Interactive training of speech articulation for hearing impaired using a talking robot

M Kitani, Y Hayashi and H Sawada

Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University, 2217-20, Hayashi-cho, Takamatsu-city, Kagawa, 761-0369, JAPAN

sawada@eng.kagawa-u.ac.jp

http://www.eng.kagawa-u.ac.jp/~sawada/index_e.html

ABSTRACT

This paper introduces a speech training system for auditory impaired people employing a talking robot. The talking robot consists of mechanically-designed vocal organs such as a vocal tract, a nasal cavity, artificial vocal cords, an air pump and a sound analyzer with a microphone system, and the mechanical parts are controlled by 10 servomotors in total for generating human-like voices. The robot autonomously learns the relation between motor control parameters and the generated vocal sounds by an auditory feedback control, in which a Self-organizing Neural Network (SONN) is employed for the adaptive learning. By employing the robot and its properties, we have constructed an interactive training system. The training is divided into two approaches; one is to use the talking robot for showing the shape and the motion of the vocal organs, and the other is to use a topological map for presenting the difference of phonetic features of a trainee's voices. While referring to the vocal tract motions and the phonetic characteristics, a trainee is able to interactively practice vocalization for acquiring clear speech with an appropriate speech articulation. To assess the validity of the training system, a practical experiment was conducted in a school for the deaf children. 19 subjects took part in the interactive training with the robotic system, and significant results were obtained. The talking robot is expected to intensively teach an auditory impaired the vocalization skill by directing the difference between clear speech and the speech with low clarity.

1. INTRODUCTION

Speech is one of the important media to communicate with each other. Only humans use words for verbal communication, although most animals have voices or vocal sounds. Vocal sounds are generated by the relevant operations of the vocal organs such as lung, trachea, vocal cords, vocal tract, tongue and muscles. The airflow from the lung causes the vocal cords vibration and generates a source sound, then the sound is led to a vocal tract to work as a sound filter as to form the spectrum envelope of a particular sound. The voice is at the same time transmitted to the human auditory system so that the vocal system is controlled for the stable vocalization. Various vocal sounds are generated by the complex articulations of vocal organs under the feedback control mechanisms using an auditory system.

Infants have the vocal organs congenitally, however they cannot utter a word. As infants grow they acquire the control methods pertaining to the vocal organs for appropriate vocalization. These get developed in infancy by repetition of trials and errors concerning the hearing and vocalizing of vocal sounds. Any disability or injury to any part of the vocal organs or to the auditory system might cause an impediment in vocalization. People who have congenitally hearing impairments have difficulties in learning vocalization, since they are not able to listen to their own voice.

Auditory impaired patients usually receive a speech training conducted by speech therapists (ST) (Boothroyd, 1973; Boothroyd, 1988; Erber and de Filippo, 1978; Goldstein and Stark, 1976), however many problems and difficulties are reported. For example, in the training, a patient is not able to observe his own vocal tract, nor the complex articulations of vocal organs in the mouth, then he cannot recognize the validity of his articulation nor evaluate the achievement of speech training without hearing the voices. Children take training at school during a semester, however it is not easy to continue the training during vacation and they

get to forget the skill. The most serious problem is that the number of ST is not enough to give speech training to all the subjects with auditory impairment.

The authors are developing a talking robot by reproducing a human vocal system mechanically based on the physical model of human vocal organs. The robot consists of motor-controlled vocal organs such as vocal cords, a vocal tract and a nasal cavity to generate a natural voice imitating a human vocalization. For the autonomous acquisition of the robot's vocalization skills, an adaptive learning using an auditory feedback control is introduced.

In this study, the talking robot is applied to the training system of speech articulation for the hearing impaired children, since the robot is able to reproduce their vocalization and to teach them how it is improved to articulate the vocal organs for generating clear speech. The paper briefly introduces the mechanical construction of the robot first, and then the analysis of the autonomous learning will be described how the robot reproduces the articulatory motion from hearing impaired voices by using a self-organizing neural network. An interactive training system of speech articulation for hearing impaired children is presented, together with an experiment of speech training conducted in a school for deaf children.

2. HEARING IMPAIRED AND THE SPEECH TRAINING

Currently, a speech training for hearing impaired is conducted by speech therapists. They give specially-designed training programs to each patients by carefully examining the symptoms of impairment. In Japan there are about 360,000 hearing impaired people who are certified by the government, however by counting patients with mild symptoms and aged people with auditory disabilities, the number will be doubled to 600,000. On the contrary, the number of ST is approximately 10,000, which is far less than the number of the patients. Conventionally the training by a ST is conducted face-to-face by using a mirror to show the articulatory motions of inner mouth. Schematic figures to conceptually show the mouth shapes and articulatory motions are also employed for intuitive understandings of the speech articulations.

