Speech Planning of an Anthropomorphic Talking Robot for Consonant Sounds Production

Kazufumi Nishikawa*1, Akihiro Imai*1, Takayuki Ogawara*1, Hideaki Takanobu*2, *4, Takemi Mochida*5, Atsuo Takanishi*1, *2, *3

*1 Department of Mechanical Engineering, Waseda University
*2 Humanoid Robotics Institute, Waseda University
*3 Advanced Research Institute for Science and Engineering, Waseda University
*2 Department of Mechanical Systems Engineering, Kagakuin University
1-24-2 Nishi-shinjuku, Shinjuku-ku, Tokyo 163-8677, JAPAN
*2 Human and Information Science Laboratory, NTT Communication Science Laboratory
3-1 Morinosato Wakamiya, Atsugi-Shi, Kanagawa 243-0198, JAPAN

Abstract

This paper describes the speech planning of the anthropomorphic talking robot WT-1R (Waseda Talker-No.1 Refined) for the production of consonant sounds.

WT-1R has articulators (the tongue, lips, teeth, nasal cavity and soft palate) and vocal organs (the lungs and vocal cords), and can reproduce human vocal movement. Its total DOF (degrees of freedom) is 15.

The vocal movement of WT-1R for vowels is steady. We produced Japanese vowels (/a/, /i/, /u/, /e/, /o/) using the first robot WT-1 in 2000. However, the vocal movement for consonant sounds is transient. We must control the 15-DOF talking robot coordinately in the space and time to reproduce the complicated phenomena of the consonant sounds. Therefore, because the Japanese voice generally consists of two phonemes of the first consonant sound and the last vowel, we proposed the speech planning of WT-1R by considering the phenomenon of the voice as three parts (steady consonant sound, transient consonant sound and vowel). WT-1R could produce Japanese vowels (/a/, /i/, /u/, /e/, /o/) and some consonant sounds (/s/, /h/, /m/, /p/ and /waseda/).

Key Words: Humanoid Robot, Speech Production, Vocal Movement, Voice, Sound

1. Introduction

Communication is extremely important in being able to keep our human society. We usually use spoken language as the form of 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 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 hasn’t been understood very clearly. Therefore, in 1998, NTT (Nippon Telegraph and Telephone Corporation) and nine research organizations (medicine, acoustics and engineering) started to develop and to analyze a dynamic model of the human speech as a project of CREST (Core Research for Evolutional Science and Technology) of JST (Japan Science and Technology). In CREST, we take charge of developing a talking robot that reproduces human vocal movement.

The purpose of this research is to clarify the human vocal mechanism from engineering viewpoints by reproducing human vocal movement using a robot, and to create a dynamic model. This model will lead to the production of cellular phones that can compress data by transmitting human vocal movement instead of the human voice. Furthermore, the model will lead to developing medical training devices for vocally handicapped people and learning devices for foreign languages.

There have been several research projects concerning the development of voice synthesis machines. These machines were developed 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 using these machines.

At the beginning of the research, we developed a talking robot WT-1 (Waseda Talker-No.1) [6]. WT-1 could speak Japanese vowels (/a/, /i/, /u/, /e/, /o/), however its vowels didn’t sound natural. Furthermore, it was difficult to say whether WT-1 could produce consonant sounds, because WT-1 didn’t have enough mechanisms to do so.

Therefore, we developed another talking robot WT-1R that was improvement over WT-1 in the production of natural vowels and consonant sounds in 2001, as shown in Fig. 1. WT-1R could speak Japanese vowels (/a/, /i/, /u/, /e/, /o/) and some consonant sounds (/s/, /h/, /m/, /p/ and /waseda/).

This paper describes the speech planning of the second robot for the production of the consonant sounds.

2. Talking Robot WT-1R’s Mechanism

WT-1R is a talking robot that was an improvement over WT-1’s mechanism in the production of natural vowels and consonant sounds. WT-1R has vocal organs (the 1-DOF lungs and 2-DOF vocal cords) and articulators (the 6-DOF tongue, 4-DOF lips, 1-DOF teeth, nasal cavity and 1-DOF soft palate) like humans. The total DOF of WT-1R is 15. Its dimensions are about 1.2-1.3 times larger than an adult male’s. The mechanism and the control parameters of each organ are shown in Fig. 2-4. The details of the mechanisms are described in the paper [6].

We must control the fifteen control parameters for positioning, as shown in Fig. 2-4 and Table 2. The conventional block diagram of servomotors for position control is shown in Fig. 5. We use Titech Driver PC-0121-2 made by OKAZAKI SANGYO CO., LTD. as a velocity feedback controller.

