are estimated based on the LPC method. Firstly, the LPC coefficients are calculated for each segment, and then the formant detection technique was applied. The number of formant frequencies varies across the sound wave due to the characteristics of a different phonemes. For this study, the LPC based estimation technique will be used to estimate the sound elements formant frequencies [48]. The sound element spectrum is calculated using the Fourier transform function. The author uses 8th order LPC to obtain the coefficients for the spectrum signal. The roots of the polynomial function from these LPC coefficients have real and imaginary parts. The angular frequencies and bandwidths are then calculated from the root's value. The formant frequencies then determine if the pole is close to the unit circle. These steps are referred to in (4.9) to (4.15).

Equation (4.1) can also be expressed as (4.9) with *e(n)* as an error signal.

$$s(n) = Z + e(n) \tag{4.9}$$

Taking the z-transform of (4.9) as filtering operation, we have (4.10) with *H(z)* is a linear predictive filter.

$$E(z) = H(z).S(z) \tag{4.10}$$

where,

$$H(z) = \sum_{i=1}^{N} a_k.z \tag{4.11}$$

The denominator of the transfer function can be factored into (4.12). Where, C_j are a set of complex numbers, with each complex conjugate pair of poles representing a resonance at frequency F_k and bandwidth B_k as shown in (4.13) and (4.14), respectively.

$$1 + \sum_{i=1}^{N} a_k.z^{-1} = \prod_{j=1}^{N} (1 - C_j.z^{-1}) \tag{4.12}$$

$$F_j = \frac{F_s}{2\pi} \tan^{-1}\left[\frac{\mathrm{Im}(C_j)}{\mathrm{Re}(C_j)}\right] \tag{4.13}$$

$$B_j = -\frac{F_s}{\pi} \ln\left|C_j\right| \tag{4.14}$$

$$r_k = \sqrt{\mathrm{Im}(C_j)^2 + \mathrm{Re}(C_j)^2} \geq 0.7 \tag{4.15}$$

For simplicity, the first three formant frequencies were taken for comparison of the similarity between human and robot sounds. The sound wave and its formants frequencies of the phrase */Arigatou/* from the robot is shown in Figure 4.3. The formant frequency is determined by (4.15).

4.2.2 Effects on formant frequencies of motor position

For an opened at one end and closed at the other end tube with the uniform cross-sectional area, the natural frequency or formant frequency is given by equation 4.16, where c is the velocity of sound (roughly 340 m/s), and l is the length of the tube.

$$F_1 = \frac{c}{4l} \tag{4.16}$$

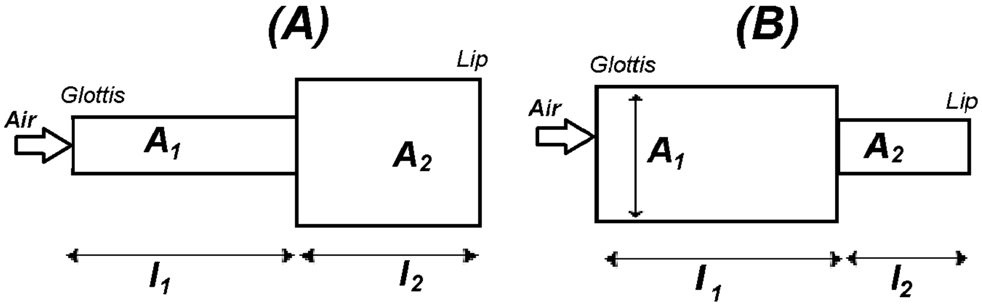


Figure 4.5: Typical vocal tract shapes

The frequency will be changed if there is a change in the cross-sectional area of the tube. The formant frequencies corresponding to the vocal tract shape in the case of Figure 4.5 (A) are the combination of formant frequencies of two opened at one end and closed at the other end tubes. The vocal tract of sound */a/* has this shape. The formant frequencies depend on the length of l_1 and l_2; the shorter the length, the higher frequency is generated. Thus, if the motors at the middle of the vocal tract change the position, the formant frequency would also change accordingly. Assuming $l_1 > l_2$, the first and second formant frequency changes are given by equation (4.17) and (4.18).

$$\Delta F_1 = F_1' - F_1 = \frac{c}{4(l_1+\Delta l)} - \frac{c}{4l_1} = -\frac{cl_1\Delta l}{4l_1(l_1+\Delta l)} \tag{4.17}$$

Similarly,

$$\Delta F_2 = -\frac{cl_2\Delta l}{4l_2(l_2+\Delta l)} \tag{4.18}$$

The formant frequencies in the case of Figure 4. 5(B) is the combination of formant frequency of one opened at one end and closed at the other end and closed at both end tube. The vocal tract of sound */i/* and */e/* have this shape. The Helmholtz resonant frequency effect of closed at both end tube is put into consideration; thus, the first formant frequency is given by equation (4.19).

$$F_1 = \frac{c}{2\pi\sqrt{\frac{l_1 A_1 l_2}{A_2}}} \tag{4.19}$$

When the lengths of l_1, l_2, and the areas A_1, A_2 are changed due to motor movements, the first formant frequencies are also changed. The second formant frequency change is similar to the case of Figure. 4.4 (A) as described in equation (4.18). The next section investigates the change in formant frequencies due to the change of motor positions in the real system.

4.3 Adjustment Strategy Based on Formant Frequency

In order to predict the adjustment values, the relationship between the sound characteristics and the adjustment value needs to be investigated. The robot vocal tract was divided into 3 sections: the lip-end section, the middle section, and the throat-end section. The lip-end section includes motors number 1 to 3, the middle section includes motors number 3 to 6, and the throat-end section includes motors number 6 to 8. For this experiment, the authors programmed the robot to produce the basic Japanese vowels */a/, /i/, /u/, /e/, /o/*. During these vowel productions, the motor positions were automatically adjusted to increase 2mm (125 in motor parameter value) in each step for 5 steps. The average results of five trials were taken, and the relationships between formant frequencies shift with motor increments are shown in Figure 4.6. This information gives a user a rough estimation for which part of the vocal tract and how much in value it needs to be adjusted in order to obtain a more precise result for each sound segment by comparing the formant frequencies of a human voice to the robot voice.

For the lip-end motors group, the relationship between the motor increment and formant frequencies increment is roughly:

- 1st formant almost the same
- 2nd formant: F2 = +50Hz per 1mm motor increment
- 3rd formant: F3 = +40Hz per 1mm motor increment

For the middle motors group, the relationship between motor increment and formant frequencies increment is roughly:

· 1st formant almost the same

· 2nd formant: F2 = +20Hz per 1mm motor increment

· 3rd formant: F3 = +20Hz per 1mm motor increment

For the throat-end motors group, the relationship between motor increment and formant frequencies increment is roughly:

· 1st formant F1 = +15 Hz per 1mm motor increment

· 2nd formant: F2 = +10 Hz per 1mm motor increment

· 3rd formant F3 almost the same

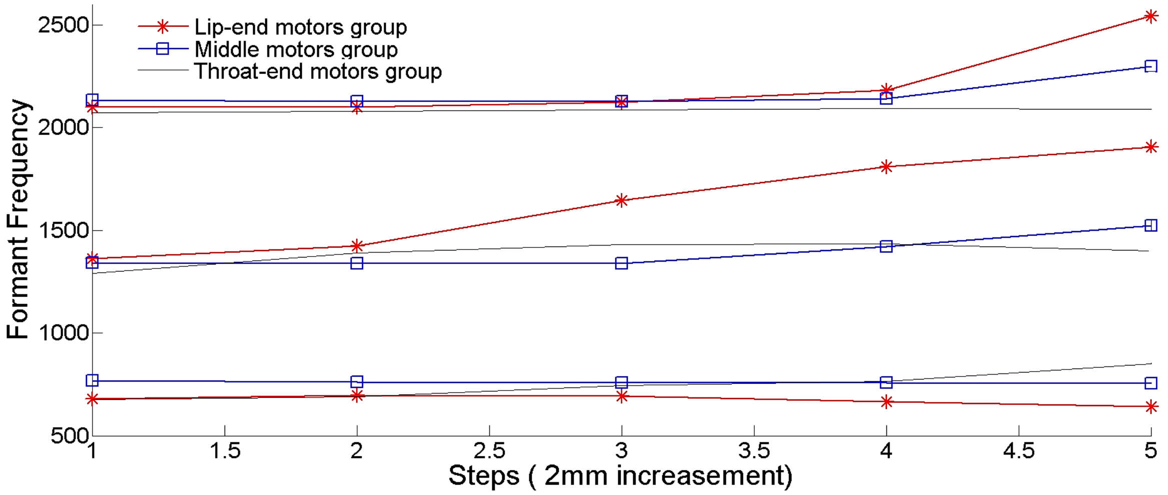


Figure 4.6: Formant frequencies change vs. motors increments

From these relationships, the author noticed that it was easier to adjust the 2nd formant frequency because it had the highest changing value according to the motor value adjustments. Therefore, the strategy is to adjust the 2nd formant frequency first if possible. The current vocal tract shape also influences which motors should be adjusted. The pitch motor value is proportional to the sound wave amplitude, and the adjustment for sound wave amplitude can be judged based on the amplitude difference between human sound and robot sound.

4.4 Performance of Interactive Modification System

As reported in our previous research [3] - [7], the adaptive control system helped the robot speak a wide range of words and phrases; however, the precision of these words and phrases was not as good as expected. Motor control vectors of vowel and consonant sounds were learned from a stable state of the vocal tract shapes, which was different from the dynamic action. For example, the robot can clearly reproduce basic Japanese sounds such as */ri/* by combining the motor vectors for the consonant */r/* and vowel */i/*; nevertheless, when the robot is in the process of reproducing the term */Arigatou/,* which means ”thank you” in Japanese, the sound */ri/* is different due to the connection with the adjacent sounds. Therefore, a real-time interaction system between the robot and human is built to visualize robot actions and allows a user to quickly adjust the sound output to get more accurate results. By combining the basic set of motor vectors corresponding to consonant and vowel sounds, 6 phrases in Japanese, German, and English were programmed, which are */A-I-U-E-O/, /Konnichiwa/, /Gutten Tag/, /Mamimumemo/, /Arigatou/*, and */Bye bye/*. The phrases were recorded at a sampling frequency of 8 kHz using the Matlab recording function. To investigate the physical factors that determine the human speech to be compared with robot speech, a set of utterances was recorded from one adult male. Figure 4.7 shows the sound wave and formant frequency of */A-I-U-E-O/* sound from human (A) and the robot (B). It is easy to perceive that the formant frequency change during phonation from human sound and robot sound are pretty similar.

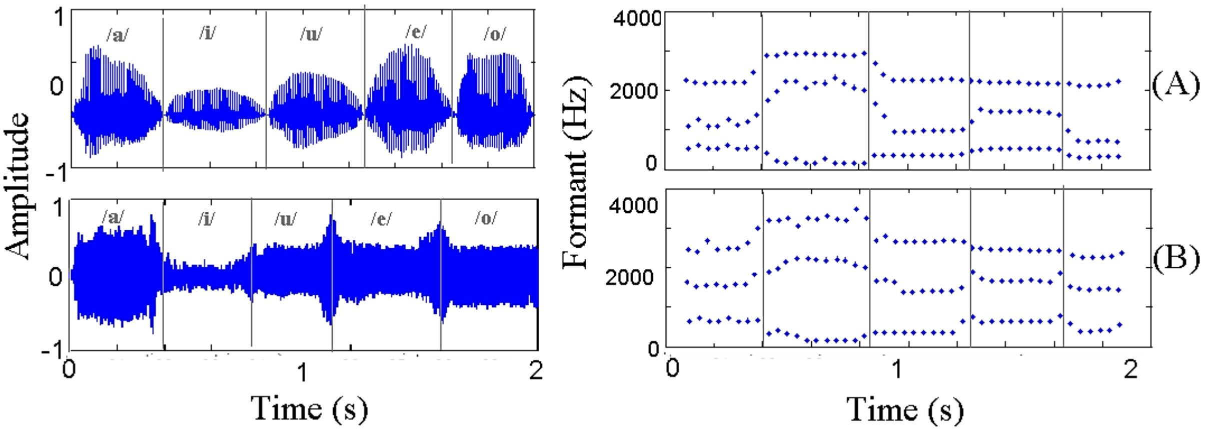


Figure 4.7: Sound */A-I-U-E-O/* from human (A) and robot (B)

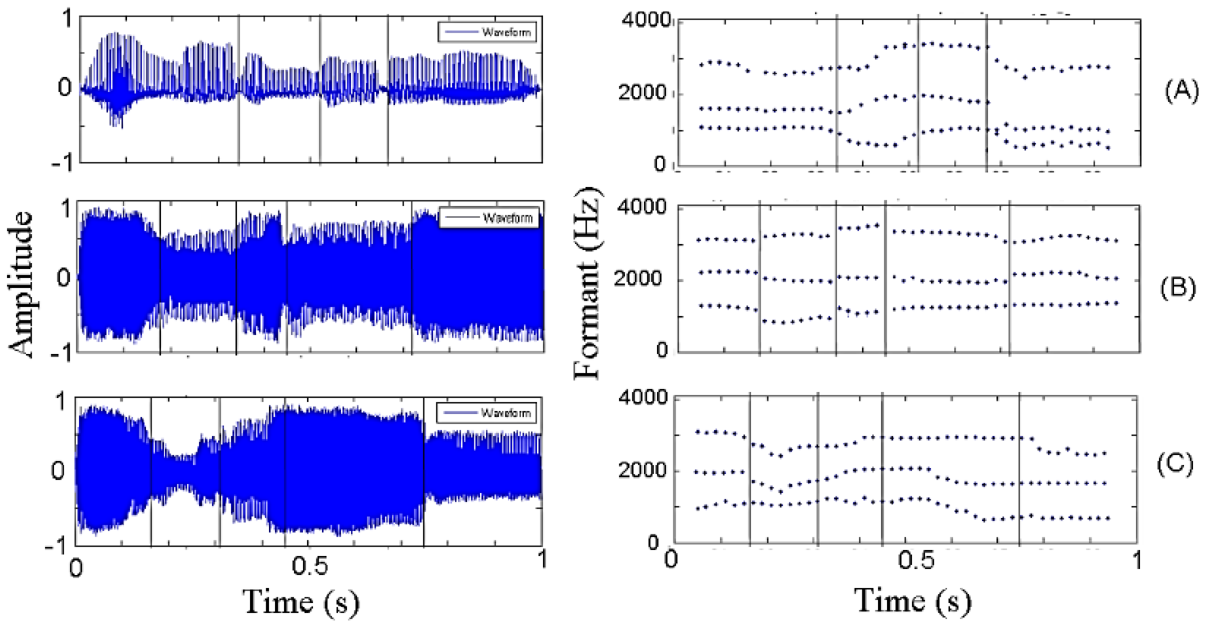


Figure 4.8: Sound wave */Arigatou/* and its formant frequencies from human (A), robot before (B), and after (C) adjustment

By combining consonants and vowels as mentioned in section 2.3, the robot was able to speak the Japanese phrase. Comparing the waveform of Figure 4.8 (A), (B), (C), and the phrase */Arigatou/* was well reconstructed. Adjusting the slider simply added clearness. However, when the robot was set to speak the German phrase */Gutten Tag/*, the produced sound was unrecognizable. Manually adjusting the sliders to change the robot motor values resulted in a recognizable */Gutten Tag/* phrase. Speech waveform of */Gutten Tag/* phrase and its formant frequencies for a male human and the talking robot with and without human adjustment are respectively shown in Figure 4.9 (A), (B), (C). When a foreign term, which was German in this case, was initially introduced, the sound reproduction performance was considerably lower. In terms of frequencies and amplitude, the waveform from Figure 4.9 (C) showed more similarity to the human waveform in Figure 4.9 (A) while the waveform from Figure 4.9 (B) displayed differences. The output robot voice for */Gutten Tag/* before adjustment sounded like */A-ui-i-ack/,* thus, it was necessary to adjust the motor values in order to obtain a recognizable output voice.

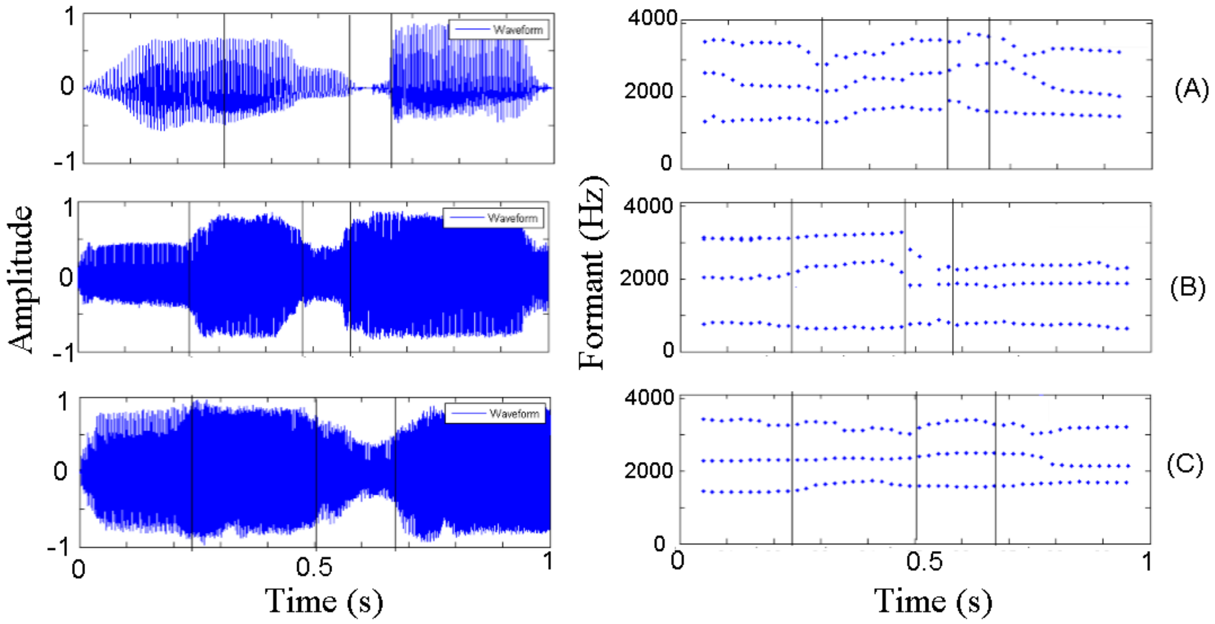
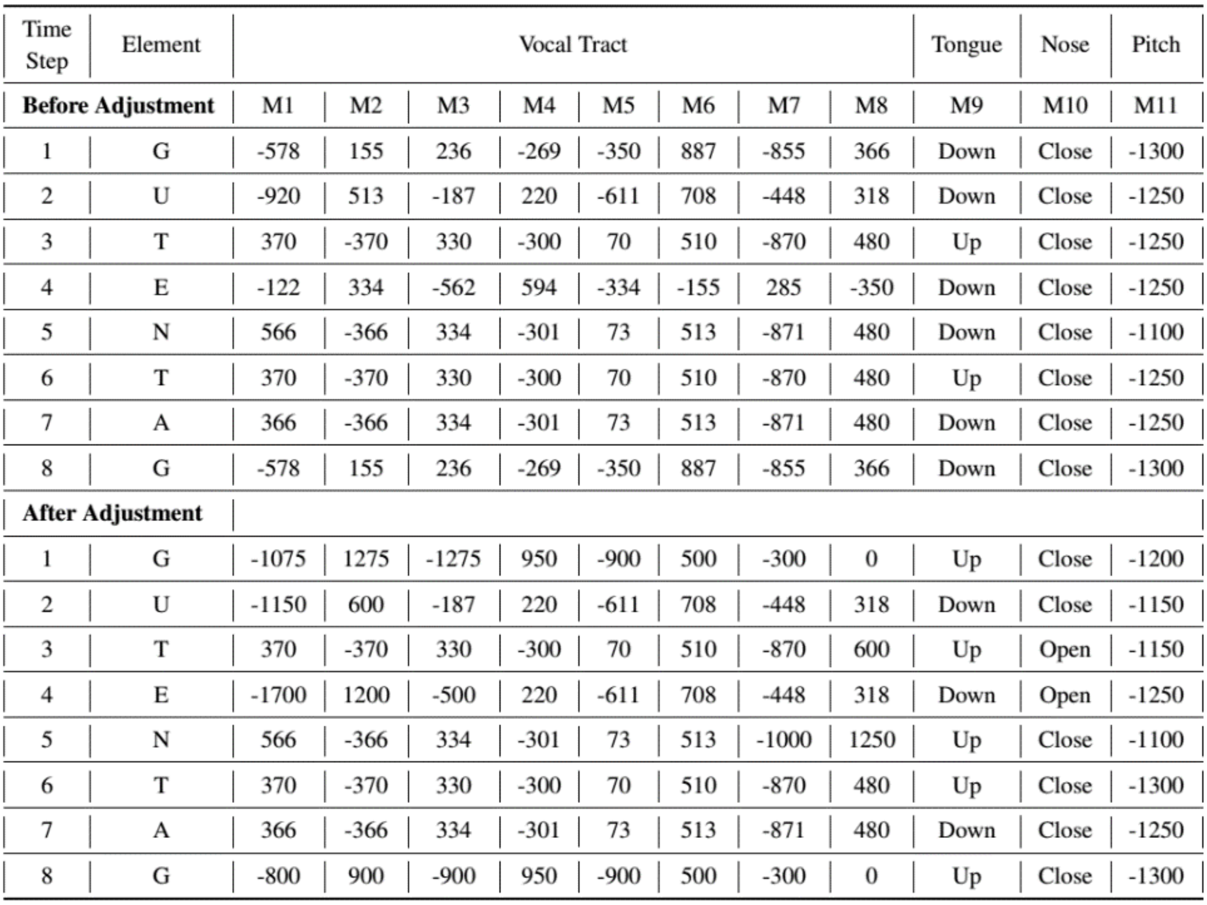


Figure 4.9: Sound wave*/Gutten-tag/* and its formant frequencies from human (A), robot before (B), and after (C) adjustment

Comparing Figure 4.9 (A) and (B), the author observed the 2nd formant frequency of the human voice for */Gutten Tag/* was about 500 Hz above the one spoken by the robot. The value of the lip-end motors group was adjusted to increase the formant frequencies of the output voice to roughly 500 Hz for the */Gut/* segment. For the second segment */Ten/*, the 2nd formant frequency was about the same, but the 1st formant frequency and 3rd formant frequency of the talking robot were respectively about 800 Hz and 500 Hz lower than the 1st formant frequency and 3rd formant of a human. Therefore, the motor values should be increased so the output formant frequency could match the human one. The detailed value of which motor groups should be changed also depends on the current shape of the robot vocal tract.

Table 4.1: Motor parameters before and after adjustment



Time Step	Element	Vocal Tract								Tongue	Nose	Pitch
Before Adjustment		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	G	-578	155	236	-269	-350	887	-855	366	Down	Close	-1300
2	U	-920	513	-187	220	-611	708	-448	318	Down	Close	-1250
3	T	370	-370	330	-300	70	510	-870	480	Up	Close	-1250
4	E	-122	334	-562	594	-334	-155	285	-350	Down	Close	-1250
5	N	566	-366	334	-301	73	513	-871	480	Down	Close	-1100
6	T	370	-370	330	-300	70	510	-870	480	Up	Close	-1250
7	A	366	-366	334	-301	73	513	-871	480	Down	Close	-1250
8	G	-578	155	236	-269	-350	887	-855	366	Down	Close	-1300
After Adjustment												
1	G	-1075	1275	-1275	950	-900	500	-300	0	Up	Close	-1200
2	U	-1150	600	-187	220	-611	708	-448	318	Down	Close	-1150
3	T	370	-370	330	-300	70	510	-870	600	Up	Open	-1150
4	E	-1700	1200	-500	220	-611	708	-448	318	Down	Open	-1250
5	N	566	-366	334	-301	73	513	-1000	1250	Up	Close	-1100
6	T	370	-370	330	-300	70	510	-870	480	Up	Close	-1300
7	A	366	-366	334	-301	73	513	-871	480	Down	Close	-1250
8	G	-800	900	-900	950	-900	500	-300	0	Up	Close	-1300

The adjusted motor values are shown in Table 4.1; M1 to M11 represent motor #1 to motor #11. For simplicity, the motor control unsound voice input is ignored in this experiment. Almost all the motors were adjusted to reproduce a recognizable */Gutten Tag/* sound. This was due to the non-linearity of the aerodynamics of continuous language speaking, which is different from word-by-word speaking.

To compare the output with the human voice before and after adjustment, the normalized Cross-correlation coefficients between human sound and robot sound were calculated and plotted as shown in Fig. Figure 4.10 (A) and (B). Normalized cross-correlation coefficients between two discrete signals *a(n)* and *b(n)* were calculated based on equation (4.12). If two signals are different in length (number of sampling), the shorter signal will have zero value added to have the same length with the longer one (zero padding). In equation (4.12), *m* is the sampling lag of two signals, a_{ave} and b_{ave} denote the mean of *a(n)* and *b(n),* respectively.

$$C_{(m)} = \frac{\sum[b[n]-b_{ave}][a[n-m]-a_{ave}]}{\sqrt{\sum[b[n]-b_{ave}]^2 \sum[a[n-m]-a_{ave}]^2}} \tag{4.12}$$

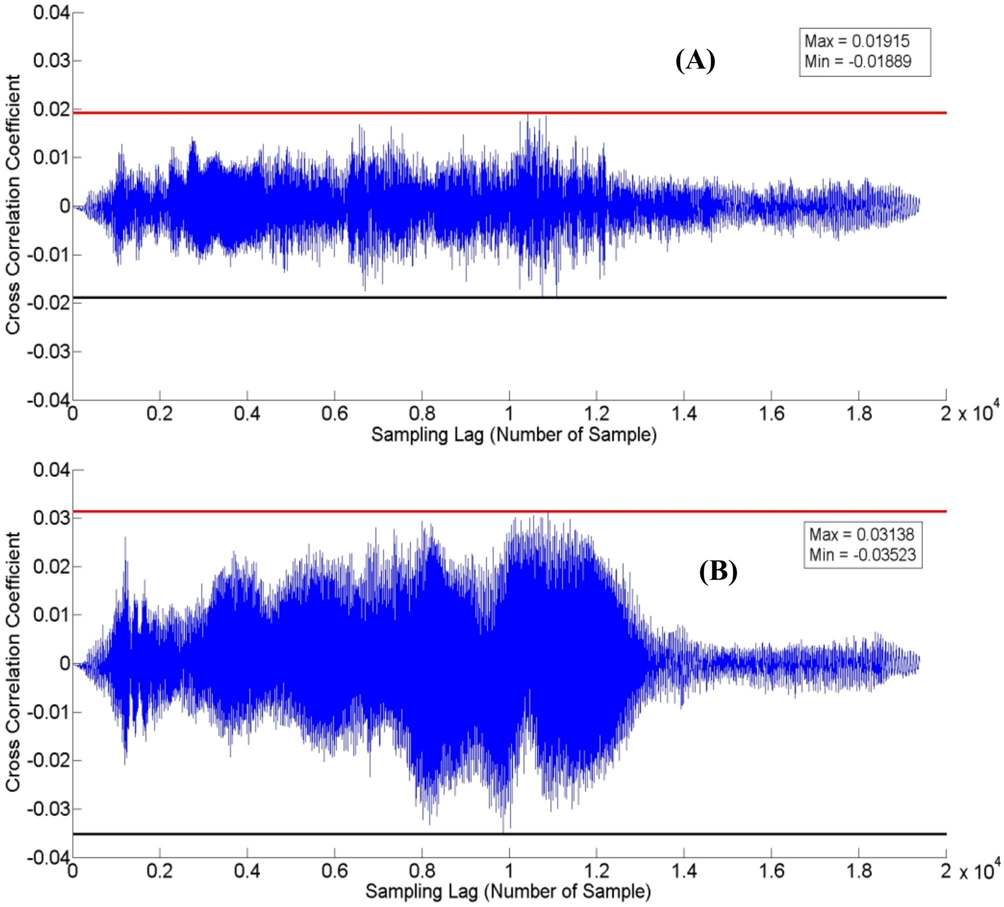


Figure 4.10: Cross-correlation between the human voice and robot voice for */Gutten Tag/* spoken before (A) and after adjustment (B).