Figure 1 shows an example of an electronic speech training system WH-9500 developed by Matsushita Electric Industrial Co., Ltd. It is equipped with a headset with a microphone, and shows the difference of sound features together with an estimated vocal tract shape on the display, so that a trainee could understand his own vocalization visually. The system is large and requires technical knowledge and complex settings, and it is difficult for an individual patient to settle it at home. By examining the problems of the conventional training mentioned above, the authors are constructing an interactive training system, by which a patient engages in a speech training in any occasion, at any place, without special knowledge, as shown in Figure 2.

We are constructing a training system employing a talking robot. By using a self-organizing neural network, the robot reproduces an articulatory motion by listening to a subject's voice, and the phoneme characteristics are visually shown in a display, so that a trainee could recognize his own phoneme characteristics and the corresponding vocal tract shape by comparing with a target voice. Besides, to realize an interactive training and easy manipulation, the robotic training system is executed by a simple user interface.

3. CONSTRUCTION OF A TALKING ROBOT

Human vocal sounds are generated by the relevant operations of vocal organs such as the lung, trachea, vocal cords, vocal tract, nasal cavity, tongue and muscles. In human verbal communication, the sound is perceived as words, which consist of vowels and consonants.

The lung has the function of an air tank, and an airflow through the trachea causes a vocal cord vibration as the source sound of a voice. The glottal wave is led to the vocal tract, which works as a sound filter as to form the spectrum envelope of the voice. The fundamental frequency and the volume of the sound source is varied by the change of the physical parameters such as the stiffness of the vocal cords and the amounts of airflow from the lung, and these parameters are uniquely controlled when we speak or utter a song.

In contrast, the spectrum envelope, which is necessary for the pronunciation of words consisting of vowels and consonants, is formed based on the inner shape of the vocal tract and the mouth, which are governed by the complex movements of the jaw, tongue and muscles. Vowel sounds are radiated by the relatively stable configuration of the vocal tract, while the short time dynamic motions of the vocal apparatus produce consonants generally. The dampness and viscosity of organs greatly influence the timbre of generated sounds, which we may experience when we have a sore throat. Appropriate configurations of the

vocal tract for the production of phonemes are acquired as infants grow by repeating trials and errors of hearing and vocalizing vocal sounds

(a) Appearance of WH-9600

(b) Headset with microphone

Figure 1. *An example of electronic speech training system.*

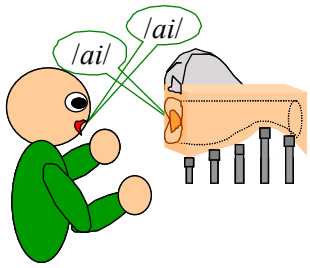


Figure 2. *Interactive training.*

The talking robot mainly consists of an air compressor, artificial vocal cords, a resonance tube, a nasal cavity, and a microphone connected to a sound analyzer, which correspond to a lung, vocal cords, a vocal tract, a nasal cavity and an auditory system of a human, as shown in Figure 3.

An air from the pump is led to the vocal cords via an airflow control valve, which works for the control of the voice volume. The resonance tube is attached to the vocal cords for the articulation of resonance characteristics. The nasal cavity is connected to the resonance tube with a rotational valve between them. The sound analyzer plays a role of the auditory system, and realizes the pitch extraction and the analysis of resonance characteristics of generated sounds in real time, which are necessary for the auditory feedback control. The system controller manages the whole system by listening to the generated sounds and calculating motor control commands, based on the auditory feedback control mechanism employing a neural network learning. The relation between the sound characteristics and motor control parameters are stored in the system controller, which are referred to in the generation of speech and singing performance.

3.1 Artificial Vocal Cords and Its Pitch Control

Vocal cords with two vibrating cords molded with silicone rubber with the softness of human mucous membrane were constructed in this study. Two-layered construction (a hard silicone is inside with the soft coating outside) gave the better resonance characteristics, and is employed in the robot (Higashimoto and Sawada, 2003). The vibratory actions of the two cords are excited by the airflow led by the tube, and generate a source sound to be resonated in the vocal tract.

