SPEECH SYNTHESIS USING A PHYSIOLOGICAL ARTICULATORY MODEL WITH FEATURE-BASED RULES

Jianwu DANG [1,2] Jiping SUN[2], Li DENG[2] and Kiyoshi HONDA [1]
[1] *ATR•Human Information Processing Research Labs, Kyoto, Japan,* [2] *Univ. of Waterloo, Waterloo, ON, Canada*

ABSTRACT

A 3-D computational model of speech articulators has been developed for human-mimetic speech synthesis. The model geometry was derived from volumetric MRI data that were collected from one male speaker. A multipoint control strategy is developed to control the model, which involves three points of the articulators: the tongue tip, tongue dorsum, and the jaw. To control these points in the geometric space of the vocal tract independently, a set of weight coefficients is defined for each muscle in a specific control point. A dynamic muscle workspace is proposed to predict muscle force vectors for a control point in any arbitrary position. Muscle activation signals are generated via the dynamic workspace, and fed to the muscles to drive the model. To develop a speech synthesis system using the physiological model, this study explores some feature-based phonological rules, which provides temporally overlapping articulatory targets from a given sequence of phonetic segments. Examples of the synthetic sounds are given using the model.

1. RECONSTRUCTION OF SPEECH ORGANS

To replicate human speech organs in a computer, speaker-specific customization of the model was carried out by duplicating the anatomical information that was obtained from volumetric MRI data of a male Japanese speaker. The shape of the speech organs was constructed based on volumetric MR images.

The tongue shapes were extracted from volumetric MR images in the midsagittal and parasagittal planes. The basic structure of the tongue tissue model approximates the fiber orientation of the genioglossus muscle. The central part of the tongue that includes this muscle is represented by a 2-cm-thick layer bounded by three sagittal planes. Each plane is divided into six sections with nearly equal intervals in the anterior-posterior direction and ten sections along the tongue surface. Altogether, the tongue model consists of 120 polyhedrons with eight vertices. This modeled tongue is capable of forming the midsagittal groove and the side airway, and thus can simulate the essential behavior of the tongue in producing vowels and consonants.

To form a vocal tract shape, this study also reconstructs the surrounding organs based on physiological data. The outlines of the vocal tract wall were extracted from MRI images in the midsagittal plane, and the parasagittal planes of 0.7 and 1.4 cm apart from the midsagittal plane on the right side. With an assumption that the left and right sides are symmetric, 3D surface models of the vocal tract wall and the mandibular symphysis are reconstructed using the outlines with 0.7 cm intervals in the left-right direction. Because of the geometrical complexity, it is difficult to describe the surface walls using an analytic function. For this reason, the surfaces of the tract wall and the mandibular symphysis are approximated using small triangular planes, 432 planes for the tract wall, and 192 for the mandible (see [3] for details).

To produce a realistic vocal tract shape for speech production, the articulatory model should include the tongue, lips, teeth, hard palate, soft palate (the velum), pharyngeal wall, and the larynx. At the present stage, the lips and the velum are not yet modeled physiologically. Their rules are taken into account in speech synthesis empirically by incorporating their areas into an acoustic model. The movements of the larynx are also not taken into account in this model yet.

2. MODELING OF THE SPEECH ORGANS

Although the finite element method is commonly used in modeling the tongue tissue [1,2], it is not good for simulating the large and fast deformation such as the deformation of the tongue body. To develop a method that is adequate to this situation, a mass-spring network is employed in this model. Use of the mass-spring network can integrate the soft tissue and rigid organs in the same motion equation. This treatment reduces the computational cost significantly. Each mass-point connects to the adjacent nodes in the same plane and corresponding nodes in the neighbor plane(s) by viscoelastic springs.

In addition to the tongue shape, outlines of the rigid organs (*i.e.*, the jaw and hyoid bone in the present work) were also traced from the MRI data for the target speaker. The contours of the bony organs are identifiable in MR images when they are surrounded by soft tissue. According to the extracted contour, the mandible is modeled by four mass-points on each side, which form two triangles using five rigid beams including one shearing-beam [4]. The mandible model is combined with the tongue model at the mandibular symphysis. The temporomandibular joint is designed to produce two types of motions; rotation and translation. The model of the hyoid bone has three segments corresponding to the body and bilateral greater horns, which also has rotation and translation motions. Each segment of the hyoid bone is modeled by two mass-points connected by a rigid beam.

