2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

# Control Methods Based on Neural Network Forward and Inverse Models for a Biomechanical Structured Vocal Cord Model on an Anthropomorphic Talking Robot

Kotaro Fukui, *IEEE Member,* Eiji Shintaku, Akihiro Shimomura, Nana Sakakibara, Yuma Ishikawa
Masaaki Honda, *IEEE Member*, Atsuo Takanishi, *IEEE Member*

***Abstract*—We have developed a vocal control method, based on forward and inverse models, to allow the anthropomorphic talking robot Waseda Talker No. 7 (WT-7) to produce various kinds of voices. The control parameters of the vocal cords on WT-7 are pressure, vocal cord tension and glottal opening, and the acoustic parameters are sound pressure, sound pitch and spectrum slope. The relationships among these parameters are complicated and difficult to model using conventional methods. Here we present a neural network (NN) control method. The learning process consists of creation of the NN forward model by back propagation methods and optimization of the inverse model using the forward model. In addition, a real-time auditory feed-back mechanism is used to reduce the error between the target and the generated acoustic parameters. Using this method, the control parameters can be adjusted to follow the target voice well.**

## I. Introduction

SPEECH has a very important role in human communication. While considerable research has been conducted to clarify the mechanism for the production and control of speech, the complex aero-acoustics of the throat, the tongue and the nasal cavity are still not well understood. Furthermore, since speech is ability unique to humans, understanding the speech acquisition process will help us understand humanity itself.

Since 1998, we have been studying these problems by developing a human-like talking robot as a mechanical model of human speech production and control. This research is collaboration between researchers in robotics and acoustics: we use robotics to mimic the human vocal mechanism and acoustic theory to understand voice production. The voice production mechanism of anthropomorphic talking robot is the same as that of a human. The airflow from mechanical lungs causes the vocal cords to vibrate, producing a source sound. The vocal tract resonance characteristics are controlled by articulating the tongue, the jaw, the lips, and the velum. Other voice synthesis machines have been developed, including those by Kempelen [1], Umeda [2], Kawamura [3] and Sawada [4].

In 2005, Waseda Talker No. 5 (WT-5) was developed with a two-dimensional tongue mechanism that was able to produce a transition in the vocal tract area in the same manner as a human to produce the Japanese vowels (/a/, /i/, /u/, /e/ and /o/) and consonant sounds. WT-5 also had a new vocal cord mechanism that mimicked the human biomechanical structure. Its voice was more human-like than to those of previous robots. In 2006, we developed Waseda Talker No. 6 (WT-6), which has a 3D tongue driven by a release mechanism [5], to reproduce human-like tongue shape.

In another aspect of our research, we are trying to reproduce the human speech acquisition process. For a human baby, mimicking sounds is a very important part of speech acquisition. To reproduce this process, we developed methods to optimize the articulation parameters for vowel production and methods to optimize consonant production using sensory parameters [6].

In the development of talking robot, we had two problems. One is insufficient pitch range, and another is the control method for the human-based vocal cord.

The pitch control mechanism of WT-5 pulled the front side of the vocal cord model, changing the length of the glottis. However, this change only resulted in a 15[Hz]-pitch range. The range was too small to reproduce human conversation. To resolve this problem, WT-7 has a pair of discs directly attached vibrating part of its vocal cords to effectively control the tension.

In the view of the vocal organs control (the lungs and the vocal cords), we had not developed effective control methods for new model. Human vocal cords are consisted very complex mechanism, and the relation between the muscle movement and acoustic parameters, such as pitch or power, are also complex. However, human can control acoustic

Manuscript received 13th Sep. 2007. This work was supported in part by a Grant-in-Aid for Scientific Research (A), 16200015 from MEXT, Japan.

K. Fukui, E. Shintaku, A. Shimomura, N. Sakakibara, Y. Ishikawa and A. Takanishi are with the Department of Modern Mechanical Engineering, School of Creative Science and Engineering, Waseda University, Tokyo, 196-8555, Japan (Corresponding author, Phone: +81-3-5286-3257; Fax: +81-3-5273-2209; e-mail: kotaro@toki.waseda.jp).

K. Fukui is a JSPS research fellow, and A. Takanishi is a member of the Humanoid Research Institute and the Advanced Research Institute for Science and Engineering of Waseda University, Japan.