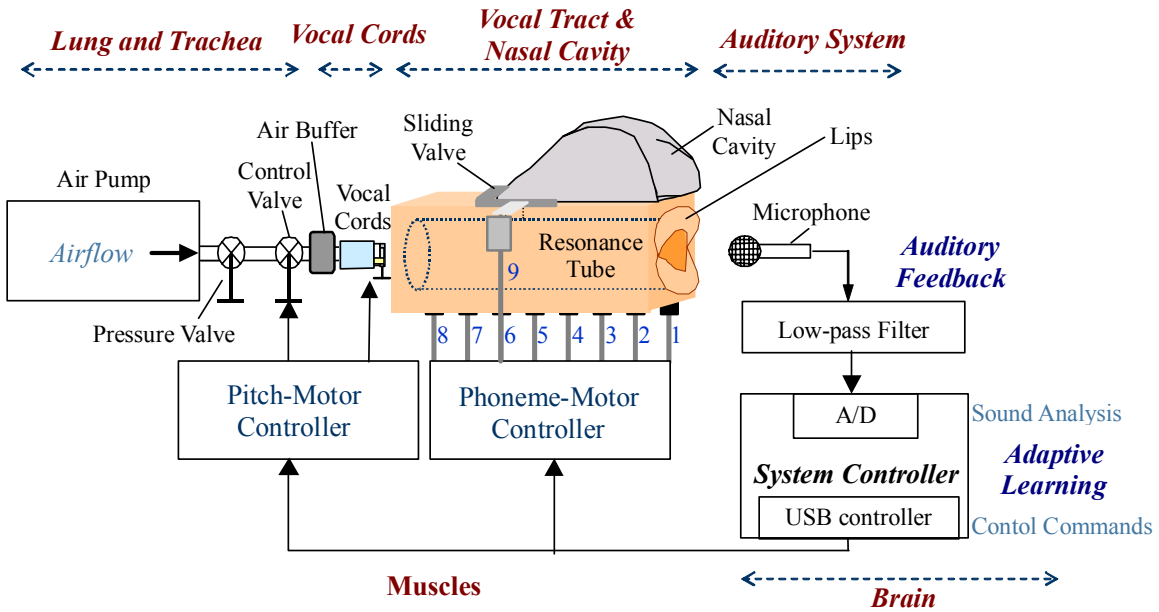


Figure 3. *Construction of the talking robot.*

The tension of vocal cords can be manipulated by applying tensile force to them. By pulling the cords, the tension increases so that the frequency of the generated sound becomes higher. The relationship between the tensile force and the fundamental frequency of a vocal sound generated by the robot is acquired by the auditory feedback learning before the singing and talking performance, and pitches during the utterance are kept in stable by the adaptive feedback control (Sawada and Nakamura, 2004).

3.2 Construction of Resonance Tube and Nasal Cavity

The human vocal tract is a non-uniform tube about 170 mm long in man. Its cross-sectional area varies from 0 to 20 cm^2 under the control for vocalization. A nasal cavity with a total volume of 60 cm^3 is coupled to the vocal tract. Nasal sounds such as /*m*/ and /*n*/ are normally excited by the vocal cords and resonated in the nasal cavity. Nasal sounds are generated by closing the soft palate and lips, not to radiate air from the mouth, but to resonate the sound in the nasal cavity. The closed vocal tract works as a lateral branch resonator and also has effects of resonance characteristics to generate nasal sounds. Based on the difference of articulatory positions of tongue and mouth, the /*m*/ and /*n*/ sounds can be distinguished with each other.

In the mechanical system, a resonance tube as a vocal tract is attached at the sound outlet of the artificial vocal cords. It works as a resonator of a source sound generated by the vocal cords. It is made of a silicone rubber with the length of 180 mm and the diameter of 36 mm, which is equal to 10.2 cm^2 by the cross-sectional area as shown in Figure 4. The silicone rubber is molded with the softness of human skin, which contributes to the quality of the resonance characteristics. In addition, a nasal cavity made of a plaster is attached to the resonance tube to vocalize nasal sounds like /*m*/ and /*n*/.

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. Compact servo motors are placed at 8 positions x_j (j = 1-8) from the intake side of the tube to the outlet side, and the displacement forces $P_j(x_j)$ are applied according to the control commands from the phoneme-motor controller.

A nasal cavity is coupled with the resonance tube as a vocal tract to vocalize human-like nasal sounds by the control of mechanical parts. A rotational valve as a role of the soft palate is settled at the connection of 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 lead the air into the nasal cavity.