3. Acoustic Simulator for Vowels

We developed an acoustic simulator for the improvement of WT-1R’s hardware and for the
The simulator can be guessed the WT-1R’s vowels by inputting the control parameters. That is, the simulator can display the movement of each articulator. In addition, it can calculate the cross sectional area of the vocal tract, the area function, the sound spectrum and the formants (F1 and F2), and can output the sound. We searched the formants of all the parameters by moving the movable area using the simulator, as shown in Fig. 7. These ellipses show the ranges of

more efficient search of the parameters the position (Table 1) when speaking vowels, as shown in Fig. 6. The simulator can be guessed the WT-1R’s vowels by inputting the control parameters. That is, the simulator can display the movement of each articulator. In addition, it can calculate the cross sectional area of the vocal tract, the area function, the sound spectrum and the formants (F1 and F2), and can output the sound. We searched the formants of all the parameters by moving the movable area using the simulator, as shown in Fig. 7. These ellipses show the ranges of

X_vo: Position of the vocal cords mechanism to change the pitch (A)
X_vl: Position of the vocal cords mechanism to switch between voiced and voiceless sounds (B)

Fig. 4 Vocal cords mechanism and the control parameters

<table>
<thead>
<tr>
<th>Control Parameters for Positioning</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongue 6 X_t, Y_t, X_th, Y_th, θ_t, θ_th</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lips 4 Y_ι, Y_ιh, X_ι, Y_ιl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teeth 1 Y_ι</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soft Palate 1 θ_sp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vocal Cords 2 X_vo, X_vl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lungs 1 Y_ιl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total 15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Y_ι: Position of the teeth (Fig. 1)
θ_sp: Orientation of the soft palate (Fig. 1)
Y_ιl: Position of the diaphragm of the lungs (Fig. 1)

Fig. 5 Block diagram of servo motor control

Fig. 6 Acoustic simulator for vowels

Fig. 7 Formants using the acoustic simulator

Fig. 8 Search for the parameters of vowels
Japanese adult males’ formant. We confirmed that WT-1R, which we designed, could speak Japanese vowels from the result of the formants using the simulator.

4. WT-1R’s Control Systems

WT-1R’s dimensions are about 1.2-1.3 times larger than a human’s, and the mechanisms and materials are different from humans’. It is difficult to say whether WT-1R can produce voices that can be easily recognized or not. Therefore, this chapter describes the control systems for the production of vowels and consonant sounds.

4.1 Vowels

We searched for the suitable parameters for positioning the robot motion to produce recognizable vowels, as shown in Table 1 and Fig. 8. At first, we determined the initial parameters by referring to the MRI images (given by Dr. Kiyoshi Honda at ATR) when a person speaks Japanese vowels. Next, we simulated the movement of each articulator using the simulator described in chapter 3. Then, we produced WT-1R’s vowels, and made fine adjustments to each parameter so that WT-1R’s formant pattern imitates a human’s using FFT.

4.2 Consonant Sounds

Although the vocal movement for vowels is steady, the vocal movement for consonant sounds is transient, as shown in Fig. 9. The acoustic phenomena of consonant sounds are complicated. It is difficult to reproduce these phenomena.

We must control the 15-DOF talking robot coordinately in the space and time to reproduce the short phenomena. Therefore, the Japanese voice generally consists of two phonemes of the first consonant sound and the last vowel. We considered the phenomenon of the voice as three parts, as shown in Fig. 10. The position during the transient consonant sound is interpolated. The target trajectory of each actuator is determined by the parameters.

We searched for the parameters of the consonant sounds, as shown in Fig. 11, which was a very difficult process. At first, we determined the initial parameters for the position and the time by referring to the parameters of Japanese vowels and the acoustic characteristics of the consonant sounds. Next, we simulated the movement of each articulator using the simulator described in chapter 3. Then, we produced WT-1R’s consonant sounds, and made fine adjustments to each parameter so

\[
q = \begin{cases} 
q_1, & (0 \leq t \leq t_1) \\
q_1 + (q_2 - q_1) \cdot \cos\left(\frac{\pi}{2} \left(1 - \frac{t-t_1}{t_2-t_1}\right)\right), & (t_1 \leq t \leq t_2) \\
q_2, & (t_2 \leq t \leq t_3)
\end{cases}
\]

Fig. 9 Trajectory of a stop: (1) Closure of the vocal tract, (2) Open of the closure, (3) Transition of the articulator to the next sound

Fig. 10 Speech planning for consonant sounds

Fig. 11 Search for the parameters of consonant sounds
that WT-1R’s consonant sound imitated a human’s using FFT, we could then recognize WT-1R’s consonant sound as that of a human’s by hearing.