In Figure 4.10, the horizontal axes indicate the sampling lag of the two signals, and on the vertical axes is the normalized cross-correlation coefficients. The maximum and minimum values of the cross-correlation coefficients between human sound and robot sound before adjustment were 0.01915 and -0.01889 while these values after adjustment were 0.03138 and -0.03523. The higher the maximum and minimum value, the more similarity between the two signals. The normalized cross-correlation coefficients, which is the similarity of the two signals, increased by more than 50% after adjustment indicating an intense increase of performance from a quick adjustment of the robot motors from the sliders on the GUI interface. The cross-correlation between human sound and robot sound after adjustment showed that after quick adjustment by the sliders, the robot sound performed better and was more recognizable to humans.

The main contribution of this study is to build a real-time interaction system between humans and a talking robot, which is an extremely useful system for setting the robot to speak different languages. In addition, based on the acoustic resonance theory, a novel formula about the formant frequency change due to vocal tract motor movements is derived, and a strategy on tuning motors value with respect to formant frequency difference is proposed. A real-time interaction system, which allowed the user to change the robot motor values by GUI sliders based on formant frequency comparison strategy was built, tested, and proven to improve the output voice of the talking robot when reproducing foreign language. It is a quick technique to get a reasonable result.

Chapter 5

Sentence Reproduction System

As mentioned in the introduction, almost all of the mechanical speech synthesis systems did not have the algorithm that allows them to repeat a sentence. For better understanding about human speech mechanism, the prosodic features are important factors that cannot be ignored in the study process. Prosodic features such as intonation, rhythm, and tempo exist only in sentences or groups of word speech. Thus, developing a system that allows the robotic speaking system to repeat a sentence from human speech is very important.

Therefore, in this chapter, an automatic vowel sequence reproduction system for the talking robot is introduced. A sound analysis system is developed to record a sentence spoken by a human (mainly vowels sequence in the Japanese language) and then analyze that sentence to give the corrected command packet for the talking robot to repeat it. An algorithm based on the short-time energy method is developed to separate and to count sound phonemes. Then, several phoneme detection methods including the direct cross-correlation analysis, the linear predictive coding (LPC) association, the partial correlation (PARCOR) coefficients analysis, and the formant frequencies comparison are applied to detect the similar voices in the talking robot's database with the spoken voice. Combining the sound separation and counting result with the detection of the vowel in human speech, the talking robot can reproduce the similar vowel sequences spoken by a human. Finally, experiments to compare these techniques and verify the working behavior of the robot are performed [44], [45].

Most speech synthesis systems record and extract the features of the human input sound. The sound features are then compared with the sound feature in the database to determine the output sound for the system. The output sound is usually generated by a speaker. However, this kind of sound regeneration system can only work with single phoneme sound For a sentence reproduction or group of words mimicking; it is required to analyze the recorded sentence and separate the sentence into individual sound elements first before the phoneme recognition and identification process. The phoneme extraction in a sentence is usually done with the zero-crossing rate or neural network technique. However, in this study, the author only focuses on the sequence of vowels generation system; thus, a simple approach using short-time energy analysis is developed for phoneme extraction and counting in this system. The block diagram of the vowels sequence reproduction system is shown in Figure 5.1. The interface for controlling the system of the robot is shown in Figure 5.2.

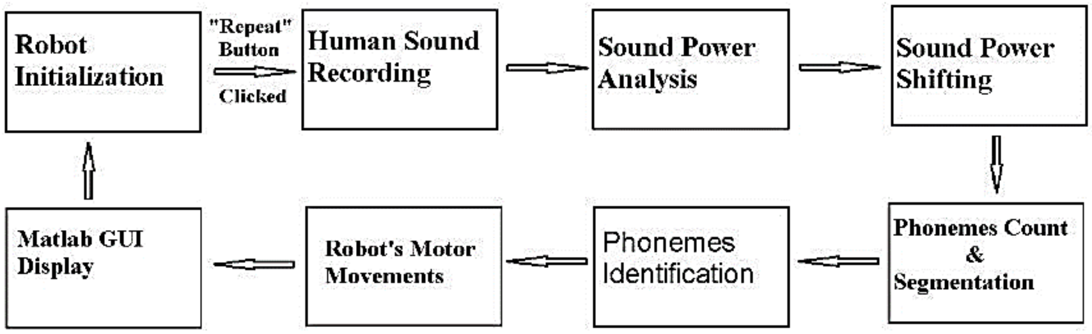


Figure 5.1: Program flow chart of vowels sequence reproduction

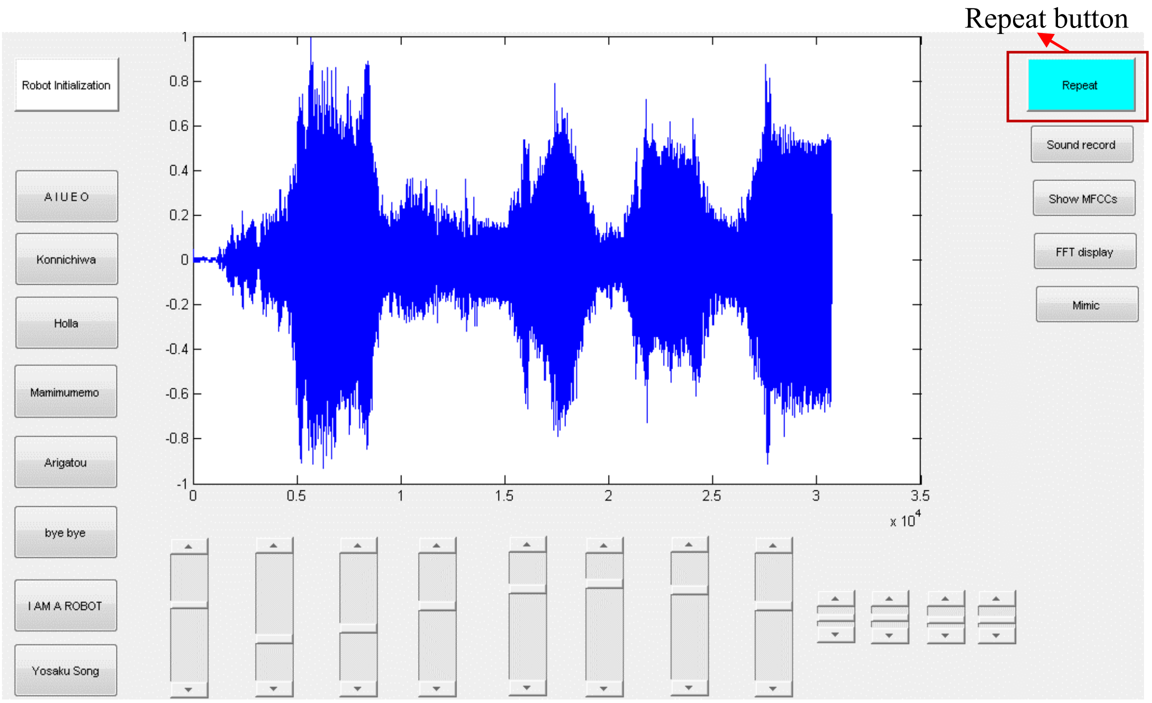


Figure 5.2: Robot sounds repeat program interface

The motor parameter vectors set of 5 vowels, which are */a/,/i/,/u/,/e/,/o/*, is used to output the command for the talking robot to regenerate respective vowels for a certain sound. As indicated in Figure 5.1, there are eight steps in the repeating algorithm. In the robot initialization step, all robot articulatory motors are set to the original position, and the communication with RS485 is opened to be ready to send the command to the articulation motors. The detail of other steps is explained below. For the sound recording module, the control interface of the talking robot is built using the Matlab Graphic User Interface (GUI) to record and display a human sound, and the talking robot reproduced sound as shown in Figure 5.2. The sound is recorded with a sampling rate of 8 kHz - single channel using built-in Matlab command; the sound wave data is then saved in Matlab workspace for further analysis.

5.1 Sound Power Analysis Module

5.1.1 Short-Time Energy (STE)

Human speech is produced from a time varying vocal tract system with time varying excitation. Due to this, the speech signal is naturally unstable. A speech signal is considered stationary when it occurs in a small duration frame of 10-30 milliseconds. Thus, it is possible to apply the signal processing tools of the stationary signal for estimating a short duration frame of a speech signal. This process is referred as short-term processing (STP). STP divides the input speech signal into short analysis segments that are isolated and processed with non-time varying properties. These short analysis segments called analysis frames always overlap one another. Short Term Processing of speech is performed both in the time domain and frequency domain.

Speech is time varying in nature. The energy associated with voiced speech is great when compared to unvoiced speech [49]. Silence speech will have the least or negligible energy when compared to unvoiced speech [50]. Hence, STE can be used for voiced, unvoiced and silence classification of speech.

Short Term Energy is derived from the following equation,

$$E_T = \sum_{m=-\infty}^{\infty} s^2(m) \tag{5.1}$$

Where, E_T is the total energy and $s(m)$ is the discrete time signal.

For STE computation, speech is considered in terms of short analysis frames whose size typically ranges from 10 to 30 milliseconds.

Consider the samples in a frame of speech as m=0 to m=N-1, where N is the length of frame, equation (5.1) can be written as

$$E_T = \sum_{m=-\infty}^{-1} s^2(m) + \sum_{m=0}^{N-1} s^2(m) + \sum_{m=N}^{\infty} s^2(m) \tag{5.2}$$

The speech sample is zero outside the frame length. Thus, equation (5.2) is written as,

$$E_T = \sum_{m=0}^{N-1} s^2(m) \tag{5.3}$$

The relation in equation (5.3) gives the total energy present in the frame of speech from m=0 to m=N-1. Short Term Energy is defined as the sum of squares of the samples in a frame, and it is given by,

$$e(n) = \sum_{m=-\infty}^{\infty}\left[s_{n(m)}\right]^2 \qquad (5.4)$$

After framing and windowing, the n^{th} frame speech becomes $s\,n\ (m) = s\,(m).\,w\,(n\ - m)$ and hence Short Term Energy in equation (5.4) is given by

$$e(n) = \sum_{m=-\infty}^{\infty}[s(m).w(n-m)]^2 \qquad (5.5)$$

Where, *e(n)* is the short-time energy, *s(m)* is the signal value of m, n is window duration n= 0,1T,2T,…,

T is the frame-shift (100 samples in this study).

w() is the Hamming window function.The Hamming window is represented by the following equation,

$$w(t) = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi t}{T}\right) & if\ \ |t| \le \frac{T}{2} \\ 0 & if\ \ |t| > \frac{T}{2} \end{cases} \qquad (5.6)$$

Or in frequency domain as

$$w(f) = \frac{0.54}{\pi(fT-1)}si\,n(\pi fT) + \frac{0.23T}{\pi(fT-1)}si\,n\,\pi(fT+1) + \frac{0.23}{\pi(fT-1)}si\,n\,\pi(fT-1) \qquad (5.7)$$

Figure 5.3 shows an example of the sound energy of a sequence of vowel calculated by STE method in time domain.

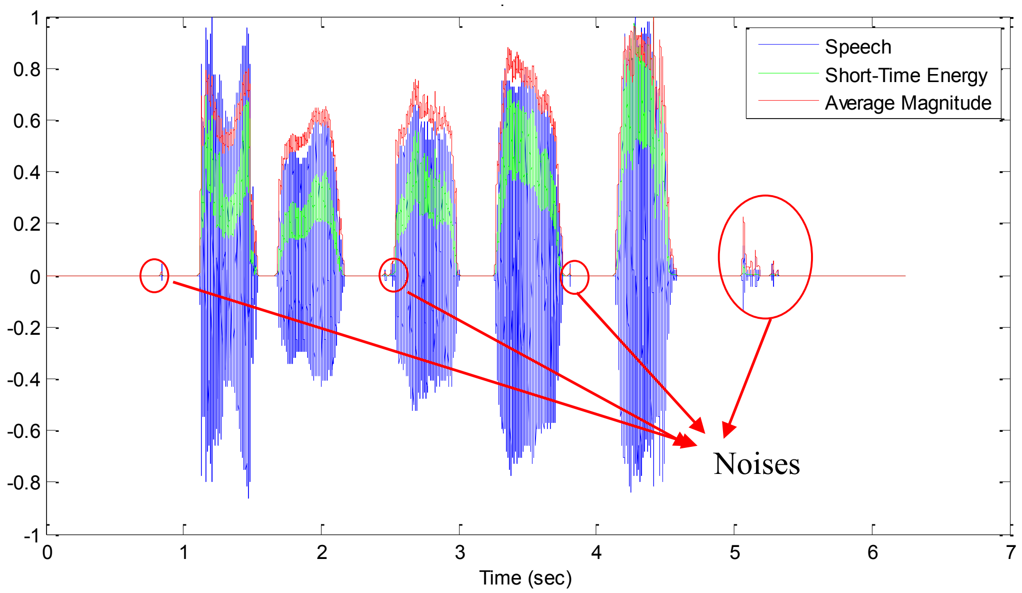


Figure 5.3: Speech, sound energy (STE), and average magnitude of sequence of vowels

5.1.2 Short Term Zero Crossing Rate (STZCR)

Zero Crossing Rate (ZCR) is defined as the number of times the zero axes is crossed per frame. If the number of zero crossings is high for a given signal, then the signal is changing quickly and accordingly the signal may contain high-frequency information which is indicated as unvoiced speech [49]. In contrast, if the number of zero crossings is low, then the signal is changing slowly, and consequently, the signal may contain low-frequency information which is indicated as voiced speech. Thus, ZCR gives indirect information about the frequency content of the signal.

Similarly to STE, STZCR is computed using a typical frame size of 10-30msec with half the frame size as frame shift. STZCR is defined as the weighted average of number of times the speech signal changes sign within the time window [49], and it is given by

$$Z(n) = \sum_{m=-\infty}^{\infty} |sgn[(s(m)] - sgn[s(m-1)]|.\ w(n-m) \tag{5.8}$$

where

$$sgn[s(m)] = \begin{cases} -1 & ,\ s(m) < 0 \\ 1 & ,\ s(m) \geq 0 \end{cases} \tag{5.9}$$

and

$$w(n) = f(x) = \begin{cases} \frac{1}{2N} & ,\ 0 \leq n \leq N-1 \\ 0 & ,\quad otherwise \end{cases} \tag{5.10}$$

If the value of samples $s(m)$ and $s(m-1)$ have different algebraic signs then, $\left|sgn[s(m)] - sgn[s(m-1)\ ]\right|$ in equation 5.8 becomes 1; and STZCR is counted. If the value of samples samples $s(m)$ and $s(m-1)$ have same algebraic signs, then the value of $\left|sgn[s(m)]-sgn[s(m-1)]\right|$ is 0 and STZCR is not counted. The values of STZCR of voice and unvoiced speech are different. Therefore, information from STZCR is used for discriminating voiced and unvoiced speech.

5.2 Filtering Module

The recorded sound is first amplified and filtered by an amplifier/low-pass filter (Entropy FLT-02) with the gain of 40 dB and the cut-off frequency of 4.05 kHz. However, there are still noises within the recorded signal as shown in Figure 5.3.; hence, the noises are also processed by the STE. To filter out these noisy parts in the STE signal, the author applies a simple approach which letting the STE signal compares with a constant θ_{th}, which is equal to the maximum value of the calculated short-time energy divided by 10 as shown in (5.11). This constant θ_{th} is called threshold number. Figure 5.4 shows an example of sound energy calculation and the threshold line. Several experiments to investigate the effect of the threshold number's value have been done. For the vowel separation part, the value of threshold number θ_{th} equaled to the maximum value of the calculated short-time energy divided by 10, is sufficient for separating the phonemes in human speech. The phoneme separation step is used to separate each individual phoneme and detect how many phonemes there are in a speech [42].

$$\theta_{th} = \frac{E_{\max}(m)}{10} \tag{5.11}$$

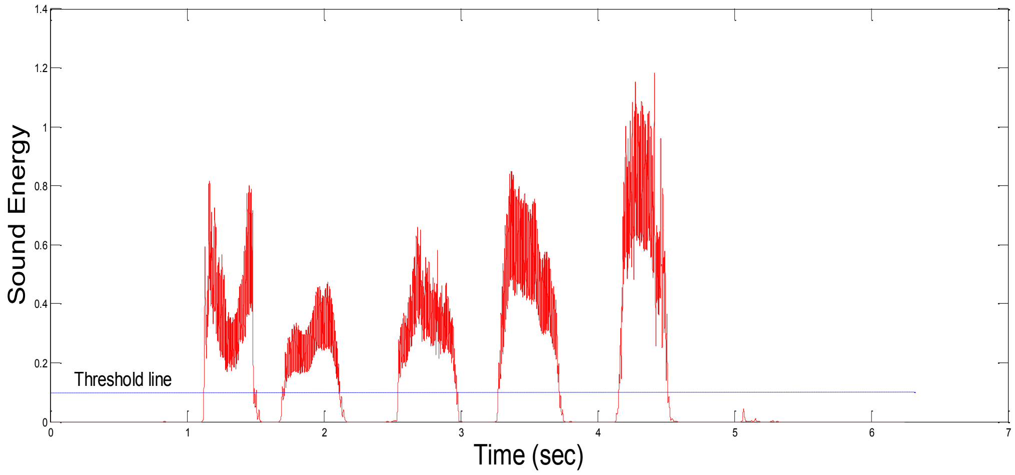


Figure 5.4: Sound energy (STE) of sequence of vowels

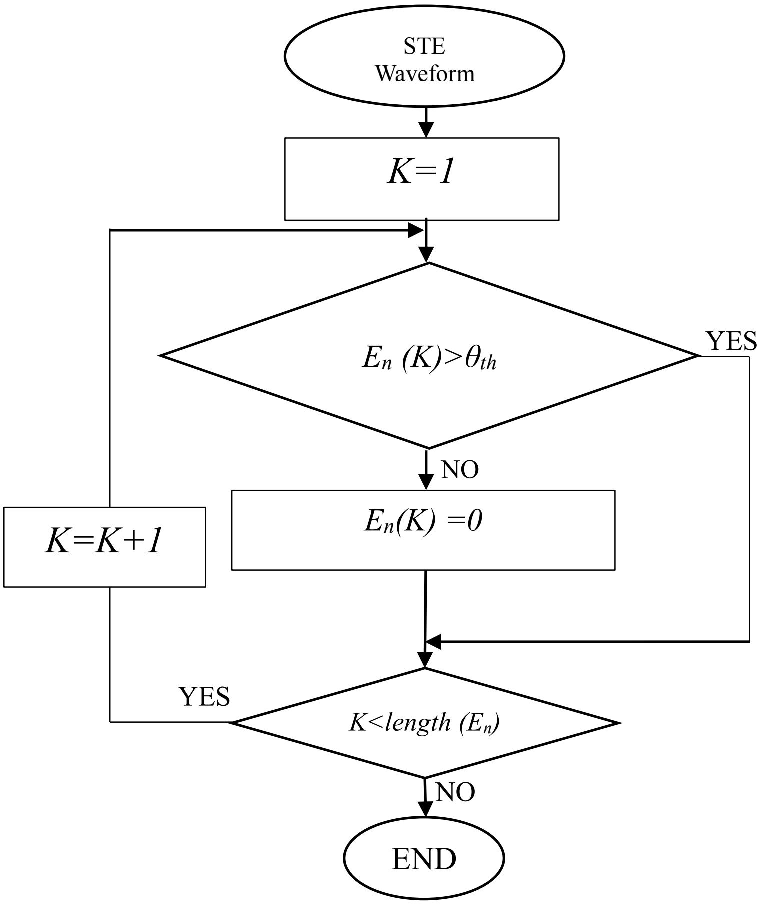


Figure 5.5: Flowchart of filtering out STE noise signal

Noise cancellation is also implemented in the system; if the length of the segment is too short, it will be noise so the system should ignore that segment. The separated phonemes are saved to the Matlab workspace for vowel detection analysis using four different acoustic features mentioned above. The flow chart of the algorithm for filtering out the noise and counting the number of phonemes using the STE waveform is shown in Figure 5.5.

The filtering for the STE signal is done by first comparing each element in the dataX(m) with the threshold number θ_{th}; if it is greater than the threshold number θ_{th}, the value of that element is kept; otherwise, its value is replaced by zero.

5.3 Separation and Counting Module

After that, the number of phonemes is counted, and the phonemes are separated by using tracing technique. The algorithm traces data X(m) from left to right, and it determines how many phonemes in a speech based on the number of times the data changes from zero value to non-zero value.

One phoneme segment is the range of X(m) that has a non-zero value.

The algorithm traces data $E_f(m)$ from left to right, and it determines how many phonemes in a speech based on the number of times the data changes from zero value to non-zero value. One phoneme segment is the range $E_f(m)$ that has a non-zero value. Noise cancellation is also implemented in the system; if the length of the segment is too short, it will be noise so the system should ignore that segment. The separated phonemes are saved to the Matlab workspace for vowel detection analysis using four different acoustic features mentioned above[42]. The flowchart for phoneme count and separation is shown in Figure 5.6 below.

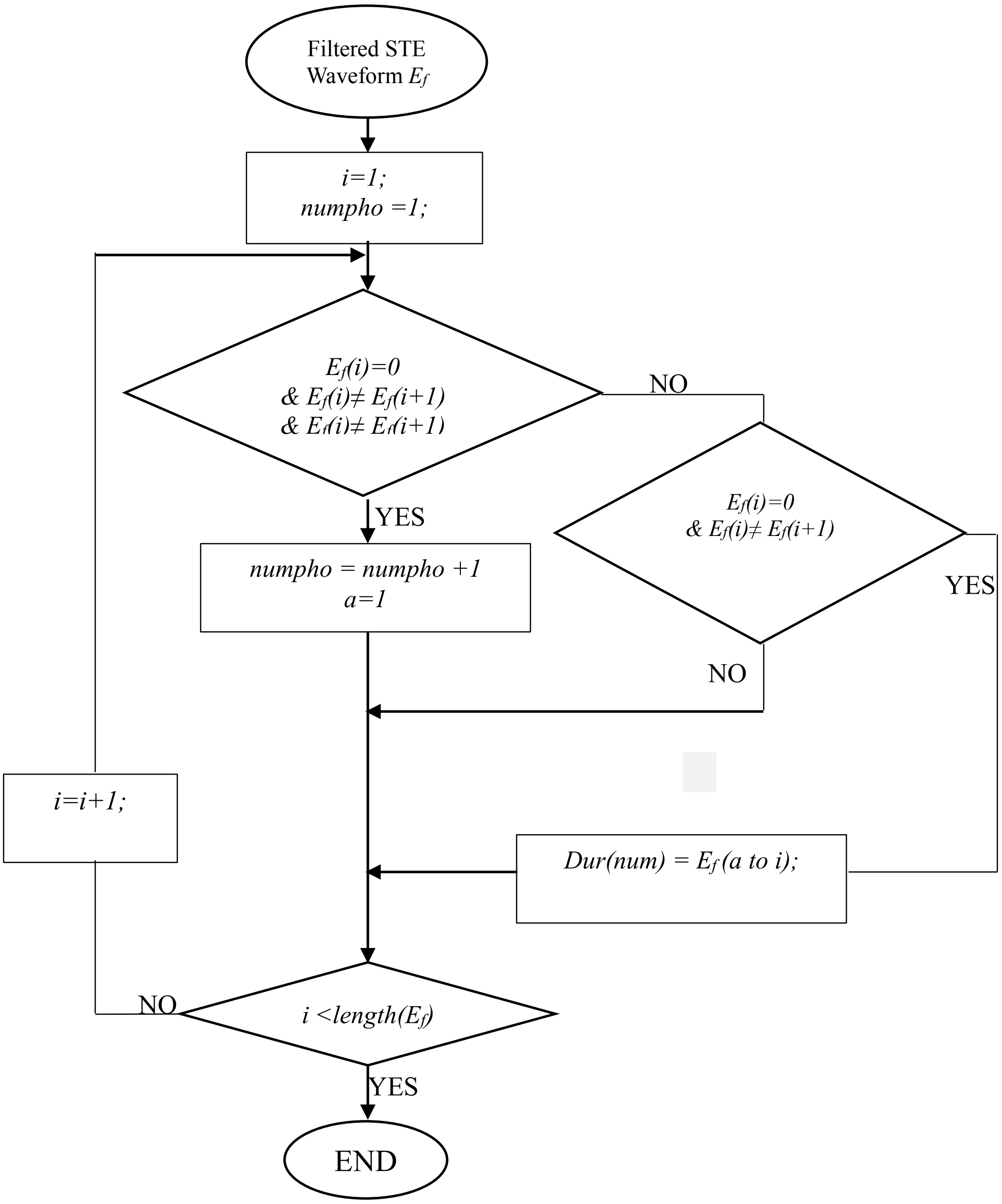


Figure 5.6: Flowchart of phoneme count and separation

Figure 5.7 shows an illustration of this phoneme separation and counting technique. Firstly, the sound of a human, which contain a sequence of vowels, is recorded (Figure 5.7(A)). Next, the short-time energy of the recorded voice is calculated as given in Figure 5.7 (B),. The filtering step, which is a comparison with the threshold number, is followed. The noise is eliminated after the filtering step as shown in Figure 5.7 (C),. Then, the phonemes count and the segmentation algorithm are applied. As a result, the number of phonemes and the new phonemes are obtained and saved in the Matlab workspace. This information is used as a part of an input data for the talking robot to regenerate a sequence of vowels spoken by a human.