By closing the middle position of the vocal tract and then releasing the air to speak vowel sounds, /*n*/ consonant is generated. For the /*m*/ consonants, the outlet part is closed to stop the air first, and then is open to vocalize vowels. The difference in the /*n*/ and /m/ consonant generations 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 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. Then the released air generates plosive consonant sounds like /*p*/ and /*t*/.

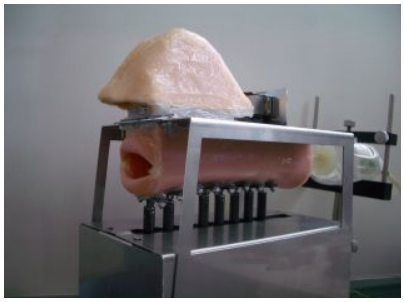


Figure 4. *Talking robot.*

4. METHOD OF AUTONOMOUS VOICE ACQUISITION

We pay attention to the ability of a neural network (NN) to associate sound characteristics with the vocal tract shape. By autonomously learning the relation, it will be possible to estimate the articulation of vocal tract, so that the robot can generate appropriate vocal sounds. The NN is expected to work for associating the sound characteristics with the control parameters of the motors as shown in Figure 5. In the learning phase, the NN learns the motor control parameters by inputting power spectra of sounds as teaching signals. The

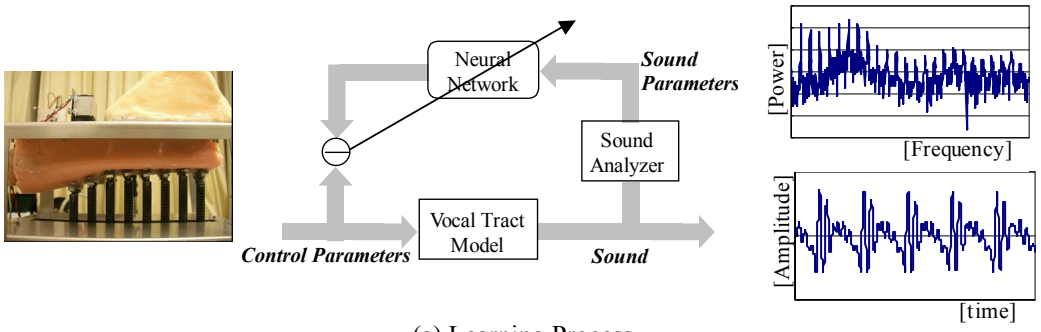
network acquires the relations between sounds and the cross-sectional areas of the vocal tract (Figure 5(a)). After the learning, the NN is connected in series into the vocal tract model as shown in Figure 5 (b). By inputting the sound parameters of desired sounds to the NN, the corresponding form of the vocal tract is obtained.

A Self-Organizing Neural Network (SONN), which consists of an input layer, a competition layer, a hidden layer and an output layer, is employed in this study to adaptively learn the vocalization skill, as shown in Figure 6. The links between the layers are fully connected with learning coefficient vectors $\{V_{ij}\}$, $\{W^1_{jk}\}$ and $\{W^2_{kl}\}$. The number of the cells in the input layer is set to 10, in accordance with the number of the sound parameters consisting of 10^{th} order cepstrum coefficients (Sawada, 2007) extracted from vocal sounds generated by random articulations of the robot mouth. The number of the output layer cells is 8, which is the number of the motor-control parameters to manipulate the vocal tract. The number of the cells in the hidden layer and the competition layer is determined by considering the number of learning patterns.

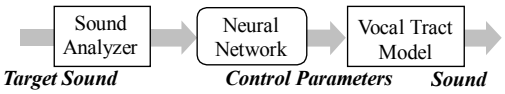
In the learning phase, the relations between the sound parameters and the motor control parameters are established. In the speech phase, motor control parameters are recalled by inputting target voices. In this study, the learning of the sound parameters in the competition layer is called "upward learning", and a topological map is expected to be established in the competition layer by the self-organizing map (SOM) learning. The learning of the relation between the SOM and the motor control parameters is called a "downward learning", which associate phonetic features with vocal tract shapes.

5. ANALYSIS OF ACQUIRED SOUNDS

In the learning phase, sounds randomly vocalized by the robot were mapped on the map array. After the learning of the relationship between the sound parameters and the motor control parameters, we inputted human voices from microphone to examine whether the robot could speak autonomously by mimicking human vocalization. Same vowel sounds were mapped close with each other, and five vowels were well categorized according to the differences of phonetic characteristics. Two different sounds having large difference of phonetic features are located far with each other. In this manner, topological relations according to the difference of phonetic features were autonomously established on the map.



