Mechanical Design of a Talking Robot for Natural Vowels and Consonant Sounds

Kazufumi Nishikawa, Kouichirou Asama, Kouki Hayashi, Hideaki Takanobu, *Atsuo Takanishi

Humanoid Robotics Institute, Waseda University
Department of Mechanical Engineering, Waseda University
#59-308, 3-4-1 Ookubo, Shinjuku-ku, Tokyo 169-8555, JAPAN
*takanisi@mn.waseda.ac.jp

Abstract

Vocal movement isn’t only a movement of the vocal organs. It is also a movement that produces acoustic signals received by hearing as linguistic information through hydroacoustic phenomena along with the formation of the vocal way.

The purpose of this research is to make clear the human vocal mechanism from the view of engineering by simulating the vocal movement with a robot, and to create the dynamic model.

Therefore, the authors developed an anthropomorphic talking robot WT-1 (Waseda Talker-No.1) in 1999. It simulates human vocal movement, and has articulators (the 6-DOF tongue, 4-DOF lips, 1-DOF teeth, a nasal cavity and 1-DOF soft palate) and vocal organs (the 1-DOF lungs, 1-DOF vocal cords); Total DOF of the robot is 14.

We experimented with it on vowels. The formant patterns of humans and the robot with Japanese vowel /i/ are shown in Fig.2. As a result, F1 and F2 frequencies of the vowel /i/ were similar to the human average. The other vowels were confirmed to be similar to it, as well. WT-1 could utter single vowels. However, its voice wasn’t natural.

In this paper we describe the improvement of the mechanisms for the realization of natural vowels and consonant sounds.

Key words: Humanoid Robot, Speech Production, Voice, Vocal Movement

1. Introduction

Communication is extremely important in being able to keep our human society. We usually use spoken language as communication in our society. In addition, gestures and the use of pictures exist as other forms of communication. However, when considering the efficiency of communication, spoken language is considered to be the best method in general. Concerning the production of the spoken language (speech production), a lot of research has been done. However, at present, there isn’t comprehensive research on the control systems for vocal movement, and vocal movement itself hasn’t been made clear. Therefore, CREST (Core Research for Evolutional Science and Technology) of JST (Japan Science and Technology), NTT (Nippon Telegraph and Telephone Corporation) as the leader and fourteen research organizations (medicine, acoustics and engineering), started to develop and to analyze the human speech’s dynamic model in 1998.

In CREST, we take charge of developing an anthropomorphic talking robot that reproduces human vocal movement.

The purpose of this research is to clarify the human vocal mechanism from the viewpoint of engineering by reproducing the vocal movement using a robot, and to create the dynamic model. This model will lead to the production of cellular phones that have a very narrow the bandwidth of radio wave due to the transmission of the human vocal movement instead of the actual human voice itself, and to develop training methods for vocally handicapped people and acquisitive methods of foreign languages.

There were several related works concerning the development of voice synthesis machines. Research concerning mechanical voice devices was performed by Kempelen,W.V. [1] in 1771, Umeda [2], Kawamura [3], Martin Riches [4] and Osuka [5]. However, it has been difficult to say whether any research can faithfully reproduce human speech by using mechanical devices.

Therefore, We developed an anthropomorphic talking robot WT-1 (Waseda Talker-No.1) that simulates human articulators and vocal organs, as shown in Fig.1.

We experimented with it on vowels. The formant patterns of humans and the robot with Japanese vowel /i/ are shown in Fig.2. As a result, F1 and F2 frequencies of the vowel /i/ were similar to the human average. The other vowels were confirmed to be similar to it, as well. WT-1 could utter single vowels. However, its voice wasn’t natural.
In this paper, we describe the improvement of the mechanisms for the realization of natural vowels and consonant sounds.

2. Human vocal mechanisms and WT-1’s mechanisms

Humans have articulators (the tongue, lips, teeth, nasal cavity and soft palate) and vocal organs (the lungs and vocal cords).

The talking robot WT-1 has articulators (the 6-DOF tongue, 4-DOF lips, 1-DOF teeth, nasal cavity and 1-DOF soft palate) and vocal organs (the 1-DOF lungs, 1-DOF vocal cords) like a human’s; Total DOF of the robot is 14. The dimensions of it are about 1.2-1.3 times larger than human’s size. The general view and the mechanisms of WT-1 are shown in Fig.1.

The typical characteristics of each human organ for the development of the mechanisms are as follows. We compared between human vocal mechanisms and WT-1’s.

2.1 Vocal organs

WT-1 has lungs and vocal cords like those of a human.

2.1.1 Lungs

1) Human

The adult male lung capacity is 3450 [ml] on average [6]. A human’s breathing pressure during speech is usually 6-10[cmH₂O] [7]. Breathing flow of adult male during speech is 136 [ml/s] on average [8]. Their specifications are shown in Table 1.

2) WT-1

We developed a mechanical breathing device using...
bellofram (Fig.1). The lung capacity of WT-1 is 7740 [ml]. The breathing pressure of WT-1 during speech is usually 11 [cmH₂O]. The breathing flow during speech of WT-1 is usually 774 [ml/s]. The lung of WT-1 supplies their specifications and simulates a human’s (Table 1).