M. Honda is with the Department of Sport Medical Science, School of Sport Sciences, Waseda University, Saitama, Japan.

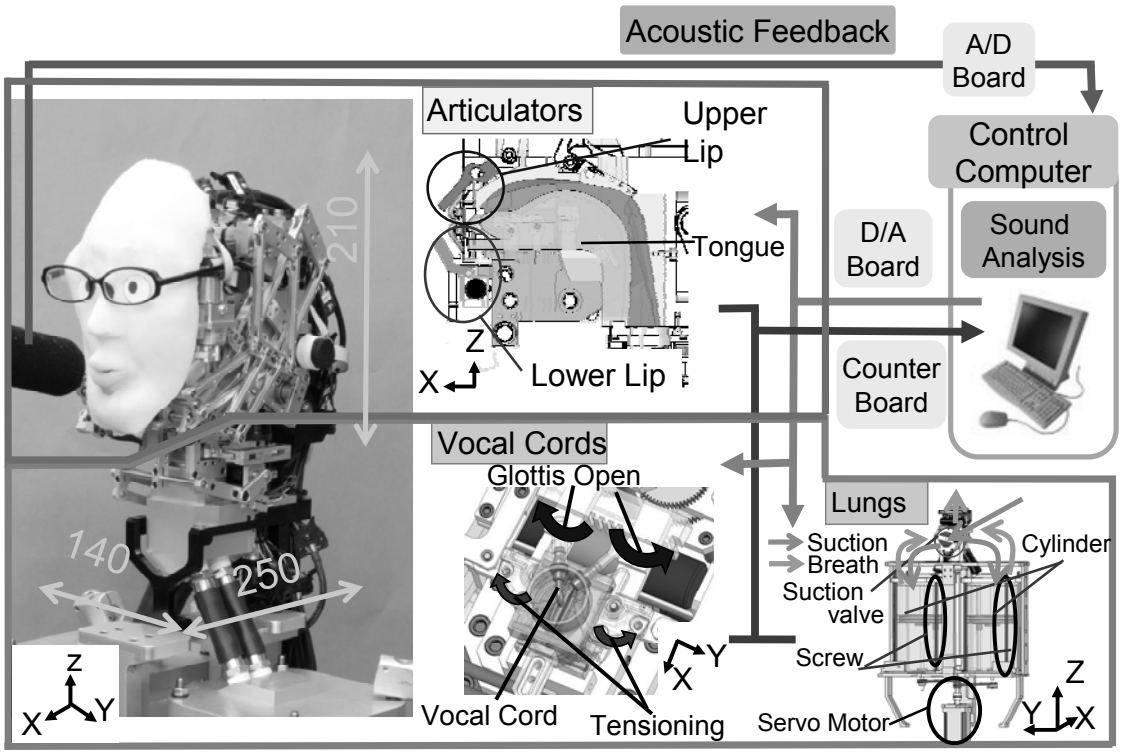



Fig. 1 Mechanical overview and control systems of talking robot WT-7

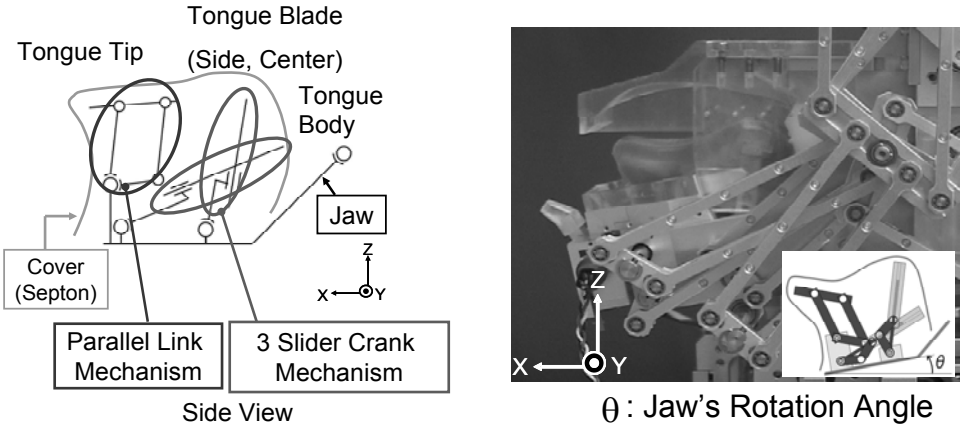



Fig. 2 Tongue Mechanism of WT-7

parameter of produced voice. Human must have control methods for this complex vocal cord, and we aimed to develop such control method for WT-7's human-based vocal cords.

