Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2008, Article ID 768232, 7 pages
doi:10.1155/2008/768232

Research Article

A Robotic Voice Simulator and the Interactive Training for Hearing-Impaired People

Hideyuki Sawada, Mitsuki Kitani, and Yasumori Hayashi

Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University, Japan

Correspondence should be addressed to Hideyuki Sawada, sawada@eng.kagawa-u.ac.jp

Received 31 August 2007; Accepted 23 January 2008

Recommended by Daniel Howard

A talking and singing robot which adaptively learns the vocalization skill by means of an auditory feedback learning algorithm is being developed. 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. In this study, the robot is applied to the training system of speech articulation for the hearing-impaired, because the robot is able to reproduce their vocalization and to teach them how it is to be improved to generate clear speech. The paper briefly introduces the mechanical construction of the robot and how it autonomously acquires the vocalization skill in the auditory feedback learning by listening to human speech. Then the training system is described, together with the evaluation of the speech training by auditory impaired people.

1. INTRODUCTION

A voice is the most important and effective medium employed not only in daily communication but also in logical discussions. Only humans are able to use words as means of verbal communication, although almost all animals have voices. Vocal sounds are generated by the relevant operations of the vocal organs such as a lung, trachea, vocal cords, vocal tract, tongue, and muscles. The airflow from the lung causes a vocal cord vibration to generate a source sound, and then the glottal wave is led to the vocal tract, which works as a sound filter as to form the spectrum envelope of a particular voice. The voice is at the same time transmitted to the auditory system so that the vocal system is controlled for the stable vocalization. Different vocal sounds are generated by the complex movements of vocal organs under the feedback control mechanisms using an auditory system.

As infants grow they acquire these 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 may result in 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. A speech therapist helps themtotrain their speech by teaching the vocal organs to learn vocalization and clear speech [1–4].

We are developing a talking robot by reproducing a human vocal system mechanically and based on the physical model of the vocal organs in the human. The fundamental frequency and the spectrum envelope determine the principal characteristics of a voice. Fundamental frequency is a characteristic of the voice source that is generated by the vibration of vocal cords. The resonance effects that get articulated by the motion of vocal tract and nasal cavity cause the spectrum envelope. For the autonomous acquisition of vocalization skills by the robot, an adaptive learning using an auditory feedback control is introduced, like the case for a human baby.

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 [5–8]. By introducing auditory feedback learning with an adaptive control algorithm of pitch and phoneme, the robot is able to autonomously acquire the control skill of the mechanical system to vocalize stable vocal sounds imitating human speech. In the first part of the paper, the construction of vocal cords

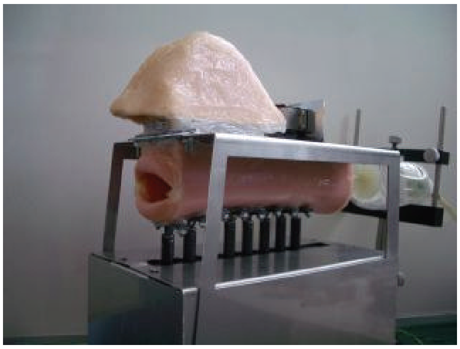


FIGURE 1: Structural view of talking robot.

and vocal tract for the realization of the robot is briefly presented, and then the analysis of the autonomous learning of how the robot acquires the vocalization skill by using the neural network will be described. Then, a robotic training system for the hearing-impaired people is introduced, together with the evaluation of the interactive speech training conducted in an experiment.

2. CONSTRUCTION OF A TALKING ROBOT

The talking robot mainly consists of an air pump, artificial vocal cords, a resonance tube, a nasal cavity, and a microphone connected to a sound analyzer, which, respectively, correspond to a lung, vocal cords, a vocal tract, a nasal cavity, and an audition of a human, as shown in Figure 1.

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 as a vocal tract is attached to the vocal cords for the modification of resonance characteristics. The nasal cavity is connected to the resonance tube with a sliding valve between them. The sound analyzer plays a role of the auditory system. It 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 vocalized sounds and calculating motor control commands, based on the auditory feedback control mechanism employing a neural network learning. The relation between the voice characteristics and motor control parameters is stored in the system controller, which is referred to in the generation of speech and singing performance.

2.1. *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 [7]. 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.