(a) Learning Process



(b) Running Process

Figure 5. *Neural network in mechanical model.*

Figure 7 shows the results of acquired spectra, in comparison with actual human voices. By comparing the robot voices with human, phonetic characteristics of Japanese vowels were well reproduced by the topological relations on the feature map. Human vowel /*a*/ has the first formant in the frequency range from 500 to 900 Hz and the second formant from 900 to 1500 Hz, and the robotic voice also presents the same formants. In the listening experiments, most of the subjects pointed out that the generated voices have similar phonetic characteristics to the human voices. These results show that the vocal tract made by silicone rubber has the tolerance of generating human-like vocalization, and the neural network learning of the voice acquisition was successfully achieved.

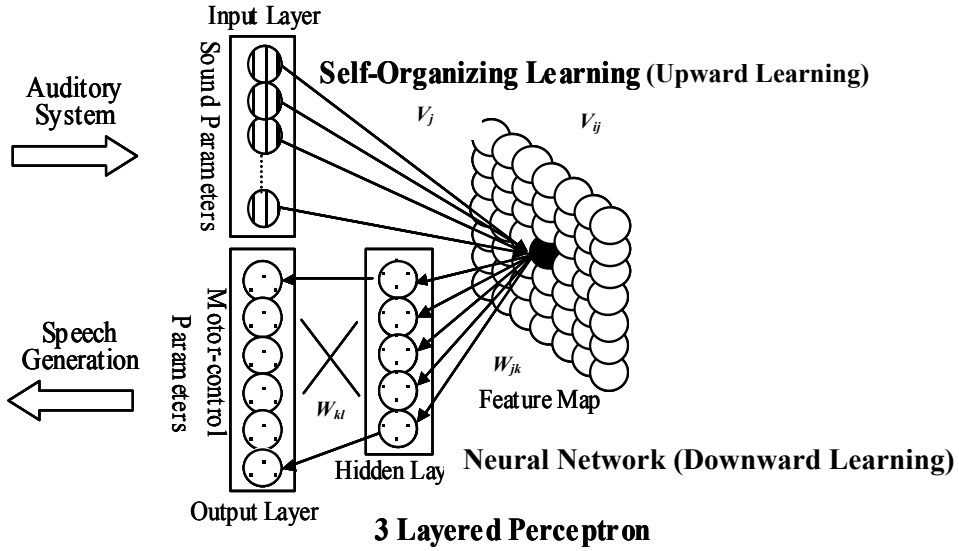
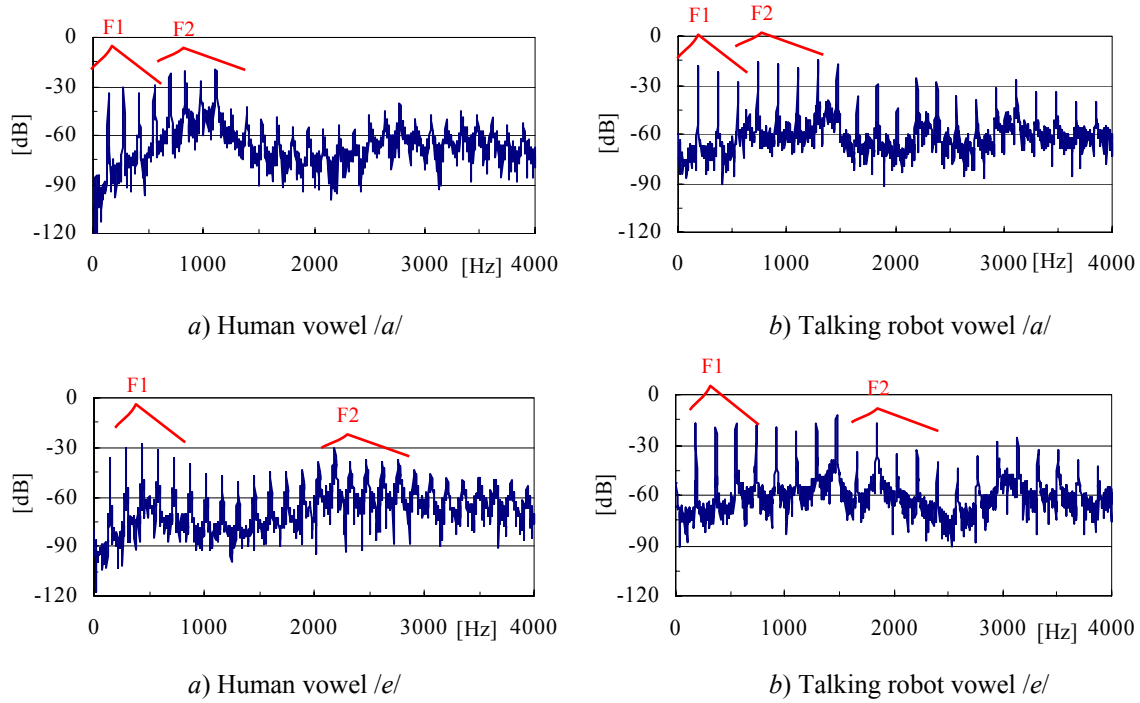


