Reproduction of Speech Feedback mechanism in an Anthropomorphic Talking Robot

Kotaro Fukui[1,5], Shunsuke Ikeo[1] Eiji Shintaku[1] Yuma Ishikawa[1]
Atsuo Takanishi[1,2,3] Masaaki Honda[4†]

[1]Department of Mechanical Engineering, School of Science and Engineering,
[2]Humanoid Robot Institute, [3]Advanced Research Institute for Science and Engineering
[4]Department of Sport Medical Science, School of Sport Sciences

Waseda University
3-4-1 Ookubo, Shinjuku-ku, Tokyo, 169-8555 Japan

[5]JSPS Research Fellow

kotaro@toki.waseda.jp

Abstract. *We reproduced the feedback mechanism of human speech for an anthropomorphic talking robot WT-5 (Waseda Talker No. 5). The Waseda Talker series mimics human speech organs such as the vocal cords, tongue and lips, and WT-5 could produce various vowels and consonant sounds. We elucidate the mechanisms of human speech and speech acquisition using this robot. In human speech acquisition, humans build phoneme categories through mimic speaking. In mimic speaking, infants reproduce adult voices by a feedback mechanism. We have developed a vowel mimic speaking mechanism using auditory feedback that optimized sound pressure and the first and second formants (F1, F2). We also developed consonant mimic speaking using sensory information such as tactile and pressure sensors. Moreover, we developed consonant mimic speaking by auditory feedback. In the feedback mechanism, we constructed acoustic features from MFCCs (Mel-Frequency Cepstral Coefficients).*

[†]Supported by a Grant-in-Aid for Scientific Research (A), 16200015 from MEXT, Japan.

1. Introduction

Given the importance of speech in human communication, a great deal of research has been devoted to determining the mechanisms of human speech. Most of studies have used computer simulation to investigate vocal tract aero-acoustics and movement of speech organs, but it is still difficult to completely simulate the complicated physical phenomenon in speech production. Another approach is to use mechanical model to investigate the speech production mechanisms. Since 1998, we have been developing mechanical models for human speech organs in order to elucidate speech mechanisms. These models have been named the Waseda Talker Series. Anthropomorphic talking robots were developed in order to produce vowels and consonant sounds in a manner similar to human speech by considering changes in the area of the vocal tract. In 2004, WT-4 (Waseda Talker No. 4) was developed. WT-4 is able to produce all 50 Japanese sounds (consonants and vowels) (Fukui et al., 2005). We also developed an auditory feedback mechanism for vowel production that optimized robot parameters using acoustic parameters (Nishikawa et al., 2004). Our aim was to elucidate the human voice acquisition mechanism by reproducing it using a robot. However, in the case of consonant sounds, acoustic parameters are unstable and highly complex. Thus, optimizing consonant speaking using auditory feedback is difficult.

Vocal synthesis machines have been developed by many other researchers (see Flanagan 1972, Umeda 1965, Izawa 1993, Riches 1998, and Sawada 2004). However, none of these machines could produce the various consonant sounds. Only WT-5 could produce various consonant sounds, and this is necessary to reproduce human consonant sound acquisition. We research to reproduce the acquisition using WT-5.