The tension of 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 [8].

2.2. *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. 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 1. 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*/. A sliding 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 motor-controlled sliding valve is open to lead the air into the nasal cavity.

By actuating displacement forces with stainless bars from the outside of the vocal tract, 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 $(j = 1\text{–}8)$ from the lip side of the tube to the intake side, and the displacement forces $P_j(x_j)$ are applied according to the control commands from the motor-phoneme controller.

3. LEARNING OF VOCALIZATION SKILL

An adaptive learning algorithm for the achievement of a talking and singing performance is introduced in this section. The algorithm consists of two phases. First in the learning phase, the system acquires two maps in which the relations between the motor positions and the features of generated voices are established and stored. One is a motor-pitch map, which associates motor positions with fundamental frequencies. It is acquired by comparing the pitches of vocalized sounds with the desired pitches, which cover the frequency range of speech [8]. The other is a motor-phoneme map, which associates motor positions with phonetic features of vowel and consonant sounds. Second in the performance phase, the robot speaks and sings by referring to the obtained maps, while pitches and phonemes of generated voices are adaptively maintained by hearing its own output voices.

3.1. Neural network learning of vocalization

The neural network (NN) works to associate the sound characteristics with the control parameters of the nine motors settled in the vocal tract and the nasal cavity. In the learning process, the network learns the motor control commands by inputting 10th-order linear predictive coding (LPC) cepstrum coefficients [9] derived from vocal sound waves as teaching signals. The network acquires the relations between the sound parameters and the motor control commands of the vocal tract. After the learning, the neural network is connected in series into the vocal tract model. By inputting the sound parameters of desired sounds to the NN, the corresponding form of the vocal tract is obtained.

In this study, the self-organizing neural network (SONN) was employed for the adaptive learning of vocalization. Figure 2 shows the structure of the SONN consisting of two processes, which are an information memory process and an information recall process. After the SONN learning, the motor control parameters are adaptively recalled by the stimuli of sounds to be generated.

The information memory process is achieved by the self-organizing map (SOM) learning [10], in which sound parameters are arranged onto a two-dimensional feature map to be related to one another.

Weight vector $\mathbf{V}_j$ at node j in the feature map is fully connected to the input nodes x_i $[i = 1, \dots, 10]$, where 10th-order LPC cepstrum coefficients are given. The map learning algorithm updates the weight vectors $\mathbf{V}_j$-s. A competitive learning is used, in which the winner c as the output unit with a weight vector closest to the current input vector $\mathbf{x}(t)$ is chosen at time t in learning. By using the winner c, the weight vectors $\mathbf{V}_j$-s are updated according to the rule shown below;

$$\mathbf{V}_j(t+1) = \mathbf{V}_j(t) + h_{cj}(t)\left[\mathbf{x}(t) - \mathbf{V}_j(t)\right],$$
$$h_{cj}(t) = \begin{cases} \alpha(t) \cdot \exp\left(-\dfrac{\|r_c - r_j\|^2}{2\sigma^2(t)}\right) & (i \in N_c), \\ 0 & (i \notin N_c). \end{cases} \quad (1)$$

Here, $\|r_c - r_j\|$ is the distance between units c and j in the output array, and N_c is the neighborhood of the node c. $\alpha(t)$ is a learning coefficient which gradually reduces as the learning proceeds. $\sigma(t)$ is also a coefficient which represents the width of the neighborhood area.

Then, in the information recall process, each node in the feature map is associated with motor control parameters for the control commands of nine motors employed for the vocal tract deformation, by using the three-layered perceptron. In this study, a conventional back-propagation algorithm was employed for the learning. With the integration of the information memory and recall processes, the SONN works to adaptively associate sound parameters with motor control parameters.

In the current system, 25 × 25 arrayed map $\mathbf{V} = [V_1, V_2, \dots, V_{25\times25}]$ is used as the SOM. For testing the mapping ability, 200 sounds randomly vocalized by the robot were mapped onto the map array. After the self-organizing learning, five Japanese vowels vocalized by six different people were mapped onto the feature map. Same vowel sounds given by different people were mapped close with each other, and five vowels were roughly categorized according to the differences of phonetic characteristics. We found that, in some vowel area, two sounds given by two different speakers fell in a same unit in the feature map. It means that the two different sounds could not be separated, although they have close tonal features with each other. We propose a reinforcement learning algorithm to optimize the feature map.

