IEIE Transactions on Smart Processing and Computing, vol. 5, no. 3, June 2016
http://dx.doi.org/10.5573/IEIESPC.2016.5.3.215

Automatic Vowel Sequence Reproduction for a Talking Robot Based on PARCOR Coefficient Template Matching

Vo Nhu Thanh[1,2] and Hideyuki Sawada[1]

[1] Graduate School of Engineering, Kagawa University / Japan thanhvous@gmail.com, sawada@eng.kagawa-u.ac.jp
[2] Faculty of Mechanical Engineering, Da Nang University of Science and Technology / Vietnam thanhvous@gmail.com

* Corresponding Author: Vo Nhu Thanh

Received April 20, 2016; Accepted May 24, 2016; Published June 30, 2016

** Extended from a Conference: Preliminary results of this paper were presented at the ICEIC 2016. This paper has been accepted by the editorial board through the regular review process that confirms the original contribution.*

***Abstract*:** This paper describes an automatic vowel sequence reproduction system for a talking robot built to reproduce the human voice based on the working behavior of the human articulatory system. A sound analysis system is developed to record a sentence spoken by a human (mainly vowel sequences in the Japanese language) and to then analyze that sentence to give the correct command packet so the talking robot can repeat it. An algorithm based on a short-time energy method is developed to separate and count sound phonemes. A matching template using partial correlation coefficients (PARCOR) is applied to detect a voice in the talking robot's database similar to the spoken voice. Combining the sound separation and counting the result with the detection of vowels in human speech, the talking robot can reproduce a vowel sequence similar to the one spoken by the human. Two tests to verify the working behavior of the robot are performed. The results of the tests indicate that the robot can repeat a sequence of vowels spoken by a human with an average success rate of more than 60%.

***Keywords*:** Talking robot, PARCOR, Vowel sequence, Human speech, Short-time energy

1. Introduction

Speech is the main and most effective communication method in human society. Speech synthesis systems have drawn the attention of many researchers for a long time. There are two main approaches to speech synthesis: software-based and hardware-based systems. Software-based systems were introduced by Guenther et al. [1] (the DIVA model) and Bernd et al. [2] (the ACT model). A hardware-based system is building a mechanical system that can reproduce sound in a way similar to the way the human articulatory system works [5]. The talking robot is one of these hardware-based systems.

The main purpose of this research is to develop an autonomous vowel sequence reproduction system for a talking robot. In the second section, mechanical configuration of the talking robot is introduced. The third section briefly describes the learning process for the talking robot using a Kohonen self-organizing map (SOM) technique. The fourth and main section describes the technique for analyzing human spoken sound and for separating the human sound signal into a set of single phonemes. Then, each phoneme is matched against a template based on PARCOR coefficient vowels for vowel determination. After that, the sequence of vowel sounds is outputted from the talking robot.

When recording and analyzing human speech, a phrase is divided into a sequence of two different parts: consonants and vowels. A sample phrase, */konichiwa/*, spoken by an adult male is shown in Fig. 1. As can be seen, the consonants */k/, /n/, /ch/,* and */w/* take about the first 50ms of each phoneme's waveform, and the rest of each phoneme waveform are the vowels, */o/, /i/,* and */a/*. The talking robot is trained to reproduce the Japanese language, which has five basic vowels and 10 basic consonants. It is difficult to detect the consonant phonemes due to instability and short occurrence times. Therefore, only a technique to analyze a sequence of vowels spoken by a human is presented.

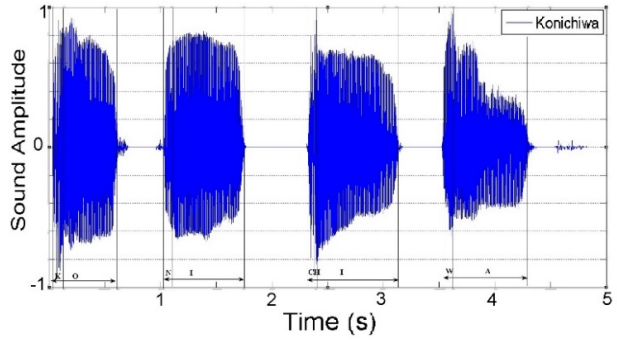

Fig. 1. */ko-ni-chi-wa/* sound wave.

