CONTROL SYSTEM FOR TALKING ROBOT TO REPLICATE ARTICULATORY MOVEMENT OF NATURAL SPEECH


*1 Communication Science Laboratories, Nippon Telegraph and Telephone, Kanagawa, Japan
*2 Department of Mechanical Engineering, Waseda University, Tokyo, Japan
*3 Humanoid Robotics Institute, Waseda University, Tokyo, Japan
*4 Advanced Research Institute for Science and Engineering, Waseda University, Tokyo, Japan
*5 Core Research for Evolutional Science and Technology, Japan Science and Technology

ABSTRACT

The ultimate goal of our study is to create a new speech production system, in which an anthropomorphic hardware talking robot is handled so as to imitate human articulatory movement. We have developed a software motion simulator, which generates control parameter sequences for the talking robot Waseda Talker No. 2 (WT-2) using the trajectories of human articulatory organs during continuous utterances measured by an electromagnetic articulograph. This paper mainly describes the motion simulator. In addition to its main function as a generator of control parameters for WT-2, the motion simulator also simulates the resultant acoustic characteristics related to WT-2’s vocal tract shape at each instance during the motion. The hardware structure of WT-2 is also described briefly.

This comprehensive approach will enable us to study speech production using the features of the vocal tract shape as speech motor tasks, instead of using acoustic features.

1. INTRODUCTION

In the process of speech production, the task planning mechanism for speech motor control is considered to have a hierarchical structure. Acoustic features (formant frequencies, formant bandwidths, etc.) may be useful in representing motor tasks, especially in the case of vowel production [1]. For consonants, however, it is difficult to extract acoustic features that represent the tasks appropriately. Furthermore, when considering the mappings between the articulators and the acoustic features, there exists some redundancy. The features of the vocal tract shape, on the other hand, are beneficial in specifying the motor tasks more appropriately for both vowels and consonants [2]. The crucial features are the constrictions or the closures in the vocal tract, which are formed by the tongue and the palate, or by the lips.

What is needed for the study of the speech motor control mechanism is a speech production process simulator that can sufficiently reproduce the behavior of human articulators and generates acoustic phenomena similar to those of natural speech. Such a simulator might be available in the form of either computer software programs or a hardware apparatus. Existing speech synthesizing techniques based on the speech production mechanisms [3], [4], however, are insufficient for the purpose. One of the reasons for this is the difficulty of simulating turbulent flow, which occurs in consonantal sounds. It would be easier to produce such a complex acoustics by driving a hardware vocal tract so as to reproduce movement like the immediate release of closure or very narrow constriction. The resultant acoustics should be closer to the natural consonants. On the other hand, developing a hardware device and driving mechanism physiologically similar to articulatory organs is also unpromising, because the accurate control of muscle-like or soft-tissue-like materials would be very difficult in terms of the complexity of their mechanical dynamics. Hence, it would be simpler to develop an excellent computational model representing the motor control mechanism.

As part of a comprehensive study of this very challenging problem, we have been developing a system comprising a hardware talking robot and its control software. For the robot, a mechanical structure that can be stably and precisely controlled is desirable. The talking robot Waseda Talker No. 1 (WT-1) [5] and its advanced version WT-1R [6] can produce CV concatenated speech with comprehensive quality. These robots have an anthropomorphic hardware mechanism consisting of a lung, vocal chords, oral and nasal cavities, which similar to human articulatory organs compared to previous hardware mechanisms [7], [8], [9]. As a result conquering some of the problems that surfaced in working with WT-1, a new hardware, Waseda Talker No. 2 (WT-2, to be published), has been developed. WT-2 has much more flexibility in changing tongue shape and much less sound pressure leakage from other parts than mouth and nostrils than WT-1R.

This paper describes a preliminary study on achieving continuous speech utterances using WT-2. A software motion simulator was developed, which generates control sequences for WT-2 to replicate an observed articulatory movement of a human. With this simulator, WT-2’s articulators move so as to imitate a human-like articulatory motion. This approach will enable us to study speech production using the features of the vocal tract shape as the motor tasks, instead of using acoustic features.
2. SYSTEM OVERVIEW

In this study, the talking robot and its motion simulator make up an entire system. The specifications of the talking robot WT-2 are briefly summarized in this section. The motion simulator generates control parameters for WT-2 based on observed articulatory movements of a human subject. Also this simulator is used for the simulation of the resultant time-varying vocal tract acoustic transfer characteristics of WT-2.