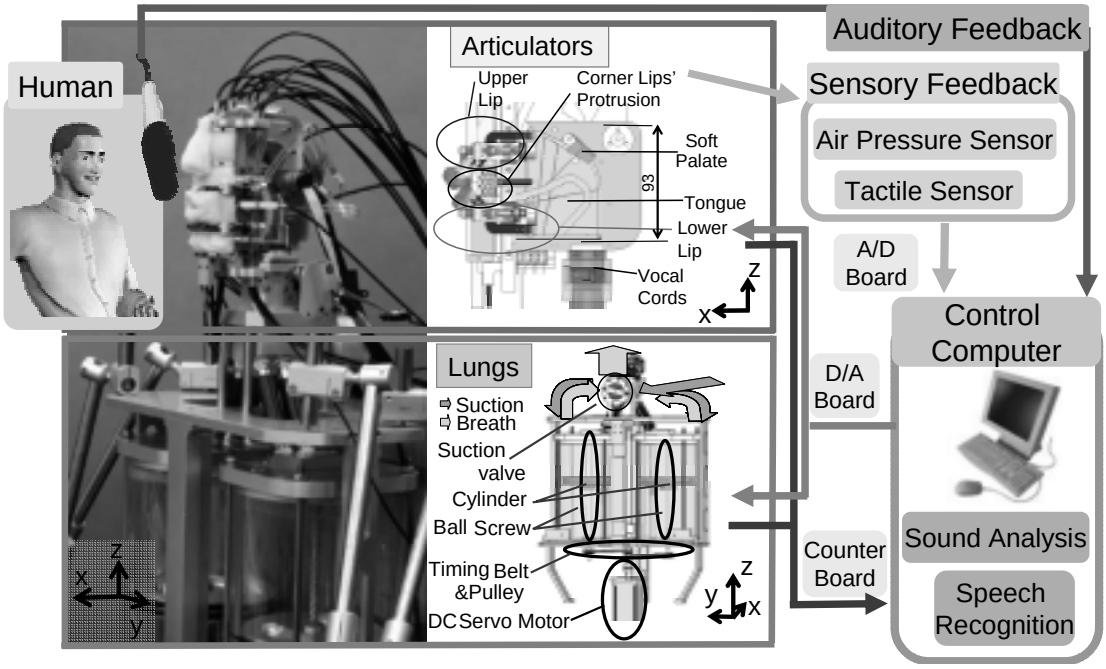


Figure 1. The anthropomorphic talking robot WT-5 and its feedback mechanism, which reproduces human speech mechanism.

In 2005, we developed WT-5 (Waseda Talker No. 5), which, in addition to an auditory feedback, had sensory feedback for consonant mimic speaking, as shown in Figure 1. Using this mechanism, the robot could optimize consonant production. However, we needed an auditory feedback for consonant sounds to reproduce human speech acquisition. We developed an auditory feedback mechanism by developing acoustic parameters from MFCCs, the most widely used parameters for speech recognition. In the present paper, we describe the development of a feedback mechanism in a talking robot for mimic speaking.

2. Feedback Mechanism in Human Speech

2.1. Three-Layer Model of Human Speech Feedback

We are developing a feedback mechanism to reproduce mimic speaking, an important process for acquiring speech. Humans use feedback mechanisms to reduce acoustic error. These feedback mechanisms are not necessary in normal situations. Rather, they are used when humans acquire language or speak in an unstable situation (Borden, 2002).

This feedback mechanism consists of three components namely, auditory feedback, sensory (tactile and pressure) feedback and proprioceptive feedback, as shown in Figure 2. The proprioceptive feedback mechanism perceives the error between the actual movement of vocal organs' muscles and the target. The proprioceptive feedback corresponds to the encoder on the actuator. The sensory feedback mechanism perceives the closure or narrowing of the vocal tract by the tactile sensor or intraoral pressure. The auditory feedback reduces the error between the target voice and the produced sound. Humans use feedback mechanisms to reduce acoustic error when communicating with each other. These feedback mechanisms are not necessary in normal situations. Rather, they are used when humans acquire language or speak in an unstable situation.