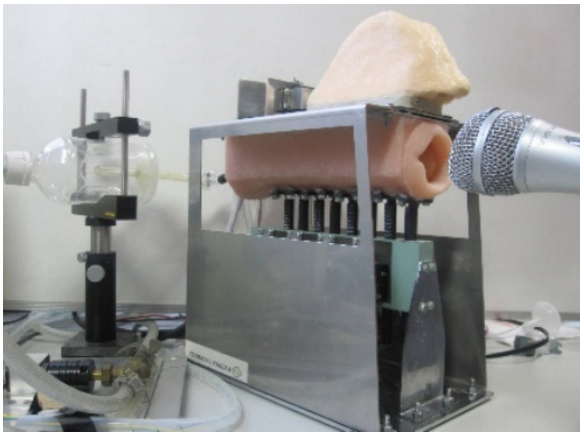

Fig. 2. Overview of the talking robot.

2. Mechanical Construction of the Talking Robot

The talking robot consists of an air pump, an artificial vocal cord, a resonance tube, an artificial nasal cavity, and a microphone connected to a sound analyzer, which respectively represent the lungs, the vocal cords, the vocal tract, the nasal cavity, and auditory feedback of a human. An overview of the talking robot structure is shown in Fig. 2.

The air compressor provides airflow for the talking robot. The airflow is directed to the vocal cords via pressure control valve and two airflow control valves, which work as the controller for the volume of both voiced and unvoiced sounds. The resonance tube functions like a vocal tract attached to the vocal cords to manipulate resonance characteristics. The nasal cavity is connected to the resonance tube with a rotary valve. The microphone and amplifier play the role of an auditory feedback system. The relationships between voice characteristics and motor control commands are stored in the system controller, which is referred to for generation of speech articulatory motion.

The characteristics of a glottal wave, which determines the pitch and the volume of the human voice, is governed by the complex behavior of the vocal cords, which is the oscillatory mechanism of the human organs (the mucous membrane and muscles) excited by airflow from the lungs.

The vibration of a thin 5mm wide rubber band attached to a plastic body creates an artificial vocal sound source [3]. The relationship between tensile force and the fundamental frequency of a vocal sound generated by the artificial vocal cord was measured. The fundamental frequency varied from 110Hz to 350Hz, depending on the pressure applied to the rubber band. The artificial vocal cords are considered suitable for the system not only because of the simple structure, but the frequency characteristics can also be easily regulated by changing the tension of the rubber and the amount of airflow to the vocal cords.

3. Robot Control System and SOM Learning

3.1 Motor Control System of the Talking Robot

As shown in Fig. 3, 12 motors control the shape of the vocal tract, the tongue motion, and the amount of air intake to the vocal tract and nasal cavity of the talking robot. Command-type servomotors (Futaba) are employed to drive the mechanisms of this robot. The advantages of the command type motor are high speed, stability, accuracy, durability, and having a 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 communications protocol. The general structure for sending a packet to control multiple motors is shown below.

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

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

Example: The command to rotate both servo motor ID 1 and servo motor ID 2 by 10 degrees, and to rotate servo ID 5 by 50 degrees is

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

By actuating displacement forces with stainless bars from the outside, the cross-sectional area of the tube is manipulated so that the resonance characteristics are changed according to the transformations of the inner areas of the resonator. Motors are placed in eight positions, from the glottis to the lips, and the displacement forces are applied according to the control command packet from the computer. A nasal cavity is attached above the resonance tube to simulate human-like nasal sounds. A rotational valve controlled by another motor is placed between the resonance tube and the nasal cavity for the selection of nasal and normal sounds. For the generation of nasal sounds */n/* and */m/*, the rotational valve is open to allow air into the nasal cavity. By closing the middle position of the vocal tract and then releasing the air to create the vowel sounds, the */n/* consonant is generated. For the */m/* consonant, the outlet part is closed to stop the air first, and then is opened to vocalize the vowels. The difference in */n/*

