Talking Robot and the Autonomous Acquisition of Vocalization and Singing Skill

Hideyuki Sawada
Kagawa University
Japan

1. Introduction

Voice is used as primary media in the human communication. It is employed not only in simple daily communication, but also for the logical discussions and the expression of emotion and feelings. Different vocal sounds are generated by the complex movements of vocal organs under the feedback control mechanisms using an auditory system. Vocal sounds and human vocalization mechanisms have been the attractive researching subjects for many researchers so far [1],[2], and computerized voice production and recognition have become the essential technologies in the recent developments of flexible human-machine interface studies.

Various ways and techniques have been reported in the researches of sound production. Algorithmic syntheses have taken the place of analogue circuit syntheses and became widely used techniques [2],[3]. Sound sampling methods and physical model based syntheses are typical techniques, which are expected to provide different types of realistic vocal sounds [4]. In addition to these algorithmic synthesis techniques, a mechanical approach using a phonetic or vocal model imitating the human vocalization mechanism would be a valuable and notable objective.

Several mechanical constructions of a human vocal system to realize human-like speech have been reported. In most of the researches [2],[5],[6], however, the mechanical reproductions of the human vocal system were mainly directed by referring to X-ray images and FEM analysis, and the adaptive acquisition of control methods for natural vocalization have not been considered so far. In fact, since the behaviours of vocal organs have not been sufficiently investigated due to the nonlinear factors of fluid dynamics yet to be overcome, the control of mechanical system has often the difficulties to be established.

The author has been developing a mechanical voice generation system together with its adaptive learning of the control skill for the realization of a talking robot which imitates human vocalization [7]-[11]. The fundamental frequency and the spectrum envelope determine the principal characteristics of a sound. The former is the characteristic of a source sound generated by a vibrating object, and the latter is operated by the work of the resonance effects. In vocalization, the vibration of vocal cords generates a source sound, and then the sound wave is led to a vocal tract, which works as a filter to determine the spectrum envelope.

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

A motor-controlled mechanical model with vocal cords, a vocal tract and a nasal cavity is constructed so far to generate a natural voice imitating a human vocalization. By introducing an auditory feedback learning with an adaptive control algorithm of pitch and phoneme, the robot is able to autonomously acquire the control method of the mechanical system to produce stable vocal sounds imitating human vocalization skill [7],[8]. In this chapter, the adaptive control method of mechanical vocal cords and vocal tract for the realization of a talking and singing robot is described, together with the singing performance with the use of acquired vocalization skill.

2. Human voice system and voice generation

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 the 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. On the other hand, the spectrum envelope or the resonance characteristics, 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 cords and vocal tract for the production of vocal sounds are acquired as infants grow by repeating vocalization and listening through trial and error.

3. Mechanical model for vocalization

3.1 Configuration of Mechanical Voice System

As shown in Figure 1, the mechanical voice system 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 audition of a human.
The air in the compressor is compressed to 8000 hpa, while the pressure of an air from lungs is about +200 hpa larger than the atmospheric pressure. A pressure reduction valve is applied at the outlet of the air compressor so that the pressure is reduced to be nearly equal to the air pressure through the trachea. The valve is also effective to reduce the fluctuation of the pressure in the compressor during the operations of compression and depression process. The decompressed air 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 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 the generated sound in real time, which are necessary for the auditory feedback control. The system controller manages the whole system by listening to the produced sounds and generating motor control commands, based on the auditory feedback control mechanism employing a neural network learning.

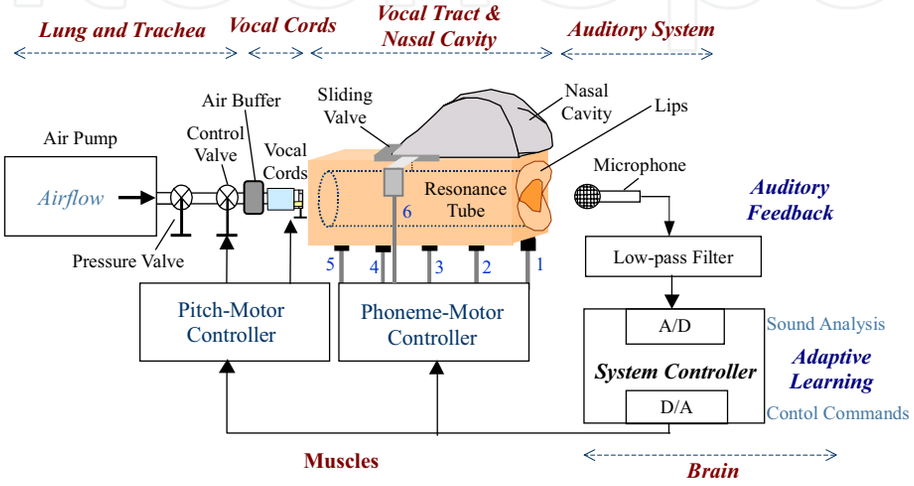


Figure 1. Configuration of the talking robot

3.2 Construction of Resonance Tube and Nasal Cavity

The human vocal tract is a non-uniform tube about 170mm long in man. Its cross-sectional area varies from 0 to 20cm^2 under the control for vocalization. A nasal tract 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 36mm, which is equal to 10.2cm^2 by the cross-sectional area as shown in Figure 2 and 3. 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 connected to the intake part of the resonance tube to vocalize nasal sounds like /*m*/ and /*n*/.