From this purpose, we developed control methods for this model using neural networks (NN). We have adopted forward and inverse models suitable for this complicated system. This paper explains the development of the WT-7's vocal cord mechanism and NN based control.

## II. Anthropomorphic Talking Robot WT-7

Anthropomorphic talking robot WT-7 possesses 19 degrees of freedom (DOF): 5 DOF lips, 1 DOF teeth, 7 DOF tongue, nasal cavity and 1 DOF soft palate as articulators; and 4 DOF vocal cords and 1 DOF lungs as vocal organs, as shown in Fig. 1. The length of the vocal tract is 180 [mm], similar to that of an average adult male. In this mechanism, airflow from the mechanical lungs vibrates the vocal cords, generating source sound. Vowel and consonant sounds are generated by articulating speech organs in a similar way to human.

### *A. Mechanism of Jaws and Tongue*

The human tongue consists of many muscles, and can form various three-dimensional (3D) shapes in oral cavity, such as "very narrow" and "perfectly closed." The movement of these muscles not only creates the static shapes, but also the dynamic ones, as when the tongue moves at high speed to

pronounce plosive sounds. The tongue is the most important part in articulation. In 2D tongue models such as WT-5 (2004), we tried to reproduce the vocal tract area function for producing vowels, consonants, and continuous speech.

However, we found that 2D model was not sufficient to reproduce the vocal tract shape with the side branches in /r/ or with the precise narrow constriction formed by the tongue groove. Therefore we are now developing 3D tongue mechanism based on human biomechanical structure.

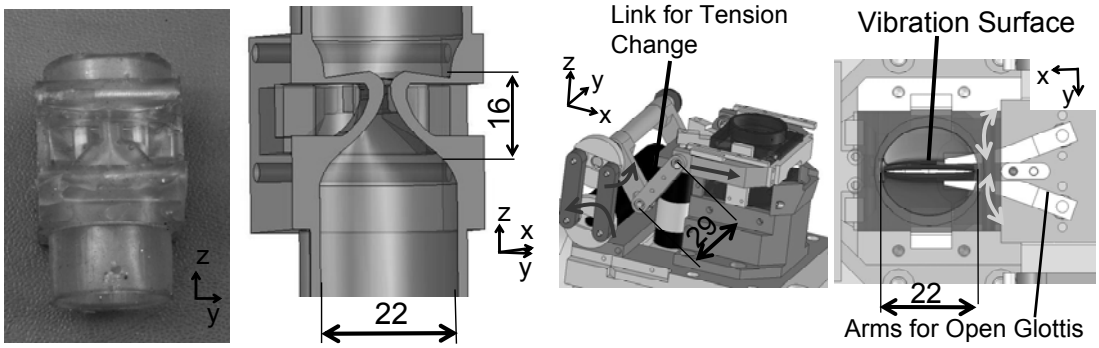
WT-7's tongue mechanism is composed of a jaw mechanism and link mechanisms set on the jaw. The tongue links have 7 DOF and are divided into 3 parts: the first link controls the shape of the tongue tip, the next controls the tongue blade, and the final link controls the tongue body, as shown in Fig. 2. These link mechanisms are covered by the thermoplastic rubber Septon [7], which can deform according to the various shapes made by the links. The front link is a 2 DOF parallel link that controls the tongue-tip position and posture. The tongue blade part, a 3 DOF pair of slide crank links, controls the rotation of the tongue and the length of the side and the center groove. The tongue body part is a set of 2 DOF slider crank links, which control the position of the back part of the tongue, near the vocal cords.

*B. Vocal cord model and control*

Human vocal cords are a pair of folds consisting of muscle and mucosa, covered by larynx cartilage. The folds are vibrated by airflow from the lungs, generating the source sounds of the voice. The vibrating pattern is different in phase in the upper and lower parts of the fold. This complex vibration is important for producing the various vibration patterns that create the human voice [8].