The anatomical arrangement of the major tongue muscles is determined based on high-resolution MR images obtained from the same target speaker. The genioglossus and geniohyoid muscles are extracted in the midsagittal plane. The superior and inferior longitudinal muscles were identified in the plane 0.6cm apart from the midsagittal. The hyoglossus and styloglossus are traced in the plane 1.5cm apart from the midsagittal. The other muscles are modeled with reference to an anatomical literature [6]. Altogether, eleven tongue muscles are included in the tongue model, and eight muscles for the mandible-hyoid bone complex, which are divided into two groups; jaw-opener and jaw-closer (see [5] for details).

3. DYNAMIC CONTROL OF THE MODEL

To simulate articulation for any arbitrary phoneme, it requires controlling several portions of the tongue to form a specific vocal tract shape. For this purpose, we propose a multi-point control strategy, which involves the tongue tip, tongue dorsum, and the jaw. The *tongue tip* is defined as the most anterior node of the tongue in the midsagittal plane. The *tongue dorsum* was represented by the average position of three midsagittal nodes around the highest point in the rest shape. The jaw is controlled using a point at the tip of the mandible incisors.

3.1. Construction of a Dynamic Muscle Workspace

The idea of this target-based control strategy is to generate muscle activation signals according to a given target for each control point, and then drive the articulatory model using the muscle activation signals. For this purpose, muscle workspaces have been developed to establish a relationship between the target and the activation signals [3]. Each of the muscle workspaces consists of muscle force vectors for a given initial position. Since the initial position moves with the jaw and tongue, the muscle force direction is expected to change with the movement. To solve the problem, this study proposes an improved method using a number of muscle workspaces for each control point in different initial positions, and then derives muscle force vectors in an arbitrary position based on the workspaces.

Figure 1 shows the muscle workspaces for three control points. Four initial positions are specified for the control points of the tongue tip and tongue dorsum, respectively. The workspaces are shown in dark lines for the tongue tip, and light lines around the tongue dorsum for the dorsum control point. The central initial position corresponds to the rest position of the tongue while the others correspond to three extreme positions. For the jaw, the initial positions are chosen in the rest position and a wide-open position. To construct the workspaces, we first input certain forces to the muscles and move the control points to a specific position, and then excite each muscle by unit activation with a fixed duration. For the given excitation, each control point moves from the initial position and to a new position. This displacement forms a vector in the geometric space, referred to as the muscle vector of a control point. All of the muscle vectors form the muscle workspaces.

Figure 2 shows an example of constructing a dynamic muscle workspace for the tongue dorsum. The thin lines show the basic workspaces with different initial positions. The dynamic muscle workspace for a given position is constructed using the following equations based on the four basic workspaces, and shown by the thick lines.

$$V = \frac{\sum_{i=1}^{4} L_i v_i}{\sum_{i=1}^{4} L_i}, \qquad L_i = \prod_{\substack{j=1 \\ j \neq i}}^{4} l_j^2$$

where V denotes the calculated muscle vector, v_i is the muscle vector in workspace i, and l_j is the distance from the current position to workspace j.

3.2. Generation of Muscle Activation Signals

Figure 3 shows an example of generating muscle activation patterns according to a given target in a simplified muscle workspace of the tongue dorsum. This simplified muscle workspace consists of four extrinsic muscles, shown by thick dark arrows. Supposing that the tongue dorsum is located in the *current position Pc* and moves backward to *target Tg*, the dashed line from *Pc* to *Tg* forms a vector, referred to as an articulatory vector. When the articulatory vector is mapped onto the muscle workspace, a set of projections is obtained for the muscle vectors. Although the obtained muscle projections can be positive or negative for each muscle vector, the positive projection alone provides an activating signal whose magnitude is proportional to the projection length. At the *current computational step* shown in Fig.3, SG and HG are the active muscles. The calculated activation signals drive the tongue dorsum to move to a new position, approaching to the target. When iterating the same procedure at each new position, the tongue dorsum finally moves from *Pc* to *Tg* along the gray path. At the same time, a set of time-varying activation signals is generated for the muscles.