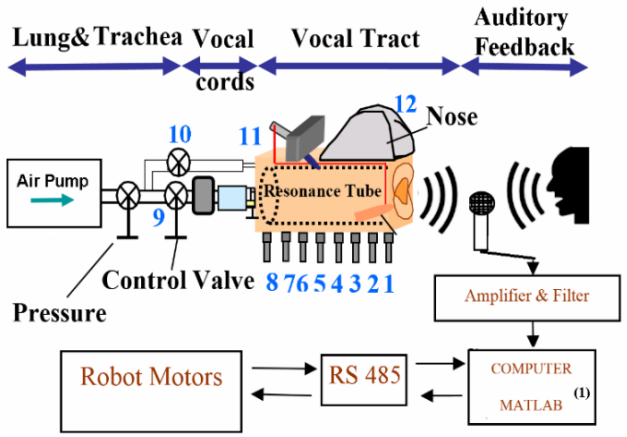

Fig. 3. System configuration.

and */m/* consonant generation is basically the narrowing positions of the vocal tract. In generating plosive sounds, such as */p/*, */b/* and */t/*, the mechanical system closes the rotational valve to not 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. Then, the released air generates plosive consonant sounds like */p/* and */t/*. The robot also has a silicone-molded tongue, which is made by referring to the shape and size of a human tongue. A string is attached to the tongue, and at the other end of the string, a servo motor is connected for manipulation of the up-down motion to vocalize the */l/* sound.

3.2 SOM-Learning of the Talking Robot

In our previous studies, a neural network (NN) was employed to autonomously associate vocal tract shapes with generated vocal sounds [5, 6]. In the learning process, the network learns the motor control commands by inputting resonance characteristics as teaching signals. By combining a 3D self-organizing map (3D-SOM) with a neural network (NN), a dual three-dimensional self-organizing neural network (dual 3D-SONN) was employed. The 3D-SONN was able to choose cells on the map and autonomously recreate voice articulations. The 3D-SONN has a three-dimensional mapping space, which allows the characteristics to be located three-dimensionally, decreasing the probability of miss locations if using a 2D-SONN.

On the 3D-SONN, the inputs are vectors of nine elements, which are nine coefficients extracted from vocal sound using Mel-frequency Cepstral coefficient (MFCC) analysis, 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 learning the three-dimensional SONN. Steps for learning are presented below. 1

1. The cell that has the minimum Euclidean distance with x_i , is selected via Eq. (1) on the 3D feature map.

$$c = arg_i * min \left| m_i - x_i \right| \tag{1}$$

2. The neighborhoods of the selected cells are calculated with Eqs. (2) and (3).

$$m_i(t+1) = m_i(t) + h_{ci}\left[x_i(t) - m_i(t) \right] \tag{2}$$

$$h_{ci} = \alpha(t) exp\left(\frac{\left| r_c - r_i \right|^2}{2\sigma^2(t)} \right) \tag{3}$$

The learning proceeds until all the phonetic characteristics are distributed properly on the feature map by repeating procedures 1 and 2, and the topological relations among different features are autonomously created. The association between vowels and consonants and the motor parameter vectors is established. This is the basic set to reconstruct the human voice, and could be used for a text-to-speech function for the talking robot. In this study, only the set of motor parameter vectors for vowels in the robot database were used for vowel sequence reproduction.

4. Sequence of Vowel Regeneration

As mentioned in Section 3.2, the motor parameter vector set of five vowels is used to output the command for the talking robot in order to regenerate the respective vowels detected by PARCOR coefficient analysis. The flow chart of the vowel sequence reproduction program is shown in Fig. 4. There are eight steps in total. In the robot initialization step, all robot articulatory motors are set to

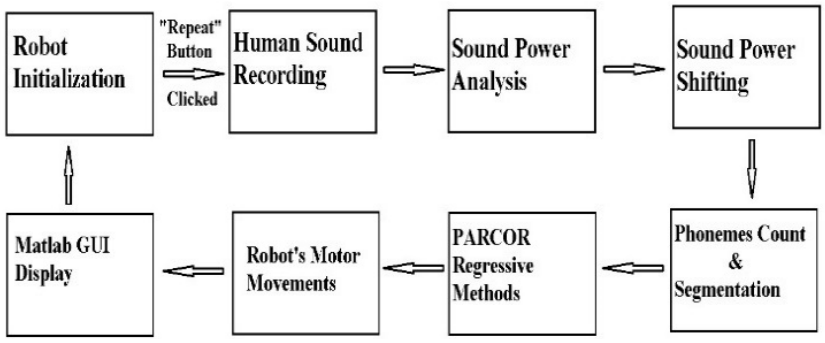

Fig. 4. System configuration.