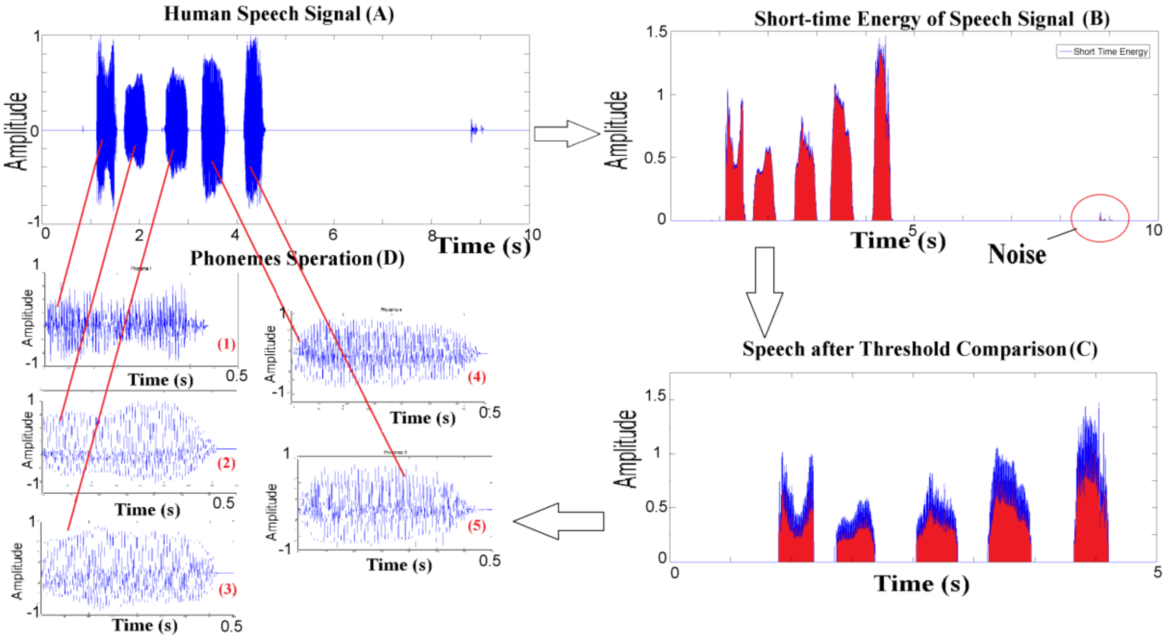


Figure 5.7: Signal flow for phoneme count and separation

5.4 Features Extraction Module

This module was developed to extract the acoustic features out of the recorded sound for the comparison with the pre-recorded sounds in the database. The acoustic features of vowels are usually extracted from a small segment of a signal, which is ranged about 50ms of the waveform. There are five pre-recorded sounds for each vowel spoken by a male adult. The central segment of each signal length of 50ms is taken to perform acoustic features extraction. In this study, the LPC coefficients, the first three formant frequencies, and the PARCOR coefficients are extracted from the recorded sounds. The average value of these acoustic features for each vowel are calculated and saved in robot database as the standard set for phoneme detection algorithm [42].

5.4.1 Cross-correlation Method

The simplest way to identify a sound element is to directly compare its signal with the vowel sound signals in the database. The highest similarity score determines the vowel sound output for a specific sound element. Thus, the author uses cross-correlation coefficient in this case. Cross-correlation coefficient *C(m)* of two discrete signals *a(n)* and *b(n),* which measures the similarity between them, is calculated using equation (5.12) with *m* is the sampling lag of two signal, and *b** denotes the complex conjugate of signal *b*. If the two signals are different in length (number of sampling), the shorter one will have zero value added to have the same length with the other. The technique to add a zero number to a signal to get the same length as the other signal is referred as zero padding.

$$C(m) = \sum_{n=-\infty}^{\infty} b^{*}[n]a[n+m] \tag{5.12}$$

5.4.2 LPC Coefficients

LPC is one of the widely used methods in the sound recognition process that attempts to predict the feature values of the input signal based on past signals. There are several techniques to solve the LPC coefficients such as covariance method, autocorrelation method, and lattice method [38], [39]. For this system, the autocorrelation method equations 4.1 to 4.8 are applied to solve 12th order LPC coefficients. The LPC coefficients of five basic Japanese vowels are plotted in Figure. 5.7.

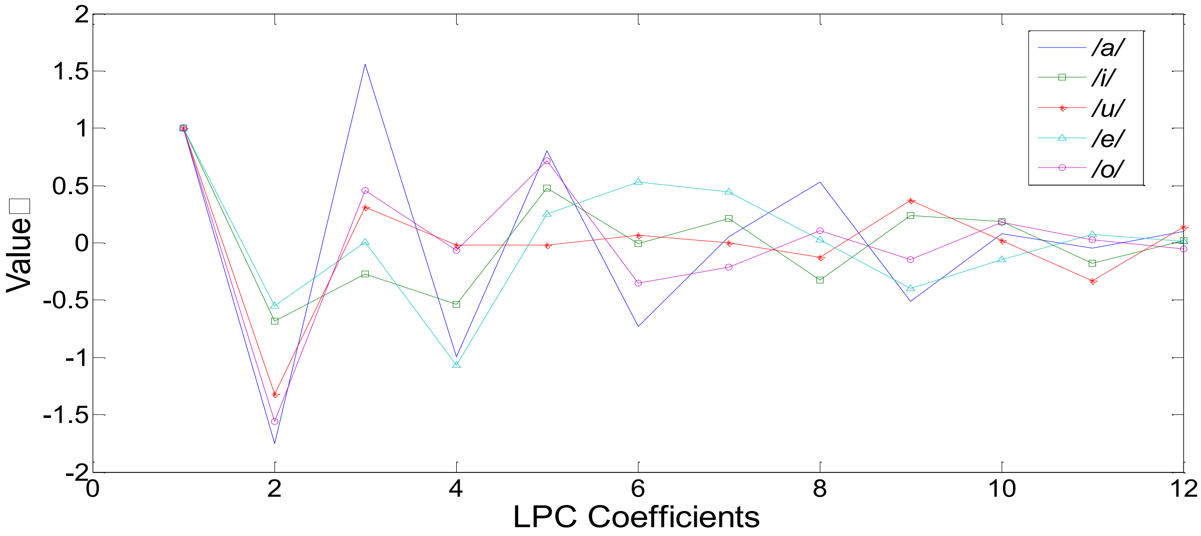


Figure 5.7: Human LPC coefficients of 5 Japanese vowels

5.4.3 Formant Frequency

Formants are the spectral peaks in the spectrum. Each formant frequencies set corresponding to certain sounds. The formant frequencies comparison technique has been widely used for vowels synthesis system for its similarity to the way human auditory system work [1]. The extraction of formant frequencies previously introduced in section 4.2.1. With this technique, the author extracted the formant frequency of five Japanese vowels from the speech of five different Japanese male adults. The average of each frequency was calculated by using equation 5.13, and the results are shown in Table 5.1. This table is used as the database set for phoneme detection.

$$F_{i\_ave} = \frac{\sum_{i=1}^{N} F_i}{N} \tag{5.13}$$

Table 5.1Formant Frequencies of Five Japanese Vowels

Formant	**F1 (Hz)**	**F2 (Hz)**	**F3 (Hz)**
/a/	460	1060	2340
/i/	120	2040	2800
/u/	180	860	2100
/e/	360	1600	2160
/o/	280	600	2100

5.4.4 PARCOR

The resonance filter can be realized by PARCOR (Partial autocorrelation) coefficients, which is a method that uses the vocal tract area function to predict the robot vocal tract shape [51]-[53]. PARCOR coefficients k_n of a waveform sample of $[x_t, x_{t-1}, ..., x_{t-n}]$ with a forward prediction error of $e^{(n-1)}{}_{ft}$, and a backward prediction error of $e^{(n-1)}{}_{bt}$ is defined in equation 5.14.

$$k_n = \frac{E\left\{e_{ft}^{(n-1)} e_{bt}^{(n-1)}\right\}}{\left[\mathrm{E}\left\{\left(e_{ft}^{(n-1)}\right)^2\right\}\mathrm{E}\left\{\left(e_{bt}^{(n-1)}\right)^2\right\}\right]^2} \tag{5.14}$$

here,

$$e_{ft}^{(n-1)} = x_t - \widehat{x_t} = x_t + \sum_{i=1}^{n-1} \alpha_i^{(n-1)} x_{t-i} \tag{5.15}$$

$$e_{bt}^{(n-1)} = x_{t-n} - \widehat{x_{t-n}} = x_{t-n} + \sum_{i=1}^{n-1} \beta_i^{(n-1)} x_{t-i} \tag{5.16}$$

PARCOR coefficients k_n can be determined by the least mean squares of x_t and x_{t-n} for forward and backward error prediction. Thus, α_I and β_t have the following relationships to reflection coefficients r_i.

$$\sum_{i=1}^{n-1} \alpha_i^{(n-1)} r_{j-i} = -r_i \qquad (i = 1,2,\cdots,n-1) \tag{5.17}$$

$$\sum_{i=1}^{n-1} \beta_j^{(n-1)} r_{j-i} = -r_{n-i} \quad (i = 1,2,\cdots,n-1) \tag{5.18}$$

With,

$$r_{j-i} = E\{x_{t-j}x_{t-i}\} = r_{i-j} \tag{5.19}$$

Hence,

$$\beta_i^{(n-1)} = \alpha_{n-i}^{(n-1)} \qquad (i = 1,2,\cdots,n-1) \tag{5.20}$$

With $\alpha^{(n-1)}{}_0$=1、$\beta^{(n-1)}{}_n$=1, (5.16) and (5.17) become

$$\sum_{i=1}^{n-1} \alpha_i^{(n-1)} r_{j-i} = 0 \qquad (i = 1,2,\cdots,n-1) \tag{5.21}$$

$$\sum_{j=1}^{n} \beta_j^{(n-1)} r_{j-i} = 0 \quad (i = 1,2,\cdots,n-1) \tag{5.22}$$

Equation 5.14 can be rewritten as equation 5.23

$$k_n = \frac{W_{n-1}}{U_{n-1}} \tag{5.23}$$

Here,

$$U_{n-1} = \left[\mathrm{E}\left\{\left(e_{ft}^{(n-1)}\right)^2\right\}\mathrm{E}\left\{\left(e_{bt}^{(n-1)}\right)^2\right\}\right]^{1/2} = \mathrm{E}\left\{\left(e_{ft}^{(n-1)}\right)^2\right\}$$
$$= \sum_{i=0}^{n-1} \alpha_i^{(n-1)} r_i = 0 \qquad \left(\alpha_0^{(n-1)} = 1\right) \tag{5.24}$$

$$W_{n-1} = \mathrm{E}\left\{e_{ft}^{(n-1)} e_{bt}^{(n-1)}\right\} = \sum_{i=0}^{n-1} \alpha_i^{(n-1)} r_{n-1} \qquad \left(\alpha_0^{(n-1)} = 1\right) \tag{5.25}$$

The following recurrence formulas are obtained when using the relationship from (5.20), (5.21), and (5.22).

$$\alpha_i^{(n)} = \alpha_i^{(n-1)} - k_n \beta_i^{(n-1)} \qquad (i = 1,2,\cdots,n) \tag{5.26}$$

$$\beta_i^{(n)} = \beta_{i-1}^{(n-1)} - k_n \alpha_i^{(n-1)} \qquad (i = 1,2,\cdots,n) \tag{5.27}$$

$$U_{n-1} = U_{n-1}(1 - k_n^2) \tag{5.28}$$

From equations (5.20) and (5.28) and $\alpha^{(n-1)}{}_{n-1} = 0$、$\beta^{(n-1)}{}_0 = 0$, the following relationship is obtained.

$$\alpha_i^{(n)} = \alpha_i^{(n-1)} - k_n \alpha_{n-i}^{(n-1)} \tag{5.29}$$

k_n is sequentially obtained from $n = 1$ by calculating W_{n-1}. In addition, $\alpha_{i,}$, $i=1,2,\ldots,n$ is also calculated. In the case k_n is obtained directly from waveform x_t, the signal is firstly transformed to z-domain,

$$A_{(z)} = \textstyle\sum_{i=0}^{n} \alpha_i^{(n)} z^{-i} \tag{5.30}$$

$$B_{(z)} = \textstyle\sum_{i=1}^{n+1} \beta_i^{(n)} z^{-i} \tag{5.31}$$

here,

$$e_{ft}^{(n)} = A_n(z) x_t \tag{5.32}$$

$$e_{bt}^{(n)} = B_n(z) x_t \tag{5.33}$$

we have,

$$A_n(z) = A_{n-1}(z) - k_n B_{n-1}(z) \tag{5.34}$$

$$B_n(z) = z^{-1}[B_{n-1}(z) - k_n A_{n-1}(z)] \tag{5.35}$$

With, $A_0(z) = 1$, $B_0(z) = \mathrm{z}^{-1}$, equations 5.34 and 5.35 can be expressed as Figure 5.8 Levinson-Durbin recursive algorithm as described in section 4.2.1 is applied to solve for PARCOR coefficients.

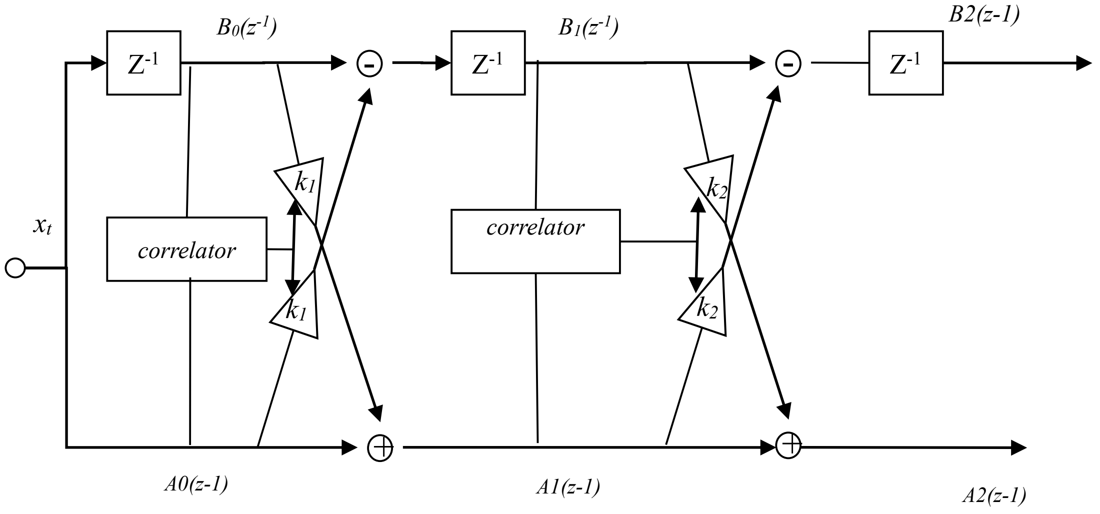


Figure 5.8: PARCOR lattice analysis model.

The average value of the 12th order PARCOR coefficients of 5 vowels in 5 different Japanese male adult voices is plotted in Figure. 5.9.

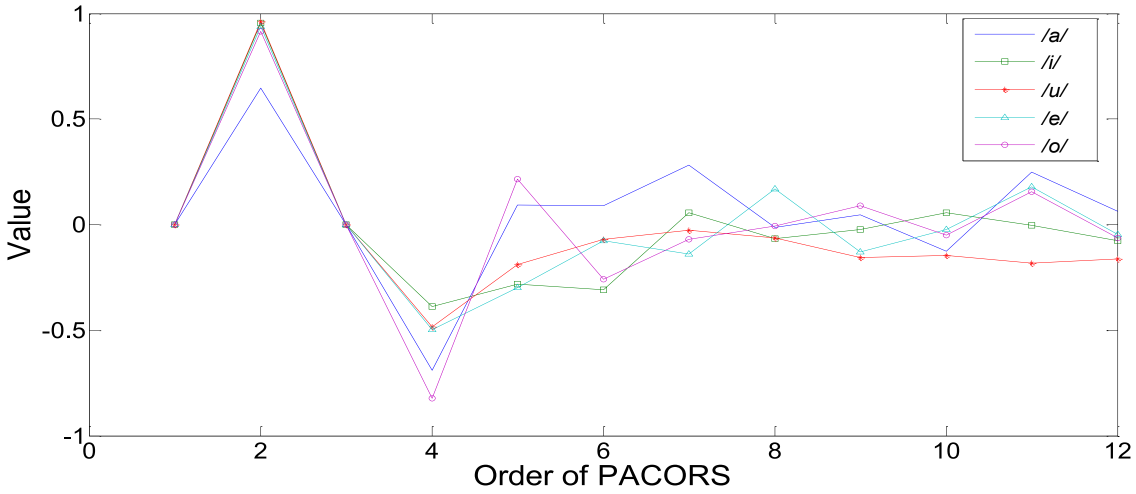


Figure 5.9: Human PARCORS coefficients

5.5 Sound Template Matching

Previously, the sets of standard acoustic features of 5 Japanese vowels, which are the LPC coefficients set, the PARCOR coefficients set, and the formant frequencies set, were established (see section 5.3). These sets are used for comparison with new human recorded sounds to determine the output speech of the robot. The sums of the difference of each acoustic feature coefficient between each vowel in the standard set of 5 Japanese and the new phoneme are calculated. The minimum difference between the new sound element coefficients and standard coefficients in the robot database indicates which vowel sound for that sound element as indicated in equation 5.36. The motor command output for each sound element in sequential order is then determined by the phoneme count and separation algorithm described above [42], [44]. The output voice of the robot will be recorded and displayed on the computer screen.

$$output\_vowel = argmin_t \|m_i - d_i\| \tag{5.36}$$

5.6 Sequence of Vowel Reproduction Experiment

Eight people, seven males and one female, were subjected to the experiment. Among them, there were four Japanese and four foreigners (French, Chinese, and Vietnamese). These people were requested to speak 5 Japanese vowels of */a/,/i/,/u/,/e/,/o* at a time for five times in a random order of their choice, and they needed to pause slightly between each vowel. Their sounds are recorded for the analysis with four different tests. The first test is for the direct cross-correlation method, the second test is for the LPC comparison, the third test is for the PARCOR analysis, and the last test is for the formant frequencies comparison.

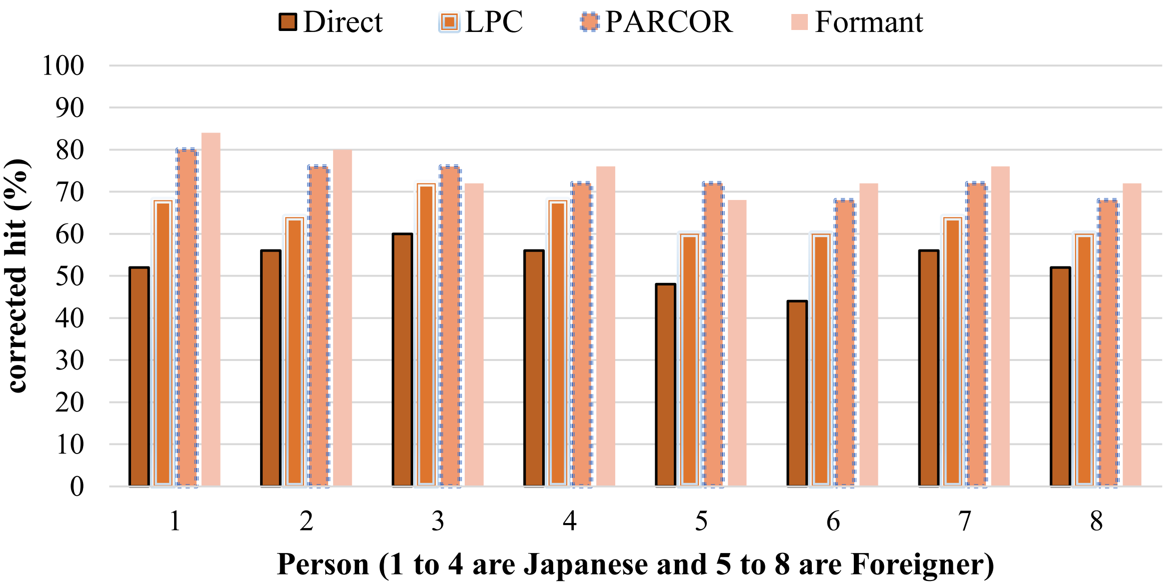


Figure 5.10: Result of reproduction of the talking robot

The average percentage of corrected vowel reproduction by the robot was then calculated. For example, if the robot can repeat 4 out of 5 vowels in one trial, it would be counted as 80% corrected hit for that trial. Then, the talking robot’s average percentage of corrected vowel reproduction for each person's sound was calculated. The result of the average corrected hits of all the tests is plotted in Fig. 10,.

The result indicated that the talking robot could regenerate a sequence of vowels with a satisfactory result when applying the PARCOR analysis and the formant frequencies comparison. However, the direct cross-correlation analysis was not working so well with around 40% to 50% accuracy. Using LPC comparison method, we had more than 60% accuracy. The PARCOR analysis method and formant frequencies comparison method delivered more than 70% accuracy, and the formant frequencies comparison method gave a little higher accuracy on average. The rate of recognizing Japanese voices was higher than for foreigner's voices since the set of vowels in the robot database was built from Japanese voices. Besides, the similarity in the characteristic of vowel */i/&/e/* and */u/&/o/* made it difficult for the robot to distinguish between these vowels, and errors usually occurred when the robot was reproducing these vowels.

The novel technique of phonemes segmentation and counting worked almost perfectly for the sequence of vowels reproduction since the people in the test were required to pause a little between vowels. The direct cross-correlation technique gave the least accurate result because a human could speak the same vowels with different sound pitch and amplitude. These sound characteristics affected the similarity score when applying cross-correlation. The LPC technique also had the effect of sound pitch and amplitude. However, since it only took a short range of sound signal for extracting the coefficients, the effect was reduced and gave better accuracy than the direct cross-correlation technique. The PARCOR analysis and the formant frequencies comparison delivered the highest accuracy since their coefficients were extracted from resonance frequencies, which only depended on the vocal tract shape. Also, a human ear distinguishes sound by detecting its resonance frequencies, which is similar to the formant frequencies detection. Therefore, using formant frequencies comparison was the most similar process to the human auditory mechanism, and it was applied by numerous speech synthesis researchers. The high accuracy outcome of applying formant frequency comparison techniques for vowel identification was verified in this study.

Chapter 6

Text-to-Speech System

Text-to-speech (TTS) is the process to transform text into sound waveforms. Most TTS synthesis systems use speakers to generate sound waveforms, and these systems depend heavily on the database of the text associated with the sound output. Our TTS system is built with a different approach, which doesn't require a big database like other systems. There are usually three parts of a phoneme, which are the consonant part, vowel part, and transition part. The consonant parts take about the first 50ms of each phoneme's waveform, followed by the transient part of roughly 50ms. The rest of each phoneme's waveform are vowel parts. Figure 6.1 shows an example of */konichiwa/* sound of the human with the timing of each part of the phonemes. Therefore, the author develops the TTS system for the talking robot by combining the consonant and vowel to make a speech. Each vowel or consonant of alphabetic character is represented by a set of the motor vector. The author also employs a tonal effect for this TTS system, which is one element of a motor vector [42].

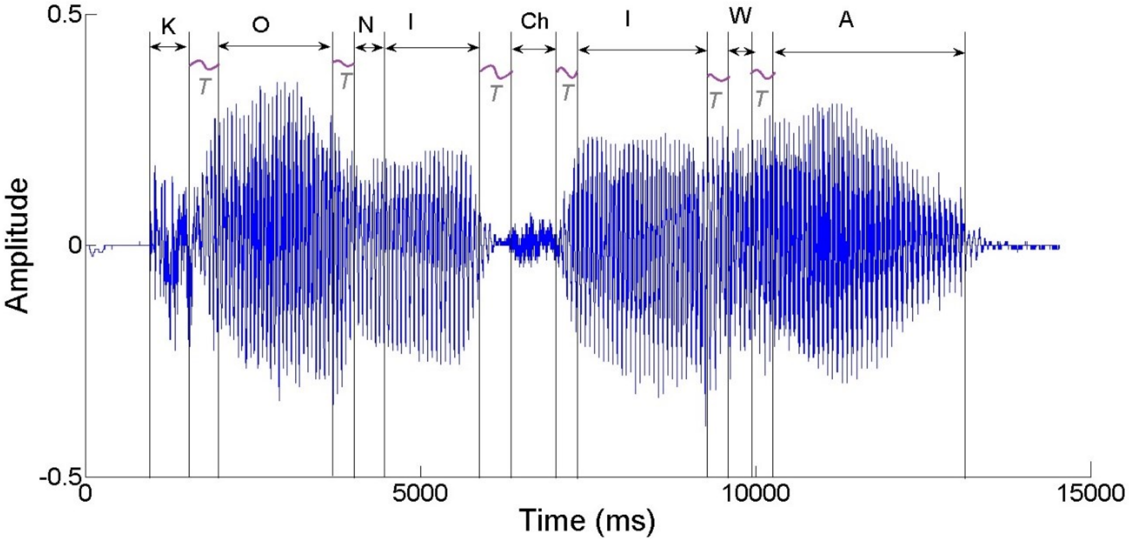


Figure. 6.1: Human */ko-ni-chi-wa/* sound wave and the timing for each part

6.1 Mapping from Alphabet Character to Robot Parameters

Based on the observations of human voice generation as introduced in Section 1.2.3, the author constructed a full mapping of vowels and consonance sound corresponding to all the English alphabet characters. This step is significant in developing the TTS system for the talking robot. Vowel sounds are generated by the static articulation of the vocal apparatus. Consonant sounds, on the other hand, are vocalized by the dynamic motions of the apparatus. Therefore the generation of consonant sounds requires a dynamic control of the mechanical system. The vowels, liquids, and fricatives generation were previously introduced in section 3. In this chapter, the mechanism of the talking robot for nasal and plosive sounds is presented.

6.1.1 Nasal Sounds of Mechanical System

For the generation of the nasal sounds */n/* and */m/,* the sliding valve is open to lead the air into the nasal cavity as shown in Figure 6.2. By closing the middle position of the vocal tract and then releasing the air to speak vowel sounds, the */n/* consonant is generated. For the */m/* consonant, the outlet part is closed to stop the air first, and then opened to vocalize vowels. The difference in the */n/* and */m/* consonants generation is basically the narrowing positions of the vocal tract. Sound spectra of nasal sounds are characterized as the first formant exists around 300 Hz, and spectrum power decreases sharply in the high-frequency range. The mechanical model's spectrum is similar to the human's and satisfies the characteristics of nasal sound.