By actuating displacement forces by 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. DC motors are placed at 5 positions x_j (j=1-5) from the intake side of the tube to the outlet side as shown in Figure 2, and the

displacement forces $P_j(x_j)$ are applied according to the control commands from the motor-phoneme 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 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 sliding valve is open to lead the air into the nasal cavity as shown in Figure 4(a). 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 /*p*/ and /*t*/, the mechanical system closes the sliding 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 as shown in Figure 4(b). Then the released air generates plosive consonant sounds like /*p*/ and /*t*/.

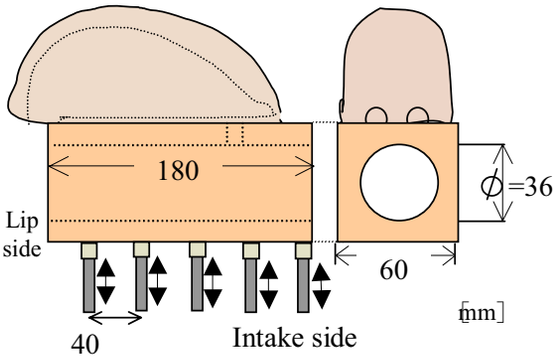


Figure 2. Construction of vocal tract and nasal cavity

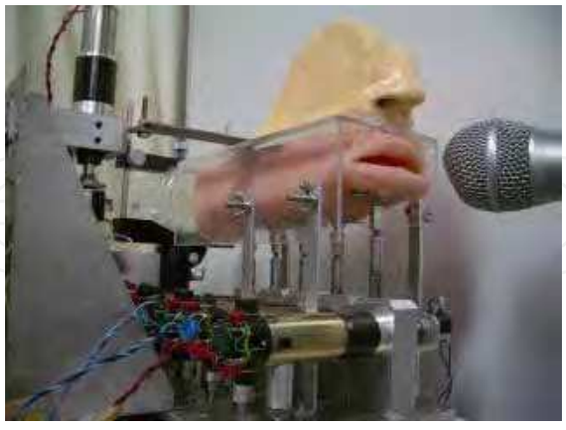
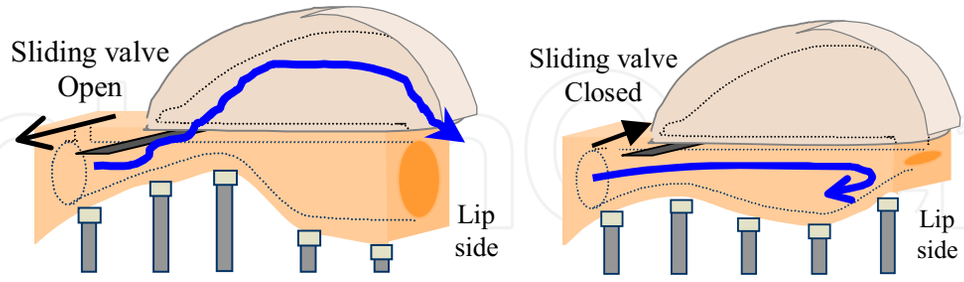


Figure 3. Structural view of talking robot



(a) Airflow Control for Nasal Sound /*n*/ (b) Airflow Control for Plosive Sound /*p*/

Figure 4. Motor control for nasal and plosive sound generation

3.3 Artificial Vocal Cords and its pitch control

(A) Construction of Artificial Vocal Cords

The characteristic of a glottal sound wave, which determines the pitch and the volume of human voice, is governed by the complex behavior of the vocal cords. It is due to the oscillatory mechanism of human organs consisting of the mucous membrane and muscles excited by the airflow from the lung. Although several researching reports about the computer simulations of these movements are available [12], the author have focused on generating the wave using a mechanical model [7].

In this study, we constructed new vocal cords with two vibrating cords molded with silicone rubber with the softness of human mucous membrane. Figure 5 shows the picture. 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.

Here, assume the simplified dynamics of the vibration given by a strip of a rubber with the length of L. The fundamental frequency f is given by the equation

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

by considering the density of the material D and the tension S applied to the rubber. This equation implies that the fundamental frequency varies according to the manipulations of L, S and D.

The tension of rubber can be manipulated by applying tensile force to the two cords. Figure 6 shows the schematic figures how tensile force is applied to the vocal cords to generate sounds with different frequencies. By pulling the cords, the tension increases so that the frequency of the generated sound becomes higher. For the voiceless sounds, just by pushing the cords, the gap between two cords are left open and the vibration stops.

The structure of the vocal cords proved the easy manipulation for the pitch control, since the fundamental frequency can be changed just by giving tensile force for pushing or pulling the cords.

We constructed three vocal cords with different kinds of softness, which are hard, soft and medium. The medium one has two-layered construction: a hard silicone is inside with the soft coating outside.