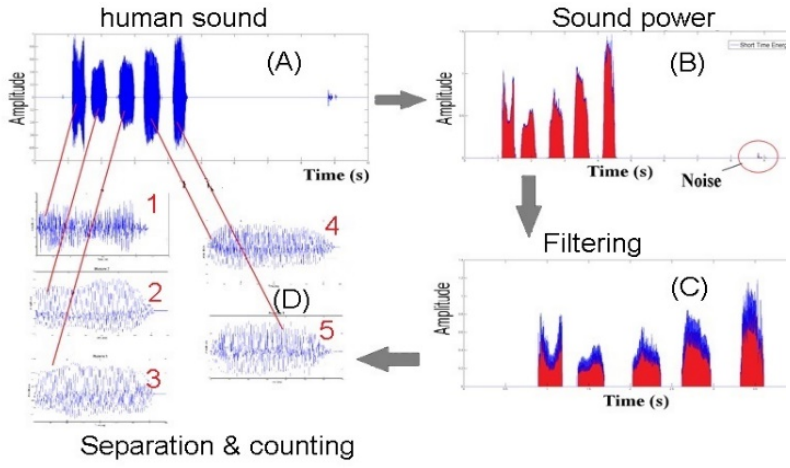

Fig. 5. Phoneme segment separation steps (a) Human recorded sound, (b) short-time energy calculation, (c) threshold comparison, (d) phoneme separation.

the original position; the details on the other steps are explained in the next sections.

4.1 Sound Recording Using Matlab

The control interface of the talking robot was built using the Matlab graphical user interface (GUI) to record and display human sound, and the talking robot reproduced the sound. The sound was recorded at a sampling rate of 8000Hz—a single channel using built-in Matlab commands—and the sound wave data were then saved in the Matlab workspace for further analysis.

4.2 Analysis of the Recorded Voice

In order to separate phonemes in speech, first, the short-time energy (STE) of the voice is calculated based on Eq. (4).

$$E_n = \sum_{m=n-N+1}^{n} \left[x(m) w(n-m) \right]^2 \tag{4}$$

In Eq. (4), E_n is the short-time energy, $x(m)$ is the signal value at m, n is the window duration, n = 0, 1T, 2T, …, T is the frame-shift (100 samples), N is the window size (101 samples), and $w()$ is the window function (Hamming).

$$\varnothing_{th} = \frac{E_{max}}{10} \tag{5}$$

$$X(m) = \begin{cases} E(m) & E(m) \geq \varnothing_{th} \\ 0 & E(m) < \varnothing_{th} \end{cases} \tag{6}$$

Then, the STE is compared to a constant, $\varnothing_{th}$, which is equal to the maximum value of the STE divided by 10. This constant, $\varnothing_{th}$, is called the threshold number, as described in Eq. (5).

Several experiments to investigate the effect of the threshold number's value have been done. For vowel separation, the value of threshold number $\varnothing_{th}$, equal to the maximum value of the calculated short-time energy divided by 10, is sufficient to separate 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 segment of speech. The separation is done by first comparing each element in the data, $E(m)$, with the threshold number, $\varnothing_{th}$; if it is greater than threshold number $\varnothing_{th}$, the value of that element is kept; otherwise, its value is replaced with zero (3). After that, the number of phonemes is counted, and the phonemes are separated by using a tracing technique. The algorithm traces data $E(m)$ from left to right, and determines the number of phonemes in the speech segment, based on the number of times the data changes from a zero value to a non-zero value. One phoneme segment is the range of $E(m)$ that has a non-zero value. Noise cancellation is also implemented in the system; if the length of the segment is too short, it is noise, so the system ignores that segment. The separated phonemes are saved to the Matlab workspace for PARCOR coefficient analysis.