2.1.2 Vocal Cords

1) Human

The pitch of the voice is decided by the frequency of the vocal cords. The higher the frequency is, the more voice is recognized. Although an adult female voice and a child’s voice are high, the adult male’s tends to be low. The range of adult voice is usually 2 [octave] [9]. Japanese adult male range is 69-587 [Hz] at the maximum and 98-123 [Hz] during normal conversation [9]. Their specifications are shown in Table 2.

Many overtones exist between 100-10,000 [Hz] [10]. The energy of human overtones is reported to be lower than 12 [dB/octave][7]. It is also known that there are voiced and voiceless sounds.

2) WT-1

Human vocal cords are very complicated and difficult to simulate by a mechanical model. Therefore, we developed mechanical vocal cords that are able to change the pitch of sound, and are capable of voiced and voiceless sounds, as shown in Fig.3. The pitch is changed by pulling the artificial glottis from side to side using a ball screw (A). Voiceless sound is made by conversely pushing and opening the artificial glottis (B).

It is ideal to have the material of the vocal cords correspond to the characteristics of the human vocal cords (Young's modulus: about 10-100) [11]. We used a super-low hard rubber TP010 (Shore hardness: 1 [JIS-A], tensile strength: 5.9 [MPa], the quality of the material: EPDM) made by TOKYO RUBBER INDUSTRIAL CO., Ltd.

We experimented with the artificial vocal cords, and quantified their specifications by using FFT analysis. As a result, the primary tone of the artificial vocal cords is 97-140 [Hz], as shown in Fig.4 and Table 2. The range of WT-1 is wider than the adult male range. Many overtones exist between 100-10,000 [Hz] like a human’s, as shown in Fig.5. The overtones of the artificial vocal cords were confirmed to be lower than 11-15 [dB/octave] on

<table>
<thead>
<tr>
<th>Table 1 Lungs of human and WT-1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Lung Capacity [m³] [6]</td>
</tr>
<tr>
<td>Breathing Pressure during Speech [Pa] [7]</td>
</tr>
<tr>
<td>Breathing Flow during speech [m³/s] [8]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2 Range of voice of human and the robot</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Range of Voice at the Maximum [Hz]</td>
</tr>
<tr>
<td>Range of Voice during Nomal Conversation [Hz]</td>
</tr>
</tbody>
</table>

![Fig.3 Vocal cords that can produce voiced and voiceless sounds](image)

![Fig.4 Pitch of the vocal cords](image)

![Fig.5 Power spectrum of overtones when the primary tone is 127[Hz]](image)
average (Fig. 5). We also conducted experiments to see the vibration of the artificial glottis using a high-speed video camera (1,000 [frame/s]). Then we confirmed that the artificial vocal cords vibrated the glottis like a human’s.

In conclusion, the artificial vocal cords that we developed are very similar to the characteristics of the human male vocal cords. They are considered to be a substitutable for a human’s. However, the artificial vocal cords have several problems. We describe these problems in section 3.1.

2.2 Articulators

WT-1 has a tongue, lips, teeth and nasal cavity like those of a human.

2.2.1 Tongue

1) Human

The tongue moves freely in the oral cavity and it is one of the most important organs for articulation. The velocity of the tongue at the tongue tip and the tongue body is shown in Table 3.

2) WT-1

We used two 3-DOF-pantograph mechanisms with high speed and high accuracy for reproducing the complicated movement of the tongue at the tongue tip and tongue body. We used super-low hard rubber TP010 of the artificial vocal cords for seals. Their links can make various shapes by pushing the rubber. The link of the body has a special mechanism which allows it to push up both sides of the tongue for realizing the production of the vowel /i/, as shown in Fig. 6. The velocity of the tongue of WT-1 supplies their specifications and simulates a human’s (Table 2). The moving area was confirmed to satisfy the human tongue from the images of MRI [13].

2.2.2 Lips and Teeth

1) Human

The lips and teeth, which are the terminal of the vocal tract, have the leading roles in articulation. The lips are known to move with especially high speed in situations with plosive sound. The velocity and moving area of the lips and teeth are shown in Table 4.

2) WT-1

Tendon mechanisms move the lips in the right and left directions, and crank mechanisms move them in the up and down directions, as shown in Fig. 7.

A crank mechanism moves the lower teeth in the up and down directions.

We used super-low hard rubber TP010 of the artificial vocal cords for seals.

We confirmed that the lips and teeth of WT-1

<table>
<thead>
<tr>
<th>Table 3 Maximum speed of the tongue [12]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tongue Tip</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Forward</td>
</tr>
<tr>
<td>Upward</td>
</tr>
<tr>
<td><strong>Tongue Body</strong></td>
</tr>
<tr>
<td>Forward</td>
</tr>
<tr>
<td>Upward</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4 Lips and teeth [12] [14]</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Moving Area [mm]</strong></td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Upper Lip</td>
</tr>
<tr>
<td>Lower Lip</td>
</tr>
<tr>
<td>Lower Teeth</td>
</tr>
<tr>
<td><strong>Maximum Speed [mm/s]</strong></td>
</tr>
<tr>
<td>Upper Lip</td>
</tr>
<tr>
<td>Lower Lip</td>
</tr>
<tr>
<td>Lower Teeth</td>
</tr>
</tbody>
</table>

Fig.6 Tongue using two 3-DOF-pantograph mechanisms

Fig.7 Lips
supply their specifications and simulate the human’s (Table 4). However, the lips have several problems. We describe these problems in section 3.2.