Figure 5. Vocal cords

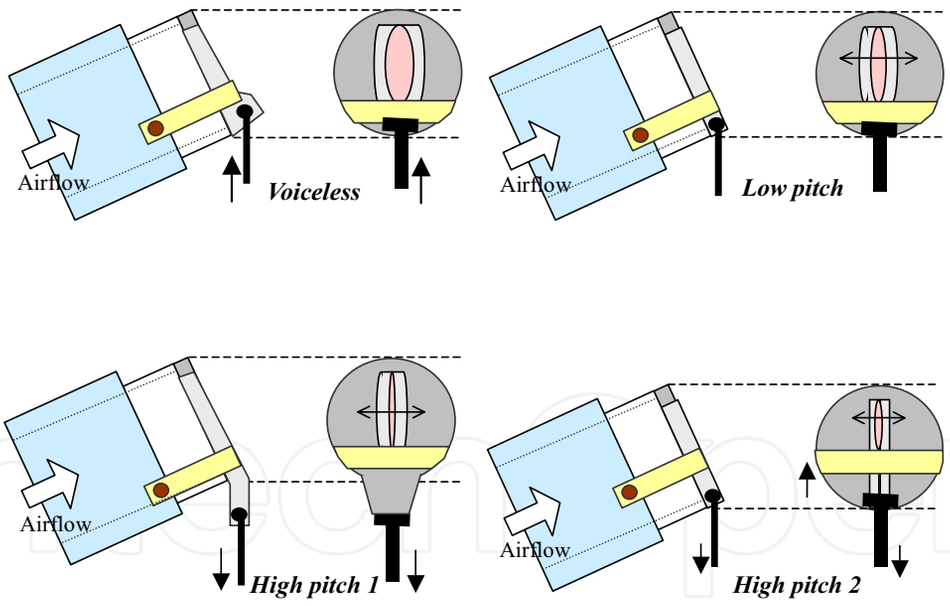
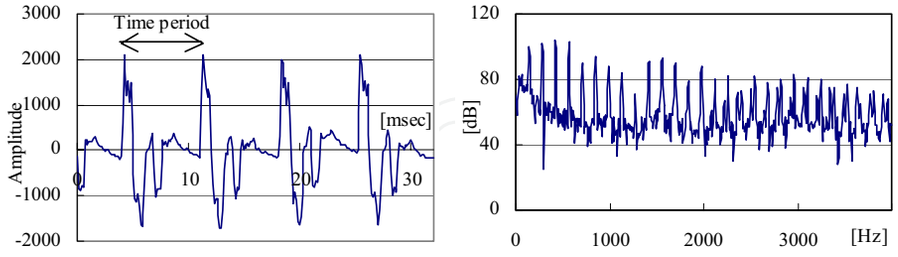
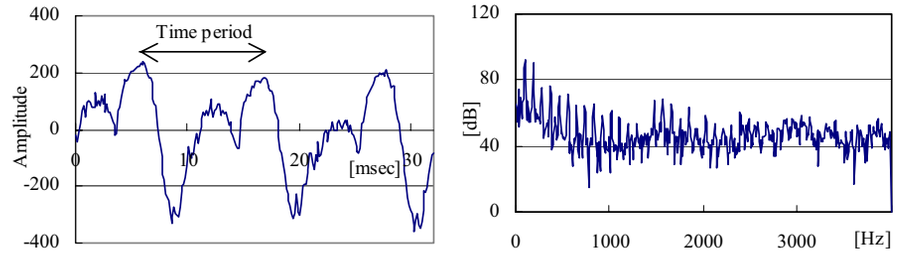


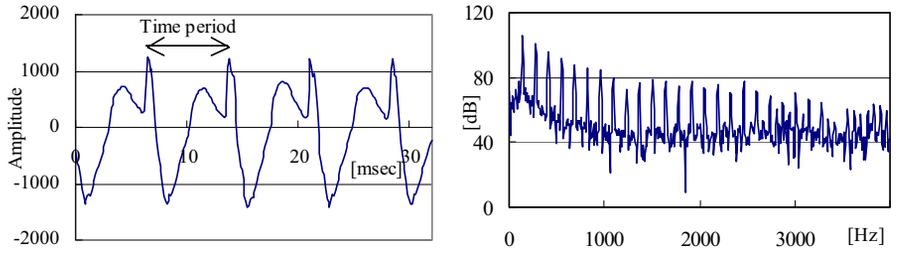
Figure 6. Different manipulations for pitch control



(a) Hard vocal cords



(b) Soft vocal cords



(c) Two-layered vocal cords

Figure 7. Waveforms and spectra of three vocal cords

Figure 7 shows examples of sound waves and its spectra generated by the three vocal cords. The waveform of the hard cords is approximated as periodic pulses, and a couple of resonance peaks are found in the spectrum domain. The two-layered cords generate an isolated triangular waveform, which is close to the actual human one, and its power in the spectrum domain gradually decreases as the frequency rises.

In this study, the two-layered vocal cords are employed in the mechanical voice system. Figure 8 shows the vocal cords integrated in the control mechanism.

(B) Pitch Control of Vocal Cords

Figure 9 shows experimental results of pitch changes using the two-layered vocal cords. The fundamental frequency varies from 110 Hz to 250 Hz by the manipulations of a force applying to the rubber.

The relationship between the applied force and the articulated frequency is not stable but tends to change with the repetition of experiments due to fluid dynamics. The vocal cords, however, reproduce the vibratory actions of human actual vocal cords, and are also considered to be suitable for our system because of its simple structure. Its frequency characteristics are easily controlled by the tension of the rubber and the amount of airflow.
For the fundamental frequency and volume adjustments in the voice system, two motors are used: one is to manipulate a screw of an airflow control valve, and the other is to apply a tensile force to the vocal cords for the tension adjustment.