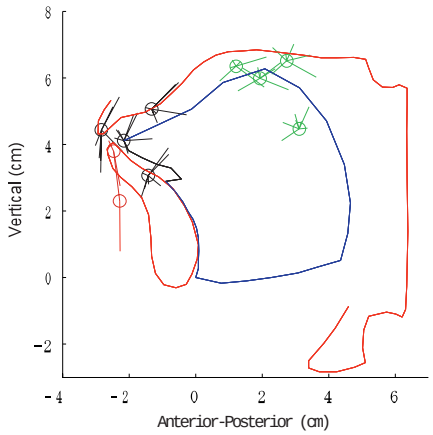

Figure 1 Basic muscle workspaces for three control points. Four muscle workspaces were built for tongue tip (dark lines) and tongue dorsum (light lines), and two for the jaw (light lines).

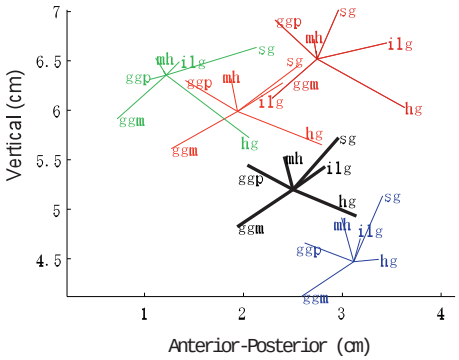

Figure 2 The constructed dynamic muscle workspace shown by the thick lines for the tongue dorsum at a given position. The thin lines show the basic muscle workspaces with different initial positions.

It is known that the extrinsic tongue muscles and the jaw muscles reign the whole tongue body including the tongue dorsum and tongue tip. To control the tongue dorsum and tongue tip independently, a weight coefficient is defined for each muscle at a specific control point. Large coefficients are defined for the extrinsic muscles to control tongue dorsum while the intrinsic muscles have large coefficients to control the tongue tip. The weight coefficients were determined by simulations using this model. Table 1 shows the determined coefficients.

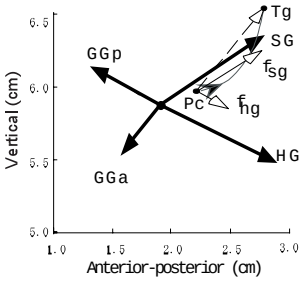

Figure 3 An example of the target-based control procedure in a simplified muscle workspace of the tongue dorsum.

Table 1 Weight coefficients of the muscles used for each control point.

	T-tip	T-dorsum	Jaw
Genioglossus Ant.	0.9	0.4	0
Genioglossus middle	0.4	0.6	0
Genioglossus Post.	0.3	1	0
Hyoglossus	0.2	1	0
Styloglossus	0.2	1	0
Longitudinal Sup.	1	0	0
Verticalis	1	0	0
Transversus	1	0.4	0
Longitudinal Inf.	1	1	0
Geniohyoid	0	1	0
Mylohyoid	0.9	1	0
Jaw-opener-muscles	0	0	1
Jaw-closer-muscles	0	0	1

3.3 Feature-Based Control Strategy

To generate articulatory targets from a text, it is necessary to develop a high-level control strategy for the articulatory model. A previous study [8] proposed a method that is able to automatically produce feature-overlapping patterns, in which multiple targets are specified simultaneously, according to physiological rules and linguistic constraints. The method in [8] provides only relative timing for the overlapping. This study extends this method and focuses on quantifying absolute timing of feature overlapping based on articulatory data obtained from the x-ray microbeam system. Briefly, the overlapping feature generator consists of 1) scanning the input string of feature bundles with high-level linguistic information; 2) matching (assigning) them to appropriate overlapping patterns; 3) executing overlapping operations by interpreting the overlapping types specified in the patterns; and 4) integrating the results of 3) to produce a state-transition network. A block diagram of the overlapping feature generator is shown in Fig. 4. The result of the generator provides detailed information about timing of several articulators' targets in an arbitrary phoneme sequence.

4. MODEL SIMULATION AND SOUND SYNTHESIS

For a given target sequence, a set of muscle activation signals is generated for the muscles at the three control points according to the above control strategy. The resultant activation pattern is the summation of the signals obtained for all the control points. As an example, this model is used to produce an English phrase "thank you". This phrase is selected because both the tongue tip and tongue dorsum in producing this phrase move from one of the extreme positions to another in the articulatory space. Figure 5 shows the activation patterns of the muscles in producing "thank you". At the first five steps, the GGm and transversus muscle act to control the tongue tip to move forward to the position of "th". Between the steps 10 and 15, HG and Jaw-opener act to move tongue body to the position for /ae/ sound. The longitudinal-inferior muscle acts in this period to draw the tongue tip back. During steps 20 to 25, a number of muscles are excited to move the tongue dorsum from /ae/ position to /k/ position. At the remaining steps, the muscles act to shape the vocal tract for the sound "you".

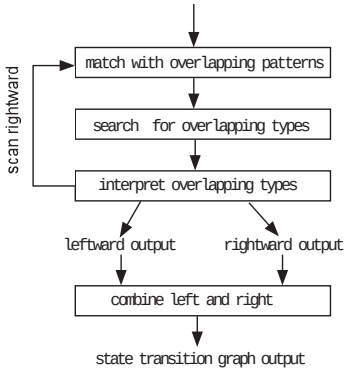

Figure 4. The overlapping feature generator from the input text.

To produce a dynamic vocal tract shape, the tongue tip, tongue dorsum and jaw are driven by the muscle activation signals to move forward to the given targets. Trajectories of the three control points are plotted in Fig. 6 using the thin lines with various marks. At the first 10 steps shown by the circle marks, the tongue tip moves forward and upward to make a constriction with the teeth while the tongue dorsum moves downward from the initial position to prepare for the next vowel. At the second 10 steps indicated by the cross marks, the tongue tip moves a long distance backward and downward while the tongue dorsum moves mainly downward. At the next 10 steps shown in squares, both the tongue tip and dorsum move in about the same direction, and then reach their targets, respectively. For the trajectory of the jaw, the movement is mainly caused by the rotation component around the rest position indicated by the arrow. When the jaw opens wider, the translation component becomes larger, and thus it makes the jaw move vertically.

As shown in Table 1, the tongue tip is mainly controlled by the intrinsic muscle, and the tongue dorsum is mainly controlled by

the extrinsic muscles. The trajectories show that both the extrinsic and intrinsic muscles demonstrate desired performance. This result suggests that this control strategy provides a practical way to drive a physiological model, although it is not fully realistic from a physiological point of view.

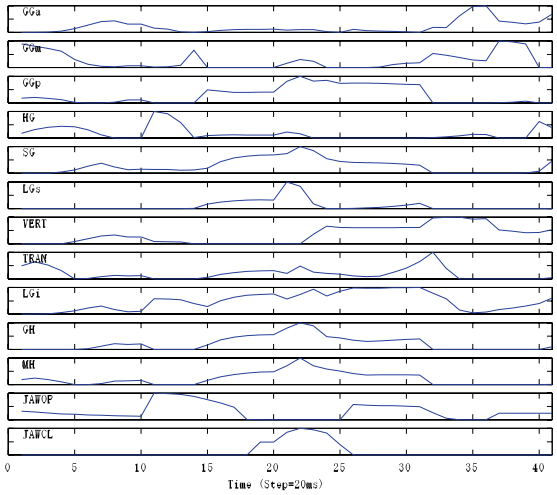

Fig. 5 The generated muscle activation patterns in producing a phrase "thank you".

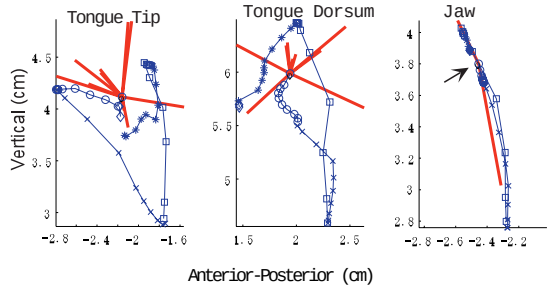

Fig. 6 Trajectories of the three control points in producing phrase "thank you". The thick lines show muscle force vectors in the muscle workspaces, and the thin lines show the trajectories. The marks indicate time steps.

From the above calculation, a series of vocal tract shapes is obtained for speech production. Since the present model does not provide a full 3D vocal tract shape, the real area function of the vocal tract has to be estimated using the information of the partial vocal tract of the model. This study first determines the vocal tract widths in the midsagittal and parasagittal planes of the model, and then estimates vocal tract area functions using the width information [3]. A series of area function is obtained in a 20-ms interval from the vocal tract shape. The nasal tract is coupled to the oral tract by a functional nasopharyngeal port, and the configuration of the lips is specified by the length and size of the lip tube. A transmission line model [7] was employed for synthesizing speech sound. Figure 7 shows the synthetic sound waveform and spectrogram of phrase "thank you". The formants in the spectrogram show realistic trajectories, especially for the first and second formants. According to an informal listening test, the quality of the synthetic sound is quite natural.

5.CONCLUSIONS

This study developed a speech synthesizer based on a physiological articulatory model that includes the tongue body and surrounding articulators. To complete the control strategy of the articulatory model, a dynamic muscle workspace is proposed to generate muscle activation signals according to articulatory targets. Using the muscle activation signals, the proposed multi-point control strategy successfully controls the tongue tip, tongue dorsum, and jaw to move to their targets. Thus, a series of dynamic vocal tract shapes is obtained for speech synthesizer. This model demonstrated some behaviors characteristic to human speech articulation in producing both consonants and vowels. This paper also discussed the rules to generate articulatory targets from a text.

ACKNOWLEDGMENTS: The synthesis program used in this study was developed from Shinji Maeda's program [7]. The authors would like to thank him for providing the program.

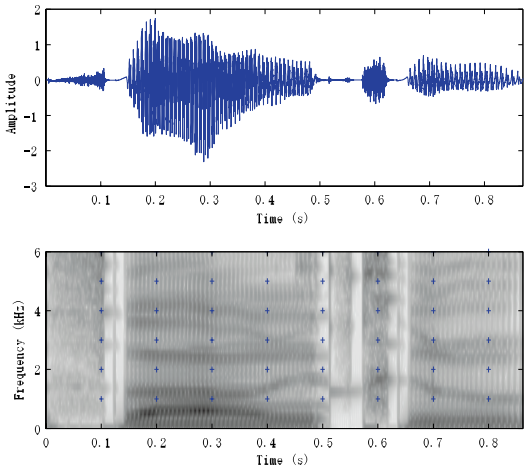

Fig. 7 The waveform and spectrogram of synthesized sound of "thank you".

REFERENCES

[1] Kakita, Y., Fujimura, O., and Honda, K. (1985). "Computational of mapping from the muscular contraction pattern to formant pattern in vowel space," In *Phonetic Linguistics*, edited by A. L. Fromkin, (Academic, New York).

[2] Wilhelms-Tricarico, R. (1995). "Physiological modeling of speech production: Methods for modeling soft-tissue articulators," J. Acoust. Soc. Am. 97, 3805-3898.

[3] Dang, J. and Honda, K. (1998). "Speech production of vowel sequences using a physiological articulatory model," Proc. ICSLP98, Vol. 5, pp1767-1770.

[4] Dang, J. and Honda, K. (1998). "A physiological model of a dynamic vocal tract for speech production," Tech. Report of ATR, TR-H-247.

[5] Dang, J. and Honda, K. (1999.3). "Speech synthesis of VCV sequences using a physiological articulatory model," 137th ASA meeting (Berlin, Germany).

[6] Miyawaki, K. (1974). "A study of the musculature of the human tongue," Ann. Bull. Res. Inst. Logoped. Phoniatrics, Univ. Tokyo, 8, 23-50.

[7] Meada, S. (1996) "Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer," Phonetica, 127-232.

[8] Sun, J. and Deng L. (1998). "Use of high-level linguistic constraints for constructing feature-based phonological model in speech recognition," Proc. ICSLP98, 73-76.