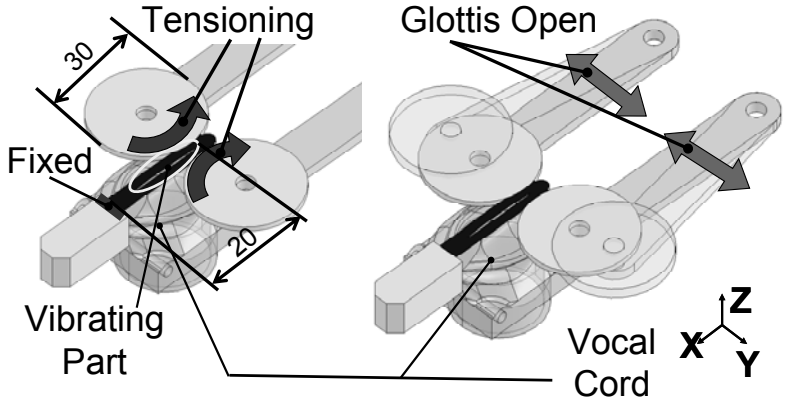
In WT-5, we developed a model of the vocal cord folds based on human biomechanical structure, as shown in Fig. 3(a). The vibration pattern became more human-like, and the sound spectrum displayed human-like attenuation. However, the pitch control mechanism could only change the glottal length, as shown in Fig. 3(b), and the pitch control range was only 15[Hz]. In contrast, an adult male can control an approximately 100[Hz]-pitch frequency range in normal speech.

To resolve this problem and to realize various voice qualities, in WT-7 we develop a new vocal cord control mechanism. In the WT-5 mechanism, the pitch control mechanism could not control tension to the vibrating point effectively. In WT-7, we have developed a pair of discs that are directly attached to the vibrating points, as shown in Fig. 4. Using FEM stress analysis, we found that the new mechanism could control the tension to the vibrating parts. Using this




(a)Vocal cords mimicking human structure (b) Pitch control and voiced/unvoiced switch mechanisms

Fig.3 WT-5's vocal cord mechanism




(a) Pitch control mechanism (b) Spectrum slope control mechanism

Fig.4 Mechanism of WT-7 vocal cords

mechanism, WT-7's vocal cord can produce pitches of 129 - 220[Hz], and it has a more human-like sound source spectrum. The pitch range is 91[Hz], near the human pitch range in normal speech utterance (approximately 100[Hz]). A separate mechanism controls the opening of the glottis. This mechanism can switch not only the voiced/unvoiced of the sound but also control the spectrum slope which is also important feature of voice quality.

## III. Forward and Inverse Modeling

In human vocal cords, the relationship between acoustic parameters and the control variables is complex and nonlinear. The new vocal cord model enabled the wide range pitch control and variable spectrum slope and had human-like complex relation. We need a new control based on neural network model for reproduce human vocal cord control and control the vocal cord model.

Control Methods based on neural network (NN) are used for such purpose. We are focused on real system includes elastic materials, and it is not rigid. From many NN based methods, we adopted the forward and inverse modeling method developed by Jordan [9]. This forward and inverse model is an acceptable model of human learning, and works in such unstable robot, because it needs fewer trials with robot to construct model.

### A. Modeling Methods

In humans, we cannot measure control data: for example, we can see how a muscle moves, but we do not know exactly how the brain controls that muscle. Therefore, we must obtain an inverse model to create a control system. However, as in humans, the inverse model of an anthropomorphic biomechanical system is nonlinear; so many researchers have used neural networks to model these systems.

For the optimization of the inverse model, the most basic method would be to use back-propagation method to reduce the error between the system input and output by inputting the system output into the inverse model. However, this method requires robot movements in the many trials and this is impossible for our vocal cord mechanism. Instead, we have adopted another method of forward and inverse modeling. First, we optimize the forward model by reducing the error between the output of the system and the forward model, and then we adjust the inverse model by using the optimized forward model, as shown in Fig. 5. The advantage of this method is that, after the forward model optimization, we need not move the robot in Jacobian calculation to optimize the inverse model.

### B. Acoustic Parameters

We selected three acoustic parameters for the vocal cord control, based on human auditory experiments that have