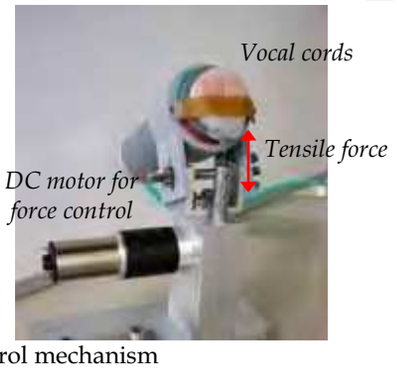


Figure 8. Vocal cords and control mechanism

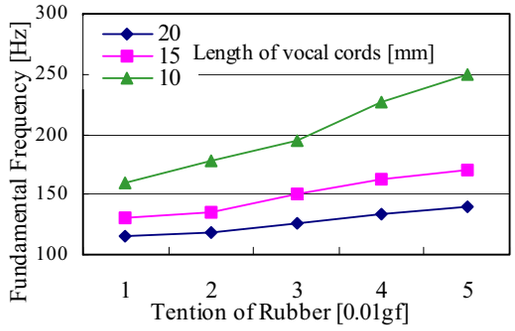


Figure 9. Relation between tensile force and fundamental frequency

(C) Adaptive Pitch Control

Not only adjusting but also maintaining the pitch of output sounds is not easy tasks due to the dynamic mechanism of vibration, which is easily disturbed by the fluctuations of the tensile force and the airflow. Stable output has to be obtained no matter what kind of disturbance applies to the system. Introducing an adaptive control mechanism would be a good solution for getting such robustness [7],[8].
An adaptive tuning algorithm for the production of a voice with different pitches using the mechanical voice system is introduced in this section. The algorithm consists of two phases. First in the learning phase, the system acquires a motor-pitch map, in which the relations

between the motor positions and the fundamental frequencies are described. It is acquired by comparing the pitches of output sounds with the desired pitches for the vocalization. Then in the performance phase, the robot utters words by referring to the obtained map while pitches of produced voices are adaptively maintained by hearing its own outputs.
Figure 10 shows a schematic diagram of the adaptive pitch learning in the learning phase. The algorithm simulates the pitch learning process of a human in practicing singing. The algorithmic process of the pitch acquisition in the system controller is shown in the dotted lines. The pitch-tuning manager manages the behaviors of all the other units presented in the boxes. The system starts its action by receiving a present-position vector v_p as a command to let the motors move. Actual values of the vector elements can be estimated by the work of calculations in the pitch-to-motor translator, which is trained in advance to output desired motor positions from pitches of produced sounds according to the relations between tensile force and fundamental frequency as shown in Figure 9.

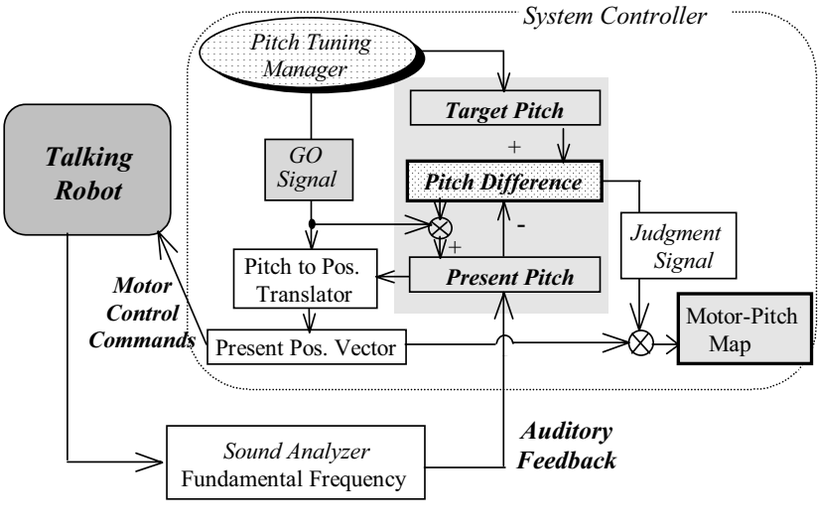


Figure 10. Adaptive pitch learning

First, the system controller starts with setting arbitrary values as a present-position vector to send to the vocal system. The fundamental frequency of the generated sound is calculated by the sound analyzer of the auditory system which realizes FFT calculations in realtime. Since the desired pitches of the produced sounds are assumed to be lower than 500 Hz in the vocal system, the sampling frequency is set to 1kHz. The analysis window of 1024 points are chosen and the frequency resolution is about 1Hz in this system. After applying the Hamming window, the produced sound wave data are fed to the FFT algorithm and the fundamental pitch is extracted. To compare with the desired pitch, the difference between the two pitches is obtained according to the tuning signal trigger generated by the tuning manager. The tuning signal, in the same instant, drives the pitch-to-motor translator to let the motors work. As the feedback process repeats, the pitch difference between the target pitch and the produced pitch decreases. When the pitch difference becomes smaller than a predetermined threshold value, which is currently set to 0.6 Hz, the judgment-signal unit arises, so that the present-position vector is associated with the target pitch and stored as the motor-pitch map.

The results of the pitch learning based on the auditory feedback are shown in Figure 11, in which the system acquired the sound pitches from C to G. The system was able to acquire vocal sounds with desired pitches.

3.4 Learning of Vowel and Consonant Vocalization

The neural network (NN) works to associate the resonance characteristics of sounds with the control parameters of the six motors equipped in the vocal tract and the nasal cavity, as shown in Figure 12. In the learning process, the network learns the motor control commands by inputting 10th order LPC cepstrum coefficients 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.