As a result, the number of phonemes and a new phoneme for each sound segment are obtained in the Matlab workspace. This information is used as part of the input data for the talking robot to regenerate a sequence of vowels spoken by the human. Fig. 5 shows an example of this sound analysis technique.

4.3 PARCOR coefficients analysis for phoneme recognition

Previously, a set of human vowels /a/, /i/, /u/, /e/, and /o/ was recorded. This set was used to make the PARCOR coefficients template for vowel recognition based on a matching technique. A PARCOR coefficient is defined as a correlation coefficient between forward and backward prediction errors in the autoregressive model of order n-1 [7]. PARCOR coefficients k_n of a waveform sample [xt,

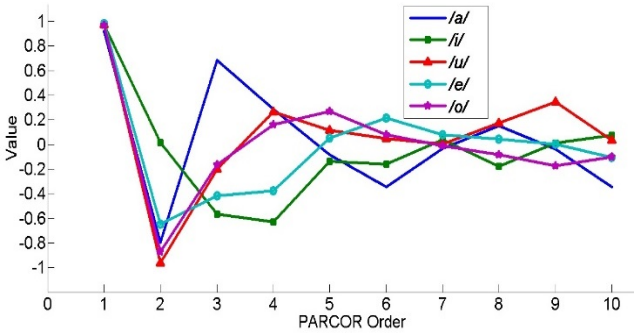

Fig. 6. Human PARCOR coefficients.

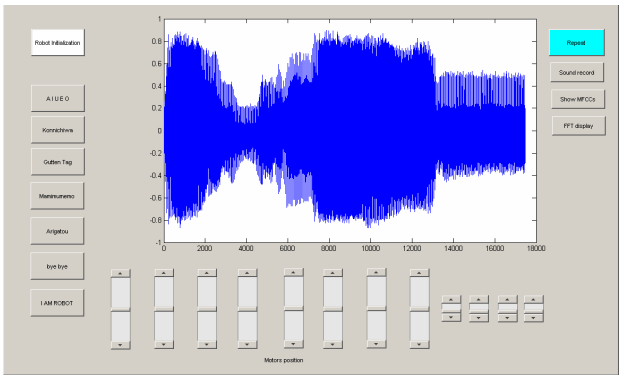

Fig. 7. Robot output sound display.

xt-1,…, xt-n] with a forward prediction error of $e_{ft}^{(n-1)}$, and a backward prediction error of $e_{bt}^{(n-1)}$ is described in Eq. (7). In Eq. (7), k_n is the PARCOR coefficient of order n, e_f is the forward prediction error signal, e_b is the backward prediction error signal, and m is the PARCOR order. In this study, the 10th order of PARCOR coefficients is applied.

$$k_n = \frac{E\left\{e_{ft}^{(n-1)} e_{bt}^{(n-1)}\right\}}{\left[\mathrm{E}\left\{\left(e_{ft}^{(n-1)}\right)^2\right\}\mathrm{E}\left\{\left(e_{bt}^{(n-1)}\right)^2\right\}\right]^2} \tag{7}$$

$$e_{ft}^{(n-1)} = x_t - \widehat{x_t} = x_t + \sum_{i=1}^{n-1} \alpha_i^{(n-1)} x_{t-i} \tag{8}$$

$$e_{bt}^{(n-1)} = x_{t-n} - \widehat{x_{t-n}} = x_{t-n} + \sum_{i=1}^{n-1} \beta_i^{(n-1)} x_{t-i} \tag{9}$$

To obtain k_n directly from waveform x_t, the signal is first transformed to the z-domain.

$$A_{(z)} = \sum_{i=0}^{n} \alpha_i^{(n)} z^{-i} \tag{10}$$

$$B_{(z)} = \sum_{i=1}^{n+1} \beta_i^{(n)} z^{-i} \tag{11}$$

$$e_{ft}^{(n)} = A_n(z) x_t \tag{12}$$

$$e_{bt}^{(n)} = B_n(z) x_t \tag{13}$$

We have the following relations:

$$A_n(z) = A_{n-1}(z) - k_n B_{n-1}(z) \tag{14}$$

$$B_n(z) = z^{-1}\left[B_{n-1}(z) - k_n A_{n-1}(z)\right] \tag{15}$$

The boundary conditions are $A_0(z) = 1$, $B_0(z) = z^{-1}$. The Levinson-Durbin recursive algorithm [8] is applied to solve the PARCOR coefficients. The human PARCOR coefficients are calculated and plotted in Fig. 6. The x-axis shows the orders of the PARCOR coefficients, and the y-axis shows the values of each sound parameter. This set was used as the standard PARCOR template to compare with new human recorded sound.