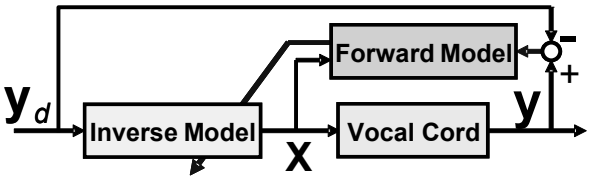


Fig. 5 Forward and inverse modeling

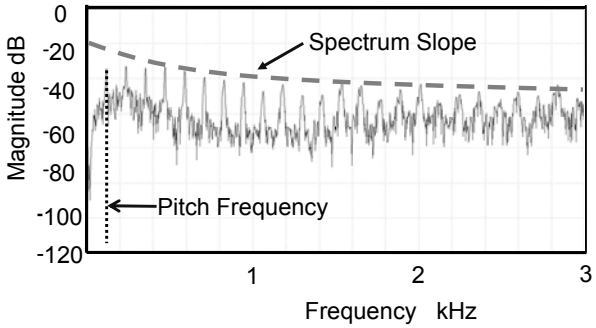


Fig. 6 Glottal source spectrum of vocal cord and sound parameters

revealed that various sounds are controlled by sound pressure, pitch frequency, spectrum slope and duration [10]. Sound duration is a dynamic control, so we focused on the other three parameters, as shown in Fig. 6. The pitch frequency obtained by using an autocorrelation method. The sound pressure is calculated as the logarithmic short time energy; the parameters are biased so 0 [dB] corresponds to no sound. The spectrum slope is calculated by a Linear Predictive Coding (LPC) method. We used the first LPC coefficient to represent the spectrum slope: when the coefficient is near 0, the spectrum slope becomes flat, and when the parameter is near 1, the spectrum slope becomes steep.

In the acoustic analysis, autocorrelation function was calculated for every speech signal segment with 30 ms Hamming window length and a sampling rate of 10 kHz. The offset on the speech waveform was removed for every segment. Pitch frequency was obtained as a time delay (the inverse of the pitch frequency) at which the autocorrelation function R(k) is maximum in the male pitch range. The short time energy, sound pressure, was calculated as the logarithm of the autocorrelation function R(0). Finally, The first linear predictive coefficient was obtained from the autocorrelation function as R(1)/R(0)

### C. Neural Network for the experiment

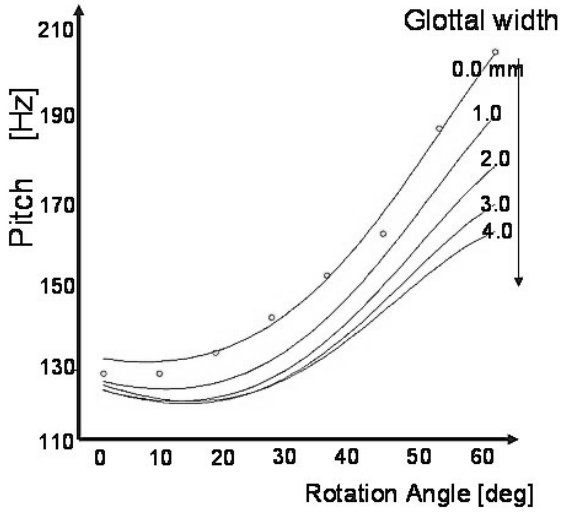
The neural network consists of three sound parameter inputs—Sound Pressure, Pitch and Spectrum Slope—and three robot control outputs—Disc Rotation, Glottis Open and Lung Pressure. We adopted 3-layer neural network for the forward and inverse networks. The number of neurons in the input, middle and output layer is 3, 2-4, and 3, respectively, to avoid over-fitting. In the optimization process, we used

Levenberg-Marqardt methods to reduce time.

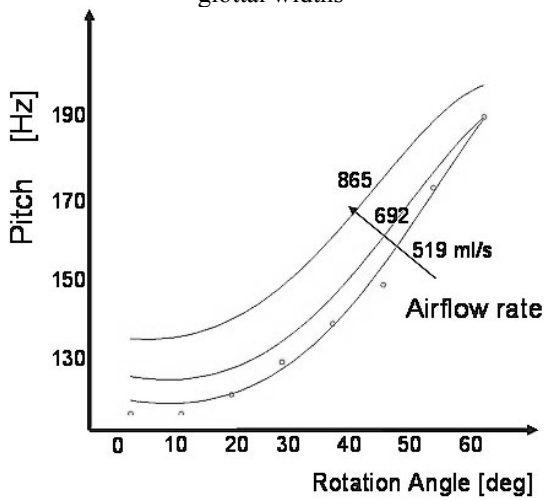
## IV. Optimization Experiments

### A. Optimization of the Forward Model

We designed the forward model of the vocal cord by changing each robotic parameter step by step and measuring the resulting acoustic parameters. The result is absolutely followed the measured data and the rule is fit to the tendency as usually said. The relationship between rotation angle and pitch frequency with changing airflow rate and glottal opening is shown in Fig. 7. This result shows that in general, high pitch is produced with high tension, and high sound pressure is produced by high lung pressure.