As the result, WT-1R could produce some consonant sounds (/s/, /h/, /m/, /p/). The time parameters are shown in table 2. We describe the acoustic characteristics and the results in chapter 5.

5. Experiments of Consonant Sounds

5.1 Fricative /s/ and /h/

A fricative is a voiceless sound without vibrating the vocal cords and is produced by the turbulent flow when breathing strongly through the narrow part of the vocal tract. The fricative /s/ is produced by approaching the tip of the tongue to the teeth and breathing strongly through the narrow part between the tongue and the teeth, as shown in Fig. 12 (a). The fricative /h/ is produced by opening the glottis of the vocal cords and breathing and vibrating the side of the glottis, as shown in Fig. 12 (b).

We experimented on the Fricative /s/ and /h/ with WT-1R. The spectrums are shown in Fig. 13 and Fig. 15. The wave pattern of /sa/ is shown in Fig. 14. We confirmed that they were similar to a human’s. We recognized WT-1R’s fricative /s/ and /h/ as a human’s by hearing.

5.2 Nasal Sound /m/

A nasal sound is a voiced sound by vibrating the vocal cords and passing air through the nasal cavity. The nasal sound /m/ is produced by closing the mouth, opening the soft palate, and passing air through the nasal cavity, as shown in Fig. 12 (c). The nasal sound has the biggest formant in low frequencies (250-300 [Hz]) [7].

We experimented on the nasal sound with WT-1R’s nasal cavity. As a result, we recognized WT-1R’s nasal sound as the human nasal sound /n/, and confirmed that it has the biggest formant in low frequencies (250-300 [Hz]) and was similar to a human’s sound, as shown in Fig. 16. We recognized WT-1R’s nasal sound /m/ as a human’s by hearing.

5.3 Stop /p/

A stop is a voiceless sound without vibrating the vocal cords and is produced by opening the closed vocal tract in a moment. The stop /p/ is produced by opening the closed mouth in a moment, as shown in Fig. 12 (d).

We experimented on the stop /p/ with WT-1R. However, because the stop has especially the short phenomena, it is difficult to estimate it using FFT. We recognized WT-1R’s stop /p/ as a human’s by hearing.

5.4 Recognition Rate of WT-1R’s Voices

The vocal movement for consonant sounds is transient. The acoustic phenomena of consonant sounds are complicated. It is difficult to estimate these phenomena using FFT.

Therefore, when we showed the robot at Nagashima Superland in Japan in April 2000 and at the Robofesta in Japan in September 2000, we surveyed 460 people on the recognition of WT-1R’s voices with the following question:

“What could you recognize in WT-1R’s speaking?” They answered by pushing the five colorful buttons from the following choices:

1. /aiueo/
2. /sasuseseo/
3. /hahihuheho/
4. /mamimumemo/
5. /papipupepo/

The result of the recognition rate is shown in Fig. 17. However, it is difficult to accurately say that WT-1R is recognized correctly. We must improve the mechanisms of WT-1R and develop control systems for the production of the human voices.

6. Conclusions and Future Works

We developed a talking robot WT-1R (Waseda Talker-No.1 Refined) that improved from WT-1 for the production of natural vowels and consonant sounds. We proposed the speech planning of WT-1R to reproduce the complicated phenomena of consonant sounds. We controlled the 15-DOF

<table>
<thead>
<tr>
<th></th>
<th>/s/</th>
<th>/h/</th>
<th>/m/</th>
<th>/p/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steady Consonant Sound [ms]</td>
<td>300</td>
<td>300</td>
<td>200</td>
<td>100</td>
</tr>
<tr>
<td>Transient Consonant Sound [ms]</td>
<td>20</td>
<td>20</td>
<td>100</td>
<td>20</td>
</tr>
<tr>
<td>Vowels [ms]</td>
<td>150</td>
<td>150</td>
<td>20</td>
<td>80</td>
</tr>
</tbody>
</table>

Fig. 12 Acoustic position of consonant sounds
talking robot coordinately in the space and time. WT-1R could speak Japanese vowels (/a/, /i/, /u/, /e/, /o/) and some consonant sounds (/s/, /h/, /m/, /p/ and /waseda/).

For future works, we aim to completely produce the human voices. We are currently in the design process of a new robot WT-2 to improve the mechanisms of WT-1R and to develop new control systems. The new robot is expected to be able to speak all consonant sounds like humans’.

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 at Okino Industries, Ltd. for helping us to develop WT-1R’s hardware.

References