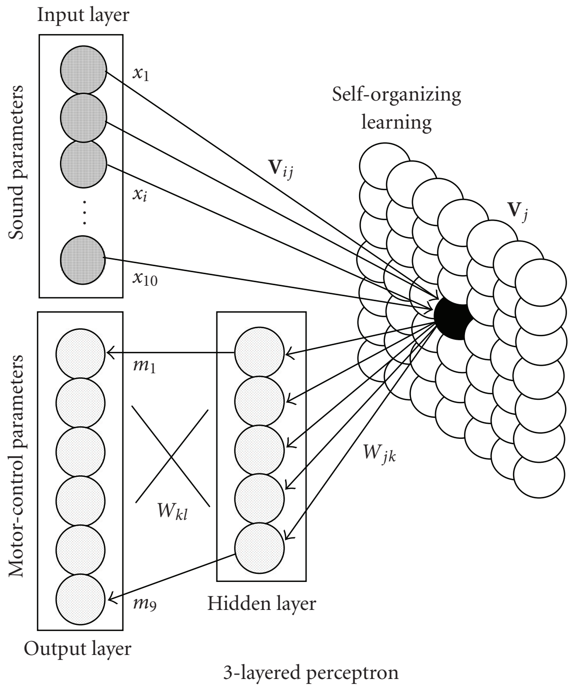


Figure 2: Structure of self-organizing neural network.

3.2. Reinforcement learning of five Japanese vowels by human voices

Redundant sound parameters which were not used for the Japanese speech were buried in the map, since the 150 inputted sounds were generated randomly by the robot. Furthermore, two different sounds given by two different speakers were occasionally fallen in the same unit. The mapping should be optimized for the Japanese vocalization.

The reinforcement learning was employed to establish the feature map optimized. After the SONN learning, five Japanese vowel sounds given by 6 different speakers with normal audition were applied to the supervised learning as the reinforcement signal to be associated with the suitable motor control parameters for the Japanese vocalization.

Figure 3 shows the result of the reinforcement learning with five Japanese vowels given by five speakers no. 1 to 5. The distribution of same vowel sounds concentrated with one another, and the patterns of different vowels were placed apart.

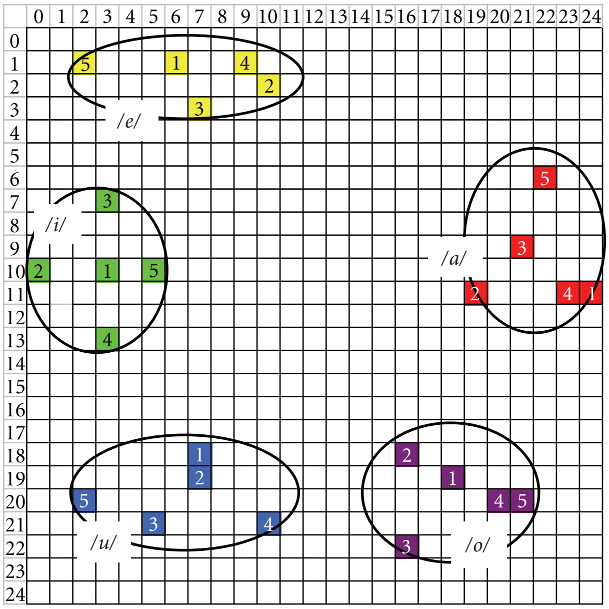


Figure 3: Result of reinforcement learning with five Japanese vowels from 5 subjects no. 1–5.