(a)Relation between rotation and pitch for various glottal widths




(b)Relation between rotation and pitch for various airflow rates

Fig.7 Acoustic characteristics - pitch frequency: the dot shows the measured data (partially) and the lines are the result of the inverse model

It should be noted that relationship between control variables and the acoustic parameters is nonlinear and almost monotonic. That means that the inverse model can be used for uniquely determining the control variables from the acoustic parameters.

After creating the forward model, the inverse model is optimized using the forward model.

### B. Tracking with the Neural Network Model

We tested the acoustic parameter tracking using the identified inverse model. In this experiment, the vocal tract area was replaced by a static vocal tract model made by acrylic resin, and the target data was simultaneously changed in pitch frequency, sound pressure and spectrum slope. The result is shown by the dash-dotted line in Fig. 9 (the target is the dotted line). From the graph, one can see that the error in pitch frequency is relatively small, 3.5[Hz]. However, the errors in sound pressure and the spectrum slope are significant: 8.1 [dB] and 0.0064, respectively.

## V. Error Reduction Using the Feedback Mechanism

The feed-forward model, including the forward and inverse modeling, has inherent modeling error; in addition, in real experiments, additional disturbance or noise occurs. To cope with this problem, we combined the feedback mechanisms with the feed-forward mechanism. The feed-forward control is fast, but not robust in the presence of noise, and the feedback control has the inverse features, so the combined methods is mainly based on a feed-forward model, however, humans also have some feedback control, as has been proved by many experiments. For example, humans find it hard to speak if they hear their own voice with a delay (this is called delayed auditory feedback, or DAF). Our feedback mechanism is simple, consisting of a low-pass filter for reducing noise, a PI controller and a limiter.

With this feedback mechanism, we tried the same target as in the feed-forward experiment described above. We used the inverse model to roughly calculate the robot control parameters, and the linear feedback control to modify the parameters in real time. The result is shown by the solid line in Fig. 9. The error in the acoustic parameters is reduced: the error in pitch is 1.7[Hz], the error in the pressure becomes

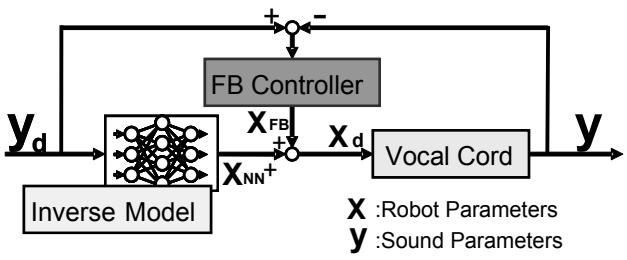



Fig. 8 Feed-back and feed-forward control model

1.7[dB] and the error in the spectrum slope becomes 0.039. We therefore conclude that the feedback mechanism is beneficial.

## VI. Conclusion and Future Work

We have constructed a new vocal cord mechanism to produce various human-like voices, and have developed control method based on forward and inverse neural network modeling. We have also developed a control method to reduce the error using a real-time feed-back mechanism. In our experiments, we match the acoustic parameters of sound pressure, pitch and spectrum slope, which are the most important voice parameters.

We used both forward and inverse modeling for this experiment. And the methods are effective for our human-based vocal cord models. However, the relationship between this type of modeling and actual human voice control is not proven. In the future, we would like to clarify this relationship.

Our ultimate aim is to clarify human speech, its control and its acquisition mechanisms in more detail through our research in the creation and control of a human-like voice model.

## Acknowledgment

The authors would like to thank the following companies: Solid Works KK for provision of CAD and FEM software; Kuraray Co. for the provision and advice about Septon; and the members of the ATR BioPhysical Imaging Project for advice about the biology of the human speech mechanism.