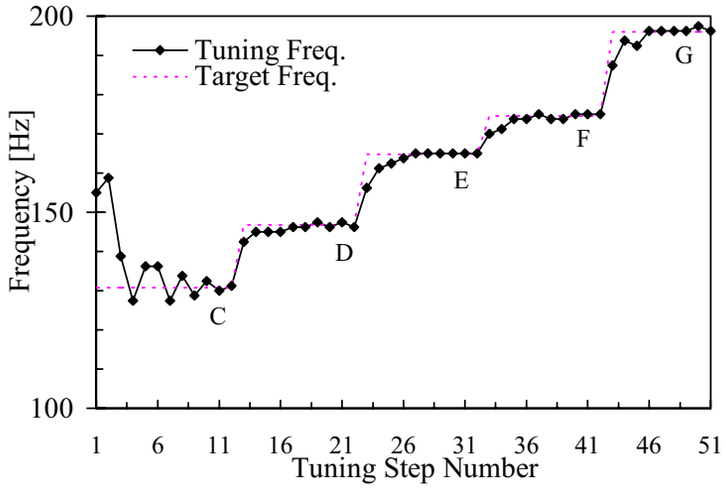
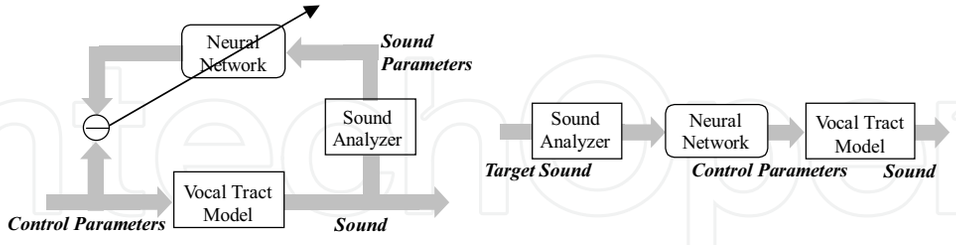


Figure 11. Experimental result of pitch tuning



(a) Learning Process (b) Running Process

Figure 12. Neural network for vocalization learning

In this study, the Self-Organizing Neural Network (SONN) was employed for the adaptive learning of vocalization. Figure 13 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, in which sound parameters are arranged onto a two-dimensional feature map to be related to one another.

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

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

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.

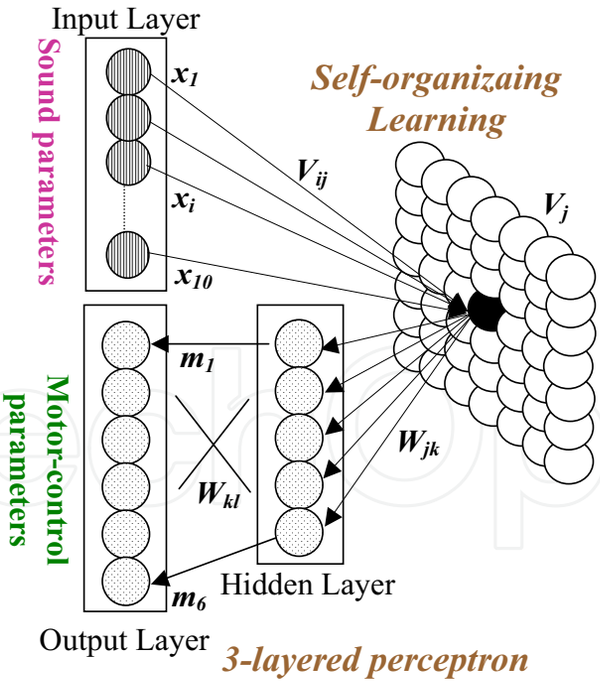


Figure 13. Structure of self-organizing neural network (SONN)

Then in the information recall process, each node in the feature map is associated with motor control parameters for the control commands of six 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, 20x20 arrayed map $\boldsymbol{V}$ =[V_1 , V_2 , ..., $V_{20\ x20}$] is used as the SOM. For testing the mapping ability, 150 sounds randomly vocalized by the robot were mapped onto the map array. After the self-organizing learning, Japanese five vowels vocalized by six different people (No.1 to 6) were mapped onto the feature map, which is shown in Figure 14. 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 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.

3.5 Reinforcement Learning of Japanese Five Vowels by Human Voices

SONN gave fairly good performance in the association of sound parameters with motor control parameters for the robot vocalization, however redundant sound parameters which are not used for the Japanese speech are also buried in the map, since the 150 inputted sounds were generated randomly by the robot. Furthermore, two different sounds given by two different speakers are 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, Japanese five vowel sounds given by 6 different speakers were applied to the supervised learning as the reinforcement signal to be associated with the suitable motor control parameters for the vocalization.
Figure 15 shows the result of the reinforcement learning with Japanese 5 vowels. The distribution of same vowel sounds concentrates with one another, and the patterns of different vowels are placed apart.

3.6 Reinforcement Learning of Japanese Five Vowels by Human Voices

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.
Figure 16 shows the comparison of spectra between human vowel vocalization and robot speech. The first and second formants F1 and F2, which present the principal characteristics of the vowels, were formed as to approximate the human vowels, and the sounds were well distinguishable by listeners.