2.2.3 Nasal Cavity and Soft Palate

1) Human

The nasal cavity has a chamber of resonance and can produce nasal sound by opening the soft palate.

2) WT-1

We developed the nasal cavity that has a chamber of resonance and can produce nasal sound by opening the soft palate using a motor like a human’s. However, the nasal cavity of WT-1 has several problems. We describe these problems in section 3.3.

3. Improvement of WT-1’s Mechanisms

3.1 Vocal Cords

The vocal cords of WT-1 can change the pitch of the sound, and are capable of voiced and voiceless sounds (Fig.3). The pitch is changed by pulling the artificial glottis from side to side using a ball screw (A). Voiceless sound is made by conversely pushing and opening the artificial glottis (B). However, because a ball screw’s mechanism drove the change in the pitch and the switch of voiced and voiceless sounds, we couldn’t keep a certain frequency and switch voiced and voiceless sounds in a moment. For example, it took more than 200 [ms] to switch from the voiced sound (130 [Hz]) to voiceless sound. Therefore, we must contrive the 2-DOF vocal cords that can change the pitch and switch from voiced sound to voiceless sound (or from voiceless sound to voiced sound) in a moment.

3.2 Lips

The lips, which are the terminal of the vocal tract, have the leading roles in articulation. Tendon mechanisms move the lips in the right and left directions, and crank mechanisms move them in the up and down directions (Fig.7). However, their mechanisms cannot make the lips protrude. It is
important in realizing the vowel /u/ to make the lips protrude.

Therefore, we have contrived the lips that can be protruded, as shown in Fig. 8. Crank mechanisms move them in the up and down directions. The tendon mechanisms drive them against the spring in the right and left directions. When the wire is loosened, the lips are protruded by the spring.

3.3 Nasal Cavity and Soft Palate

The nasal cavity has a chamber of resonance and can produce nasal sound by opening the soft palate like a human’s, as shown in Fig. 9. However, we couldn’t distinguish between the WT-1’s composite tones by hearing when it opened and closed the soft palate. It was difficult to say WT-1 could realize nasal sound. Then we considered two reasons for the problem. These are as follows.

1. The position of a human soft palate is just behind the vocal cords and helps the flow to the nasal cavity be efficient. However, the position of WT-1’s soft palate is located above the tongue and far from the vocal cords. We considered that the position doesn’t help the flow be efficient.

2. The nasal cavity of WT-1 is small, compared with the oral cavity’s volume. We considered that we couldn’t hear the big difference between their sounds.

Therefore, we referred to the images of MRI [10] and improved the design of the nasal cavity and the soft palate.

We experimented on nasal sound with the improved nasal cavity. We describe the result of the experiments in section 4.2.

4. Experiments

4.1 Fluctuation of Amplitude

Human natural voice includes the fluctuation of the amplitude and the pitch of the sound, and it is an important factor for natural voice [19]. We compared the artificial vocal cord’s sound blown by WT-1’s lung with the sound blown by human lungs and human vowel /a/, as shown in Fig. 10. As a result, the amplitude of the sound blown by WT-1’s lung wasn’t almost changed, compared with the sound blown by human lungs and human vowel /a/. The WT-1’s sound wasn’t natural. Therefore we contrive to control the breathing pressure to use the actuator (250 [W] DC servo motor, no reduction gear) instead of the actuator (90 [W] DC servo motor, reduction gear rate: 4.3) for the fluctuation of the amplitude, as shown in Fig. 11.

4.2 Nasal Sound

Nasal sound has a biggest formant in low frequencies (250-300 [Hz]) [15]. We experimented on
nasal sound with the improved nasal cavity. As a result, we recognized WT-1’s composite tone as the human nasal sound /n/, and confirmed that it has a biggest formant in low frequencies (250–300 [Hz]) and was very similar to a human’s, as shown in Fig.12.

5. Conclusions and Future Works

We developed an anthropomorphic talking robot WT-1 (Waseda Talker-No.1) that simulates human articulators and vocal organs, and experimented with it on vowels. As a result, F1 and F2 frequencies of all Japanese vowels were similar to the human averages. WT-1 could utter single vowels. However, its voice wasn’t natural.

We have been improving the mechanisms for the realization of natural vowels and consonant sounds.

For future works, we aim at the realization of natural vowels and consonant sounds, and will improve the mechanisms of WT-1 and develop control systems for consonant sounds.

Acknowledgement

This research has been supported by CREST of JST (Japan Science and Technology) and we would like to express our thanks to all researchers in our CREST project and Mr. Akihisa Okino of Okino Industries, Ltd. for helping us to develop WT-1’s hardware.

References

[12] Kaburagi, Kyushu Institute of Design