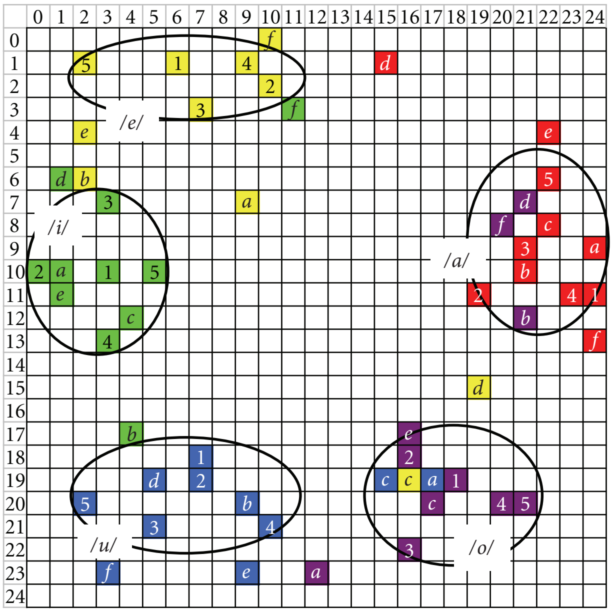


Figure 4: Mapping results of six different voices given by hearing-impaired speakers no. a–c.

4. ARTICULATORY REPRODUCTION OF HEARING-IMPAIRED VOICE

After the learning of the relationship between the sound parameters and the motor control parameters, we inputted human voices from microphone to confirm whether the robot could speak autonomously by mimicking human vocalization. With the comparison of spectra between human vowel vocalization and robot speech, we confirmed that the first and second formants F1 and F2, which present the principal characteristics of the vowels, were formed properly as to approximate the human vowels, and the sounds were well distinguishable by listeners. The experiment also showed the smooth motion of the vocalization. The transition between two different vowels in the continuous speech was well acquired by the SONN learning, which means that all the cellsoninthe SOM are associated with motor control parameters properly to vocalize particular sounds [11].

Voices of hearing-impaired people then were given to the robot so as to confirm that the articulatory motion would be reproduced by the robot. Figure 4 shows the mapping results of six different voices given by hearing-impaired speakers no. a, no. b, no. c, no. d, no. e, and no. f. The same colors indicate the vocal sounds generated by the same vowels. In Figure 5, vocal tract shapes estimated by the robot from voices of hearing-impaired person no. a are presented, together with the comparison of the vocal tract shapes estimated by the able-bodied speaker no. 1 voices.

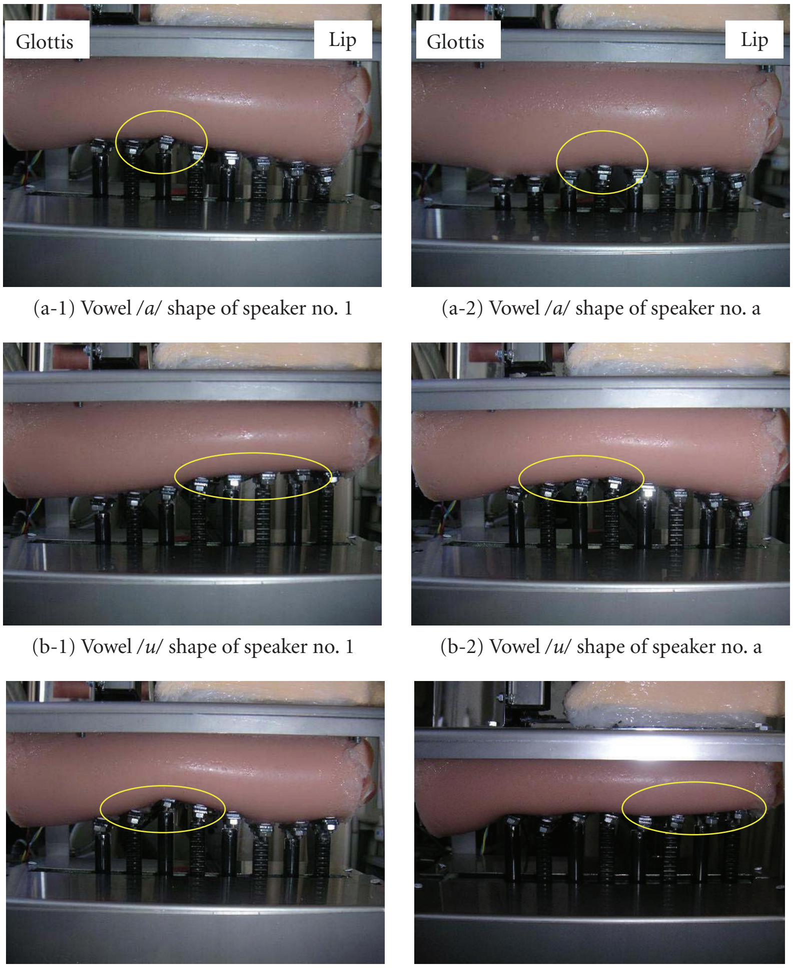
From the observation of the robot's reproduced motions of the vocal tract, the articulations of auditory-impaired people were apparently small, and complex shapes of vocal tract were not sufficiently articulated. Furthermore, in the map shown in Figure 4, /*u*/ sound given by the hearing-impaired speaker no. a is located inside the /*e*/ area of able-bodied speakers, and his /*o*/ vowel is located close to the /*u*/ area of able-bodied speakers. These articulatory characteristics also appear in the vocal tract shapes shown in Figure 5. In the figures, the vowel /*u*/ shape of speaker no. a shown in (b-2) is almost the same with the /*o*/ shape of speaker no. 1 presented in (c-1). Likewise, the /*o*/ shape shown in (c-2) appears close to the shape of (b-1). Thus, these results proved that the topological relations of resonance characteristics of voices were well preserved in the map, and the articulatory motion by the robot was successfully obtained to reproduce the speech articulation by listening arbitrary vocal sounds.