The auditory feedback mechanism is mainly used when producing vowel sounds. The sensory feedback mechanism is more important in consonant sound production, particularly in the production of obstacle consonants. It perceives the closure or

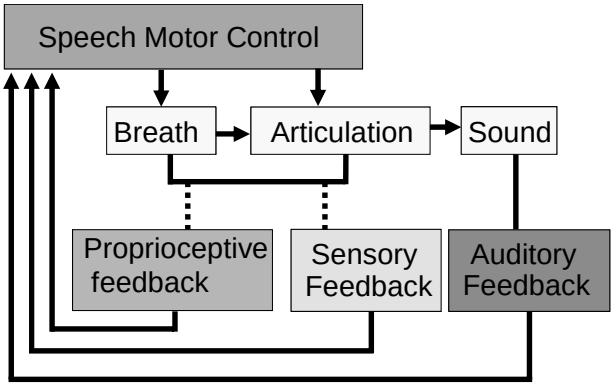


Figure 2. The three-layer feedback model of human speech developed by Fairbanks 1952.

narrowing of the vocal tract (Honda et al., 2002). However, the auditory feedback is also used for acquiring consonant sounds.

2.2. Feedback Mechanism in Human Speech Acquisition

Humans use a feedback mechanism for speech acquisition. However, while much research has been devoted to investigating the acquisition stage, very few studies have examined the mechanism involved. It is hard to observe the brain because the use of MRI (or other similar equipment) on human babies is restricted and babies cannot express their thoughts. Our talking robot could make simulations for speech acquisition. It enabled us to focus on mimic speaking. Human babies reproduce adult voices without phoneme information by using a feedback mechanism. This is mimic speaking and is an essential process for acquiring speech.

3. Anthropomorphic Talking Robots: Waseda Talker Series

The anthropomorphic talking robot WT-5 (Waseda Talker No. 5) had 1-DOF (degree of freedom) lungs, 3-DOF vocal cords, and articulators (7-DOF tongue, 1-DOF soft palate, 1-DOF teeth, 5-DOF lips and nasal cavity). WT-5 had a total of 18 DOF. The length of the vocal tract is 170 mm, which is approximately equivalent to that of an adult male. WT-5 could produce all 50 Japanese sounds, and the spectrum was close to that of human speech. The lung mechanism could control the pressure on the vocal cords by changing the lung speed. The vocal cords, shown in Figure 3, mimic the human biomechanical structure, and were shaped by the thermoplastic rubber Septon manufactured by Kuraray Co. Ltd. (http://www.septon.info/). The vocal cords vibrated in a manner similar to human vocal cords. The arm could change the opening of the glottis to produce voiced/voiceless sounds. The tongue mechanism shown in Figure 4(a) consisted of EPDM (Ethylene Propylene Diene Monomer, made by Tokyo Rubber Industrial Co., Ltd.) rubber having seven control points actuated by looped wires. The tongue could produce various shapes and degrees of vocal tract closure. The teeth are controlled by a release mechanism, and the lip mechanism could control opening and protrusion of the upper and lower lips, as well as changing of the corner protrusion, as shown in Figure 4(a). The lips were shaped by Septon (same material used for the vocal cords).