Figure 6. *Structure of self-organizing neural network.*

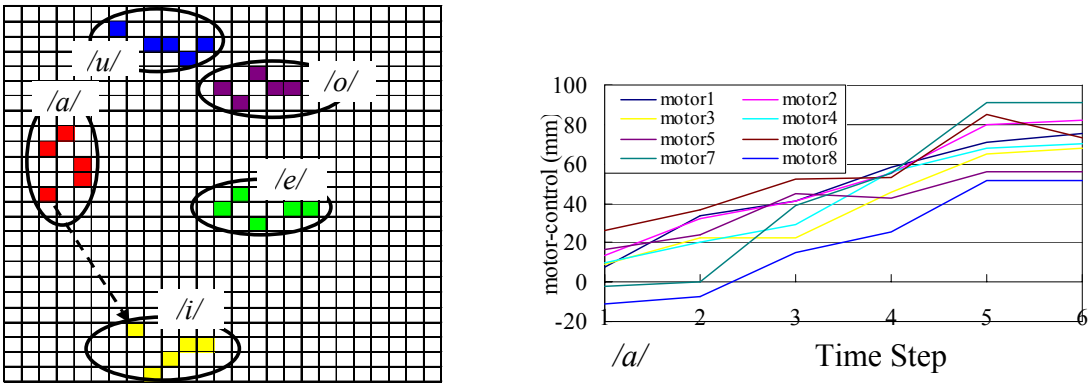
We also examined a topological structure autonomously established on the feature map by the SOM learning. Figure 8 (a) shows an example of a topological map established by the learning. By choosing 6 grids from the /*a*/ area to /*i*/ area shown by a dotted arrow, a voice transition between the two vowels was studied. Figure 8 (b) shows the transition of control values of 8 motors from /*a*/ vocalization to /*i*/ vocalization. Each value is transiting smoothly from the shape of /*a*/ to /*i*/, and this proved that the robot successfully established the topological relations of phonetic features of voices reproduced by the articulatory motions.