5. INTERACTIVE VOICE TRAINING SYSTEM FOR HEARING-IMPAIRED PEOPLE

In the speech training, the robot interactively shows the articulatory motion of vocal organs as a target to a trainee so thats/he repeats his/her vocalization and the observation of the robot motion. The trainee is also able to refer to the SOM to find the distance to the target voice. The flow of the training is summarized in Figure 6. The training of speech articulation by an auditory-impaired subject is shown in Figure 7.

Subject

An experiment of speech training was conducted by six hearing-impaired subjects: no. a–f (four males and two females), who study in a high school and a junior high school. In Figure 8, the training results of three subjects no. a, no. e, and no. f are shown by presenting the trajectories of voices appeared in the SOM during the training experiments. Figure 8(a) shows a result of successful training with less trials conducted by the subject no. a. By observing the articulatory motion instructed by the robot, this subject recognized



(a-1) Vowel /*a*/ shape of speaker no. 1

(a-2) Vowel /*a*/ shape of speaker no. a

(b-1) Vowel /*u*/ shape of speaker no. 1

(b-2) Vowel /*u*/ shape of speaker no. a

(c-1) Vowel /*o*/ shape of speaker no. 1

(c-2) Vowel /*o*/ shape of speaker no. a

Figure 5: Comparison of vocal tract shapes of the hearing-impaired (right) with the able-bodied (left).

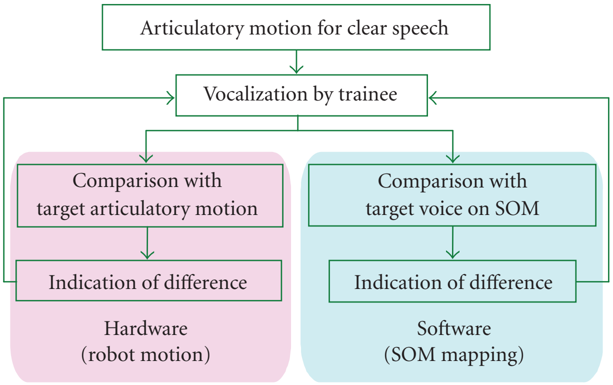


Figure 6: Flowchart of training of speech articulation.