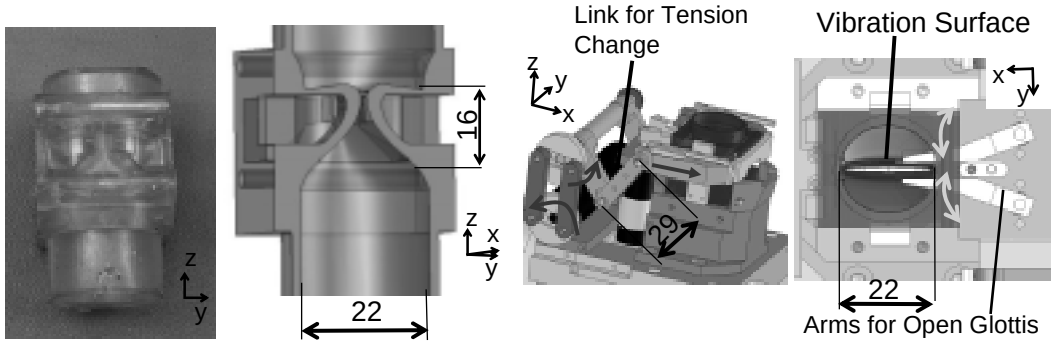
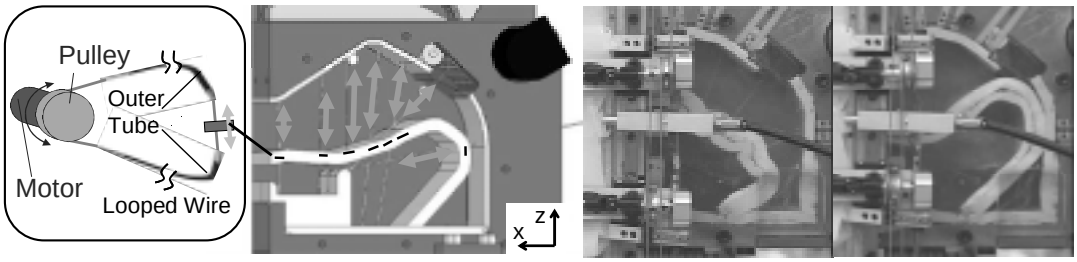
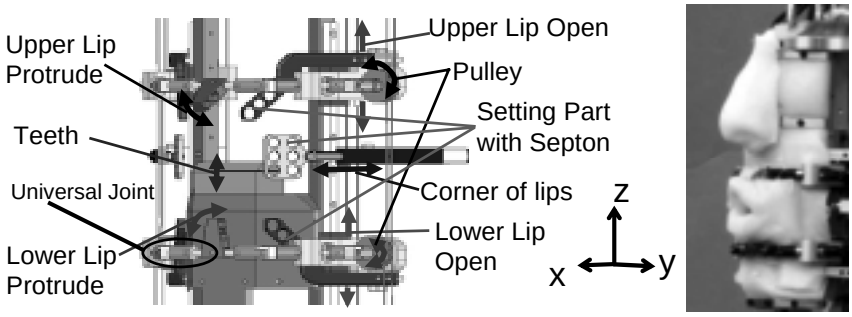


Figure 3. WT-5's Vocal Cords made of Septon, having thin folds and the mechanism for pitch control and opening of the glottis.



Mechanism /i/ shape /o/ shape

(a) Tongue mechanism



(b) Lip and Teeth mechanisms

Figure 4. The mechanisms of WT5's Tongue and Lips

4. Auditory Feedback on Vowel Speaking

WT-4 could mimic human vowel speech without using phoneme information. WT-4 obtained the acoustic parameters from the voice produced, and the Jacobian matrix was derived by observing the shift in the acoustic parameters. The algorithm changed the control parameter to obtain acoustic parameters similar to those of human speech using the Jacobian. By repeating this procedure, the parameters of the voice produced by the robot approached that of a human.

Pitch, sound pressure and the first and second formants (F1, F2) were used as acoustic parameters in the WT-4 mechanism. The formants are the peaks of the DFT (Discrete Fourier Transform) spectrum. F1 is the frequency of the peak on the lowest frequency, and F2 is the second lowest frequency. F1 and F2 are the most important parameters for recognizing vowels. Continuous mimic speaking was performed by optimizing every 50-ms frame in the speech, and the optimized speech approached that of a human.

5. Sensory Feedback on Consonant Speaking

Consonant sounds are very complex and change rapidly. They are very difficult to optimize by auditory feedback. Moreover, in the production of consonants, particularly obstacle consonants, the sensory feedback mechanism is more important because it perceives the closure or narrowing of the vocal tract. The use of sensory feedback simplifies the optimization of closure. We developed a sensory feedback mechanism using tactile and pressure sensors. As parameters for optimization, we used the pressure peak and the rate of decrease of pressure. This algorithm could be employed to optimize plosive sounds.