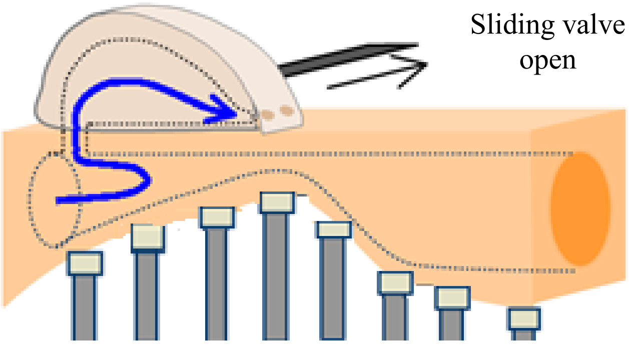


Figure. 6.2: Mechanism of nasal sound

6.1.2 Plosive Sounds of Mechanical System

In generating the plosive sounds */p/* and */t/*, the mechanical system closes the sliding valve so as not to release the air in the nasal cavity. By closing one point of the vocal tract, air provided from the lung is stopped and compressed in the tract as shown in Fig.6.3. Then the released air generates plosive consonant sounds like */p/* and */t/*. For example, the plosive sound */pa/* is vocalized by combing the dynamic motions for the plosive sound */p/* and the vowel sound */a/*. The sound wave of a plosive sound has impulse at the beginning, followed by the stable vowel sound waves.

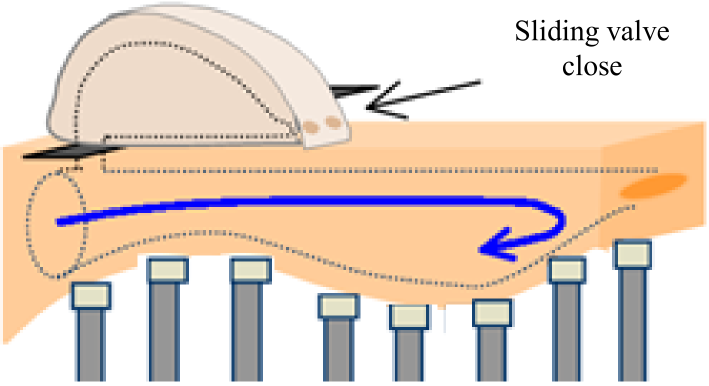


Figure. 6.3: Mechanism of plosive sound

6.1.3 Robot Parameters

Robot vector X is presented as following

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ x_8 \\ x_{tg} \\ x_{no} \\ x_{un} \\ x_{vl} \\ x_{pt} \\ x_t \end{pmatrix} \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \\ \end{array}\right\} \textbf{Vocal tract parameters} \\ \\ \\ \\ \\ \\ \\ \end{array} \tag{6.1}$$

One robot vector has 14 parameters as shown in equation 6.1.

- The first eight parameters, x_1 to x_8, represent the vocal tract shape. The value of vocal tract shape parameters ranges from -1500 to 1500, which respectively represent for close and widest open of the cross-sectional areas of the vocal tract.
- Parameter x_{tg} represents the tongue position, and it has a value of up or down indicating the position of the tongue.
- Parameter x_{no} represents nasal cavity chamber control which has a value of open or closed.
- Parameter x_{un} is for unvoiced sound pitch control; its value is either yes or no indicating the air pathway to the nasal cavity chamber is either opened or closed.
- Parameter x_{vl} controls the sound volume, and it is the amount of the airflow, which has 5 levels of low, mid-low, mid, mid-high, and high, to the vocal cords.
- Parameter x_{pt}, which is manipulated by the vocal cords motor value as introduced in section 3.2.1, controls the fundamental frequency of the sound sources.
- Parameter x_t indicates the duration of vocalization for the talking robot. This value varies between vowels and consonants. For simplicity, the author set the duration to 500 milliseconds for vowels and 50 milliseconds for consonants.

The value of x_{vo} and x_{pt} are determined by intonation input by a user via GUI interface. The relationship between sound volume and fundamental frequency was presented in Figure 3.20 in section 3.21. For this TTS system, the author divides the intonation input into 9 levels. The tonal effect was previously set at middle tone at 5th level. If the user does not input the tonal effect after inputting the text, the intonation is kept at the 5th level.

The detail of nine intonation levels is shown in Table 6.1. The default intonation is highlighted with gray color. The robot parameters corresponding to alphabetic characters is shown in Table 6.2.

Table 6.1: Intonation effect parameters

Intonation level input	**Sound volume (x_{vo})**	**Pitch (x_{pt})** **(Vocal cords motor angle)**
1	Low	-100
2	Mid-low	-80
3	Mid-low	-50
4	Mid	-30
5	Mid	0
6	Mid-high	30
7	Mid-high	50
8	High	80
9	High	100

Table 6.2: Robot parameters corresponding to alphabetic characters

	M1	**M2**	**M3**	**M4**	**M5**	**M6**	**M7**	**M8**	**Tongue**	**Nose**	**Unvoiced**	**Duration (ms)**
a	366	-366	334	-301	73	513	-871	480	Down	Close	N	500
i	-741	920	-920	855	-562	171	122	-155	Down	Close	N	500
u	-920	513	-187	220	-611	708	-448	318	Down	Close	N	500
e	-122	334	-562	594	-334	-155	285	-350	Down	Close	N	500
o	-578	155	236	-269	-350	755	-755	366	Down	Close	N	500
b	370	-370	330	-300	70	510	-870	480	UP	Close	N	50
c	-500	630	-300	450	-570	750	-770	920	UP	Close	N	50
d	70	-370	330	-300	70	510	-870	480	UP	Close	N	50
f	-100	600	-700	200	-170	450	-770	920	Down	Close	Y	50
g	-578	155	236	-269	-350	887	-855	366	Down	Close	N	50
h	-370	-370	330	-300	70	510	-870	480	UP	Close	Y	50
j	-500	630	-300	450	-570	750	-770	920	UP	Close	N	50
k	-100	600	-700	200	-170	450	-770	920	Down	Close	N	50
l	0	0	0	0	0	0	0	0	UP	Close	N	50
m	1300	1000	-500	-20	-40	-70	180	-180	Down	Open	N	50
n	-700	540	-750	800	-660	520	-400	180	UP	Open	N	50
p	370	-370	330	-300	70	510	-870	480	UP	Close	N	50
q	-920	513	-187	220	-611	708	-448	318	Down	Close	N	50
r	0	0	0	0	0	0	0	0	UP	Close	N	50
s	-500	630	-300	450	-570	750	-770	920	UP	Close	Y	50
t	370	-370	330	-300	70	510	-870	480	UP	Close	N	50
v	-920	513	-187	220	-611	708	-448	318	Down	Close	Y	50
w	-800	900	-900	950	-900	500	-300	0	UP	Close	N	50
x	0	0	0	0	0	0	0	0	UP	Close	N	50
y	-920	513	-187	220	-611	708	-448	318	Down	Close	Y	50
z	-578	155	236	-269	-350	887	-855	366	Down	Close	Y	50
Stop	0	0	0	0	0	0	0	0	Down	Close	N	500

In table 6.2, the color highlighted for each alphabet character indicates its sounding characteristic. The yellow color highlight is for vowels, no color highlight is for liquid sound, the blue color highlight with bold characters is for plosive sound, the green highlight is for nasal sound, and the gray color highlight is for fricative sound.

6.2 TTS Algorithm for Robotic Speaking System

In general, TTS systems first convert the input text into its corresponding linguistic or phonetic representations and then produce the sounds corresponding to those representations. With the input being plain text, the generated phonetic representations also need to be augmented with information about the intonation and rhythm that the synthesized speech should have [54]. In this study, the text and intonation input are anal sized and transformed to robot parameters. The robot vocal motors move according to the robot parameters, and the sounds are generated.

The block diagram for the TTS system is shown in Figure 6.4. When the program is run, the robot is initialized, and all the motors move to the initial position. A text dialog box appears for a user to input text. After the text is input, another dialog box appears for the user to input intonation information. Figure 6.5 shows the picture of these dialog boxes. The users can ignore this step if they want the system to speak at default intonation level.

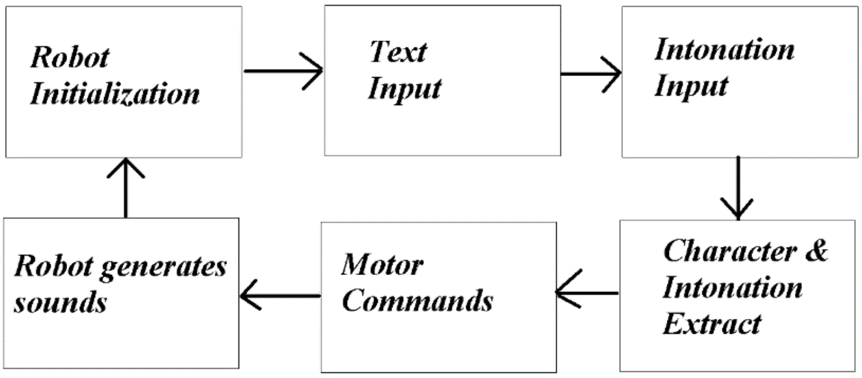
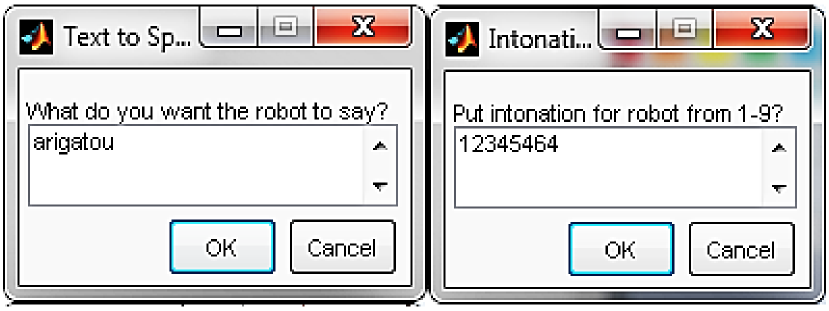


Figure. 6.4: Program outline



What do you want the robot to say?
arigatou

Put intonation for robot from 1-9?
12345464

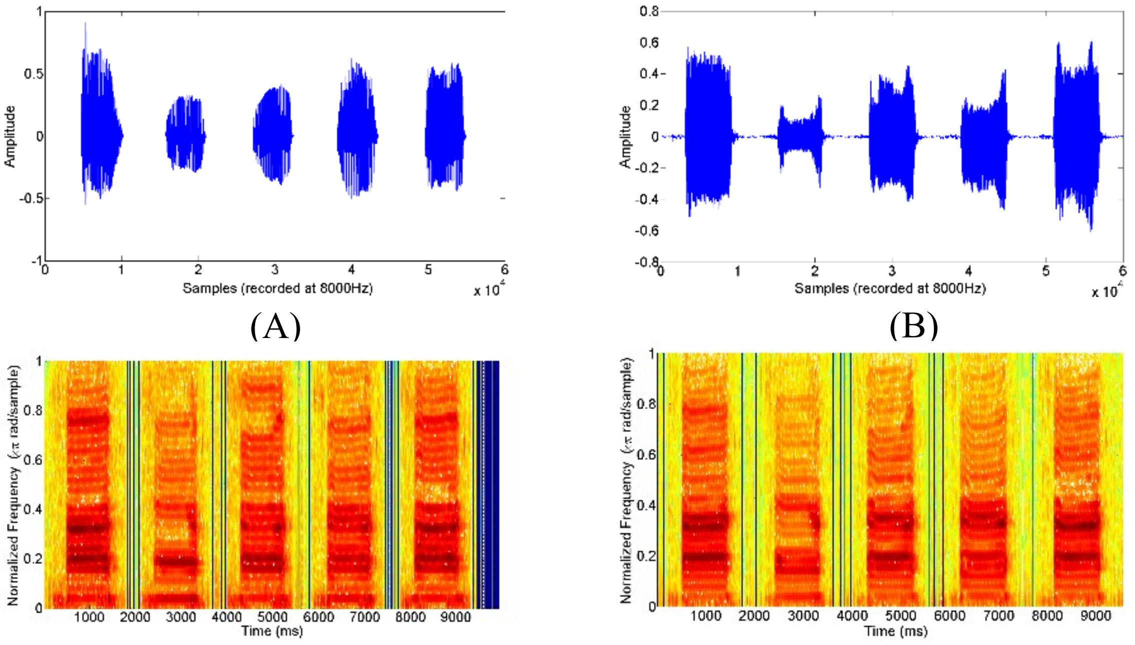
Figure 6.5: Input text and intonation dialogs

The length of input text determines how many motor command the system will generate by using a looping sequence. In each loop, an algorithm to separate the character and its tonal effect from the input data is performed. Each character and its intonation are translated to robot parameters as shown in Table 1. After generating the motor command, the talking robot controller drives the articulatory motor to make a sound. By doing that, the robot sequentially generates each character sound; thus, the entire text will be generated to speech mechanically.

For word separation in the case where the user input several words in the input dialog, an algorithm using blank characters is applied. When a blank character is detected, the system will generate a stop command to move robot articulatory to the initial position and pause for 500 milliseconds.

6.3 Robot Performance with TTS System

Using the TTS system, the robot was tested to speak some Japanese and English phrases. First of all, the text of five vowels, */a/, /i/, /u/, /e/, /o/*, was input to the dialog text for the robot to repeat. Each character was separated by a space. Thus, each vowel sound was generated for 500 ms and then paused with silence for another 500 ms. The intonation of this test was set at the default value of 5th level. The result of this test is shown in Figure 6.6 The human sound for these five vowels is shown in Figure 6.6 (A) for comparison, and the robot sound is shown in Figure 6.6 (B). The sound waveform is shown on the top, and the spectrum of the sound wave is shown on the bottom of the figure. As noted, the amplitude and the shape of the waveform of the human voice and robot voice is very similar. The sound spectrum of these vowels for both sounds are pretty well matched.



(A)

(B)

Figure 6.6. Sound wave of */a,i,u,e,o/* and its formant frequencies from human (A), robot (B)

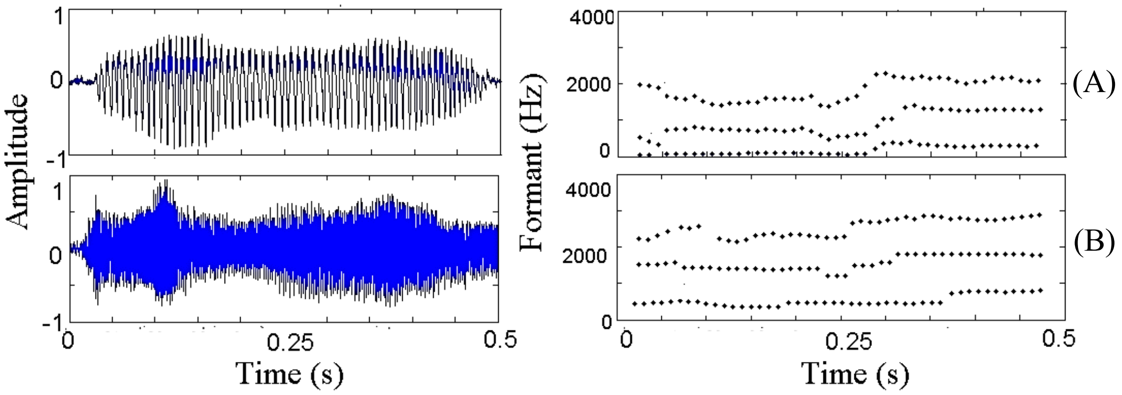


Figure 6.7. Sound wave of */hello/* and its formant frequencies from human (A), robot (B)

However, when the robot was set to speak an English phrase */hello/,* the produced sound was not as good as the Japanese sound. As shown in Figure 6.7, although the sound waveform and shape are well matched, when comparing the formant frequencies of the two sounds, the human sound, and the robot sound displayed differences. Nevertheless, the robot sound of the */hello/* phrase was still recognizable by a human. This is because the training data is taken from Japanese sound in the learning phase; thus, the production of the talking robot with Japanese sound delivers better quality. Also, some foreign languages, like German or English, require a continuous flow of sounds, often referred to as connected speech or linking, in order to produce smooth speech. Combining the character set of phonetics to reproduce foreign sounds proved to be more difficult than Japanese sounds.

CHAPTER 7

Cerebellum-like Neural Network as Short-range Timing Function

Timing is an important characteristic of human speech, which influences its rhythm, stress, duration, and intonation. In this study, the authors focus on building a cerebellum-like model that has the ability to learn and control the timing characteristic of the talking robot using a field programmable gate array (FPGA [55]) board. The details and structure of a fully functional cerebellum model are very complicated and out of the scope of this study.

The main purpose of this section is to build a cerebellum-like neural network model as a short-range timing function for the talking robot. Although the timing function for the talking robot can be straightforwardly calculated by using mathematical analysis and be implemented to the talking robot, the author employs a human-like regulatory mechanism to control a human-like mechanical system. The cerebellum has been known for its role in precision, coordination and accurate timing of motor control as indicated in a study of Lewis and Miall in 2003 [56], Koekkoek and his teams in 2003 [23], Ackermann in 2008 [13], and Ivry and Spencer in 2004 [14]. Slurred speech is clear evidence of cerebellar ataxia as described in the papers cited by Boyd [57]. Research by Ivry and his team showed that lateral cerebellar lesions patients had difficulty in discriminating sound duration interval but had no trouble in discriminating the intensity of sounds [58]. The fMRI evident given by Ghosh and his team show the timely control contribution of the cerebellum to vocalization [59]. Thus, the approach for building a timing model based on cerebellum anatomy is employed in this study. Figure 7.1 (Adapted from Schlerf et al. [60]) shows the overview of the cerebellar circuit. The excitatory input pathways are the mossy and climbing fibers. Climbing fibers synapse directly onto the Purkinje cells. Mossy fibers synapse onto granule cells, which give rise to excitatory parallel fibers which synapse onto Purkinje cells, the output of the cerebellar cortex. Stellate, Basket, Golgi, and Purkinje cells are all inhibitory. All input pathways and Purkinje cells synapse on the Deep Cerebellar Nuclei (DCN), the output of the cerebellum. Some model of cerebellar neural networks were introduced by Chapeau and Chauvet in 1991 as delay line model [61], by Bullock, Fiala, and Grossberg in 1994 as oscillator model [62], by Garenne and Chauvet as spectral timing model in 2004[63], and most recently by Yamazaki and Tanaka as an internal clock of the random projection model [64],[65]. Among these models, the random projection model is believed to be more biologically plausible as described in the research conducted by Yamazaki and Tanaka in 2007. The passage of time (POT) or timing control function of the Yamazaki model is mainly derived from the neural signal of the granular layer, and the timing characteristic is lost if the granular layer malfunctions. Therefore, an assumption of the short-range timing function derived from the eye-blink paradigm is proposed to apply to the talking robot. For hardware design and implementation, Bamford and his team built a VLSI field-programmable mixed-signal array based system in 2012 [66], and Luo et al. built an FPGA-based system which can simulate 1 second of cerebellum activities in less than 26 milliseconds [67]. Based on the model of the spiking neural network introduced by Yamazaki and Tanaka (2007) and FPGA implementation by Luo, et al. (2014), the timing function for the talking robot is modelled using System Generator (SG) software, and then a co-simulation with Xilinx SP605 FPGA board is done to get the timing analysis data as the timing function. The timing function is combined with motor position control function to fully control talking robot's speech output.

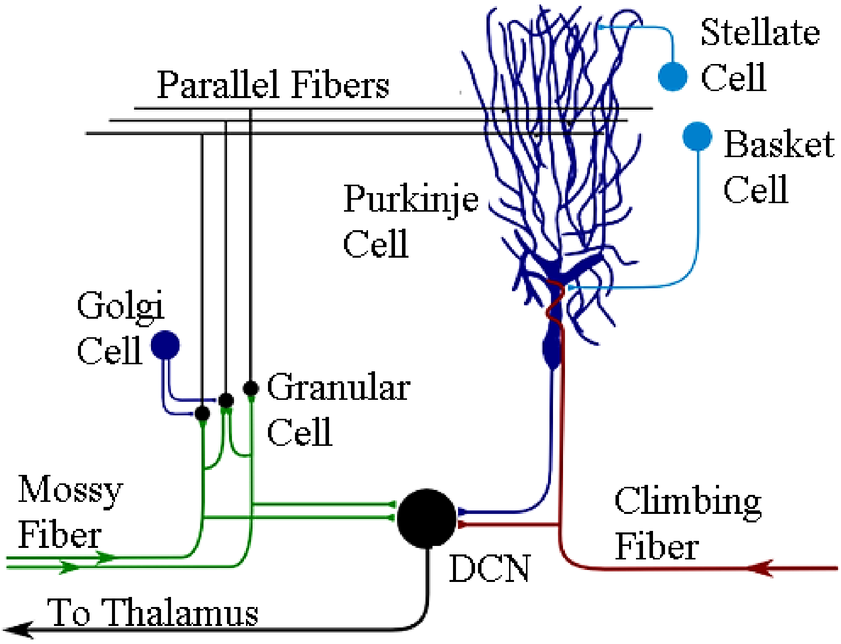


Figure 7.1 Overview of the cerebellar circuit

A regular computer CPU using a sequential computing structure makes it difficult to run a bio-realistic neural network in real-time due to the parallel processing of many neurons at the same time. FPGA technology offers programmable logic as well as high-speed operation, which potentially can create a real-time control system and are well suited for prototyping big-scale hardware models of a neural system due to its parallel processing and fast clocking cycle. The use of FPGAs for neuronal ion channel dynamics simulation was first demonstrated by Graas, Brown, and Robert (2004) [68], and then by Mak, Rachmuth, Lam, and Poon, (2006) [69]. Simulation results showed a promising application of FPGAs for bio-realistic neural network implementation. With the advantage of high-speed signal processing of FPGA, it has been applied to modeling and simulating large-scale biologically realistic neural networks [70] (Cheung, 2012). Thus, an FPGA is suitable for implementing a cerebellum neural network model as the timing function for the talking robot. In this study, to save on resources, a pipelining technique using First-In-First-Out (FIFO) blocks was applied when designing the network model [71]. The timing output is obtained from the co-simulation between System Generator software and the FPGA board. Then, timing information is combined with motor's position vectors of the talking robot to generate a sound that contains prosody characteristic.

This chapter introduces the mechanical construction and the motor control system of the talking robot. Then, the author introduces the cerebellum-like neural network as a timing function for the talking robot followed by the experimental results of short, medium and long duration vowel sounds. Overview of the system is shown in Figure 7.2

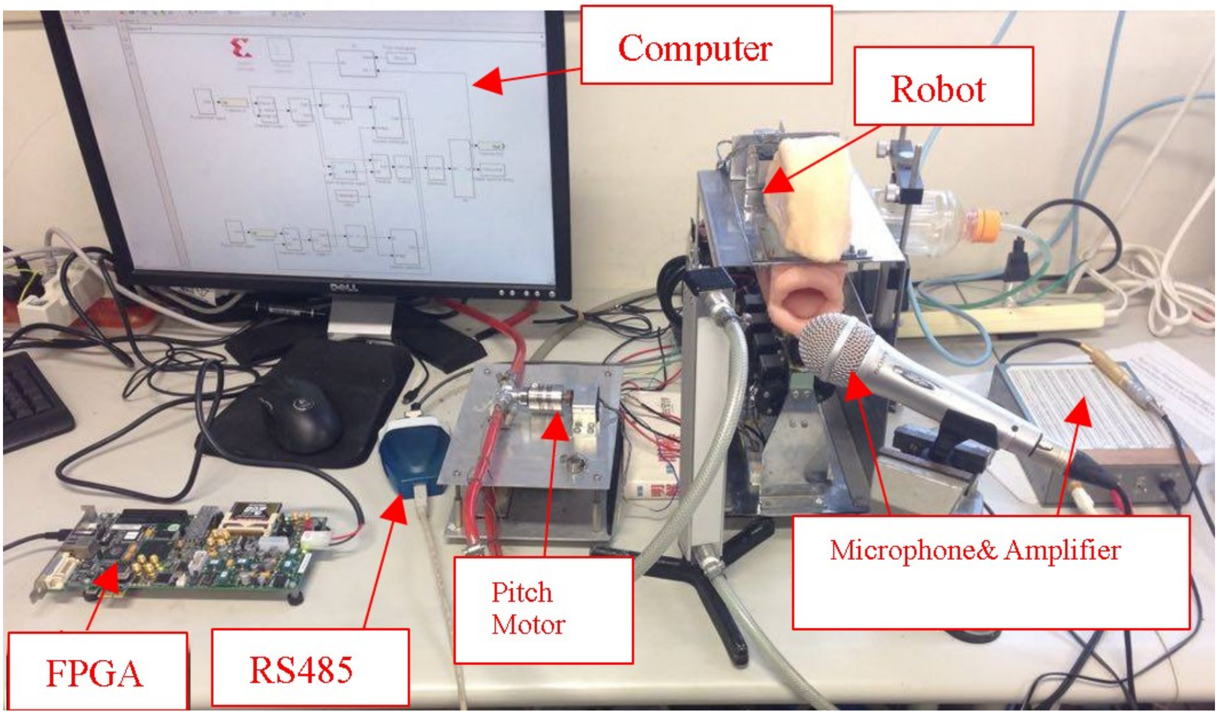


Figure 7.2: Overview of system

7.1 Cerebellum Neural Network Overview

The cerebellum is a region of the brain that plays an important role in motor control. The cerebellum is not primarily involved in movement, but it contributes to coordination, precision, and accurate timing. The location of the cerebellum in the human brain is shown in Figure 7.3 below (Figure adapted from https://jp.pinterest.com).

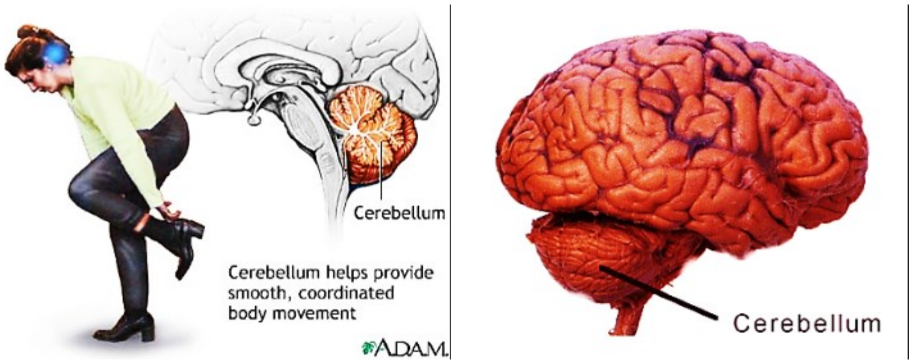


Figure. 7.3: The location of the cerebellum in the human body

At the level of large scale anatomy, the cerebellum consists of a tightly folded and crumpled layer of cortex. At the microscopic view, each part of the cortex consists of the same small set of neural elements, laid out in a highly stereotyped geometry. At an intermediate level, independently functioning modules are called microzones or microcompartments. The anatomy and the basic structure of the cerebellum are shown in Figure 7.4 and 7.5, respectively.