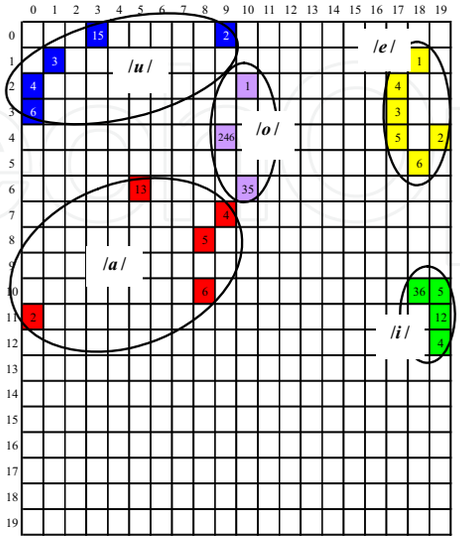


Figure 14. Experimental result of Japanese 5 vowel mapping by SONN

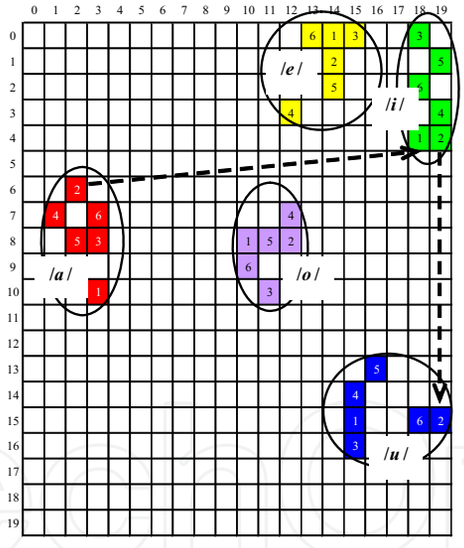
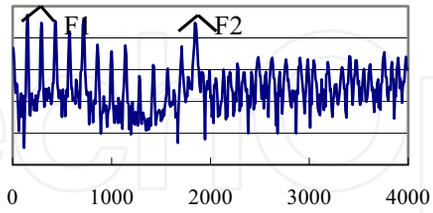
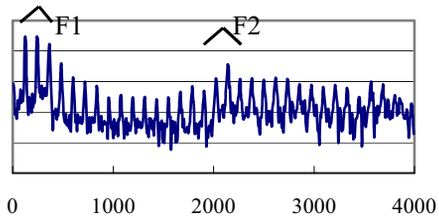


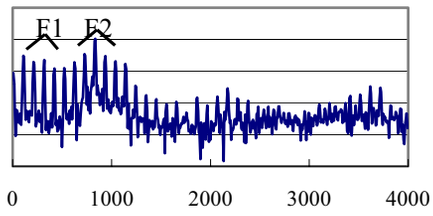
Figure 15. Result of reinforcement learning with Japanese 5 vowels



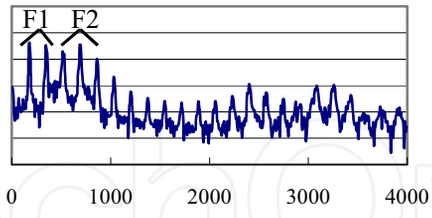
(*a*-1) Robot Speech /*i*/



(*a*-2) Human Speech /*i*/



(*b*-1) Robot Speech /*o*/



(*b*-2) Human Speech /*o*/

Figure 16. Comparison of spectra of Japanese /i/ and /o/ vowel

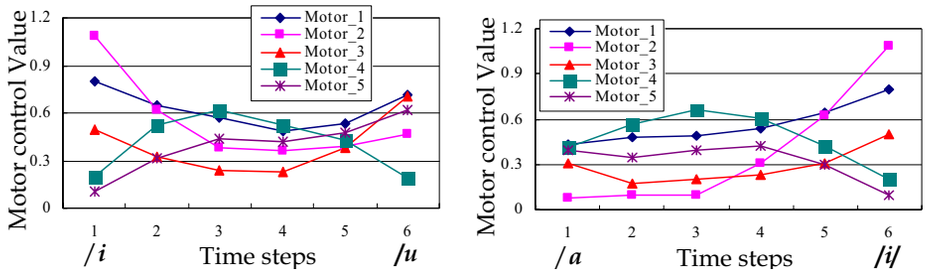
We also verified the robot's vocalization motion. In the /*a*/ vocalization, for example, the glottal side was narrowed while the lip side was open, which was the same way as a human utters the /*a*/ sound. In the same manner, features for the /*i*/ pronunciation were acquired by narrowing the outlet side and opening the glottal side.

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.

Figure 17 shows two experimental results of the temporal motor control values of the vocal tract in the speech /*ai*/ and /*iu*/, where the motor values are autonomously generated by SOM as shown by the dotted arrows in Figure 15. The /*a*/ vocalization was transited to /*i*/ vocalization, then /*u*/ speech smoothly.

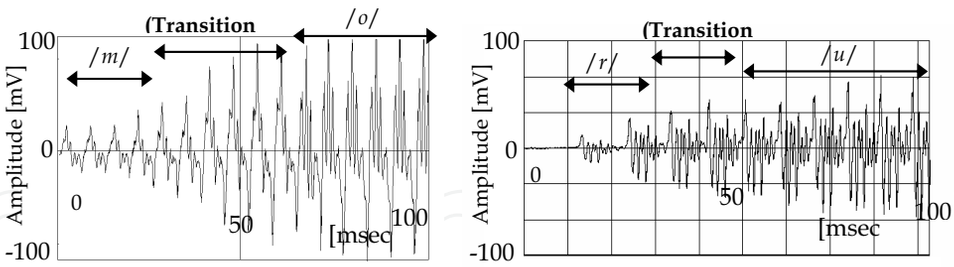
Nasal sounds such as /*m*/ and /*n*/ are generated with the nasal resonance under the control of the valve between the nasal cavity and the vocal tract. A voiced sound /*m*/ was generated by closing the lip and leading the sound wave to the nasal cavity, then by opening the lip and closing the nasal valve, the air was released to the mouth to vocalize /*o*/ sound.

Figure 18 shows generated sound waves in the vocalization of /*mo*/ and /*ru*/, where the smooth transition from consonant sounds to vowel sounds was achieved by the acquired articulations.