Human PARCOR coefficients are usually calculated on a small segment of a signal, which ranges within about 50ms of the waveform [3]. There were 10 recorded signals for each vowel spoken beforehand by an adult male. The center segment of each signal length of 50ms was taken to perform PARCOR coefficient analysis. The average value of 10 PARCOR coefficients for each vowel was calculated and saved in the robot database as the standard set. The average value of these PARCOR coefficients is plotted in Fig. 6. PARCOR coefficient analysis is suitable for recognizing vowels, because a vowel is usually a stable signal. However, a consonant is an unstable signal, and only lasts for about the first 50ms of an utterance; thus, it is difficult to determine its PARCOR coefficient correctly. Hence, the phonemes after separation shown in Fig. 5(d) are calculated for PARCOR coefficients using only 50ms of the middle segment of the waveform.

The calculated PARCOR coefficients from each phoneme segment are then compared with the standard set of PARCOR coefficients in the robot database, and the motor command output for each phoneme segment is determined. A template-matching method was employed to identify the vowel for each phoneme segment. Euclidean distances between the templates and a new phoneme were obtained. The minimum Euclidean distances between the recordings indicated the vowel sound for that recorded segment.

4.4 Talking Robot Sound Reproduction and Display

To visualize and evaluate the performance of the control system, a display interface using Matlab's GUI is shown in Fig. 7. The interface has a button to initialize the robot motors to their original positions. It also has buttons for testing robot sound quality, so the user can adjust the parameters for adequate sound output. After motor commands for each separated phoneme segment are determined, it is sent to the robot's motors. Then, the robot's motors move to regenerate a sequence of vowels spoken by the human. The spoken voice from the robot is recorded by the microphone and displayed on a computer screen. The display interface was built using the Matlab GUI in Fig. 7.

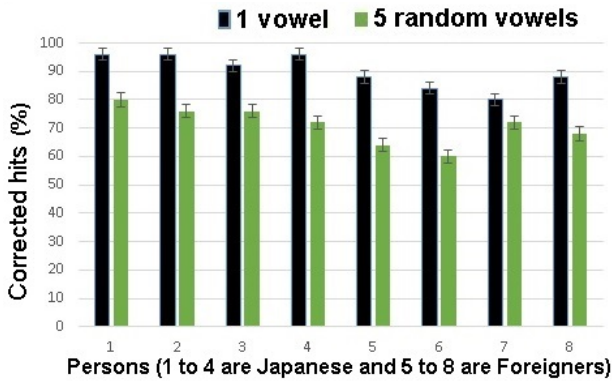

Fig. 8. Results for reproduction of robot sounds.

5. Testing and Analysis

Eight people (seven males and one female) were subjects for testing. Among them were four Japanese and four foreigners (French, Chinese, and Vietnamese). When they clicked the "Repeat" button on the GUI, their sounds were recorded and analyzed, and then the talking robot automatically reproduced the sounds.

The subjects were asked to say five Japanese vowels (*/a/, /i/, /u/, /e/, /o/*) in two separate tests. In the first test, they were asked to say only one vowel at a time and wait for the robot to repeat the vowel. They were asked to utter each vowel five times in a random order of their choice in this test. The total number of samples for this test was 25. The average percentage of correct vowel reproduction by the robot was then calculated. This value was defined as the "corrected hits" of the robot, as shown in the y-axis of Fig. 8.

For the second test, the subjects were asked to utter five vowels at a time, for five times, in a random order of their choice, with a short pause between each vowel. For this test, the percentage of correct vowel reproduction by the robot for each trial was calculated. For example, if the robot repeated four out of five vowels in a trial, it would be counted as 80% correct hits for that trial. Then, the average percentage for correct vowel reproduction by the robot for all trials was calculated. The results for average correct hits from both tests are plotted in Fig. 8.