2.1. Talking robot

WT-2 consists of a lung, vocal chords, and oral and nasal cavities. It has a 15-DOF (degrees of freedom) structure: 1 DOF for the lung, 3 for the vocal chords, 5 for the tongue, 4 for the lips, 1 for the lower jaw, and 1 for the pharyngeal port. The head of WT-2 is shown in Fig. 1. The oral cavity was designed by deforming the midsagittal profile of a human vocal tract, but the cross section of the vocal airway has rectangular shape, except for the mouth opening. The width of the vocal tract cross section is 30 mm. The average vocal tract length is approximately 175 mm.

The tongue is super soft rubber (TP010, Tokyo Rubber Industrial Co., Ltd). The thickness of the rubber is 7 mm. The shape of tongue is controlled by five points using looped wires, as shown in Fig. 2. These five control points are at 7, 16, 35, 48, and 87 mm from the tip of tongue along the surface.

For both the upper and the lower lips, opening and protrusion degrees are controlled using looped wires. The function of the lower jaw is approximated by linear movement of the lower teeth in vertical direction. The structures of the lips and the lower teeth sections are shown in Fig. 3.

The vocal chords are made from the same material as the tongue. A pair of symmetric rubber plates forms a slit like the glottis. These plates are stretched or loosened by a rotor. When loosening them fully, the air from the lung directly flows into the vocal tract. By adjusting the degree of stretching, the fundamental frequency of the vocal chords vibration can be controlled with a range of approximately 120 to 230 Hz.

2.2. Motion simulator

The motion simulator 1) generates control parameter sequences for WT-2 based on observed articulatory movement of a human subject and 2) calculates the time-varying vocal tract acoustic transfer characteristics by specifying the vocal tract area function on each instantaneous vocal tract shape.

2.2.1. Generating control parameters

Articulatory movements of a human subject were observed with an AG100 electromagnetic articulograph (EMA) [10]. Seven receiver coils were placed on the lower lip (LL), upper lip (UL), lower jaw (LJ), four points on the surface of tongue from the tip to the blade (T1, T2, T3 and T4) in the midsagittal plane. The LJ receiver was attached to the center of the lower incisors. Positions of these receiver coils were scanned at a sampling frequency of 250 Hz. Figure 4 shows observed trajectories during an /aïueo/ utterance. Also, the shape of hard palate was extracted by moving the receiver coil along the hard palate by hand. Speech sound was simultaneously recorded at a sampling frequency of 16 kHz using a dynamic microphone.

Figure 5 shows the main window of the motion simulator. The observed positions of the seven EMA receiver coils are denoted LL, UL, LJ, T1 – 4. This window can animate the observed sequences. The human palate shape is replaced with that of WT-2 by overlapping the neck of central incisors, denoted PIVOT. In generating suitable tongue movement for WT-2, the
angle of the palate on the observation coordinate is very important. The appropriate angle is defined manually by rotating the palate around the pivot, so that the crucial closure or constriction related to the tongue and palate is formed during the entire utterance. Also the set of the four tongue-related points, T1 – 4, is shifted along the direction of the hard palate for fine adjustment of closure and constriction. In order to calculate the length of the five wires related to WT-2’s tongue control, the tongue surface line is generated by interpolation of those observed points. To acquire an appropriate tongue surface shape, the very tip of the tongue, V1, is defined by referring to T1 and T2, because the T1 receiver coil is placed a little behind the tongue tip for a tight fit. Point V3 is the WT-2’s tongue root position. Point V2 is defined by referring to T4 and V3. Finally, the tongue surface line is generated by interpolating the seven points (V1, T1 – 4, V2 and V3). Then, the length of five wires on WT-2’s tongue surface are extracted as the control parameters. The distance between the upper and lower lips (the double-headed arrow) and the downward opening degree of the lower teeth (the downward-pointing arrow) directly correspond to the WT-2’s control parameters.