(*a*) Vocalization /*a*/ to /*i*/ (*b*) Vocalization /*i*/ to /*u*/

Figure 17. Transition of motor control values in vocalization



(*a*) /*ma*/ vocalization (*b*) /*ru*/ vocalization

Figure 18. Results of /*mo*/ and /*ru*/ vocalizations

4. Listening experiments

For the assessment of speech vocalized by the robot, listening experiments were conducted and voices were evaluated by questionnaires. 14 able-bodied subjects (9 males and 5

females) listened to 11 Japanese words and 1 English word given by the robot, while they were watching the motion of the robot vocalization, and answered what they thought it said. The results are shown in Table 1.

The words /Uma/ and /Ai/ were perfectly recognized by all the listeners, and 70.2 % recognition was achieved in average. The clarity of /*r*/ consonants was not enough to be distinguished, since the robot is not equipped with a flexible tongue to generate fricative sounds, and the sounds were often mis-recognized by the listeners as nasal sounds. Some listeners commented that the actual motion of the speech by the robot helped them to estimate what it speaks, although the speech itself was not clear enough.

Words	**Correct #**	**Listened as**
MonoMane	8	Mamomame, Mamamama
Raion	8	Airon, Naiyo, Nyanyane
Inu	9	Nna, Ushi
Uma	14	-
Momo	12	Oh-oh-
Remon	5	Imo, Ranran, Meron
Meron	9	Remon
Umi	9	Suna, Ume
Nami	12	Hami, Ii
Marimo	10	Mambo, Menzu, Mamma
Ai	14	-
Good-bye	8	Umai, Good-night
Average	**9.8**	**Recognition rate: 70.2%**

Table 1. Results of listening experiments

5. Singing Performance with Adaptive Control

The robot is able to sing a song by referring to the acquired maps for controlling its vocal organs. The schematic diagram of the singing performance is shown in figure 19. The performance-control manager takes charge of two tasks; one is for the performance execution presented by bold lines in the figure, and the other is for the adaptive control of pitches and phonemes with the auditory feedback during the performance. The score-information unit stores melody lines as sequential data of pitches, durations, phonemes and lyrics. Figure 20 shows one of the user interfaces , by which a user inputs musical score information for the singing performance.

The singing performance is executed according to the performance signals generated by the performance-control manager. The manager has the internal clock and makes a temporal

planning of note outputs with the help of the duration information in the score-information unit. The note information is translated into present-position vectors by referring to the motor-pitch map and the motor-phoneme map.

During the performance, unexpected changes of air pressure and tensile force cause the fluctuations of sound outputs. The adaptive control with the auditory feedback is introduced in this mechanical system by hearing the own output voices. The auditory units observe errors in the pitch and the phoneme so that the system starts fine tuning of produced sounds by receiving the tuning-signal trigger under the control of the performance manager. The motor-pitch / motor-phoneme maps are also renewed by the map-rewrite signal. The system is able to realize a stable singing performance under the adaptive mechanical control using the auditory feedback.

An experimental result of the singing performance is presented in Figure 21. In spite of the unexpected disturbances being applied or the drop of an air pressure in an air pump, the system was able to maintain the target pitch.

The system autonomously performs singing as a robot by generating performance signals which are governed by the internal clock in the computer. The robot also makes mimicking performance by listening and following a human singer. The auditory system listens to a human singer and extracts pitch and phonemes from his voice in realtime. Then the pitch and phonemes are translated into motor-control values to let the robot follow the human singer.