6. Conclusion

Based on the results in Section 5, the robot had above 90% for the success rate in repeating one vowel from the Japanese speaking voices and above 80% for the foreign voices. In the second test of a random order vowel-sequence reproduction, the talking robot could repeat the sequence of vowels with a success rate above 70% for the Japanese speaking voices and above 60% for foreign speaking voices. It could be this resulted from the set of vowels in the robot database being built from Japanese voices, and thus, the rate for recognizing the Japanese voice would be higher than with a foreigner's voice. Besides, the similarity in characteristics of vowels */i/ & /e/* and */u/ & /o/* made it slightly difficult for the robot to distinguish between these vowels.

In conclusion, the authors built a talking robot to automatically repeat a sequence of vowels from human speech segments. The tests showed that the system worked well for one vowel repeated, but it still needs some improvement in reproducing a random sequence of vowels. Future plans include increasing the recognition precision of the system by automatically updating the PARCOR coefficients in the database with any new recorded voice; this would reduce the risk of misrecognition from new people with a different voice.

Acknowledgement

This work was partly supported by the Grants-in-Aid for Scientific Research, the Japan Society for the Promotion of Science (No. 15K01459).

References

[1] F. H. Guenther, et al., "A neural modelling and imaging of the cortical interactions underlying syllable production", Brain and Language, Vol 96(3), pp. 280-30, 2006. Article (CrossRef Link)

[2] H. Bernd, et al., "Associative learning and self-organization as basic principles for simulating speech acquisition", Speech Production, and Speech Perception. EPJ Nonlinear Biomedical Physics. pp. 2-28, 2014. Article (CrossRef Link)

[3] M. Kitani, H. Sawada, et al, "A talking robot and its singing performance by the mimicry of human vocalization", Human-Computer Systems Interaction: Backgrounds and Applications. Advances in Intelligent and Soft Computing, Vol 99, pp. 57-73, 2012. Article (CrossRef Link)

[4] H. Sawada, "Talking robot and the autonomous acquisition of vocalization and singing skill", Robust Speech Recognition and Understanding, Vol 22, pp.385-404, 2007. Article (CrossRef Link)

[5] K. Fukui, E. Shintaku, et al, "Mechanical vocal cord for anthropomorphic talking robot based on human biomechanical structure", The Japan Society of Mechanical Engineers, Vol 73, pp. 112-118, 2007. Article (CrossRef Link)

[6] Flanagan. J.L, Speech Analysis Synthesis and Perception, Springer-Verlag, 1972. Article (CrossRef Link)

[7] Atal. B.S, Hanauer. S.L, Speech analysis and synthesis by linear prediction of the speech wave, JASA,50, 637-655,1971 Article (CrossRef Link)

[8] J.Durbin, The fitting of time-series models, Rev. Inst. Int. de Stat., Vol.28, No.3, pp.233-244, 1960 Article (CrossRef Link)

Vo Nhu Thanh received his bachelor degree in mechanical engineering from Cal Poly Pomona, USA, in 2008, and a master's degree in mechatronics and sensor system technology from Karlsruhe University of Applied Science, Germany, in 2012. He is a lecturer in the Faculty of Mechanical Engineering, Da Nang University of Science and Technology, The University of Da Nang. He was a Japanese Government Scholarship (MEXT) recipient for his PhD research in 2014. He is currently a PhD student at Kagawa University, Japan. He is a student member of IEEE, and vice chair of the IEEE Kagawa University student brand. His research interests include sound and image processing, robotics.

Hideyuki Sawada is a Professor in the Department of Intelligent Mechanical Systems Engineering in the Faculty of Engineering, Kagawa University. He received a B.Eng., a M.Eng., and a PhD in applied physics from Waseda University in 1990, 1992, and 1999, respectively. His current research interests include sound and image processing, neural networks, robotics, human interfaces, and tactile sensing and display. He is a member of IEEE, RSJ, IPSJ, IEICE, SICE and JSME.