**2.2.2. Calculation of acoustic transfer characteristics**

The motion simulator can also calculate the acoustic transfer characteristics corresponding to the vocal tract shape at each instance during a motion. To calculate the transfer characteristics, the vocal tract area function is extracted from the vocal tract shape first. The method for extracting the area function is similar to the conventional one [11]. Figure 6 shows the simulator’s window drawing the midsagittal profile of the determined vocal tract area function. The mid-line of the vocal tract is defined, at first, by concatenating a series of mid-point between the palate and the tongue surface. These mid-points are found in each segmental region separated at a specified interval along the palate line. Then, a series of lines perpendicular to the defined mid-line is extracted at the interval that specifies the sampling frequency for acoustic simulation. Finally, a series of cross-sectional area are calculated by multiplying the width of vocal tract to the length of those perpendicular lines. The acoustic transfer characteristics corresponding to the vocal tract area function are calculated using an acoustic propagation model that takes into account the effect of acoustic loss due to viscosity and heat conduction of air, and the tract’s wall vibration.

On the other hand, the acoustic transfer characteristics also can be extracted from speech signal that was recorded simultaneously with the articulatory movement. An ordinary LPC analysis technique is used for the extraction. The simulator draws these two power spectrum envelopes in a window, and calculates the distance between the two spectrum envelopes as the cepstrum distance.
The whole procedure for determining the vocal tract area function and calculating its acoustic transfer characteristics is completely automatic. The time-varying features, therefore, can be monitored by animation.

The calculation procedure for the vocal tract transfer characteristics is, of course, applicable to both a human subject and WT-2.

3. ACOUSTIC EVALUATION

The acoustic transfer characteristics calculated from the vocal tract shape and that extracted from speech signal were compared using the simulator. The data set consisted of the simultaneously measured articulatory movement and speech signal for one adult male subject. The utterances were eleven types of CVCVCCVCV sequences, such as /babibubebo/ and /kakikukeko/. The subject uttered each sequence three times.

The distance was calculated in frame-by-frame comparison manner. The rate of the frame shift was 4 ms according to the sampling frequency for the observation of the articulatory movement. The number of analysis frames was 6420. The cepstrum distance values for the whole 6420 frames of spectrum envelopes were calculated. Table 1 shows the comparison between human speech signal and human vocal tract shape. Table 2 shows the comparison between human speech and WT-2’s vocal tract shape. In the case of the human subject, as shown in Table 1, the length from the glottal port to the glottis (L in Fig. 6) was varied on determining the vocal tract area function. There were no significant differences depending on the position of glottis within 10 mm up and down. In the case of robot, the length L is 75 mm. Though the distance was increased by changing the palatal shape from the human one to the WT-2 one, the minimum value for WT-2 still remained small. WT-2’s motion based on the observed natural trajectories of human is, consequently, expected to reproduce natural speech-like acoustic phenomena.

<table>
<thead>
<tr>
<th>Length from pharyngeal port to glottis [mm]</th>
<th>Cepstrum Distance [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>70</td>
<td>4.274</td>
</tr>
<tr>
<td>75</td>
<td>4.278</td>
</tr>
<tr>
<td>80</td>
<td>4.266</td>
</tr>
</tbody>
</table>

Table 1: Spectral distance on vocal tract transfer characteristics (between human speech and human vocal tract area function). Length from pharyngeal port to glottis (L in Fig. 6) was varied on determining the vocal tract area function.

<table>
<thead>
<tr>
<th>Length from pharyngeal port to glottis [mm]</th>
<th>Cepstrum Distance [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>75</td>
<td>4.918</td>
</tr>
</tbody>
</table>

Table 2: Spectral distance on vocal tract transfer characteristics (between human speech and WT-2’s vocal tract area function).

4. CONCLUSIONS

A software motion simulator has been developed that generates control sequences for a talking robot to replicate an observed articulatory movement of a human.

We have already carried out some preliminary experiment using WT-2 and the motion simulator and confirmed that WT-2 reproduces consonantal sounds, such as fricatives or plosives, in a continuous utterance. Future work includes verifying the accuracy of WT-2’s articulatory positions in motion using the control parameters generated from the motion simulator.

5. REFERENCES