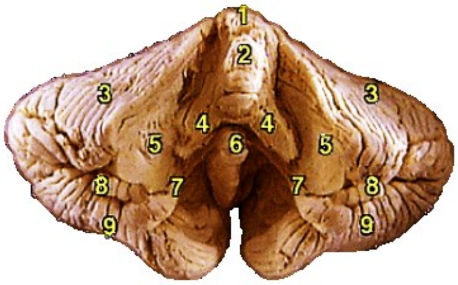


Figure 7.4: The anatomy of the cerebellum (1. Vermis, 2. Central lobule, 3. Anterior lobe, 4. Superior cerebellar peduncle, 5. Middle cerebellar peduncle, 6. A nodule of vermis, 7. Inferior cerebellar peduncle, 8. Flocculus, 9. Posterior lobe.) (Adapted from https://jp.pinterest.com)

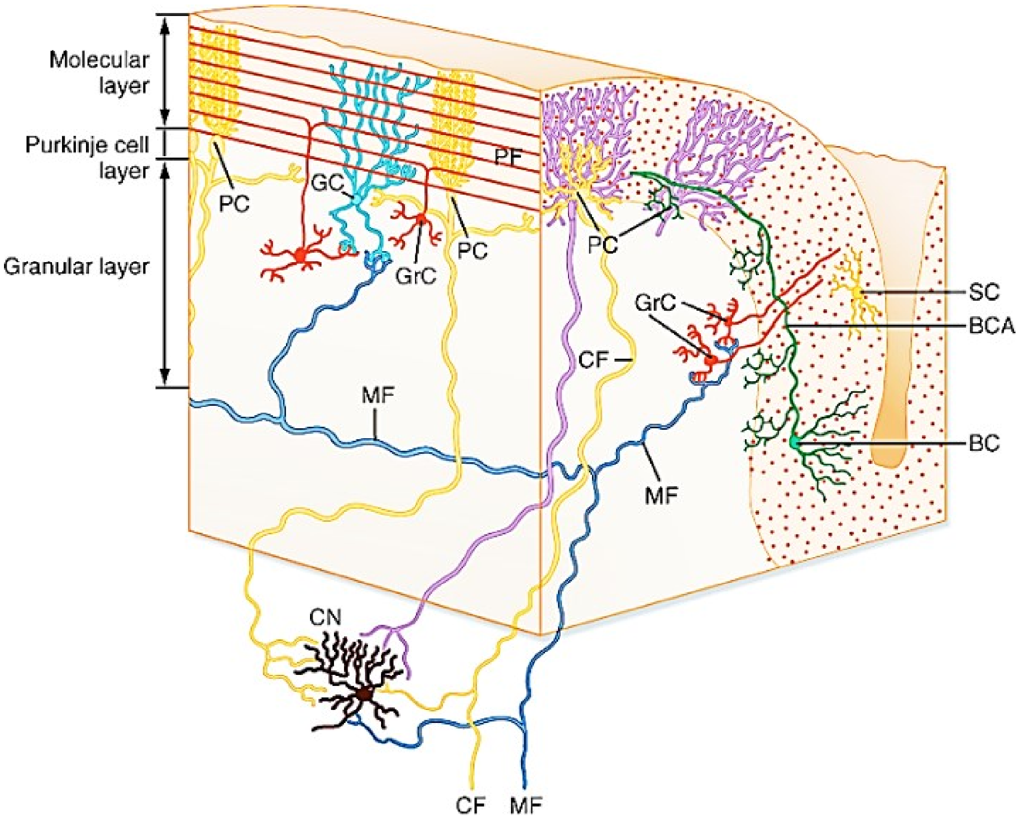
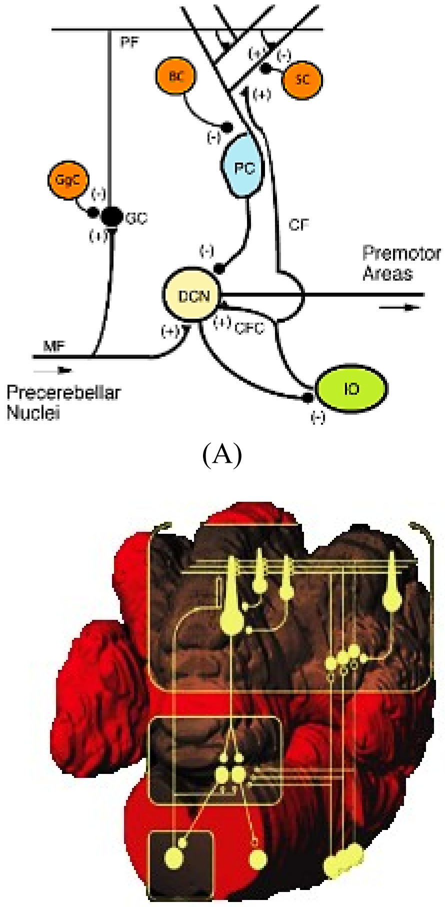


Figure 7.5: Basic structure of the cerebellum cortex (Adapted from [72])

Climbing fibers (CF), which are the full ramifications of the olive cerebellar axons, make direct excitatory contact with Purkinje cells (PKJ), and mossy fibers (MF) make excitatory synaptic contacts with granule cells (GR) and with Golgi cells (GO). Each axon of a granule cell branches to the two ends of parallel fibers (PF), which create the excitatory synaptic contacts with PKJ, the molecular layer interneurons, and GO. PFs extend for several millimeters along individual cerebellar folia [72].

Purkinje cells and granule cells are two types of neuron that have dominant roles in the cerebellar circuit. Three kinds of fibers that have significant roles MF, CF, and PF. There are two main pathways in the cerebellar circuit, originating from mossy fibers and climbing fibers, both ultimately terminating in the deep cerebellar nuclei (DCN). MF project directly to the DCN, but also rise to the pathway: mossy fiber – granule cells – parallel fibers – Purkinje cells – deep nuclei. Climbing fibers project to Purkinje cells and also send collaterals directly to the deep nuclei. The microcircuit of the cerebellum is shown in Figure 7.6 below.



(A)

(B)

Figure 7.6: Microcircuit of the cerebellum (Adapted from Wikipedia.com)

In Figure 7.6 (A), +: excitation, -: inhibition, MF: mossy fiber, GC: Granule cell; GgC: Golgi cell; PF: parallel fibres; BC: Basket cell, SC: Stellate cell, PC: Purkinje cell, CF: climbing fiber, DCN: deep cerebellar nuclei, IO: Inferior olive. As shown in Figure 7.6 (B), cerebellar neurons are arranged in a regularly iterating, geometrical array to form a huge set of regularly repeating microcircuits. The anatomical and physiological similarity of these microcircuits suggests a consistent type of information processing in the cerebellar [72].

For studying about the mechanism of the cerebellum, Hofstötter and his teams proposed a model of the cerebellum in action in 2002 [73]. This model aims to study the effect of changes in the strength of the synapses between parallel fiber and Purkinje cells. Five assumptions are defined in this model. Basically, long time depression (LTD) and long time potential (LTP) in synapse plasticity decide whether there is a central response (CS) elicited by the conditional stimulus (CS). Neuron model is based on the generic type of integrate-and-fire model. Moreover, the key part is the synapse plasticity. The model circuit is shown in Figure 7.7, and the learning mechanism is shown in Figure. 7.8 (A), (B), (C) which respectively show the pathways, signals process before training, and signals process after training.

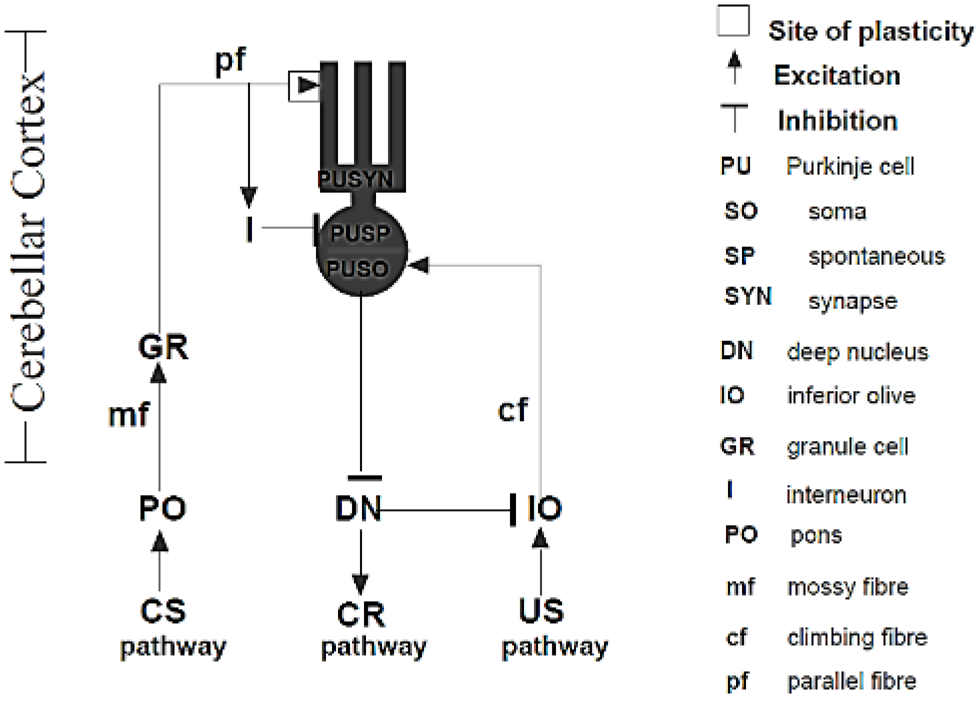


Figure 7.7: Cerebellar cortex model circuit in Constanze study [73]

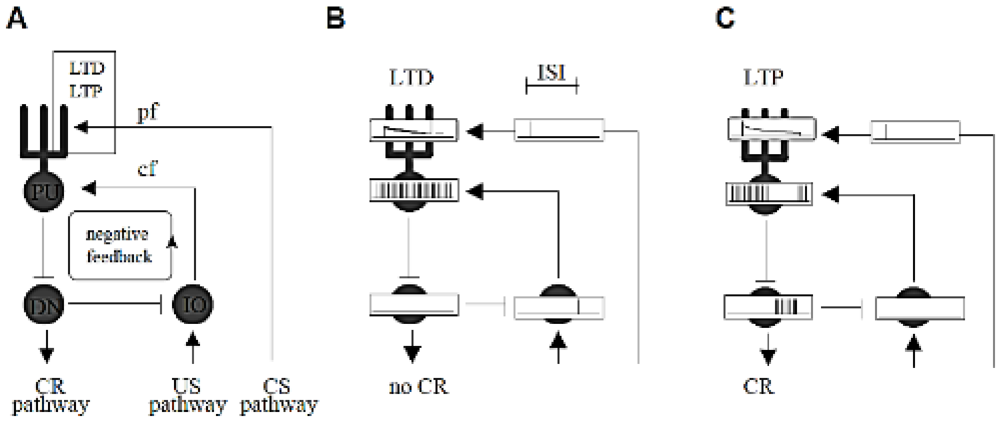


Figure 7.8: The learning mechanism is embedded in the model [73]

The simulation results in Hofstötter's study suggested that the higher the chosen value of LTP, the stronger the LTP and the faster the extinction of a CR; while the lower the selected value of LTD, the stronger the LTD and the faster the acquisition of a correctly timed CR.

7.2 Bio-Realistic Cerebellar Neuron Network Model

7.2.1 System Configuration [66]-[71]

A field-programmable gate array (FPGA) is an integrated circuit that can be configured by a user for a computational purpose. An FPGA device contains an array of programmable logic blocks whose interconnections are reconfigurable. Therefore, the logic gates can be interconnected in many different configurations. Each logic block can be used as a simple logic gate function (such as AND or OR logic operation) or as a complex function. The logic blocks also include memory components, which may be simple flip-flops or a complex memory microcircuit.

The FPGA device is usually programmed using a hardware description language (HDL) or (VHDL). In this study, the author used System Generator software in combination with Matlab Simulink to program the FPGA board. System Generator is a block diagram programming software working under the Matlab Simulink environment. Thus, people can take advantage of signal simulation of an FPGA design before employing it in the FPGA device. The block diagram of this system is shown in Figure 7.9 below.

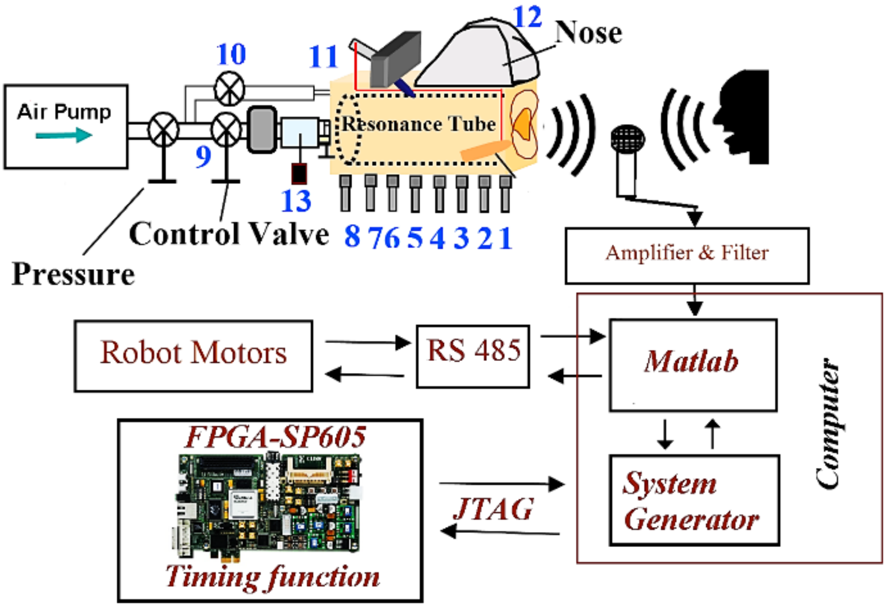


Figure 7.9: System configuration with FPGA as timing function

The computer communicates and controls the talking robot via the RS485 port. The auditory and motor position feedback of the talking robot is also sent to the computer via the RS485 port. The timing function is employed in an FPGA-SP605 board by System Generator software via the JTAG port.

The sound input from the microphone after amplification and filtering is used as the training input signal for the timing function. The training is a co-simulation process between the FPGA board and the computer via the JTAG port. The output signals after the co-simulation process are decoded to extract the timing information for the vocal motor movements. This timing data is combined with other robot parameters to let the talking robot generate speech that contains prosodic features.

7.2.2 Assumption of Short-Time Learning Capability

While visual and auditory signals have sensory input processing in the brain, there is no clear sensory information for timing. However, timing encoded in the temporal pattern of the neuron's activities is plausible. The human sense of time is around 2-3 seconds according to Fraisse (1963) [74]. However, Lewis and Miall (2003) debated that short-range timing was instinctive and associated with skilled movement's production [56]. Long-range timing is stated to be cognitive and related with brain memory. From the eyeblink conditioning experiments of Pavlovian, several studies (Bao, Chen, Kim, and Thompson, 2002 [75]), (Gerwig et al., 2003 [76]), (Garenne and Chauvet, 2004 [63]) have indicated that the cerebellum is the main organ of the brain responsible for this short-range timing function. Yamazaki and Tanaka built a simulation model based on a real cerebellum structure to prove the working mechanism of this classical conditioning [65]. Based on these findings, the author hypothesizes that the short-range timing in speech is also a function of the cerebellum.

Each Granular neuron has its own unique temporal pattern, which is active or inactive for a short period. The range can vary from 100ms to 1000ms due to the random recurrent neural network between Granular cells (GR) and Golgi cells (GO). The author assumes that the input signal from the Mossy fiber (MF) is the predictive timing, which is the conditional stimulus in other cerebellum network studies. The predictive timing in our model is a fixed 5-second 30Hz Poisson spike signal. The actual sound input is pre-processed and transformed into a 30Hz Poisson signal with the same duration to serve as a climbing fiber (CF). Long-term depression (LTD) at the parallel fiber (PF) adjusts the synaptic weight between the Granular cells (GRs) to the Purkinje cells (PKJs). The adjustment coefficient in our model is big, and the cerebellum has a super-fast learning rate. Due to LTD, the synaptic weight reduces the input signal from the GRs to PKJs at the range of time when the sound is active at CF. Thus, the PKJ would not fire a spike at that range of time. Because of the inhibitory signal input from PJK to DCN, the DCN is released to fire a spike signal at the same range of time input from the climbing fiber. This is the timing signal for the talking robot to regenerate the specific duration of a vowel.

The inhibitory connection from DCN to IO prevents the cerebellum from over-training. LTP restores the weight connection of GRs to PKJs to the initial value if the learning signal is off for a certain time. However, due to the fact that most Granular cells will be active or inactive during a specific short duration, if the sound signal input is a long duration signal (more than 2 seconds for example), the output of the network would be the same with predictive timing since all synaptic weight from Granular cells to PKJ are reduced. So the author hypothesizes that this network would work well for short-time duration learning. For long duration, the network would not be able to learn the timing. The author also conducts the experiment to verify our short time only learning capability in section.

7.2.3 Experimentation for Verifying Human Short-Timing Function

In order to verify our hypothesis, the author conducts a short experiment with a group of 7 people. They were requested to listen to different durations of the same sound in 3 cases. In the first case, they listened to 2 x 500 millisecond and 2 x 600 millisecond sounds. The second case was listening to 2 x 1200 millisecond and 2 x 1300 millisecond sounds, and the last case was listening to 2 x 3000 millisecond and 2 x 3500 millisecond sounds. They were asked to distinguish which sound they heard had the shorter duration, and how difficult for that case. The result showed that they could determine the difference between the 500ms and 600ms sounds in case 1. They felt a little difficulty in case 2 but still could guess the difference. However, they said it was very difficult for case 3 and almost couldn't tell which sound had the longer duration. The result of this experiment is shown in Table 7.1.

Questionnaire detail:

1. *Which sound has the longest duration?*

   1, 2, 3, 4 (please arrange in order)

2. *Do you think any sound has the same duration? (Yes/No) If (Yes) which one and which one?*

3. *Which case is easiest and most difficult?*

Table 7.1: Short-time learning experiment result

	Case 1	**Case 2**	**Case 3**
Accuracy	83%	67%	33%
Same duration	50%	50%	17%
Difficulty	1	2	3

The author also conducted another experiment for short-timing distinction. He let this group listen to 4 different sound durations from 400 milliseconds to 700 milliseconds with 100 millisecond increasing steps in random order and asked them to rearrange these sounds from the shortest to longest duration sounds. After around 3 to 5 trials, all members of the group could exactly rearrange the sound. From these experiments, the author assumes that humans can learn short-timing with the range of 250 milliseconds to 1500 milliseconds, and the most accurate range for learning is from 400 to 800 milliseconds. This also verified our hypothesis of the short-timing learning ability of humans.

7.3 Neuron Model Structure

The Cerebellar cortex is a gigantic neural network with a moderately simple and orderly structure of well-defined input/output terminals as described by Apps and Garwicz [33]. Major neuron types in the cerebellar cortex include Purkinje cells (PKJs), Granular Cells (GRs), Golgi cells (GOs) and deep cerebellar nuclei (DCN). Fibers, acting as cables to conduct excitatory or inhibitory signals from one neuron to others, include Mossy fibers (MF), Climbing fibers (CF), and Parallel fibers (PF) (Ito, 1984).

PKJs have the most synaptic inputs compared to other types of neurons and create the output of the cerebellar cortex. The density of PKJs is roughly 300 to 1000 cells per mm^2 in mammals, including rats, cats, monkeys and humans (Harvey and Napper, 1988) [78], (Mwamengele, Mayhew, and Dantzer, 1993) [79], (Palkovits, Magyar, and Szentagothai, 1971)[80]. PKJs receive excitatory synaptic inputs from PFs and CFs. In addition, a PKJ cell has up to 200,000 PF connections. GRs form a densely packed layer lying below the Purkinje layer and form a heavy signal-processing unit in the cerebellum cortex. They receive excitatory input from MFs but also receive inhibition input from GOs, which are inhibitory neurons within the granular layer of the cerebellum. The connections between MF and GO are claimed to exist [93], however, according to Jakab and Hamori [81], GO dendrites of rat cerebellum are zero in the granular layer; thus, the author ignores this connection in this study.

In order to assure the implementation of the cerebellar model to fit our FPGA board (Xilinx-SP605), the author uses the cerebellum model introduced by Yamazaki and Tanaka [65] scaled down with some parameter adjustments. Their model contains a granular layer, which has 320 x 320 GRs and 32 x 32 GOs, 16 PKJs, IO, CN (or VN), MFs and PFs. In their model, they claimed that the passage of time (POT) representation is mainly due to the neural activity between the granular layer and Purkinje cells.

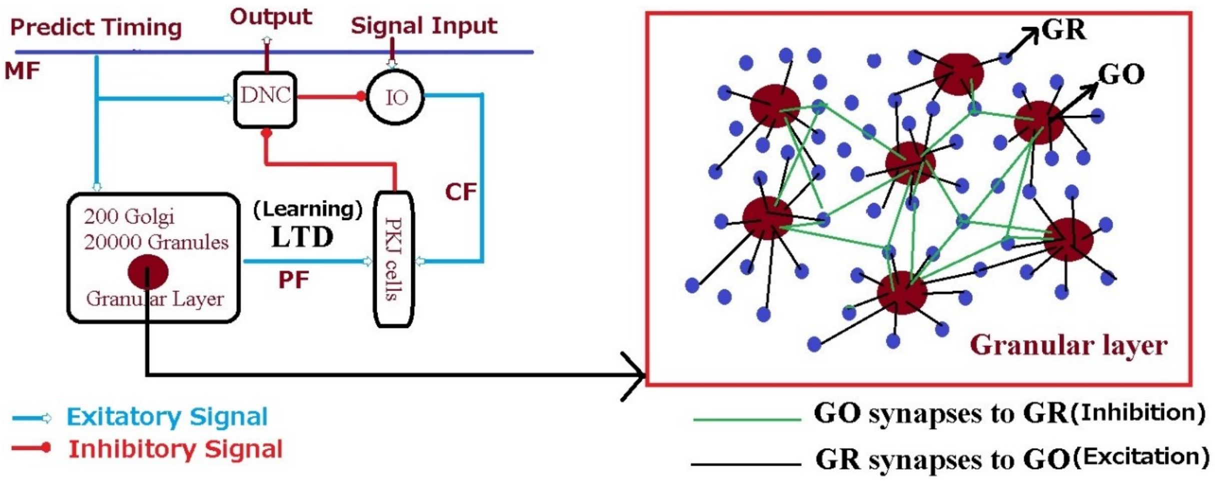


Figure 7.10: Schematic of cerebellar neural network for short range timing function

Our cerebellum model is shown in Figure 7.10. The model contains 20000 GRs, which are 200 GR clusters with 100 cells in each cluster. GRs receive random recurrent inhibitory input from up to 4 nearby GO cells with a probability of 0.025, a synaptic weight of 2 and excitatory input from MFs with a synaptic weight of 0.18. One GO cell receives random excitatory input from about 10 x10 nearby GR cells with a probability of 0.05 and a synaptic weight of 0.02. The synaptic weight from GR cells to PKJ cells is chosen to be 0.0058 with a probability of 0.75 in order to have a firing rate around 90 spikes/second at the initiation stage. Synaptic weight from PKJ cells to DCN is chosen to be 0.5. This gives DCN enough inhibition to prevent it from firing all the time. The weighting needs to be larger because the firing rate of PKJ cells decreases rapidly. Synaptic weight from DCN to IO is 45, which gives IO enough inhibition from DCN to prevent it from over-training. The author only pays attention to the timing control mechanism of the cerebellum, which he assumes to be the function within the GR layer. Basket and stellate cells are inhibitory interneurons that do not appear in the granular layer. They do not affect the timing function of the cerebellum, so the author omits these cells in our model. Neuron parameters such as synaptic weight, leak voltage, and threshold voltage were chosen according to the biological nature of the cerebellum. Due to limited hardware resources and our hypothesis of the GR layer effect on short-timing learning, all the other types of neurons such as basket cells, stellate cells, and unipolar brush cells are ignored.

7.4 Neuron Model Implementation

The neuron model as shown in figure 7.10 is implemented in the System Generator (SG) software as shown in Figure 7.11. There are three inputs, including two Poisson signals to a GR cluster, and one pre-processed sound from the workspace as a 30 Hz Poisson spike signal to IO. The output from

DCN indicates the firing pattern encoded timing information for the talking robot. This output is saved to the Matlab workspace for timing analysis.

There is an initialization step lasting about 20000 clock cycles to establish the neural network behavior to a steady state before the learning begins. The Granular cluster block (1 and 2) is modeled to contain the neuron signals of 100 Granular clusters using the pipelining technique with a latency of 1000 clock cycles represented for 1 cycle of human brain activity. This is necessary because the SP605 FPGA board is limited and this pipelining technique saves a lot of hardware resources. The Splitter block is used to separate the signal of 100 clusters for data analysis and verification. The Golgi block contains signals of 100 Golgi cells. The output from Golgi blocks changes to Random distribution blocks as random recurrent inhibitory input then back to Granular cluster blocks. Inputs from IO blocks and Granular cluster blocks are sent to Plasticity blocks for learning. The weight distribution of Granular cluster blocks to PKJ blocks is updated at Plasticity blocks. The signal is later led to PJK blocks, and then to the DCN block as the output signal. Also, the output signal from the DCN blocks is inhibitory feedback to the IO block for preventing overtraining. Details of these blocks are shown in the later section.