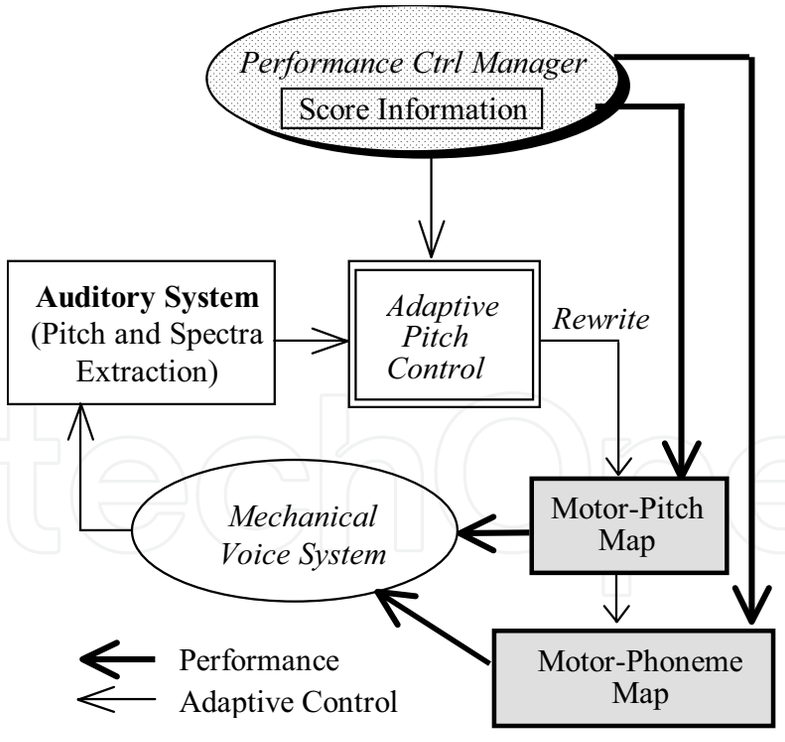


Figure 19. Singing performance with adaptive control

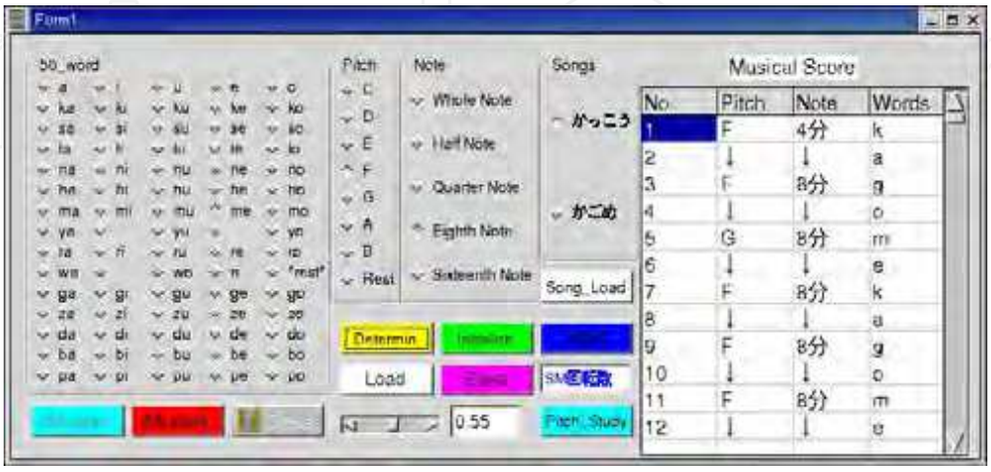


Figure 20. An interface for singing performance

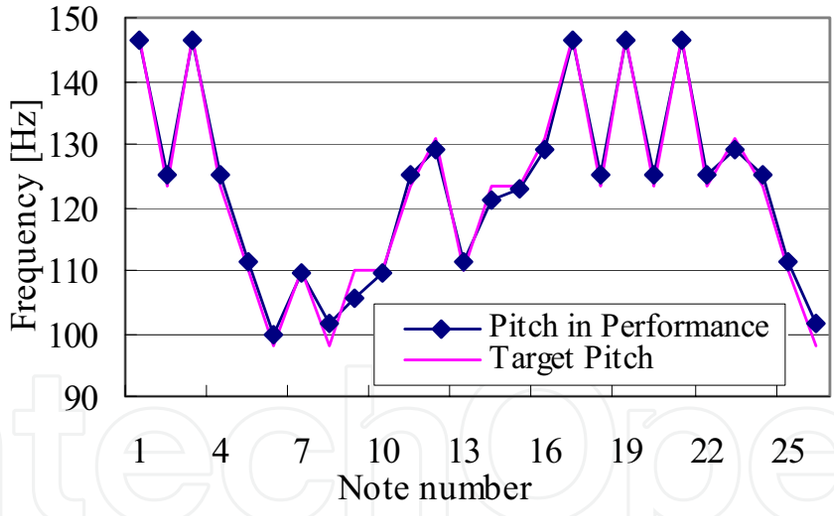


Figure 21. Result of singing performance

6. Conclusions

In this paper a talking and singing robot was introduced, which is constructed mechanically with human-like vocal cords and a vocal tract. 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 when he glows up, and generate vocal sounds whose pitches and phonemes are uniquely specified.
The authors are now working to develop a training device for auditory impaired people to interactively train the vocalization by observing the robot motion. The mechanical system reproduces the vocalization skills just by listening to actual voices. Such persons will be able to learn how to move vocal organs, by watching the motions of the talking robot.
A mechanical construction of the human vocal system is considered not only to have advantages to produce natural vocalization rather than algorithmic synthesis methods, but also to provide simulations of human acquisition of speaking and singing skills. Further analyses of the human learning mechanisms will contribute to the realization of a speaking robot, which learns and sings like a human. The proposed approach to the understandings of the human behavior will also open a new research area to the development of a human-machine interface.

8. 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 author would like to thank my students Mr. Toshio Higashimoto, Mr. Mitsuhiro Nakamura, Mr. Yasumori Hayashi and Mr. Mitsuki Kitani for their efforts for this research and study.

9. References

Y. Hayashi, "Koe To Kotoba No Kagaku", Houmei-do, 1979

J. L. Flanagan, "Speech Analysis Synthesis and Perception", Springer-Verlag. 1972

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

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

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

K. Fukui, K. Nishikawa, S. Ikeo, E. Shintaku, K. Takada, H. Takanobu, M. Honda, A. Takanishi: "Development of a Talking Robot with Vocal Cords and Lips Having Human-like Biological Structure", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp2526-2531, 2005.

H. Sawada and S. Hashimoto, "Adaptive Control of a Vocal Chord and Vocal Tract for Computerized Mechanical Singing Instruments", International Computer Music Conference, pp. 444-447, 1996

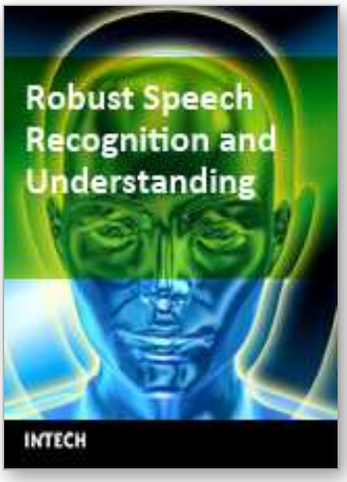
T. Higashimoto and H. Sawada, "Vocalization Control of a Mechanical Vocal System under the Auditory Feedback", Journal of Robotics and Mechatronics, Vol.14, No.5, pp. 453-461, 2002

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

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

M. Nakamura and H. Sawada, "Talking Robot and the Analysis of Autonomous Voice Acquisition", IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4684-4689, 2006

K. Ishizaka and J.Flanagan, "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Chords", Bell Syst. Tech. J., 50, 1223-1268, 1972



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen