Vocal Cord Model to Control Various Voices for Anthropomorphic Talking Robot

Kotaro Fukui[1, 5], Yuma Ishikawa[1], Eiji Shintaku[1], Keisuke Ohno[1],
Nana Sakakibara[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
[5]Research Fellow of Japan Society for the Promotion of Science
E-mail: kotaro@toki.waseda.jp

Abstract

We have developed a biomechanical vocal cord model for an anthropomorphic talking robot—the Waseda Talker Series. The model mimics the biomechanical structure of human vocal cords and is made of a soft rubber called Septon. It is able to reproduce vocal cord vibrations of the normal (modal) voice in a form similar to the human. We also developed a neural network based model for controlling the vocal cord parameters with acoustic parameters of the speech signal. The control model consists of a forward and inverse model, and a simple feedback model. We also describe voice quality control for the vocal cord model which is important for generating fluent and emotional speech.

1 Introduction

Vocal cord vibration is oscillated by air flow from the lung. The vibration pattern is closely related to the voice quality of the generated speech sound. Most studies on voice quality have been done based on the acoustic analysis of the speech sounds and direct observation of the vocal cord vibration. The aero-acoustic phenomenon in generating a glottal sound source related to various voice qualities is very complicated, and it is hard to examine the phenomenon by direct measurements.

We have been developing a human-like talking robot and in 2005, WT-5 (Waseda Taker No. 5), we developed a vocal cord mechanism to mimic the human biomechanical structure, as shown in Fig. 1 [1]. The model is made with a thermoplastic rubber Septon [2], which deforms easily and is strong enough to allow it to stretch. This model could reproduce the vibrations of the modal voice and its sound source spectrum has a human-like spectral slope. We developed WT-7 (Waseda Talker No. 7), which has a pair of discs directly attached to the vibrating part of its vocal cords to effectively control the tension [3], as shown in Fig. 2. WT-7's vocal cords can produce pitches of 129–220 [Hz]. It also has a separate mechanism to control glottal opening and closing.

The mechanical vocal cord model provides an efficient tool for investigating the vocal cord vibration, associated aero-acoustic phenomenon, and laryngeal control. In this paper, we describe an acoustic goal-oriented control method for the vocal cord model. The relationship between the robot's parameters and the acoustic parameters (pitch frequency, spectral slope, and sound power) of the sound source is not simple. We propose a forward-inverse model combined with a real-time feed-back control. We also describe sound source generation having various voice qualities (modal, creaky, and breathy). These voice qualities are controlled by adjusting the robot's parameters in a manner similar to human laryngeal control. The vocal cord vibration pattern, the glottal flow, and the spectral characteristics for various voice qualities are examined in the experiments.

2 Vocal Cord Control Model

We implemented an acoustic goal-oriented control of the vocal cord model based on a neural network (NN) model. We adopted a forward and inverse modeling method developed by Jordan [4]. This model works in complex control, because it needs few learning trials to construct the model. Our purpose is to construct human vocal control model

and this model shows one possibility of the human vocal cord control. However, in experiments, the forward-inverse model could not track the desired acoustic parameters because of an additional disturbance on the acoustic output by the robot. Therefore, in addition to the forward-inverse model, we used a feedback mechanism to achieve robust control.

2.1 Forward and Inverse Modeling

We adopted three acoustic parameters—sound pressure, pitch frequency, and spectrum slope—in the vocal cord control, based on human auditory experiments [5]. We used the first LPC coefficient to represent the spectrum slope. The neural network consists of three sound parameter inputs and three robot control outputs—disc rotation, glottal opening and lung pressure. We adopted a 3-layer NN for the forward and inverse networks. After teaching the forward model, the inverse model was optimized using the forward model.

We tested the acoustic parameter tracking using the identified inverse model. In this experiment, we used a static vocal tract tube made with acrylic resin, and the target data was simultaneously changed in terms of three acoustic parameters. The result is shown by the dash-dotted line in Fig. 3 (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.

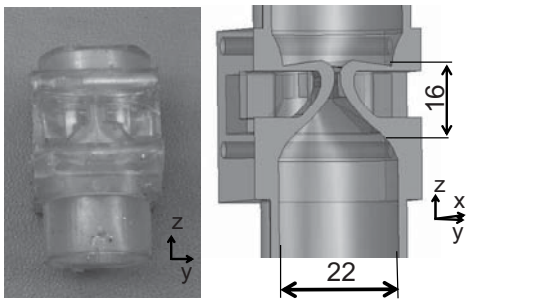
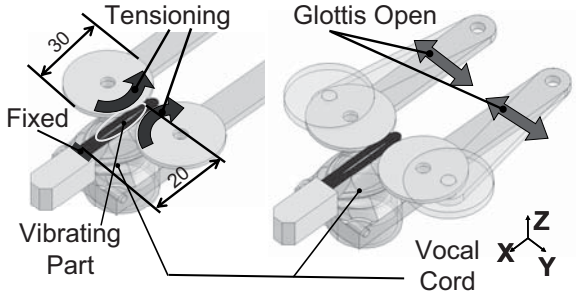


Fig. 1 Mechanical vocal cord model



(a) Pitch control mechanism (b) Glottal opening control mechanism

Fig. 2 Mechanism of the WT-7's vocal cords

2.2 Improvement by Feedback Mechanism

The feed-forward model includes inherent modelling error and, in fact, the feed-forward experiment remained in error. To cope with this problem, we combined a real-time feedback mechanism with the feed-forward model. We then tested this feedback mechanism with the same acoustic targets. The result is shown by the solid line in Fig. 3. The error in the acoustic parameters was reduced: the pitch error was 1.7 [Hz], the pressure error 1.7 [dB], and the error in the spectrum slope became 0.039. We, therefore, concluded that the feedback mechanism was effective.

3 Various Voice Quality

In human speech, various voice qualities are employed in expressing emotional and paralinguistic information. These are important parameters in

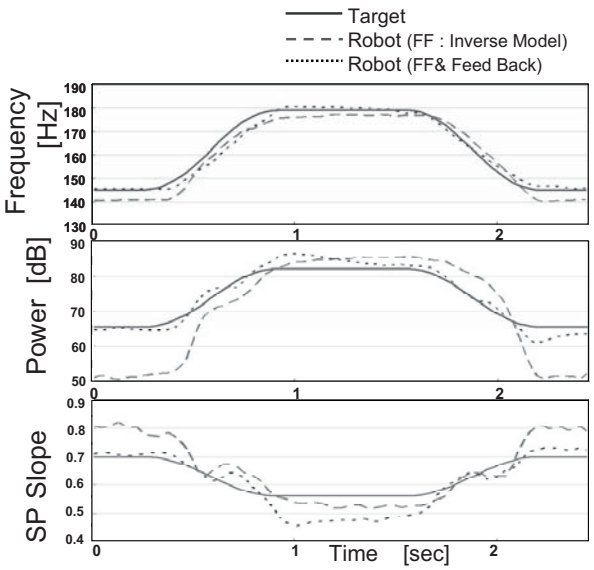


Fig. 3 Experimental result of acoustic target tracking control of the vocal cord model

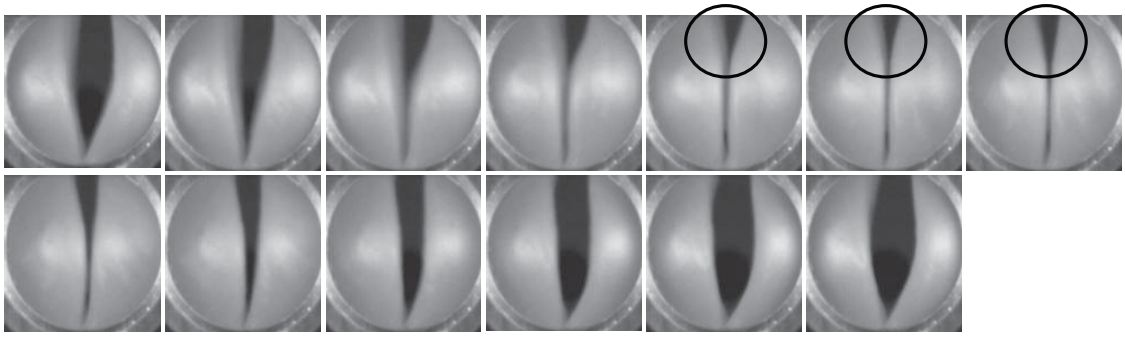


Fig. 4 Vibration of a breathy voice taken by high-speed camera 1000 [fps]

reproducing fluent and emotional speech with a talking robot.

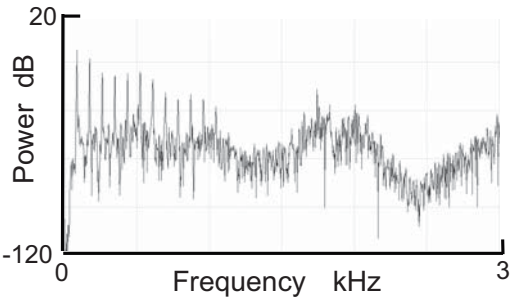
We tried sound generation with various voice qualities using the mechanical vocal cord model. We reproduced the breathy voice and creaky voice, because their vibration is peculiar. In these experiments, we examine similarities between the mechanical vocal cord model and that of a human.

3.1 Production of a breathy voice

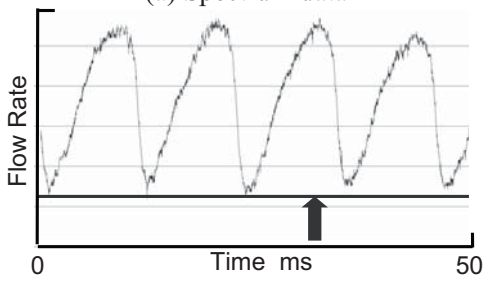
Breathy voice is the voice quality which compounds breathiness and modal voice. The characteristics of breathiness are voice mixed with breath.

In our experiment, we adjusted the glottal opening of the model for generating the breathy voice, which is similar to human control. The glottal opening is adjusted by changing the distance between the discs attached to the ends of the vocal folds. This glottal opening mechanism is similar to the adjustment of the arytenoid cartridge.

One cycle of the glottal vibration of the model, taken by high-speed camera, is shown in Fig. 4. It is shown that, in the vibration cycle, the upper side of the glottis is kept open, even in the closed phase of the vibration. The DFT spectrum of the generated sound and glottal flow, measured using a Rothenberg mask [7], show similar characteristic to those for a human, as shown in Fig. 5. The spectrum shows unclear harmonics in the higher frequency region. The glottal flow wave has an offset from zero level, which means that there exists a flow even in the closed phase of the vibration. These results show that, with control similar to that of a human, the mechanical model could produce a breathy voice.



(a) Spectrum data



(b)Airflow data

Fig. 5 Spectrum and airflow of a breathy voice

3.2 Production of a creaky voice

Creaky voice is the voice quality which compounds creaky and modal voices. A creaky voice is characterized by an extremely low frequency periodic vibration, or double pitch vibration.

We generated the creaky voice in a similar way to the human control, namely using weak cord tension. The tension control is done by adjusting the rotation angle of the discs attached to the ends of both vocal folds. However, tension control alone cannot generate a creaky voice. For extremely weak tension, the glottis remains open and does not vibrate. We adopted an additional side push mechanism at the

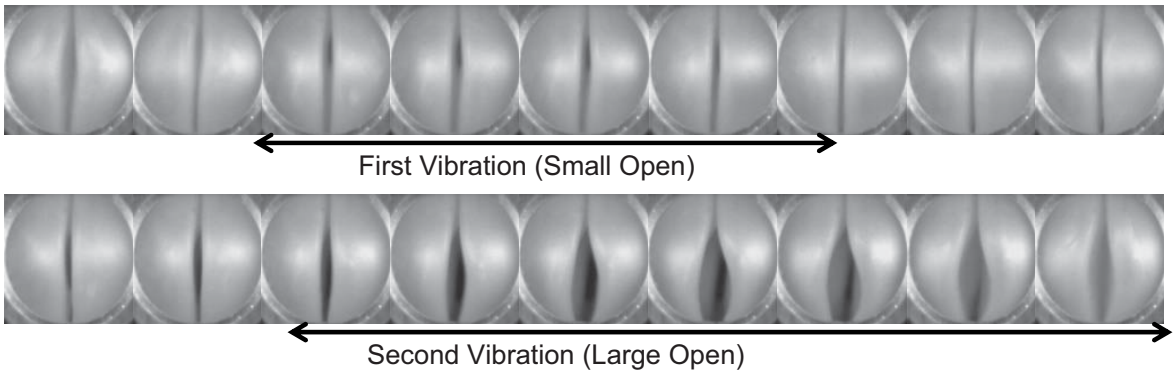
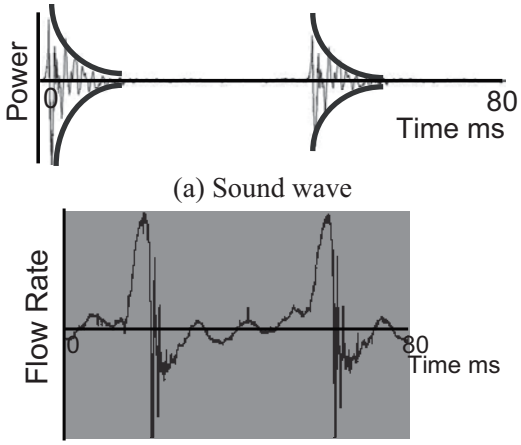


Fig. 6 Vibration of creaky voice taken by high-speed camera 1000 [fps]



(a) Sound wave

(b) Airflow data

Fig. 7 Sound wave and airflow of a creaky voice

lower part of the vocal folds, to stabilize the vibration. With this mechanism, the vocal cord successively vibrates with the extremely low pitch cycle, even for weak cord tension. One cycle of the glottal vibration, captured by high-speed camera, is shown in Fig. 6. The vibration shows a double pitch vibration, and the glottis is barely opened once it is in the vibration cycle, so the closing interval is much longer than the opening interval. In analysis of the sound wave shape and flow measurement, the produced sound has similar characteristics to those of a human, as shown in Fig. 7.

4 Conclusion and future work

We have developed a control method consisting of a forward and inverse neural network model, with a real-time feedback mechanism for a mechanical vocal cord model. The experiments showed that the desired acoustic trajectories are well tracked by the control model. We also generated breathy and creaky voices by adjusting the mechanical vocal cord model. The experiments showed that the model could reproduce these voice qualities by using a method similar to that of a human. Acoustic goal-oriented vocal cord control for generating various voice qualities is planned as future work.

Acknowledgment

The authors would like to thank the following companies: Solid Works KK, Kuraray Co., and the members of the ATR BioPhysical Imaging Project.

References

[1] K. Fukui, M. Honda, and A. Takanishi: Development of a Human-like Sensory Feedback Mechanism for an Anthropomorphic Talking Robot, 2006 IEEE ICRA, pp101-106, 2006

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

[3] K. Fukui, M. Honda, and A. Takanishi: New Anthropomorphic Talking Robot having a Three-dimensional Articulation Mechanism and Improved Pitch Range, 2007 IEEE ICRA, pp.2922-2927, 2007.

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

[5] H. Kido, and H. Kasuya: Representation of voice quality features associated with talker individuality, 5th ICSLP, 1998

[6] J. Laver: The phonetic description of voice quality, Cambridge University Press, pp 109-135, 1980

[7] http://www.glottal.com