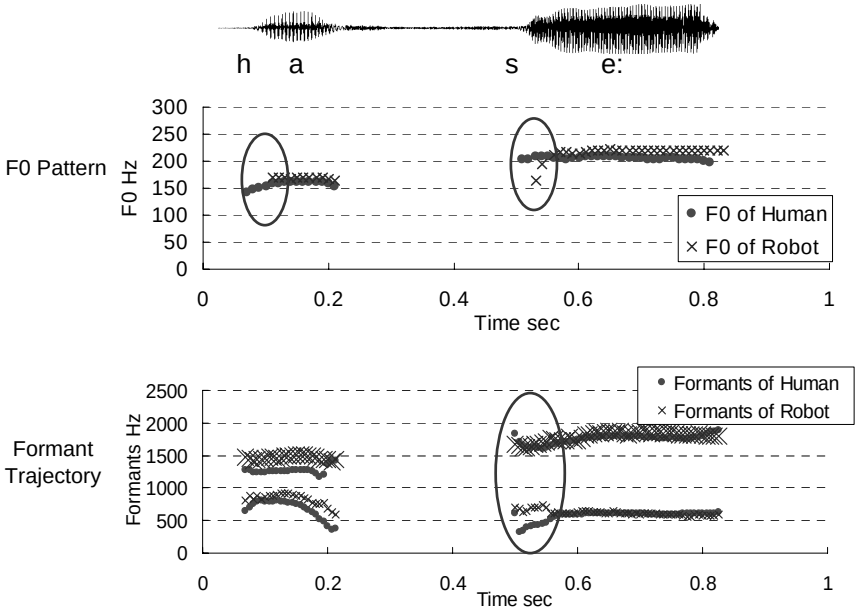
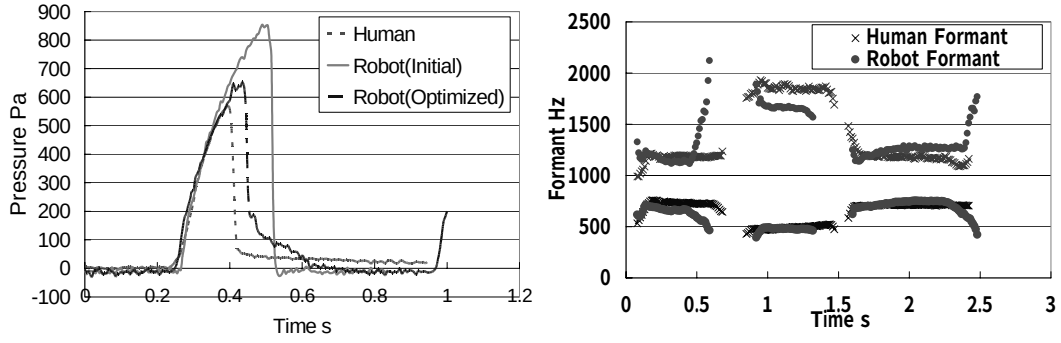


Figure 5. Organized experiment to mimic the human speech sounds "hassei", in which vowel production is optimized by auditory feedback.



(a) Optimization of inner pressure when producing /ta/

(b) Optimization of continuous speech /Waseda/

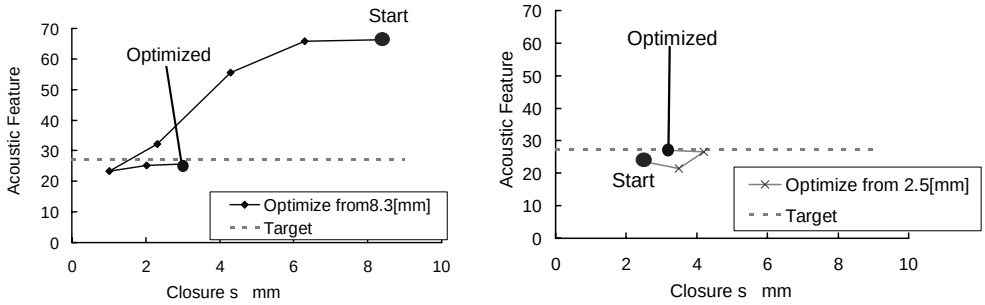
Figure 6. Consonant sound optimization by sensory feedback.

6. Auditory Feedback on Consonant Speaking

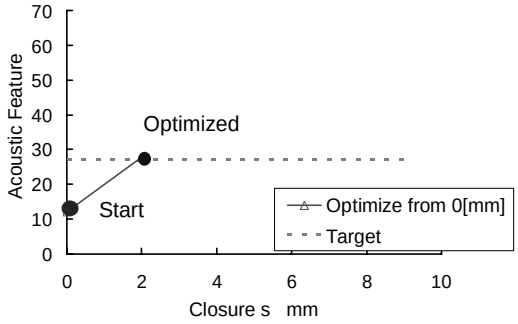
In human speech acquisition, humans can only use acoustic information because mimic speech is mimicking the adult voice, not the sensory information. We need to simulate auditory feedback of consonant sounds in addition to the sensory feedback. However, consonants are difficult to handle because the acoustic features differ depending on the manner of articulation. Moreover, certain consonant sounds are noisy, which makes them hard to model.

The characteristic parameters of consonants are unstable and hard to detect by recognition. Thus, it is impossible to optimize them based on the individual characteristics of each consonant. We focused on the fricative /s/ and the plosive /t/ sound. In terms of articulation space, /s/ and /t/ have the same articulation point, and /s/ is narrow, while /t/ is close the vocal tract. By optimizing the closure, the consonant may be optimized.

To be able to perform this optimization, we need acoustic parameters that vary linearly with to the vocal tract closure. We seek various parameters. However, fricative consonants are louder than plosive sounds. Thus, many of the parameters had peaks in fricative sounds. The 4^{th} and 5^{th} of the MFCCs show parallel to closure of the vocal tract. We optimized the closure using these parameters. Figure 7 shows how the optimization experiment varied the initial closure.



(a) Initial Closure 8.3 [mm] (b) Initial Closure 2.5 [mm]



(c) Initial Closure 0 [mm]

Figure 7. Optimization of the vocal tract closure using the acoustic feature

7. Conclusion and Future Work

We developed an auditory feedback mechanism for vowels, a sensory feedback mechanism for consonants and an auditory feedback mechanism for consonants in an anthropomorphic talking robot. These mechanisms enable the robot to mimic production of human vowel and consonant sounds.

In future work, we will develop a robot capable of mimicking voices as humans do. Using this robot, we hope to elucidate the human speech control mechanism. We also hope to develop more human-like robot control in order to reproduce the human speech control mechanism more precisely.

References

Borden, G. J., Harris, K. S., and Raphael, L. J. *Speech Science Primer, 4th edition*, Lippincott Williams& Wilkins, 2002

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

Fukui, K., Nishikawa, K., Kuwae, T., Takanobu, H., Mochida, T., Honda, M. and Takanishi, A., Development of an Human-like Talking Robot for Human Vocal Mimicry, *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp1449-1454, 2005

Honda, M., Fujino, A., and Kaburagi, T. Compensatory responses of articulator to unexpected perturbation of the palate shape, *Journal of Phonetics*, Vol. 30 pp281-302, 2002

Izawa, A., Hattori, K., Matsuoka, Y. and Kawamura, S. Speech Synthesis by Mechanical System Control, *Journal of Robotics society of Japan*, pp. 273-278, 1993

Nishikawa, K., Kuwae, T., Takanobu, H., Mochida, T., Honda, M. and Takanishi, A., Mimicry of Human Speech Sounds using an Anthropomorphic Talking Robot by Auditory Feedback, *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems.* pp272-278, 2004

Riches, M. MOTORMOUTH, A Speaking Machine, *Journal of Experimental Musical Instrument*, pp.20-23, 1998

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

Umeda, N. and Teranishi, R., 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