a) Human vowel /*a*/ *b*) Talking robot vowel /*a*/

a) Human vowel /*e*/ *b*) Talking robot vowel /*e*/

Figure 7. *Comparison of spectra.*



a) Result of Japanese vowel mapping

b) Speech articulation from */a/* to /i/

Figure 8. *Acquired topological map and voice transition.*

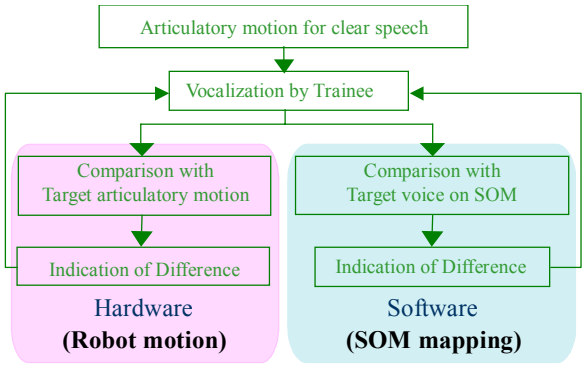


Figure 9. *Flow of Speech Training.*

6. INTERACTIVE TRAINING OF SPEECH ARTICULATION FOR HEARING IMPAIRED

6.1 Training Methods

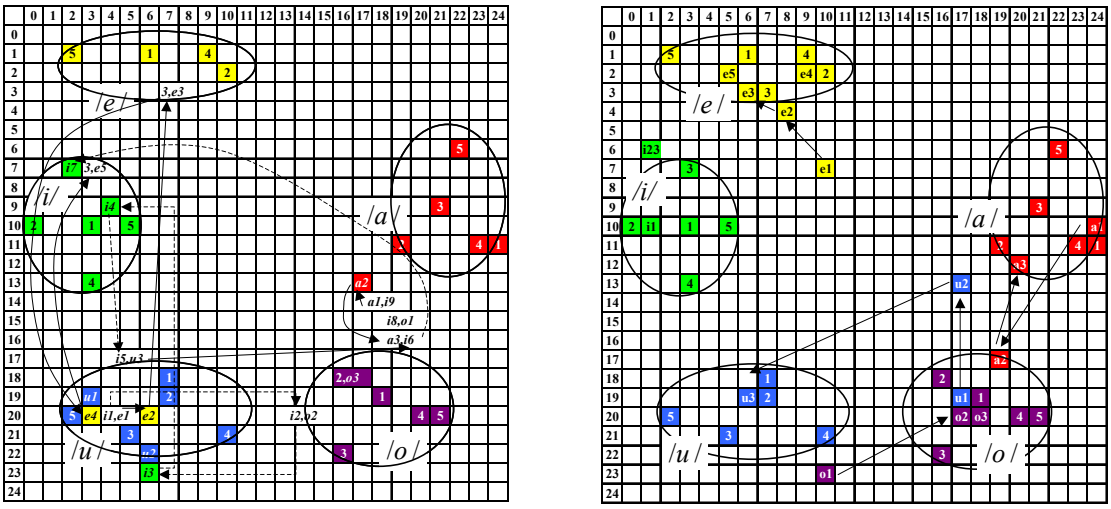
The talking robot is able to reproduce an articulatory motion by only listening to a voice, and we are developing a training system which teaches auditory impaired children how clear speech is generated by interactively directing articulation of inner mouth.

The training is given by two approaches; one is to use the talking robot for showing the shape and the motion of the vocal organs (hardware training), and the other is to use a topological map for presenting the difference of phonetic features of a trainee's voices (software training). Figure 9 shows the flow of the training. At first, an ideal vocal tract shape is presented to a trainee by the talking robot, and the trainee tries to articulate the vocalization referring to the robot. Then, by listening to the trainee's voice, the robot reproduces the trainee's estimated vocal tract shape, and directs how the trainee's voice would be clarified by the change of articulatory motions, by intensively showing the different artiulatory points. The trainee compares his own vocal tract shape and the ideal vocal tract shape, both of which are shown by the articulatory motions of the robot, and tries to reduce the difference of the articulations. The system also presents phonetic features using the topological map, in which the trainee's voice and the target voices are displayed. During the repetition of speech and listening, the trainee recognizes the topological distance between his voice and the target voice, and tries to reduce the distance. In the training, a trainee repeats these training processes for learning 5 vowels.

A training experiment was conducted in a school for the deaf children. 12 high school students and 7 junior-high school students (19 students in total) were engaged in the experiment.

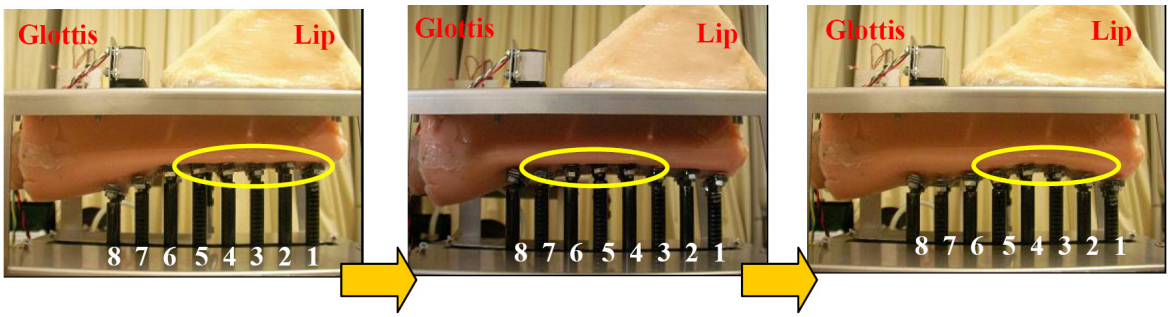
6.2 *Training Results*

Figure 10 shows the result of speech training conducted by two subjects #a and #b. Labels with numbers 1 to 5 show the vowels vocalized by able-bodied subjects #1 to #5, respectively, and the grids indicated by the numbers encircled with vowel names present the area of clear phonemes. During the training, the trainees practiced the vocalization to try to bring their voices fall into the circles of each vowel. "*a*1" means the first vocalization of vowel /*a*/ by the subject, and the arrows show the transition of trials to achieve the clear vocalization. A label "*a*123" in one grid, for example, means that the vocalization stayed in the same phonetic characteristics during the first to the third trials.