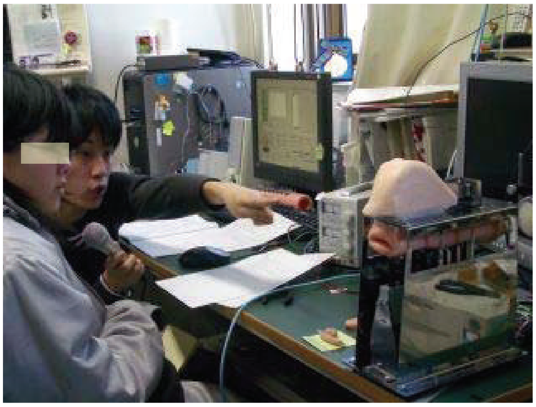
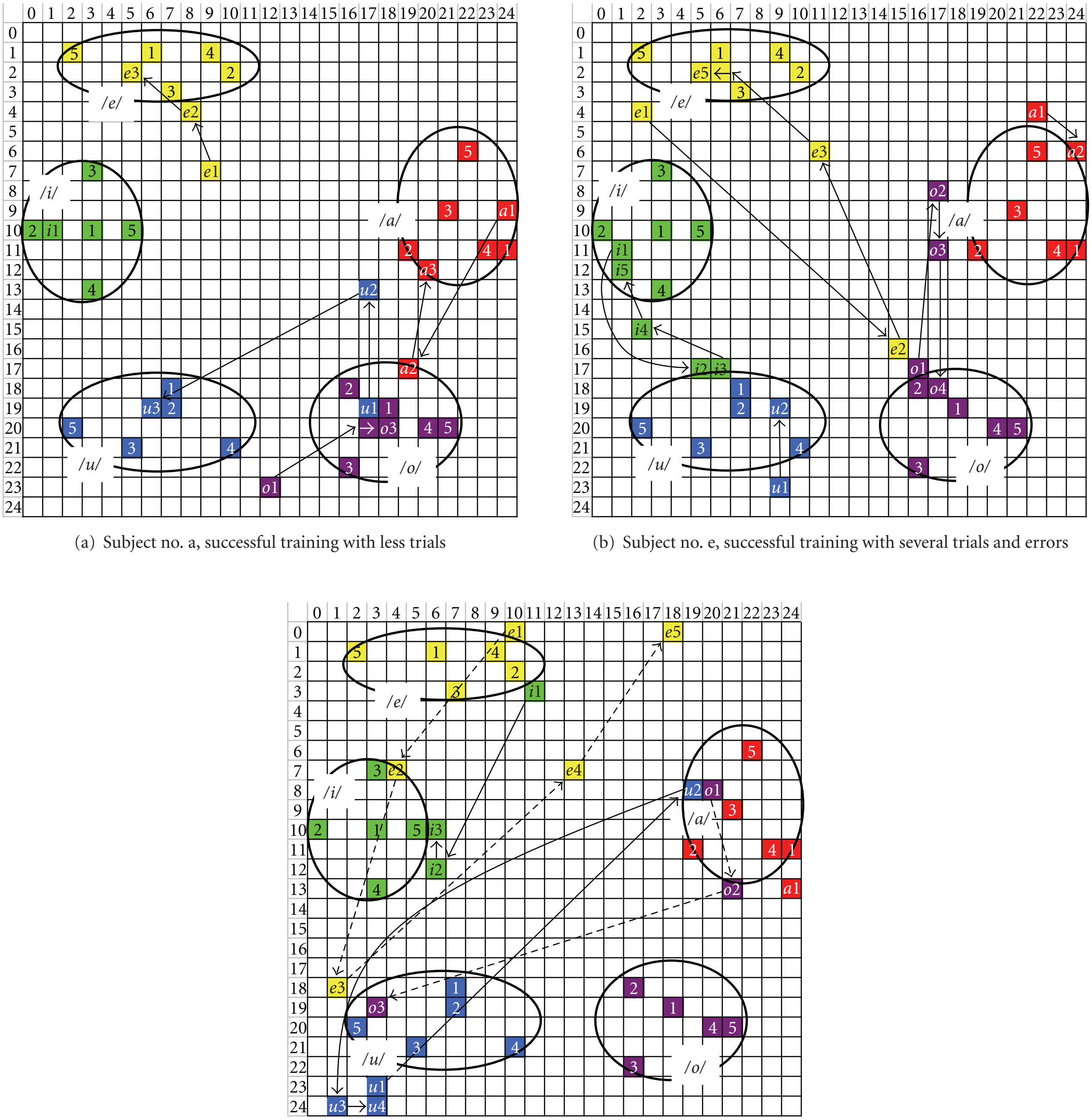


Figure 7: Training of speech articulation by auditory-impaired people.

the difference in his articulation and effectively learned the correct motion. Figure 8(b) also shows the successful training results by the subject no. e, however, he had achieved the vocalization by repeating several trials and errors, especially for the vowels /*i*/ and /*e*/ as presented by the arrows from *i1* to *i5* and *e1* to *e5*, respectively.