## References

[1] J. L. Flanagan: *Speech Analysis Synthesis and Perception 2nd ed.*, Springer, pp205-206, 1972

[2] N. Umeda, and R. Teranishi: "Phonemic Feature and Vocal Feature -Synthesis of Speech Sound, using an Acoustic Model of Vocal Tract-", *Journal of Acoustical Society Japan,* Vol.22, No.4, pp195-203, 1965

[3] A. Izawa, K. Hattori, Y. Matsuoka and S. Kawamura: "Speech Synthesis by Mechanical System Control", *Journal of Robotics Society of Japan,* pp. 273-278, 1993

[4] H. Sawada, M. Nakamura, T. Higashimoto: "Mechanical Voice System and Its Singing Performance", *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems.* pp1920-1925, 2004

[5] K. Fukui, Y. Ishikawa, T. Sawa, E. Shintaku, M. Honda and A. Takanishi: New Anthropomorphic Talking Robot having a Three-dimensional Articulation Mechanism and Improved Pitch Range, *2007 IEEE International Conference on Robots and Automations*, pp.2922-2927, 2007.

[6] K. Fukui, K. Nishikawa, S. Ikeo, M. Honda and A. Takanishi: Development of a Human-like Sensory Feedback Mechanism for an Anthropomorphic Talking Robot, *2006IEEE International Conference on Robotics and Automation*, pp101-106, 2006

[7] http://www.septon.info/

[8] I. R. Titze: *Principles of Voice Production*, Prentice Hall, 1994

[9] M. I. Jordan and D. E. Rumelhart:Forward models: Supervised learning with a distal teacher, Cognitive Science, vol.16, pp.307–354,1992

[10] H. Kido and H. Kasuya: Representation of voice quality features associated with talker individuality, *5th International Conference on Spoken Language Processing*,1998

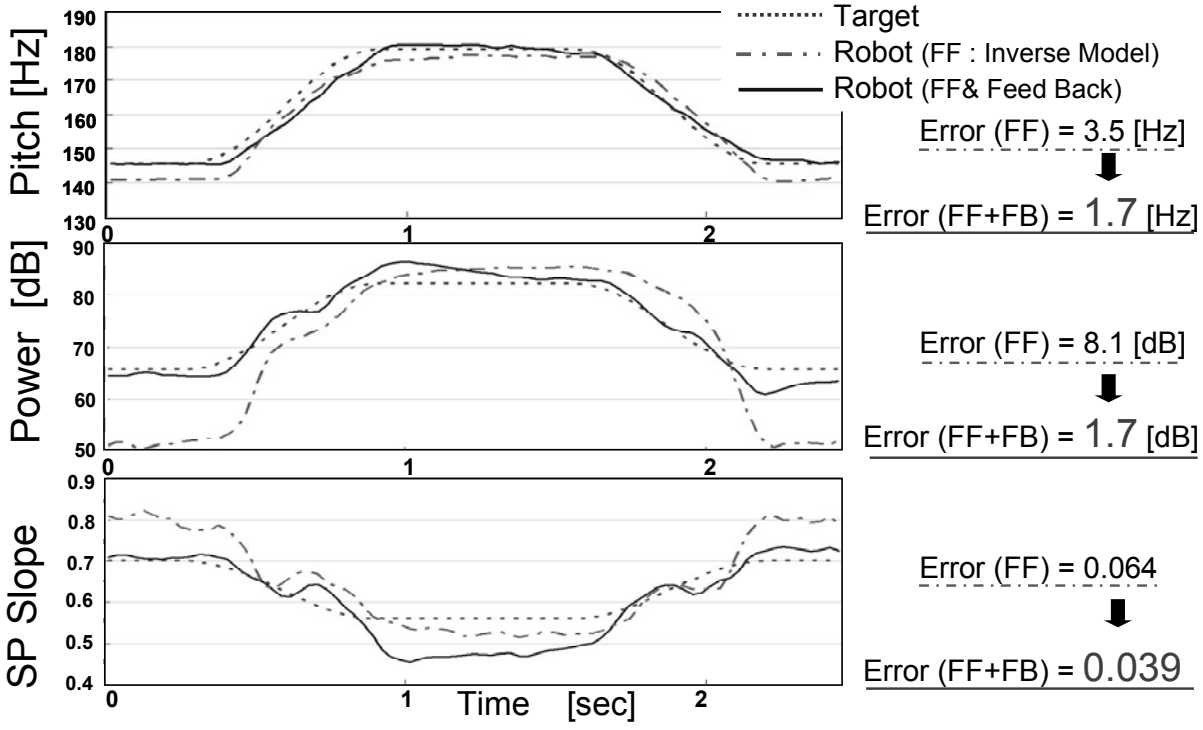



Fig. 9 Experimental result of target tracking of vocal cord model