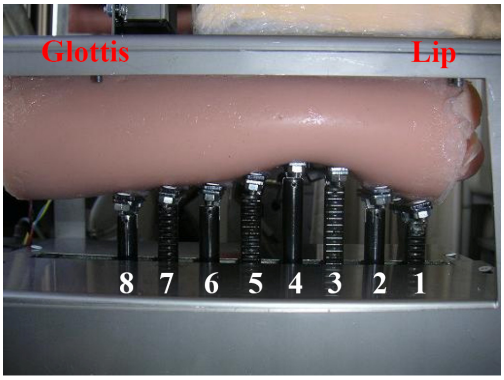


(b) Subject #a (c) Subject #b

Figure 10. *Training results of two subjects #a and #b.*



(a) Step 1 (b) Step 2 (c) Step 3



(d) Vowel /*e*/ by able-bodied speaker

Figure 11. *Progress of /e/ Vocalization by Subject #b*

In the experiment, subject #a could not learn all the ideal vowels. In the training of vowel /*a*/, for example, his first voice fell in the location between the /*a*/ and /*o*/ vowel area. He made trials to bring his voice to the /*a*/ area by referring to the robot vocalization, however he could not achieve it. On the other hand, subject #b could successfully achieve the training to acquire the vocalization skill of five Japanese vowels, after the several trials. Figure 11 shows the progress of the vocal tract shapes of the vowel /*e*/ vocalization in the training of Subject #b. The circles show the articulation points for the vocalization of /*e*/, which the subject intensively tried to articulate during the training. After several trials, he could successfully acquire the vocalization, which is almost the same with the vocalization given by an able-bodied speaker.

Through the training, 13 students out of 19 could achieve the clear vocalization, and all the students at least learned better vocalization than that before the training. Most of the subjects reported that they enjoyed the training using the talking robot, and wanted to continue it in the future.

7. CONCLUSIONS

A talking robot and its articulatory reproduction of voice of hearing impaired was described in this paper. By introducing the adaptive learning and controlling of the mechanical model with the auditory feedback, the robot was able to acquire the vocalization skill as a human baby does in a speech training. The robot was applied to introduce a new training system for auditory impaired children to make an interactive training of speech articulation for learning clear vocalization. The robotic system reproduces the articulatory motion just by listening to actual human voices, and a trainee could learn and know how to move the vocal organs for the clear vocalization, by observing the motions directed by the talking robot.

Acknowledgements: This work was partly supported by the Grants-in-Aid for Scientific Research, the Japan Society for the Promotion of Science (No. 18500152). The authors would like to thank Dr. Yoichi Nakatsuka, the director of the Kagawa Prefectural Rehabilitation center for the Physically Handicapped, Mr. Tomoyoshi Noda, the speech therapist and teacher of Kagawa Prefectural School for the Deaf, and the students of the school for their helpful supports for the experiment and the useful advice.

8. REFERENCES

A Boothroyd (1973), Some experiments on the control of voice in the profoundly deaf using a pitch extractor and storage oscilloscope display, IEEE Transactions on Audio and Electroacoustics, Vol.21, No.3, pp. 274-278.

A Boothroyd (1988), Hearing Impairments in Young Children, A. G. Bell Association for the Deaf.

N P Erber and C L de Filippo (1978), Voice/mouth synthesis and tactual/visual perception of /pa, ba, ma/, Journal of the Acoustical Society of America, Vol.64, No.4, pp.1015-1019.

M H Goldstein and R E Stark (1976), Modification of vocalizations of preschool deaf children by vibrotactile and visual displays, Journal of the Acoustical Society of America, Vol.59, No.6, pp.1477-81.

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

H Sawada and M Nakamura (2004), Mechanical Voice System and its Singing Performance, IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1920-1925.

H Sawada (2007), Talking Robot and the Autonomous Acquisition of Vocalization and Singing Skill, Chapter 22 in Robust Speech Recognition and Understanding, Edited by Grimm and Kroschel, pp. 385-404.