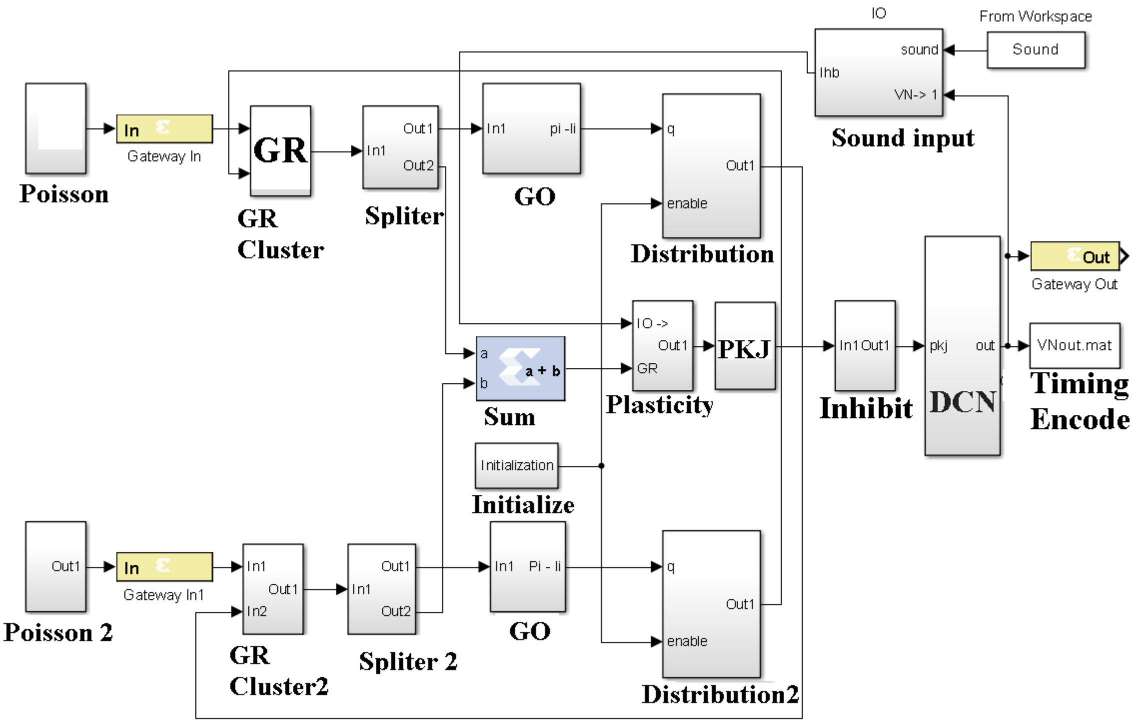


Figure 7.11: Neuron model implementation in System Generator

The structure of neural interconnections in FPGA is shown in Figure 7.12. The external input signals are summed up by an adder before entering the system. The summed up signal is processed in the neuron model with FIFO structure which is the first-in-first-out block to process the input data for each neuron in the system. Then the output signal from the neuron model is separated by the time division demultiplexer as the individual signal for each neuron. These signals are saved in ROM structures for further analysis. The number of ROM structures depends on the number of neurons in the model.

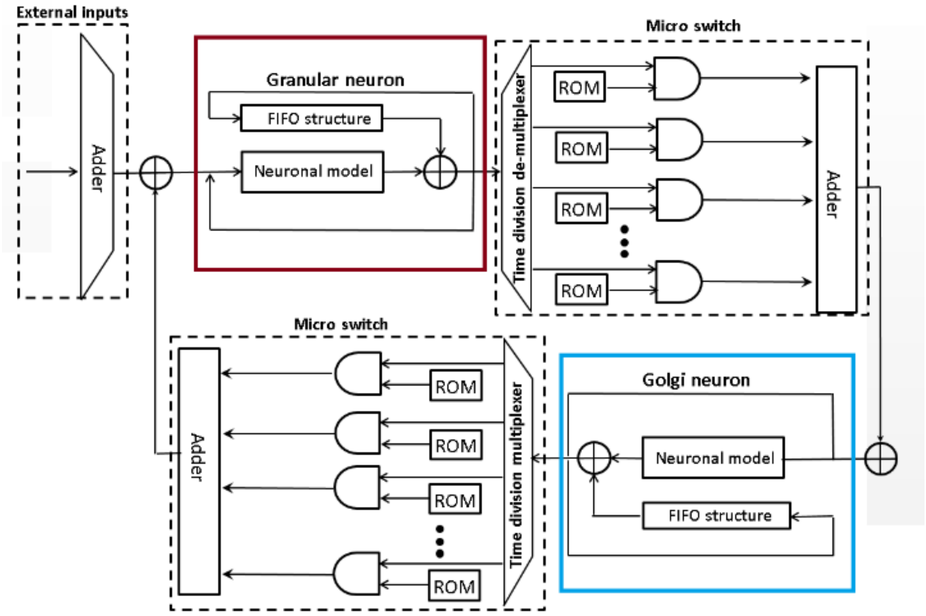


Figure 7.12: Structure of neural interconnections in FPGA

The model in SG is first simulated with the Matlab Simulink environment for testing and debugging. After verifying that the neural network is working properly, it is implemented on FPGA, and a hardware co-simulation is performed to generate the timing information encoded in the neural temporal pattern. The timing information is saved in the Matlab workspace for timing decoding using a short-time energy analysis. As the motor control of the talking robot is written in Matlab, co-simulation is sufficient for combining the motor's vector data and timing information to generate a sound with the same characteristics and duration of the human speech.

The model in this study uses the conductance-based leaky integrate-and-fire neuron model as shown in equation (1). Detailed parameters are shown in Table 7.3. The data is taken from table 1 and 2 of [65] with adjustments for optimal FPGA implementation.

$$C\frac{dV}{dt} = -g_{leak}(V(t) - E_{leak}) - g_{ex:short}(t)(V(t) - E_{ex}) - g_{ex:long}(t)\,(V(t) - E_{ex}) - g_{inh}(t)(V(t) - E_{inh}) - g_{ahp}(t - \hat{t}\big(V(t) - E_{ahp}\big) + I_{External} \tag{7.1}$$

in which

$$g_i(t) = \bar{g}_i \sum_j w_j \int_{-\infty}^{t} \alpha(t-s)\delta_j\,(s)ds \tag{7.2}$$

and

$$g_{ahp}(t-\hat{t}) = \exp(-\frac{(t-\hat{t})}{\tau_{ahp}}) \tag{7.3}$$

- $V(t)$ is the membrane potential of neuron at time t.
- C is the capacitance.
- $I_{External}$ is the external input current(for example from Mossy fiber).
- E is the reversal potential.
- $g_{leak}(t)$ is the conductance function of leakage current.
- $g_{ex\_short}(t)$ is the conductance function of short time excitatory current.
- $g_{ex\_long}(t)$ is the conductance function of long time excitatory current.
- $g_{inh}(t)$ is the conductance function of inhibitory input current.
- $g_{ahp}(t)$ is the conductance function of after-hyperpolarization input current.
- $\bar{g}_i$ is the maximum conductance of *i* (*i* = leak, ex_short, ex_long, inh, ahp).
- w_j is the synaptic weight .
- δ_j(t) is the presynaptic neuron j activity at time t (1 or 0).
- τ_{ahp} is the after-hyperpolarization time constant.
- α(t) is the alpha function of a neuron as shown in Table 7.2.

Table 7.2 Alpha function

Neuron	Alpha function
PKJ	$\alpha_{exsht} = e^{-\frac{t}{8.25}}$
GR	$\alpha_{exsht} = e^{-\frac{t}{1.25}}$ $\alpha_{exlg} = e^{-\frac{t}{52}}$ $\alpha_{inh} = \frac{7}{16}e^{-\frac{t}{7}} + \frac{9}{16}e^{-\frac{t}{59}}$
GO	$\alpha_{exsht} = e^{-\frac{t}{1.5}}$ $\alpha_{exlg} = \frac{3}{8}e^{-\frac{t}{31}} + \frac{5}{8}e^{-\frac{t}{170}}$
DCN	$\alpha_{exsht} = e^{-\frac{t}{10}}$ $\alpha_{exlg} = e^{-\frac{t}{30}}$ $\alpha_{inh} = e^{-\frac{t}{43}}$
IO	$\alpha_{exsht} = e^{-\frac{t}{10}}$ $\alpha_{inh} = e^{-\frac{t}{10}}$

Based on the conductance-based leaky integrate-and-fire neuron model, the author applied the neuron model to SG. A neuron generates a spike at a time when its membrane potential exceeds a threshold, and then after-hyperpolarization would follow. The implementation of equations from (1) to (3) for GR is shown in Figure 7.13. GOs, PKJs, DCN, and IO have a similar structure with different parameters as shown in Table 7.2 and 7.3. Due to the parallel processing behavior of the FPGA device, delay (latency) needs to be considered so that each individual path is performed consistently. Overflows would occur in the case where a mismatch happened in just one path. The FIFO router block in figure 5 is built with 1000 elements of FIFO. Thus, for completing 1 cycle simulation of a real-time signal, the network needs to take 1000 cycles within the FIFO block. However, due to the very fast clocking signal of the FPGA board, the computation for simulating 1 second of real-time action can be completed in 62 milliseconds.

Table 7.3 Neuron parameters

Parameter		Neuron type				
	Unit	PKJ	GR	GO	DCN	IO
C	pF	107	3	28	122	10
g_leak	nS	2.5	0.45	2.5	1.75	0.625
E_leak	mV	-68	-58	-55	-56	-60
g_exshort	nS	0.75	0.175	45	50	1
E_exshort	mV	0	0	0	0	0
E_exlong	nS	-	0.025	30	26	-
E_exlong	mV	-	0	0	0	-
g_inh	nS	-	0.025	-	30	0.175
E_inh	mV	-	-82	-	-88	-75
After-hyperpolarization						
g_ahp	nS	0.125	1	20	50	1
E_ahp	mV	-70	-82	-73	-70	-75
T_ahp	ms	5	5	5	2.5	10
Threshold	mV	-55	-35	-52	-39	-50

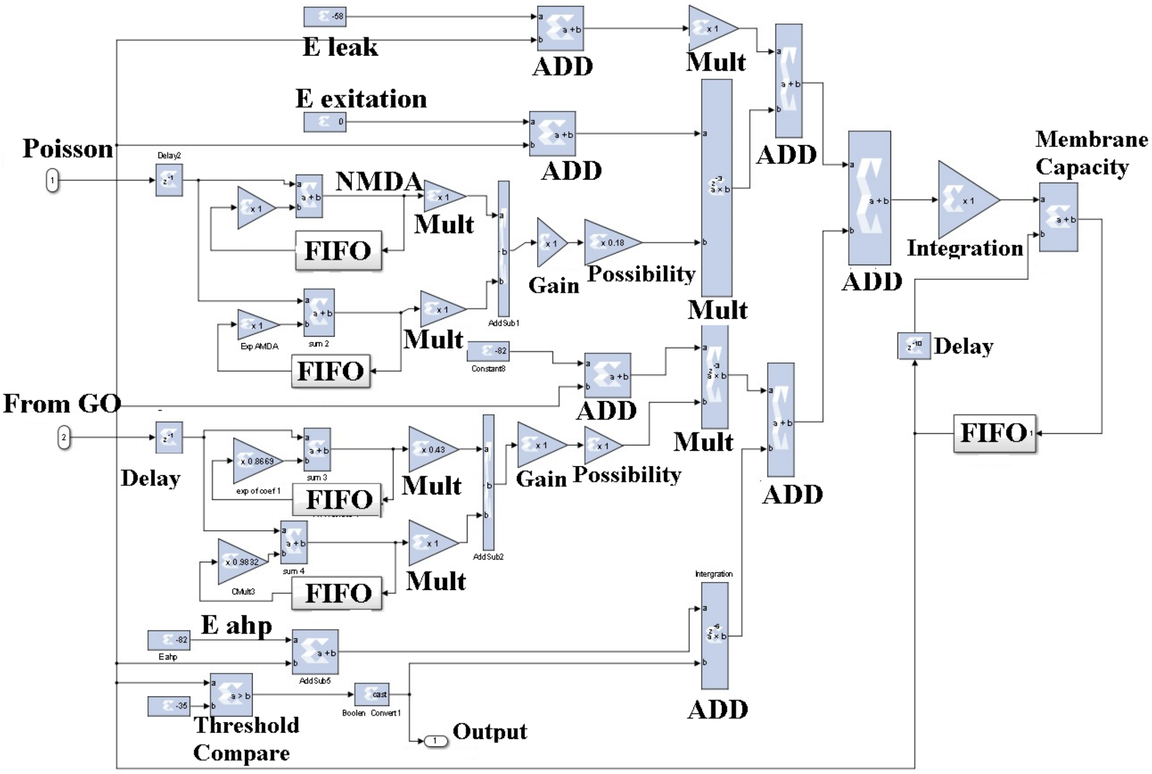


Figure 7.13: Neuron model implementation

7.5 Neural Spiking Pattern and Analysis

The author conducts a simulation using the Matlab Simulink environment to verify the working behavior of the cerebellum neural network. The author inputs the external signal of a Poisson spike with a transient portion of 200Hz lasting for 1 millisecond at the beginning of the simulation period followed by a 5Hz signal and a sustain portion of 30Hz lasting for the rest of simulation period. The author only tests the working behavior of the neural network, so the learning signal from IO was not used in this simulation.

The neuron activities for a 1200 millisecond simulation of a GR cell at 4 random Granular clusters are shown in Figure 7.14. The results in this figure indicate that each GR cell generates a unique neuron activity pattern. Some neurons fire at the beginning of the period and then stop; some neurons fire at the end of the simulation; some other neurons fire at the beginning and at the end of simulation; and some neurons fire in the middle of the simulation period. The period, in which a GR neuron fires or not, is about 300 milliseconds to 900 milliseconds. In contrast, GO cells fire more regularly as shown in Figure 7.15. This is reasonably due to the high probability connection from GRs to GOs, and the random recurrent inhibitory network from GOs to GRs within the Granular layer. The mechanism to create such a unique pattern is explained below.

Initially, all Granular neurons fire due to a 200Hz transient signal. Then, all signals from GRs are transmitted to GOs with a random distribution. In other words, some GOs have a strong firing rate while others have a weaker firing rate. The GOs inhibit the firing of GRs. Therefore; GRs that received strong inhibitory input would stop firing. Due to the inactivation of the GRs, the GOs that receive input from inactivated GRs would have a weaker firing rate. In other words, the inhibitory signals from these GOs to some GRs are weakened and due to these weakened inhibitory signals from such GOs, some of the GRs, which were not active, can produce an action potential at this moment. The firing pattern of Figure 7.14 (B) is an example of the neurons that stop firing at the beginning. New activations from these GRs gives a stronger firing rate for other GOs which in turn have strong inhibitory inputs to other GRs causing them to stop firing. Figure 7.14 (A) and (C) are the examples of neurons that were initially firing then stopped due to this mechanism. This action continuously occurs inside the GR layer, and a unique pattern of each GR cell is created. Therefore unique patterns of GRs with short-range active or inactive states guarantees a short-time learning ability of this bio-realistic neural network.

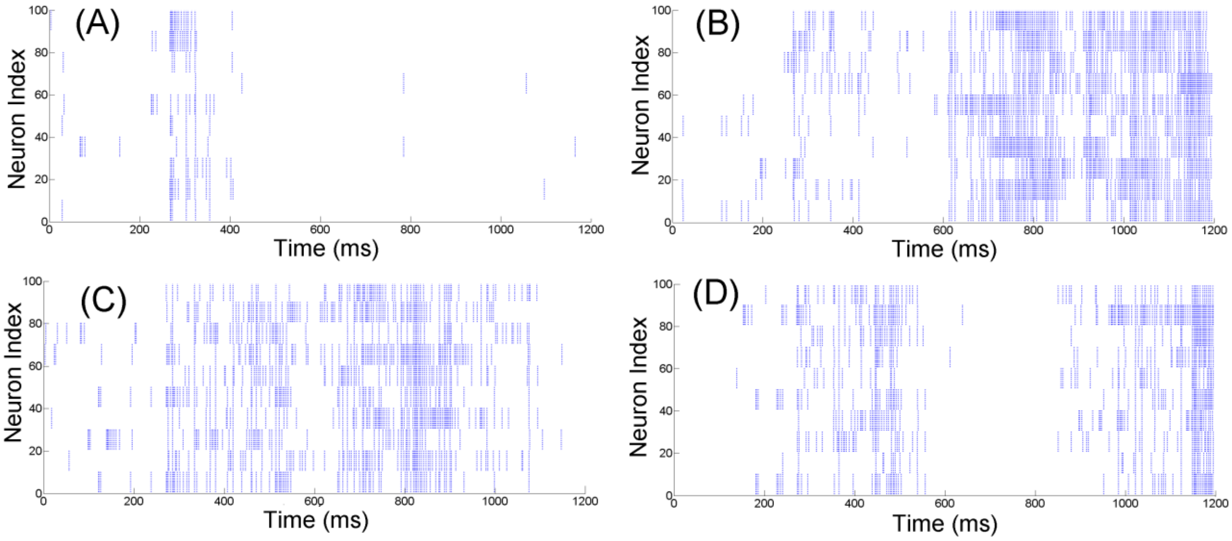


Figure 7.14: Granular cluster activities of 4 random clusters

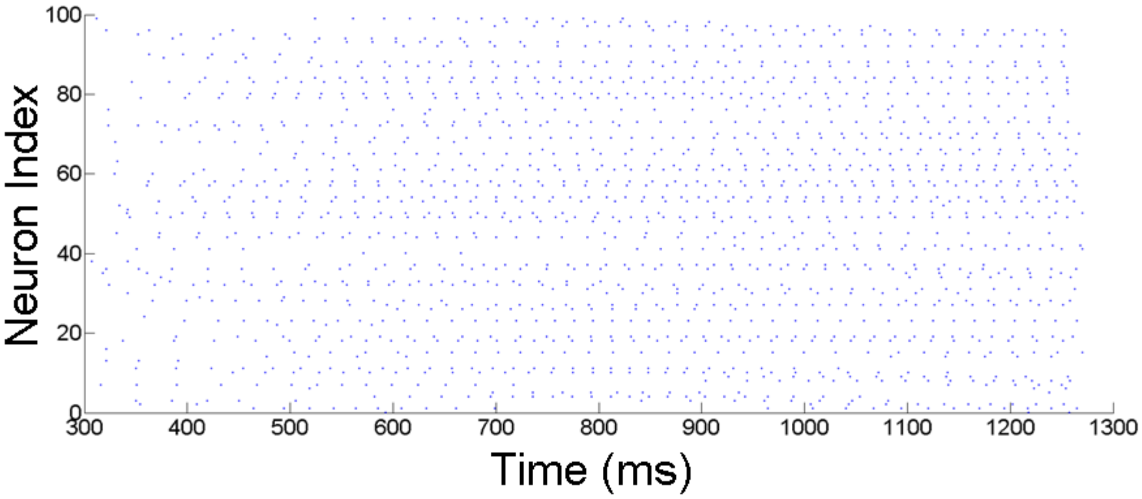


Figure 7.15: Golgi cells activities of 100 random neurons

To evaluate the network, the author calculates the similarity index of the neuron activity. Similarity index S (Δt) is the averaged autocorrelation of GR neuron activities with respect to time shift Δt as shown in equations from (4) to (6). The result of the similarity index is shown in Figure 7.16. The similarity index decreased as Δt increased indicating that the activity pattern changed with time and each granule-cell that was active, was unique. The minimum similarity index in our case was 0.52 at a time shift of 400 milliseconds. Compared to the 320x320 GRs model of Yamazaki and Tanaka (2007), which has a GR layer five times larger, the difference is less than 28%. This indicates that our neural network is sufficient for modeling a cerebellar mechanism. In equations (7.4) to (7.6), $z_i(t)$ is the average activity of a GR cluster i,τ is the decay time constant of PKJ, N is the number of Granular cells in a cluster (100 in our case), γ_j(t) is the spike signal at time t (1 if firing or 0 if not), and C is the correlation coefficient between $z_i(t)$ and $z_i(t+\Delta t)$.

$$z_i(t) = \frac{1}{\tau}\sum_{s=0}^{t} e^{\left(-\frac{t-s}{\tau}\right)} \times \left(\frac{1}{N}\sum_j \gamma_j(s)\right) \tag{7.4}$$

$$C(t, t+\Delta t) = \frac{\sum_i z_i(t) z_i(t+\Delta t)}{\sqrt{\sum_i z_i^2(t)}\sqrt{\sum_i z_i^2(t+\Delta t)}} \tag{7.5}$$

$$S(\Delta t) = \frac{1}{T}\sum_{t=0}^{T} C(t, t+\Delta t) \tag{7.6}$$

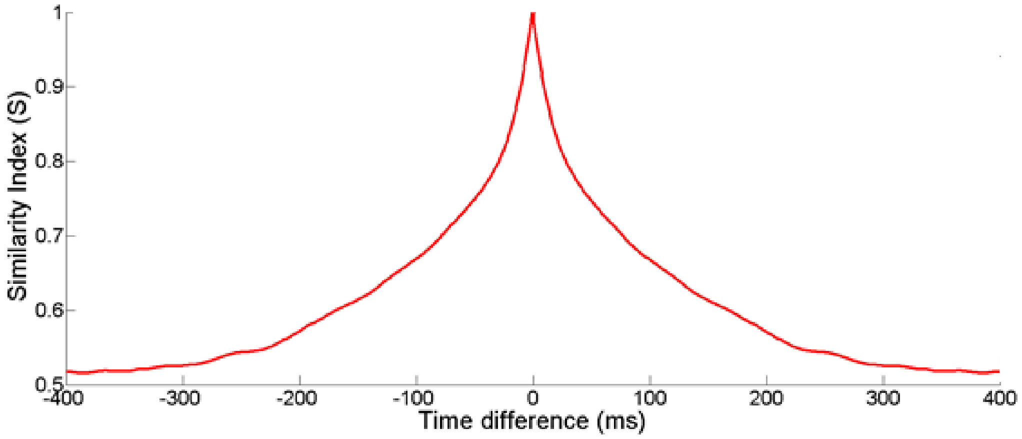


Figure 7.16: Similarity index

7.6 Long-Term Potentiation/Depression and Timing Learning Mechanism

Long-term depression (LTD) [82] and long-term potential (LTP) [83] in the cerebellar cortex are two forms of synaptic plasticity that play important roles in the learning mechanism (Koekkoek et al., 2003) [84]. According to D'Angelo, et al. (2016) [85], there are more than 15 types of plasticity in the cerebellum that are responsible for tasks such as cognition, coordination, and precision control. Other plasticity types such as those from MF neurons to CN neurons represent the amplitude of CR not the timing of CR.

However, since this study mainly focuses on the timing learning capability (POT presentation) of the Granular layer and the Yamazaki and Takana model which only employed LTD and LTP for timing encoding, the author omits other types of plasticity and uses only LTD and LTP for short-timing learning in our model. The author uses LTD and LTP equations from [65] with adjusted coefficients to save FPGA hardware resources. LTD occurred when there was co-activity of PF and CF at a time (t) as described in equation (7). For LTP, the synaptic weight from GR to PJK through PF is updated each time as described in (8). In equations (7.7) and (7.8), $w_{PF_i \to PKJ}$ is the synaptic weight from a GR cells to PKJ cells through PF, $CF(t)$ is the signal on the Climbing fiber at time t (1 or 0), and $PF(t)$ is the signal on the Parallel fiber at time t (1 or 0).

$$w_{PF_i \to PKJ} = w_{PF_i \to PKJ} - 0.75 w_{PF_i \to PKJ}\big(CF(t)PF(t)\big) \tag{7.7}$$

$$w_{PF_i \to PKJ} = w_{PF_i \to PKJ} + 0.01(w_{init} - w_{PF_i \to PKJ})\big(PF(t)\big) \tag{7.8}$$

LTP, particularly the postsynaptic type, is also believed to be important for motor learning through counterbalancing LTD. LTD/LTP modified the synaptic weight of granule cell populations corresponding to the output between predicted timing and learning timing. Given enough training, the response at this specific duration of neuron activity would be presented. LTDs of PF-PKJ connections are simulated by progressively decreasing the synaptic weight for every conjunctive activation of the CF and the PF. The LTD released the DCN from inhibition by PKJ cells and consequently triggered an action with the same duration of learning timing. The implementation of equations (7) and (8) is shown in Figure 7.17. Synaptic weights are first initialized in a 1000 Depth RAM block. The weights in the RAM block are updated for each cycle if the condition of CF and PF are satisfied as described above. Figure 7.18 shows a learning result of this block using LTD coefficient of 0.5.

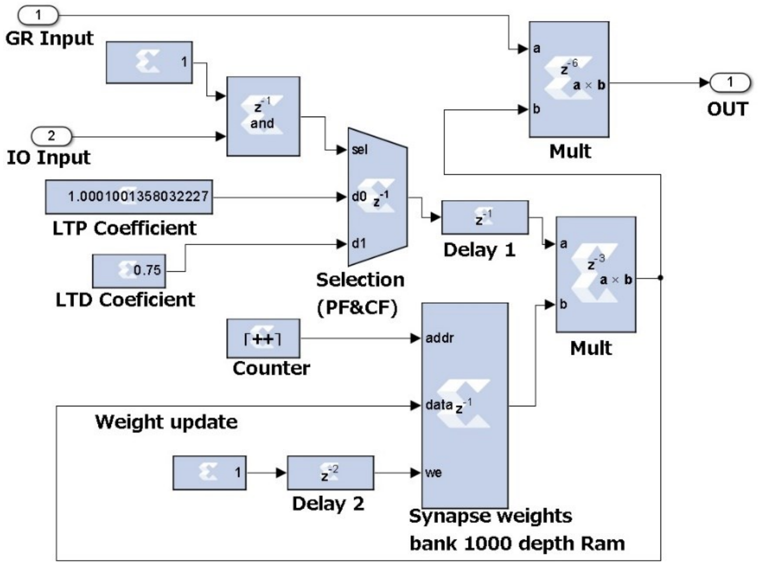


Figure 7.17: LTD and LTP implementation

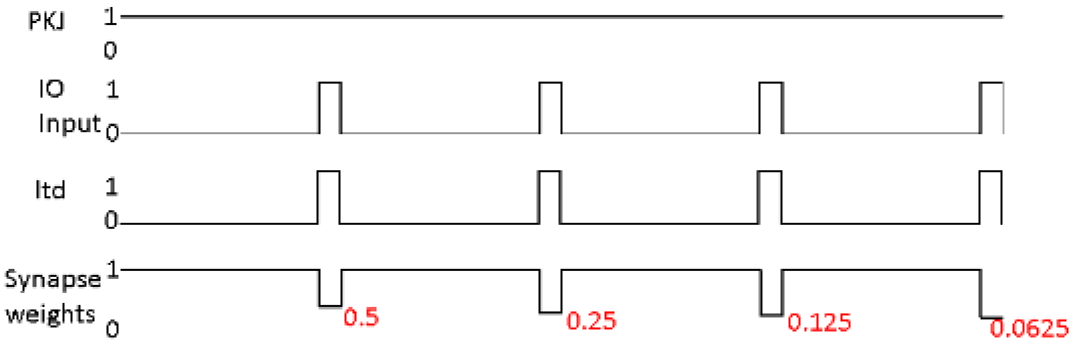


Figure 7.18: Learning with IO input signal with LTD coefficient of 0.5

The author implements the bio-realistic cerebellum neural network with the FPGA board for testing. The hardware design specification is shown in Table 7.4. After testing with the FPGA board, the author performs a hardware co-simulation in Matlab with an input signal of 0.6 seconds, 1.2 seconds and 2.5 seconds respectively as short, medium and long learning sounds from IO.