In the case of the training conducted by the subject no. f, he could not achieve the learning by the system. The clarity of



(a) Subject no. a, successful training with less trials

(b) Subject no. e, successful training with several trials and errors

(c) Subject no. f, fail of training

Figure 8: Example trajectories in training.

his voices was quite low, and the original voices were mapped far from the area of clear voices. He could not understand the shape of the robot's vocal tract, nor realize the correspondence between the robot's motion and the motion of his inner mouth. This subject tried to articulate his vocal tract following the articulatory motion indicated by the robot, however, his voice moved to the different direction in the SOM as shown by arrows in Figure 8(c). He failed the acquisition of vocalization skill and could not achieve the training. In the questionnaire after the training, he pointed out the difficulties of moving a particular part of the inner mouth so as to mimic the articulatory motion of the robot.

By the experimental training, five subjects could mimic the vocalization following the directions given by the robotic voice simulator, and acquired the better vocal sounds. In the questionnaire after the experiment, two subjects commented that the correspondence between robot's vocal tract and human actual vocal tract should be instructed, so that they could easily understand which part inside the mouth should be intensively articulated for the clear vocalization.

6. CONCLUSIONS

A robotic voice simulator and its articulatory reproduction of voice of hearing-impaired people were introduced in this paper. By introducing the adaptive learning and controlling of the mechanical model with the auditory feedback, the voice robot was able to acquire the vocalization skill as a human baby does in speech training.

The robot was applied to introduce a training system for auditory-impaired people to interactively train the speech articulation for learning proper vocalization. The robotic voice simulator reproduces the articulatory motion just by listening to actual voices given by auditory-impaired people, and they could learn and know how to move their vocal organs for the clear vocalization, by observing the motions instructed by the talking robot. The use of SOM for visually presenting the distance between target voice and trainee's voice is also introduced.

We confirmed that the training using the talking robot and the SOM would help hearing-impaired people learn the articulatory motion in the mouth and the skill of clear vocalization properly. In the next system, the correspondence between robot's vocal tract and human actual vocal tract should be established so that a subject could understand which part inside the mouth should be intensively articulated in the training. By analyzing the vocal articulation of auditory-impaired people during the training with the robot, we will investigate the factor of unclarity of their voices originated by the articulatory motions.

ACKNOWLEDGMENTS

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.

REFERENCES

[1] A. Boothroyd, *Hearing Impairments in Young Children*, Alexander Graham Bell Association for the Deaf, Washington, DC, USA, 1988.

[2] A. Boothroyd, "Some experiments on the control of voice in the profoundly deaf using a pitch extractor and storage oscilloscope display," *IEEE Transactions on Audio and Electroacoustic*, vol. 21, no. 3, pp. 274–278, 1973.

[3] N. P. Erber and C. L. de Filippo, "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, 1978.

[4] M. H. Goldstein and R. E. Stark, "Modification of vocalizations of preschool deaf children by vibrotactile and visual displays," *Journal of the Acoustical Society of America*, vol. 59, no. 6, pp. 1477–1481, 1976.

[5] H. Sawada and S. Hashimoto, "Adaptive control of a vocal chord and vocal tract for computerized mechanical singing instruments," in *Proceedings of the International Computer Music Conference (ICMC '96)*, pp. 444–447, Hong Kong, September 1996.

[6] T. Higashimoto and H. Sawada, "Vocalization control of a mechanical vocal system under the auditory feedback," *Journal of Robotics and Mechatronics*, vol. 14, no. 5, pp. 453–461, 2002.

[7] T. Higashimoto and H. Sawada, "A mechanical voice system: construction of vocal cords and its pitch control," in *Proceeding of the 4th International Conference on Intelligent Technologies (InTech '03)*, pp. 762–768, Chiang Mai, Thailand, December 2003.

[8] H. Sawada, M. Nakamura, and T. Higashimoto, "Mechanical voice system and its singing performance," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '04)*, vol. 2, pp. 1920–1925, Sendai, Japan, September-October 2004.

[9] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, Germany, 1995.

[10] J. D. Markel, *Linear Prediction of Speech*, Springer, New York, NY, USA, 1976.

[11] M. Nakamura and H. Sawada, "Talking robot and the analysis of autonomous voice acquisition," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pp. 4684–4689, Beijing, China, October 2006.