Table 7.4 FPGA Design Specification

Device Utilization Summary			
Number of	Used	Available	Utilization
LUTs Slice	4952	6822	72%
LUT Flip Flop	11652	15983	72%
MUXCY	12432	13644	91%
Ram	40	116	34%
Timing Constrains			
Maximum Period	61.663 ns		
Maximum Frequency	16.217MHz		

7.7 Implementation to the Talking Robot

The author defines the timing constraint of the talking robot as the shortest duration of a sound that the robot can create given the limitation of the motor’s response and also the influence of the sound dynamics of the system. The author uses a command packet to drive the pitch motor, which controls the airflow to the robot, to instantly open and close the air intake to the robot at its maximum capability. The robot then generated a sound that lasted 120 milliseconds. The author considers this value as the timing constraint of the robot, which shouldn't influence the actual timing function calculated by the bio-realistic cerebellar neural network. For example, if the network generates a timing output of 0.6 seconds then the actual time to generate a sound by the robot should be 0.48 seconds.

The author uses the built-in Matlab function to pause the pitch motor for a duration, which equals the cerebellum timing minus the robot timing constraint. The timing was extracted from a sound recorded at 8000Hz. The author creates three 8000Hz sound signals of 0.6 seconds, 1.2 seconds and 2.5 seconds respectively as short, medium, and long sound inputs for the cerebellum neural network. The signals were transformed to Poisson spike signals of 30Hz with 1000 times stretching for each sample using the pipelining technique. The natural activity level of neurons in rats is 5 Hz (Freeman & Muckler, 2003) [86]. The neural signal consists of two components, transient and sustained parts. The firing rate of the sustained component is approximately 30 Hz. Therefore, the author uses a 30Hz signal to represent the neural activity in the audio cortex of the brain, which is transmitted to the cerebellum as a learning signal. Co-simulation with FPGA hardware was then performed to generate an output pattern, the neuron activity. Supervised learning behavior of the cerebellum was reported by Knudsen (1994) [87], and Kawato (2011) [88]. A baby learns to speak through multiple listen and repeat trials. The learning includes vocalic sound as well as sound rhythm. The sound signal is transformed into a neuron signal in the ear and is sent to the brain for processing. In the simulation, the supervised learning signal is coming from the sound input, which is the sound that people hear.

The author hypothesizes that the human brain creates multiple calls back of the input signals thereby updating the synaptic weight at PF accordingly. In our simulation, the author uses a weight update coefficient of 0.75, which is a super-fast learning rate in comparison to the real brain model. As the similarity index shows in Figure 7.16 if the time shift is small the firing pattern of one GR cell is almost the same. This means the final learning state is not depending on the initial condition and the learning rate. The results of short duration learning of PKJ and DCN for a 0.6-second learning input are shown in Figure 7.19 (A) and (B) after the 5th trial training, respectively. The high activities region indicates the timing in the DCN output signal. After that, the signal is output to the Matlab workspace where the author uses a short-time energy function to decode the timing from this signal.

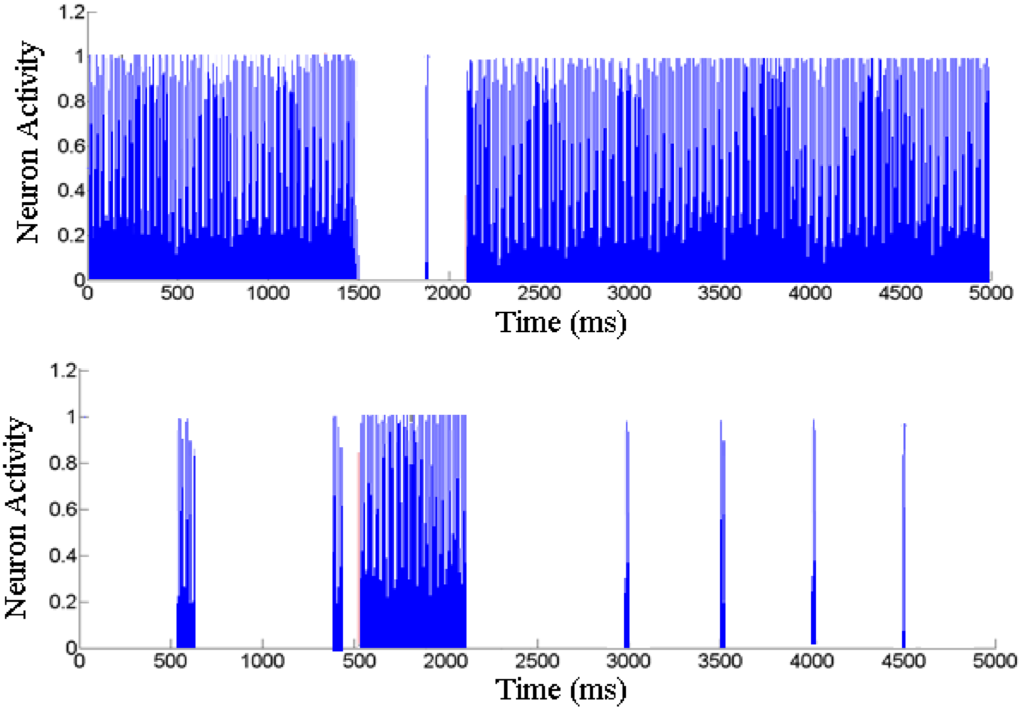


Figure 7.19: PKJ (A) and DCN (B) activities with 0.6 seconds training signal

In order to analyze the timing output pattern encoded at DCN, the short-time energy of DCN was calculated based on equation 5.11 of section 5.2. Figure 7.20 shows the short-time energy analysis for the timing output of the talking robot. If the value of E_n is greater than the threshold $\emptyset_{th}$ (equation 7.9), the value is kept; otherwise, it is replaced with zero (equation 7.10). The tracing technique was used to detect the range of E_n, which has a non-zero value indicating the duration timing of speech.

$$\emptyset_{th} = \frac{E_{max}(n)}{4} \tag{7.9}$$

$$E(n) = \begin{cases} E(n) & E(n) \geq \emptyset_{th} \\ 0 & E(n) < \emptyset_{th} \end{cases} \tag{7.10}$$

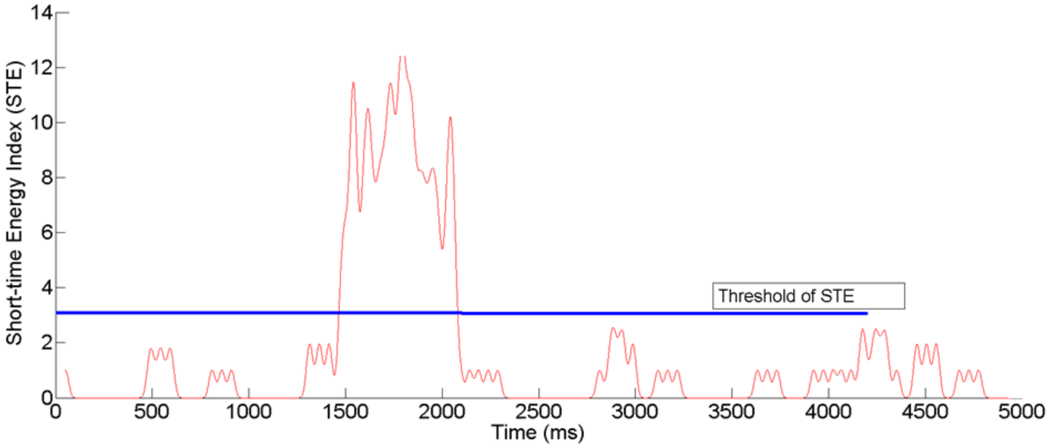


Figure 7.20: Short-time energy analysis for timing output

In Figure 7.21, the output sounds of the talking robot for both the 0.6 second and the 1.2-second input learning signals, have the same duration as the learning signals (see Figure 7.21 (A) and (B)). However, for the 2.5-second input learning signal, the talking robot generated a sound that lasted 5 seconds, shown in Figure 7.21 (C). As predicted the activated or inactivated granular cluster signal for a short duration (200 milliseconds to 1200 milliseconds) will have the same output. However, with a very long input signal of 2.5 seconds, all the weights from GR to PKJ are depressed, so there is no activity at PKJ. Due to inhibition from PKJ to DCN, the DCN will have a sustained signal, which in our case is 5 seconds. This verifies our hypothesis of short-time learning as described in section 7.2.1 above.

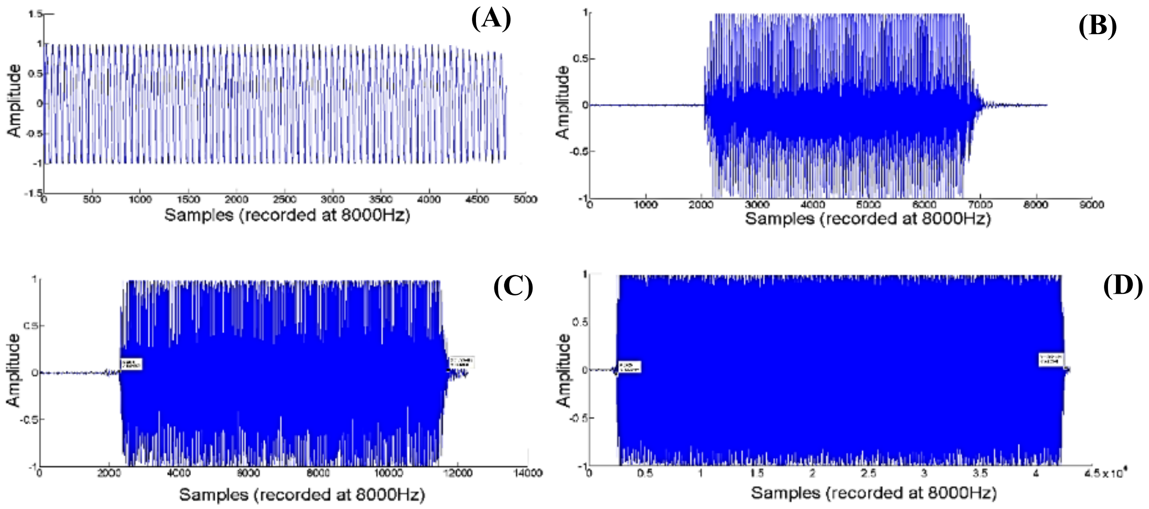


Figure 7.21. Human input voice (A) for 0.6 seconds and robot regenerated voice for (0.6 (B), 1.2 (C), 2.5 (D) seconds voice input)

The author conducts an experiment for timing perception of the talking robot using this cerebellum-like neural network. The author tests the capability of the talking robot in generating two very similar words in Japanese which are */ie/* and */iie/*. The sound */ie/* means "house" while the sound */iie/* means "disagreement." For these two words, the timing duration in the speech indicates the difference. First of all, the author records his speech of the sounds */ie/* and */iie/* as shown in Figure 7.22 and 7.24, respectively. Then, he applies the technique of phoneme separation and counting as introduced in section 5.3 to get individual sound element in speech. Each element is fed to the neural network for estimating the timing duration. The sound template matching introduced in section 5.5 is used to determine the motor parameters to generate the sound. The acoustic characteristic and timing duration for each sound element are processed in orderly looping sequence. By combining the motor parameters and timing duration, the talking robot generates an output sound as indicated in Figure 7.23 and 7.25. The timing duration for each sound in the talking robot speech is very similar to human sounds. The results indicate that the cerebellum-like works properly for short-range timing perception of the talking robot.

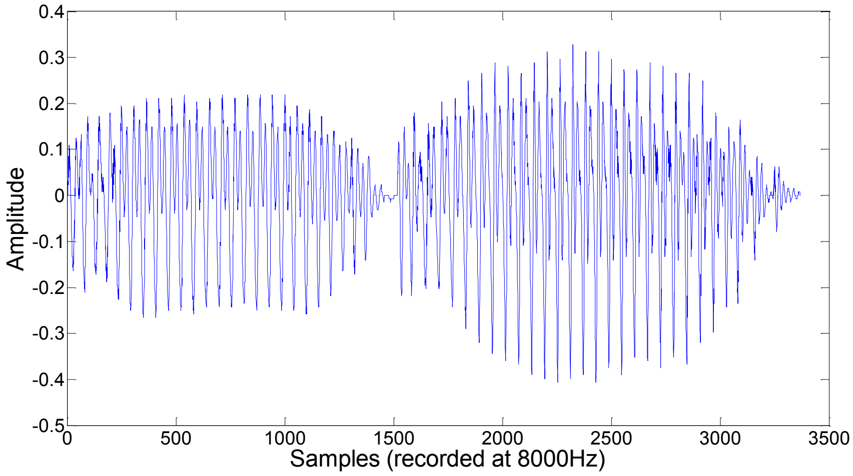


Figure 7.22: Human */ie/* sound waveform

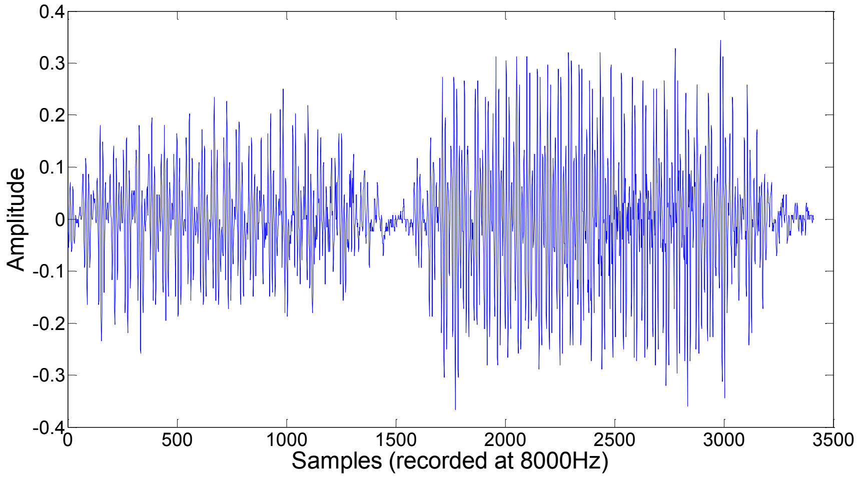


Figure 7.23: Robot */ie/* sound waveform

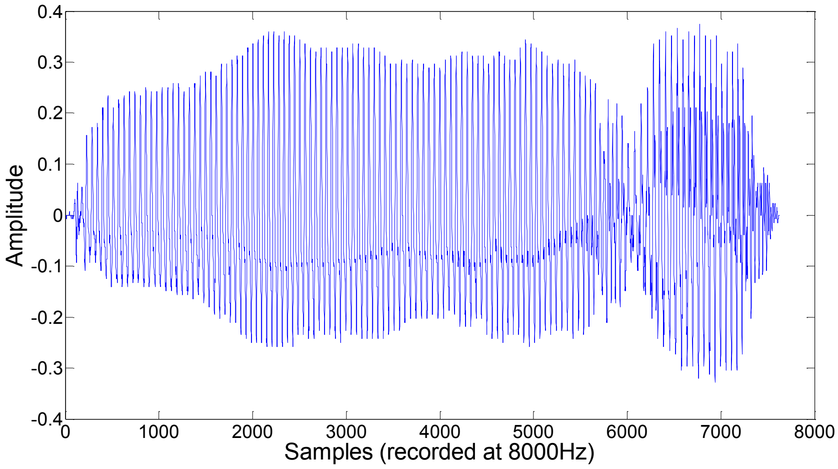


Figure 7.24: Human */iie/* sound waveform

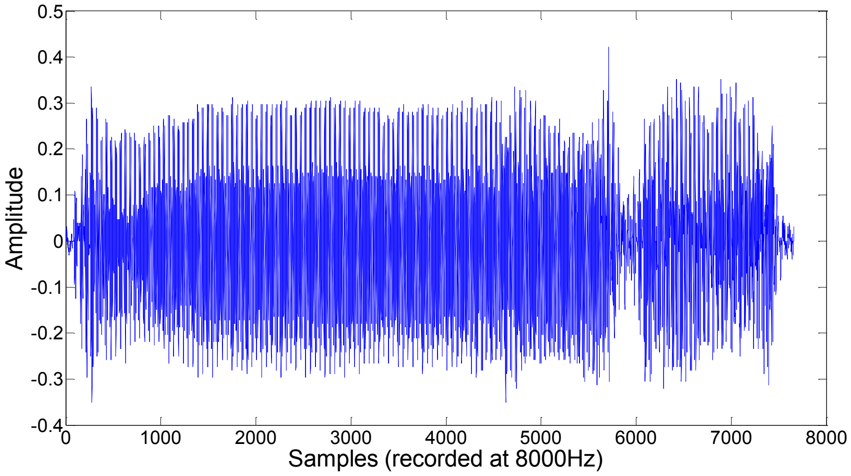


Figure 7.25: Robot */iie/* sound waveform

In this chapter, the author has proposed a short-timing function for the talking robot based on a cerebellar neural network model. The cerebellum has been known for its role in the precision, coordination and accurate timing of motor control according to various researchers such as Ivry and Spencer (2004) [14], Lewis and Miall (2003) [56], Ivry, Spencer, Zelaznik, and Diedrichsen (2002) [58]. For this reason, building a timing model based on cerebellum anatomy was used in this study. Casper, Raphael, Harris, and Geibel (2007) [89] found that the speech prosody of patients with cerebellar ataxia was significantly different from that of healthy speakers. Although the individual phonemes spoken by patients with cerebellar ataxia were similar to healthy speakers, the rhythm, and duration when they generated speech was different. Therefore, when building a speech synthesis system for long phrase generation, the timing function must be treated separately. Guenther, Ghosh, and Tourville (2006) built a software-based speech synthesis system that could determine the articulation of speech units including the time interval of the articulator's movements [25]. However, details of how the neural network generated time intervals were not clearly stated in their study. Furthermore, Bernd and his team (2014) developed a computer simulated model for speech synthesis based on action repository using a neural network mapping technique [26]. In their system, the timing was treated as one parameter while it should have been addressed as a separate function. For most of the hardware-based speech synthesis systems, the main focus was on individual phoneme vocalization. Nishikawa et al. (2002) built an anthropomorphic robot that can produce Japanese consonant and vowel sounds [90]. The motor positions and time intervals were calculated based on MRI images but the system was manually controlled by a human, the timing characteristic of prosody features is not mentioned in this study. Yoshikawa et al. (2007) built a talking robot which could produce certain vowels such as */a/, /i/, /u/,* and */e/,* but not */o/*, due to the motor's limited degrees of freedom movements. A self-organizing mapping technique had been used for the autonomous vocalization of this robot, but the time interval was not mentioned in their study [27].

Unlike other mechanical vocalization systems, the author has offered a more biologically plausible model for controlling the timing function for a mechanical vocalization system based on cerebellar autonomy. A computer simulated cerebellar neural network model was proposed by Yamazaki and Tanaka to represent Pavlovian eyeblink conditioning [65], but the simulation took a very long time due to the sequential processing of the computer algorithm. Luo and his team used the cerebellar neural network model with an FPGA board for real-time application [67]. Nevertheless, the study was only a proposal method and used Pavlovian eyeblink conditioning to verify the results. The short-time learning capability of the cerebellar neural network has not yet been clearly stated nor explained. From, the study of Ivry et al. [14], and Medina et al. [91], the author proposed that the cerebellum is responsible for the regulation of short-time learning abilities of a human. The author treated the sound input signal as a learning signal and a fixed 5 second input signal as the predictive signal. While the learning input signal for other cerebellar neural network models was a discrete spike to represent the eye blink condition experiment, our learning input signal was a sequence of spikes pre-processed from human sounds, which is important for short-range timing. The predictive signal was taken as a fixed 5 seconds in our study, however, the signal could have been a little longer or shorter and the output results would have been similar due to the temporal patterns of the granular cells and the learning signal from the sound input having a major effect on the output signal but not the predictive signal

CHAPTER 8

Conclusions and Future Research

8.1 Conclusions

This dissertation consists of a series of studies on prosodic features in a mechanical vocalization system together with a cerebellum-like neural network and its application to control timing characteristics in the output speech. Mechanical vocalization is an interesting topic in speech synthesis; however, it is also a challenging area of research, as it involves the most complex mechanism of the human body to analyze, the vocalization system. Almost all mechanical vocalization systems focused on the regeneration of a single phoneme or syllable and applied advanced methods such as SONN to train the system to get a better sound output in term of naturalness and clarity. The prosodic features in speech are not paid much attention. Thus, in this study, the authors focus on the aspect regarding the prosody of the output sound of the talking robot. The study includes two parts which are the improvement of the current talking robot and the development of new algorithms and a control system for mechanical vocalization systems to generate the output sound with the prosodic features. Some significant contributions of this study include:

In comparison with the previous version of the talking robot, the author had developed the new voiceless system for the generation of fricative sounds. The talking robot with the voices system was programmed to generate 5 basic Japanese fricative sounds. The result of fricative generation indicates that the robot with this system is able to generate fricative sounds at a certain level. This is significant in hardware-based speech synthesis, especially when synthesizing a foreign language that contains many fricative sounds. However, the limitation on the fricative sound system is that the robotic sound has shorter transition part in comparison to the human one. Thus, the result of the fricative phonetics of the talking robot is a little less clear than can be recognized. This limitation comes from the motor speed that lets the air rush through the vocal tract too fast. Also, the vocal tract of the mechanical system is not completely sealed; thus, there is air leakage during the unvoiced sound generation process, and it disturbs the output of fricative sound. In human vocalization, the tongue movements also contribute to the production of fricative sounds; however, the tongue action in this system only has two positions, up or down; thus, fricatives generated by the talking robot do not have the same characteristic as in a human yet.

The second contribution and also the originality of this study is a design and construction of new vocal cords for the talking robot. The intonation and pitch are two important prosodic features which are determined by the artificial vocal cords. In this study, the author successfully designed and implemented the new vocal cords in the talking robot which is a human-like mechanical vocalization system. The newly redesigned vocal cords greatly increase the speaking capability of the talking robot, especially for its singing performance. The fundamental frequency can be easily adjusted from a much lower frequency to a much higher frequency by just changing the shape of the rubber band. The range of fundamental frequency is from around 50Hz to 250Hz depending on the combination of rubber band shape, input pressure, and vocal tract shape. By comparing with the vocal cords of Waseda Talker, the most advanced mechanical vocalization system at the current, this vocal cords mechanism is much simple (1DOF vs. 4DOF of WT-R7II), but the fundamental frequency of our robot is much wider (50-250Hz vs. 129-220Hz of WT-R7II). The design of the artificial vocal cords is novel and unique; however, it is difficult to assemble the rubber band to stick in the rotational rod due to its small size. To fix one end of the rubber band on the rod, the author used an epoxy mix type adhesive but sometimes slipperiness occurs between the rubber band and the rod, and it reduces the accuracy of the pitch of the sound source. Due to assembly difficulty, the fundamental frequency of 50-250 Hz is the result of safe condition experiment which means the author only tested the motor angle up to 100 degrees while its maximum value is 150 degrees. The maximum fundamental frequency it can achieve is around 350 Hz under the extreme conditions, which means the motor angles can rotate up to 150 degrees and the rubber band is at its maximum tension. The significant contribution of these vocal cords to the speech synthesis field is that it provides the widest pitch range for a hardware-based system thus far. It greatly increases the synthesizing capability of the system, especially for its singing performance.

The third contribution of this study is to build a real-time interaction system between a human and a talking robot, which is a highly useful system for setting the robot to speak different languages. Also, based on acoustic resonance theory, an original formula about the formant frequency change due to vocal tract motor movements is derived, and a strategy for tuning the motors value to the formant frequency difference is proposed. This is also the first study as far as the authors are aware of using Matlab to control and receive real-time feedback from command type servo motors. This is a valuable approach because servo command type motors have built-in feedback and are very fast and stable. In this study, the robot was programmed to speak Japanese phrases and a foreign language phrase based on a fundamental set of consonants and vowel vectors. As expected, the speaking action of the robot for a Japanese term was reasonable and recognizable for a human. However, when the robot was set to speak a foreign language, the result was not clear enough for a human to recognize, due to the non-linear aerodynamics of continuously speaking that is different than single sound utterances. Therefore, the robot motor values needed to be adjusted to obtain a recognizable sound from the robot. A real-time interaction system, which allowed a user to change the robot motor values by GUI sliders based on formant frequency comparison strategy, was built, tested, and proven to improve the output voice of the talking robot when reproducing foreign language. It is a quick technique to get a reasonable result.

The fourth contribution is the development of a sentence repeating system for the talking robot. The ability to mimic human vocal sounds and reproduce a sentence is also an important feature for a speech synthesis system, but not implemented in other mechanical vocalization systems thus far. The novel technique of phoneme segmentation and counting worked almost perfectly for the sequence of vowel reproduction since the subjects in the test were required to pause a little between vowels. The direct cross-correlation technique gave the least accurate result because a human could speak the same vowels with different sound pitch and amplitude. These sound characteristics affected the similarity score when applying cross-correlation. The LPC technique also had the effect of sound pitch and amplitude. However, since it only took a short range of sound signals for extracting the coefficients, the effect was reduced and gave a better accuracy than the direct cross-correlation technique. The PARCOR analysis and the formant frequencies comparison delivered the highest accuracy since their coefficients were extracted from resonance frequencies, which only depended on the vocal tract shape. Also, a human ear distinguishes sound by detecting its resonance frequencies, which is similar to formant frequencies detection. Therefore, using formant frequencies comparison was the most similar process to the human auditory mechanism, and it was applied by numerous speech synthesis researchers. The high accuracy outcome of applying formant frequency comparison techniques for vowel identification was verified in this study.

The final contribution point is about the timing characteristic, which greatly influences the prosody of the output speech. The author built a bio-realistic neural network modeling the cerebellum to play a role as the internal timing function for the talking robot's control system. This is the novel and first study, which uses a bio-realistic timing function separately from the articulator control function of the talking robot. This is a significant contribution to building an intelligent mechanical vocalization system because the timing and the articulator's position are controlled separately by the human brain. The FPGA implementation of a bio-realistic neural network is a major step for employing a human-like controller to handle a human-like mechanism. The proposed assumption of short-time learning capability of the cerebellar neural network was demonstrated by the sounds produced by the talking robot. The similarity index of GR neuron activities with respect to time shift in our model showed a 28% difference compared with the Yamazaki and Tanaka model [65], which was five times larger. This indicated that even though our model was significantly smaller, it still displayed timing learning capability. The pipelining technique was applied to save hardware resources, and with the 1000 latency of FIFO and the maximum period of 61.663 nanoseconds, the system achieved a performance of fewer than 62 milliseconds to simulate a 1-second real-time action when connected to an FPGA device. This is sufficient for real-time application and real-time speech regeneration of the talking robot. The sounds generated by the talking robot have duration similar to the input sounds for short-range timing.

8.2 Future Researches

The author has developed two important improvements for the mechanical system and four new functions for the talking robot in order to improve the prosody of its output speech. However, the current system still has some limitations as described in section 8.1 above and require further study and enhancement in both mechanical and software portions. Therefore, the author proposed several important points in the future research in order to improve prosodic features in the output speech of the talking robot.

For the mechanical system, the tongue movement is very limited as it can only have an up or down position while in a human it has very complex movement. Thus, the tongue movement mechanism needs to be investigated and studied. One possible option is to employ a snake-like robot mechanism to the internal body of a silicone tongue so it could move with wider range and position. For the elongation and abbreviation of the tongue, which are very important to clearly generate some plosive sounds, such as */t/, /d/,* and */k/,* and fricative sounds, such as */z/* and */sh/,* the author considers applying the flexible neck mechanism of a straw to the end of a silicone tongue so the tongue can extend or retract.

The vocal tract shape of the current talking robot is a long, uniform, hollow tube which is not the same shape as a human. Thus, the author proposes the future design of the vocal tract which is built with the same silicone material and formed to the same shape as the human vocal tract. The molding for the new vocal tract plans to apply 3D-printing technology in the future. The author proposes using wax as the material of the mold for the new vocal tract. The silicone after drying inside the wax mold will be put into hot water, and the wax will be dissolved, thus the vocal tract will be formed.

As the author observed in the sound generation experiments, the motor speed is not high enough to rapidly open the vocal tract to generate some plosive sounds like the */p/* sound. Thus, the actuators for the vocal tract also need to be upgraded with more torque and a faster speed motor. This will improve the ability to generate plosive sounds of the talking robot.

For FPGA implementation of the cerebellum-like neural network, hardware limitations with the Xilinx SP-605 FPGA board used in this study meant that some parts of the cerebellum were ignored such as stellate cells and basket cells, reducing the plausibility of a bio-realistic neural network. Moreover, the robot was only tested with single vowel sounds. The next step will be to conduct a sequence of sounds or a sentence. In future, the author plans to improve the design by using more advanced FPGA boards to improve the neural network system. The success of this study will open the opportunity for the talking robot to be a test system for an articulatory speech-prosthesis system for animals or humans.

The talking robot is able to generate a sequence of vowels but not consonant sounds yet. Thus, it is necessary to improve the current sound analysis algorithms in order to detect vowels and consonants. Also, a recognition algorithm for consonants needs to be developed. Due to the highly unstable signal of the consonant sounds, it is very difficult to use template matching as applied to the vowel sound. Thus, the author proposes to use an artificial neural network for detecting the consonant sounds within speech. This will improve the capability of the talking robot in repeating human speech.

Almost all of the mechanical vocalization systems are trained to vocalize based on mother-infant or infant-caregiver approaches autonomously. In this approach, the mechanical vocalization system is treated like a baby who tries to listen to their mother's sounds, mimic the speech, and obtain their articulatory movement through many repeats and trial and error. This approach is mostly applied for the vocalization system which has an empty articulatory database. Also, SOM is the most widely used tool for training and obtaining the articulatory information. The issue of this approach is it relies on the language of the mother or caregiver, and the vocalization system can only regenerate this language, the ability of this system when it is put to the task of generating an entirely new sound is not discussed in any system so far. This approach is good for the initial learning phase in order to establish a vocabulary bank or articulation database for the system. This is also the hypothesis model for the way children learn a new language. However, it is very different from the way an adult learns to speak. The adults heavily rely on their mother tongue and vocabulary in order to construct a new articulatory motion to generate a new language sound or an unknown sound. Therefore, the author proposes a new learning type for the talking robot called the adult-self learning approach. In this approach, the robot learns to adjust its vocalization movements to produce output speech that has similar formant frequencies to the target speech of the foreign language. In detail, the articulation database of vowels and consonants is initially established by the SONN technique. Then, the robot is allowed to produce a foreign language sound by using this database. At this stage, the output speech has the native language characteristic. The formant frequencies of the output speech are then compared with those of foreign language target speech. An adaptive algorithm, which based on the vocal tract adjustment depending on formant frequency difference as introduced in section 4.3, is developed to adjust the motor movements of the talking robot in order to modify the output speech. The modified speech will have the changes in formant frequencies that come close to those in the target speech. Because the formant frequency changes are related to each vocal tract section change and the interrelationships between vocal tract sections in order to prevent the silicon resonance tube from tearing apart, the vocal tract section for adjustment must be determined. Thus, a fuzzy neural network, which is used to determine the relative position for each vocal tract section for optimal adjustment, is also developed. As a result, this adult-self learning system will train the talking robot to vocalize a foreign language in the same way an adult learns to speak a new language.

List of References

[1] J. L. Flanagan, *Speech Analysis Synthesis and Perception,* Springer-Verlag. 1972

[2] E.R.Truitt, "Talking Heads: Astral Science, Divination, and Legends of Medieval Philosophers*,"* *Medieval Robots: Mechanism, Magic, Nature, and Art,* Philadelphia University of Pennsylvania Press, pp. 69–96, 2015.

[3] V.N. Thanh and H. Sawada, “A Talking Robot and Its Real-time Interactive Modification for Speech Clarification." *SICE Journal of Control, Measurement, and System Integration*. No: 10. Pages: 251-256, 2016.

[4] T. Higashimoto and H. Sawada: "A Mechanical Voice System: Construction of Vocal Cords and its Pitch Control," *International Conference on Intelligent Technologies*, pp. 762-768, 2003.

[5] H. Sawada and M. Nakamura: "Mechanical Voice System and its Singing Performance," *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1920-1925, 2004

[6] Mitsuki Kitani, Tatusya Hara, Hiroki Hanada and Hideyuki Sawada, “Robotic Vocalization Training System for the Auditory-impaired,” *International Conference on Disability, Virtual Reality and Associated Technologies,* pp.263-272 2010

[7] M. Kitani, T. Hara, and H. Sawada, "Voice articulatory training with a talking robot for the auditory impaired,” *International Journal of Disability and Human Development*, Vol.10, No.1, pp.63-67, 2011.

[8] J. Fletcher*, "The Prosody of Speech: Timing and Rhythm,"* The Handbook of Phonetic Sciences, 2nd ed., pp. 523–602, 2010.

[9] J. B.Pierrehumbert and M. E. Beckman*, Japanese Tone Structure*, Linguistic Inquiry Monograph Series No. 15, MIT Press, 1988

[10] P. Karel , *MATLAB for Engineers - Applications in Control, Electrical Engineering, IT and Robotics*, InTech, 2011

[11] H. A. Phạm, *Vietnamese Tone – A New Analysis*, New York Routledge, *2003*

[12] V.N. Thanh and H.Sawada, “Automatic Vowel Sequence Reproduction for a Talking Robot Based on PARCOR Coefficient Template Matching," *IEIE Transactions on Smart Processing and Computing*. No: 5. pp. 215-221, 2016.

[13] H. Ackermann, “Cerebellar contributions to speech production and speech perception: psycholinguistic and neurobiological perspectives,” *Trends in Neurosciences*, Vol 31(6), pp.265-72. 2008

[14] R. B. Ivry and R. M. Spencer, “The neural representation of time,” *Current Opinion in Neurobiology*, Vol 14(2), pp. 225–232, 2004.

[15] R.T. Beyer, *Sounds of Our Times: Two Hundred Years of Acoustics*, Springer-Verlag New York,1999.

[16] H.Dudley and T.H Tarnoczy, “The Speaking Machine of Wolfgang Von Kempelen” *The Journal of the Acoustical Society of America*, Vol 22, No 2, pp. 151-166, 1950.

[17] L.Rabiner and B.H.Juang, “Fundamental of speech recognition,” Prentice-hall. Englewood Cliffs. New Jersey, 1993.

[18] K. Osuka, Hiroyuki Araki, K. Sawada and T. Ono, "For the Realization of Mechanical Speech Synthesizer - Realization of Three-Dimensional Shapes of Articulatory Organs," *Journal of the Robotics Society of Japan Vol*. 16, No. 2, pp. 189-194, 1998

[19] X. Rodet and G. Benett, "Synthesis of the Singing Voice," *Current Directions in Computer Music Research*, PIT Press, 1989

[20] K. Hirose, "Current Trends and Future Prospects of Speech Synthesis," *Journal of the Acoustical Society of Japan*, pp. 39-45, 1992

[21] Ph. Depalle, G. Garcia and X. Rodet, "A Virtual Castrato," *International Computer Music Conference*, pp. 357-360, 1994

[22] J.O. Smith III, "Viewpoints on the History of Digital Synthesis," *International Computer Music Conference*, pp. 1-10, 1991

[23] Markel, J.D., and Gray, A.H., Linear Prediction of Speech, *Springer-Verlag, New York*, 1976

[24] N. Umeda and R. Teranishi, "Phonemic Feature and Vocal Feature -Synthesis of Speech Sounds Using an Acoustic Model of Vocal Tract," *Journal of the Acoustical Society of Japan*, Vol.22, No.4, pp. 195-203, 1966

[25] F.H. Guenther, S.S.Ghosh, and J.A. Tourville, “Neural modeling and imaging of the cortical interactions underlying syllable production,” *Brain and Language*, 96, pp. 280–301,2006

[26] B.J. Kröger, J. Kannampuzha, E. Kaufmann, “Associative learning and self-organization as basic principles for simulating speech acquisition, speech production, and speech perception,” *EPJ Nonlinear Biomedical Physics* Vol 2 No.1, 1-28, 2014.

[27] Y. Yoshikawa, K. Miura, and M. Asada, “Unconscious Anchoring in Maternal Imitation that Helps Finding the Correspondence of Caregiver's Vowel Categories.” *Advanced Robotics,* Vol.21, No.13, pp.1583-1600, 2007.

[28] N. Endo, T. Kojima, H. Ishihara, T. Horii, and M. Asada, “Design and preliminary evaluation of the vocal cords and articulator of an infant-like vocal robot ‘Lingua,” *IEEE RAS Int. Conf. Humanoid Robots*, pp. 1063–1068, 2014.

[29] K. Nishikawa, H. Takanobu, T. Mochida, M. Honda and A. Takanishi, "Development of a New Human-like Talking Robot Having Advanced Vocal Tract Mechanisms," *IEEE/RSJ International Conference on Intelligent Robot and Systems*, pp. 1907-1913, 2003.

[30] J. Vreeken, “Spiking neural networks, an introduction,” *Technical Report UU-CS-2003-008 Institute for Information and Computing Sciences, Utrecht University,* pp.1-5, 2002.

[31]] Samanwoy Ghosh-Dastidar, Hojjat Adeli, “Third Generation Neural Networks: Spiking Neural Networks,” *Advances in Computational Intelligent*, Springer-Verlag Berlin Heidelberg, pp. 167–178, 2009.

[32] Sander M. Bohte, Joost N. Kok, “Applications of Spiking Neural Networks,” *Journal of Information Processing Letters*,Vol. 95, no. 6, pp. 519-520, 2005.

[33] R. Apps, M. Garwicz, “Anatomical and physiological foundations of cerebellar information processing,” *Nature Reviews Neuroscience*, Vol. 6, pp. 297-311, 2005.

[34] D. Bullock, J.C. Fiala, and S. Grossberg, “A neural model of timed response learning in the cerebellum,” *Neural Network*, 7, 1101–1114, 1994.

[35] R. FitzHugh, “Impulses and physiological states in theoretical models of nerve membrane,” *Biophysical Journa.,* Vol. 1, pp. 445–466, 1961.

[36] D. Sterratt, B.Graham and A. Gillies, “Principles of Computational Modelling in Neuroscience,” *Cambridge University Press*, 2011.

[37] W. Gerstner and W.M. Kistler, “Spiking Neuron Models. Single Neurons, Populations, Plasticity”, *Cambridge University Press*, 2002.

[38] D. Gargouri, M. A. Kammoun, and A. B. Hamida, “A Comparative Study of Formant Frequencies Estimation Techniques,” *Proceedings of the 5th WSEAS. International Conference on Signal Processing*, pp. 15-19, May 2006.

[39] D. Y. Loni, S. Subbaraman, "Formant estimation of speech and singing voice by combining wavelet with LPC and Cepstrum techniques," *9th International Conference on Industrial and Information Systems (ICIIS)*, pp. 1-7, Dec 2014.

[40] V.N. Thanh and H. Sawada, "A Spiking Neural Network for Short-range Timing Function of a Robotic Speaking System," *The 3rd International Conference Proceedings on Control, Automation and Robotics (ICCAR 2017)* , In press, 2017.

[41] V.N. Thanh and H. Sawada, "Vietnamese Language Speech Performance by the Talking Robot," *Electrical and Related Engineers Shikoku Section Joint Conference Proceedings*, pp. 82, 2016.

[42] V.N. Thanh and H. Sawada, "Autonomous Vowels Sequence Reproduction of a Talking Robot Using PARCOR Coefficients*," Proceedings of International Conference on Electronics, Information and Communication (ICEIC2016)*, pp. 14-18, 2016.

[43] V.N. Thanh and H. Sawada, "Speech Performance of Fricative Sounds by the Talking Robot," *6th Symposium between Kagawa University and Changmai University*, pp. 109, 2016.

[44] V.N. Thanh and H. Sawada, "Comparison of Several Acoustic Features for the Vowel Sequence Reproduction of a Talking Robot," *Proceedings of IEEE International Conference on Mechatronics and Automation*, pp. 1137-1142, 2016.

[45] V.N. Thanh and H. Sawada, "Automatic Vowel Sequence Reproduction for a Talking Robot Based on PARCOR Coefficient Template Matching," Journal *IEIE Transactions on Smart Processing and Computing*, Vol 5, Pp. 215-221, 2016.

[46] V.N. Thanh and H. Sawada, "A Real Time Visualization System for Articulatory Analysis of the Talking Robot," *Electrical and Related Engineers Shikoku Section Joint Conference Proceedings*, pp. 86, 2015.

[47] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 2000.

[48] C. S. Roy,"Formant Location from LPC Analysis Data," *IEEE Transactions on Speech and Audio Processing*, pp. 129–134, 1993.

[49] R.G. Bachu, S. Kopparthi, B.Adapa and B.D. Barkana, "Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy," *Advanced Techniques in Computing Sciences and Software Engineering,* pp.279-282, 2010.

[50] S. Nandhini, A. Shenbagavalli "Voiced/Unvoiced Detection using Short Term Processing" *International conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, pp. 39-43, 2014.

[51] B.S.Atal and S.L.Hanauer: "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, Vol. 50, pp. 637-655,19712

[52] J. Makhoul,"Linear prediction: A tutrial review," P*roceedings of the IEEE*, 63, 4, pp.561-580(1975)

[53] J.D. Markel andA.H.Gray, *Linear Prediction of Speech*, Springer-Verlag, 1976

[54] P. Taylor, *Text-to-speech synthesis,* Cambridge University Press, 2009

[55] C. Maxfield, *The Design Warrior's Guide to FPGAs: Devices, Tools and Flows*, Elsevier, 2004

[56] P.A., Lewis and R.C. Miall, "Distinct. Distinct systems for automatic and cognitively controlled time measurement: evidence from neuroimaging," *Current Opinion in Neurobiology*, Vol. 13, pp. 250-255, 2003

[57] C. Boyd, "Cerebellar agenesis revisited," *Brain*; vol. 133, pp. 941–944, 2010.

[58] R.B. Ivry, R.M. Spencer, H.N. Zelaznik and J. Diedrichsen, "The Cerebellum and Event Timing," *Annals of the New York Academy of Sciences*, Vol. 978, pp. 302-17, 2002.

[59] S. S. Ghosh, J. A. Tourville, and F. H. Guenther, "A Neuroimaging Study of Premotor Lateralization and Cerebellar Involvement in the Production of Phonemes and Syllables," *Journal of Speech, Language, and Hearing Research,* Vol. 51, No. 5,pp. 1183–1202,2008.

[60] J. Schlerf, T. Wiestler, T. Verstynen and J. Diedrichsen, "Big Challenges from the Little Brain — Imaging the Cerebellum," *Advanced Brain Neuroimaging Topics in Health and Disease - Methods and Applications*, Ms. Danijela Duric (Ed.), InTech, DOI: 10.5772/58266.

[61] B. F. Chapeau and G. Chauvet, “A neural network model of the cerebellar cortex performing dynamic associations,” *Biological Cybernetics,* Vol. 65, 267–279, 1991.

[62] D. Bullock, J.C. Fiala, and S. Grossberg,” A neural model of timed response learning in the cerebellum,” *Neural Network*, Vol. 7, pp.1101–1114, 1994.

[63] A. Garenne and G.A. Chauvet, “A discrete approach for a model of temporal learning by the cerebellum: in silicon classical conditioning of the eye blink reflex,” *Journal of Integrative Neuroscience,* Vol.3, pp. 301–318, 2004.

[64] T. Yamazaki and S. Tanaka, “Neural modeling of an internal clock,” *Neural Computation* , Vol. 17, pp. 1032–1058, 2005.

[65] T. Yamazaki and S. Tanaka, “A spiking network model for passage-of-time representation in the cerebellum,” *The European Journal of Neuroscience*, Vol. 26, No. 8, pp. 2279–2292, 2007.

[66] S.A. Bamford, R. Hogri, A. Giovannucci, A.H. Taub, I. Herreros, P.F.Verschure, P. D. Giudice, “A VLSI Field-Programmable Mixed-Signal Array to Perform Neural Signal Processing and Neural Modeling in a Prosthetic System,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 20, No.4, 455-67, 2012.

[67] J. Luo, G. Coapes, T. Mak, T. Yamazaki, C. Tin and P. Degenaar , “A Scalable FPGA-based Cerebellum for Passage-of-Time Representation,” *Conference of the IEEE Engineering in Medicine and Biological Society*, pp. 3102 – 3105, 2015.

[68] E. L.Graas, E.A. Brown and H.L. Robert, “An FPGA-based approach to high-speed simulation of conductance-based neuron models,” *Neuroinformatics*, Vol. 2, No. 4, pp. 417–435, 2004.

[69] T.S. Mak, G. Rachmuth, K.P. Lam and C.S. Poon, “A component-based FPGA design framework for neuronal ion channel dynamics simulations,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 14, No. 4, pp. 410–418, 2006.

[70] K.Cheung, S. R. Schultz, W. Luk, “A Large-Scale Spiking Neural Network Accelerator for FPGA Systems,” In: Villa A.E.P., Duch W., Érdi P., Masulli F., Palm G. (eds), *Artificial Neural Networks and Machine Learning – ICANN. Lecture Notes in Computer Science*, Vol. 7552. Springer, 2012.

[71] L.K. Robert, *Data Structures & Program Design*, 3rd ed. Prentice-Hall, 1994.

[72] B. Koeppen and B. Stanton, *Berne & Levy Physiology*, 6th ed., Elsevier, 2009.

[73] C. Hofstötter, M.Mintz,P.F. M. J. Verschure, "The cerebellum in action: a simulation and robotics study," *European Journal of Neuroscience*, Vol. 16, No. 7, pp. 1361–1376, 2002.

[74] P.Fraisse, *The psychology of time,* New York: Harper & Row, 1963.

[75] S. W. Bao, L. Chen, J.J. Kim and R.F. Thompson, “Cerebellar cortical inhibition and classical eyeblink conditioning,” *Proceedings of the National Academy of Sciences,* Vol. 99, pp. 1592-1597, 2002.

[76] M.Gerwig, A.Dimitrova, F.P. Kolb, M.Maschke, B. Brol, A. Kunnel, D.Timmann , “Comparison of eyeblink conditioning in patients with superior and posterior inferior cerebellar lesions,” *Brain*, Vol.126, No. 1, pp.71-94, 2003.

[77] V.N. Thanh and H. Sawada, "Cerebellum-like Neural Network for Short-range Timing Function of A Robotics Speaking System,” *Proceedings of Conference on Control, Automation and Robotics (ICCAR)*, In press, Nagoya, April 22-24, 2017.

[78] R.J. Harvey and R.M. Napper, “Quantitative Study of Granule and Purkinje Cells in the Cerebellar Cortex of the Rat,” *The journal of comparative neurology*, Vol. 274 No. 2, pp. 151-157, 1988.

[79] G. L. Mwamengele, T. M. Mayhew, and V. Dantzer, “Purkinje cell complements in mammalian cerebella and the biases incurred by counting nucleoli,” *Journal of Anatomy*, Vol. 183, pp. 155–160, 1993.

[80] M.Palkovits, P. Magyar, and J. Szentagothai, “Quantitative histological analysis of the cerebellar cortex in the cat: I. Number and arrangement in space of the Purkinje cells,” *Brain Research,* Vol. 32, pp.1–13, 1971.

[81] R. L.Jakab, and J.Hámori, “Quantitative morphology and synaptology of cerebellar glomeruli in the rat,” *Anatomy and Embryology*, Vol. 179, pp. 81-88, 1988.

[82] M. Ito, “Long-term depression,” *Annual review of neuroscience*, Vol. 12, pp. 85-102, 1989.

[83] T.Hirano, “Depression and potentiation of the synaptic transmission between a granule cell and a Purkinje cell in rat cerebellar culture,” *Neuroscience Letters*, Vol. 119, No. 2, pp.141-144, 1990.

[84] S.K. Koekkoek, H.C. Hulscher, B.R.Dortland, R.A. Hensbroek, Y.Elgersma, T.J. Ruigrok, and C.I. De Zeeuw, “Cerebellar LTD and learning-dependent timing of conditioned eyelid responses,” *Neural Science*, vol. 301, pp. 1736-1739, 2003.

[85] E. D’Angelo, L.Mapellim, C. Casellato, A. Jesus, J.A. Garrido, N.Luque, et al., “Distributed Circuit Plasticity: New Clues for the Cerebellar Mechanisms of Learning,” *Cerebellum*, 15(2):139-51, 2016.

[86] J.H.J. Freeman and A.S. Muckler, “Developmental changes in eyeblink conditioning and neuronal activity in the pontine nuclei,” *Learning & Memory,* Vol.10, pp. 337–345, 2003.

[87] E.I. Knudsen, "Supervised Learning in the Brain,” *The Journal of Neuroscience*, Vol. 14, No. 7, pp.3985-3997, 1994.

[88] M. Kawato, S. Kuroda, and N. Schweighofer, "Cerebellar supervised learning revisited: biophysical modeling and degrees-of-freedom control," *Current Opinion in Neurobiology*, Vol.21, pp.1–10, 2011.

[89] M.A. Casper, L.J. Raphael, K.S.Harris and J.M. Geibel, “Speech prosody in cerebellar ataxia,” *International Journal of Language & Communication Disorders*, Vol. 42, No. 4, pp. 407-26, 2007.

[90] K. Nishikawa, A. Imai, T. Ogawara, H. Takanobu, T. Mochida and A. Takanishi, “Speech planning of an anthropomorphic talking robot for consonant sounds production,” *IEEE International Conference on Robotics & Automation*, pp. 1830–1835, 2002.

[91] J.F. Medina, K.S. Garcia, W.L. Nores, N.M. Taylor and M.D. Mauk, "Timing mechanisms in the cerebellum: Testing predictions of a large-scale computer simulation," *Journal of Neuroscience,* Vol. 20, pp. 5516–5525, 2000.

[92] Xilinx, SP605 Hardware User Guide, UG526 (v1.8) September 24, 2012. Available online at https://www.xilinx.com/support/documentation/boards_and_kits/ug526.pdf

[93] Xilinx, Vivado Design Suite Tutorial：Model-Based DSP Design using System Generator UG948 (v2013.1) March 20, 2013, Available online at https://www.xilinx.com/support/documentation/sw_manuals/xilinx2013_1/ug948-vivado-sysgen-tutorial.pdf

List of Publications

The Publications in Journals

1. Vo Nhu Thanh and Hideyuki Sawada (2016), "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. doi:10.5573/IEIESPC.2016.5.3.215

2. Vo Nhu Thanh and Hideyuki Sawada (2016), "A Talking Robot and Its Real-Time Interactive Modification for Speech Clarification." *In SICE Journal of Control, Measurement, and System Integration*, Vol. 9, No.6, 251-256. doi:http://doi.org/10.9746/jcmsi.9.251

The Publications in International Conferences

1. Vo Nhu Thanh and Hideyuki Sawada (2016), "Autonomous Vowels Sequence Reproduction of a Talking Robot Using PARCOR Coefficients." *In proceedings of International Conference on Electronics, Information and Communication* (IEEE ICEIC2016), pp.14-18. doi:10.1109/ELINFOCOM.2016.7563013

2. Vo Nhu Thanh and Hideyuki Sawada (2016), "Comparison of several acoustic features for the vowel sequence reproduction of a talking robot." *Proceedings of 2017 International Conference on Mechatronics and Automation (ICMA 2016),* pp. 1137-1142. doi:10.1109/ICMA.2016.7558722

3. Vo Nhu Thanh and Hideyuki Sawada (2017), "Cerebellum-like Neural Network for Short-range Timing Function of a Robotics Speaking System." *Proceedings of 2016 International Conference on Control, Automation and Robotics (ICCAR 2017),* pp. 184-187.

4. Vo Nhu Thanh and Hideyuki Sawada (2017), "Text-to-Speech of a Talking Robot for Interactive Speech Training of Hearing Impaired." *Proceedings of 10th International Conference on Human System Interaction (HSI 2017),* Accepted.

5. Vo Nhu Thanh and Hideyuki Sawada (2017), "Singing Performance of the Talking Robot with Newly Redesigned Artificial Vocal Cords" *Proceedings of 2017 International Conference on Control, Automation and Robotics (ICMA 2017),* Accepted.

The Publications in Domestic Conferences

1. Vo Nhu Thanh and Hideyuki Sawada (2015), “A Real Time Visualization System for Articulatory Analysis of the Talking Robot.” *In Electrical and Related Engineers Shikoku Section Joint Conference Proceedings. SJCIEE 2015,* pp.86.

2. Vo Nhu Thanh and Hideyuki Sawada (2016), “Vietnamese Language Speech Performance by the Talking Robot.” *In Electrical and Related Engineers Shikoku Section Joint Conference Proceedings*. *SJCIEE 2016,* pp.82.

3. Vo Nhu Thanh and Hideyuki Sawada (2016), “Speech Performance of Fricative Sounds by the Talking Robot.” *In The 6th Joint Symposium between Kagawa University and Changmai University, 6th KU-CMU Joint Symposium*, pp. 109.

Award

1. Best presentation award, Vo Nhu Thanh and Hideyuki Sawada (2017), “Cerebellum-like Neural Network for Short-range Timing Function of a Robotics Speaking System.” *Proceedings of Conference on Control, Automation and Robotics (ICCAR),* pp. 184-187.