SPEECH SYTHESIS OF VCV SEQUENCES USING A PHYSIOLOGICAL ARTICULATORY MODEL

Jianwu DANG [1,2] and Kiyoshi HONDA [1]
[1] ATR Human Information Processing Research Labs,
2-2 Hikaridai Seikai-cho Souraku-gun, Kyoto, Japan, 619-02
[2] Univ. of Waterloo, Waterloo, ON N2L 3G1, Canada

ABSTRACT

A 3-D articulatory model has been male speaker. The model consists of the constructed based on volumetric MRI data for a Japanese midsagittal layer of the tongue, jaw-hyoid bone complex, and vocal tract wall that comprise the main vocal tract. This work describes a multi-point control strategy for producing vowel-consonant-vowel sequences through the generation of muscle contraction parameters based on target-reaching tasks. In this method, three control points are chosen on the mandible, tongue tip and tongue dorsum, and the corresponding target points are defined as a fixed position for each phonetic segment of the utterance. The time sequences of muscle activation signals are determined by iterating model simulation to reduce the distance between the control and target points. The muscle activation signals are obtained in a space of muscle force vectors defined for each control point, and fed to the muscles to drive the model. Generated articulatory movements of the model derive the sequence of vocal tract area functions. Examples of the synthetic sounds are demonstrated using the area functions.

1. RECONSTRUCTION OF SPEECH ORGANS

To replicate human speech organs, 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 roughly replicates 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 with 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. To the end, the tongue model consists of 120 polyhedrons with eight vertices. This modeled tongue is capable to form the midsagittal groove and the side airway, which can simulate the essential behaviors of the tongue in producing vowels and consonants.

To form a vocal tract shape, this study also reconstructed 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 were reconstructed using the outlines with 0.7 cm intervals in the left-right direction. Because of the geometrical complexities, 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 were approximated using small triangular planes, 432 planes for the tract wall, and 192 for the mandible. A 3D projection of the proposed model is shown in Figure 1.

Fig. 1 The oblique view of three-dimensional model of the speech organs. All dimensions are in cm.

To produce a 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 were not modeled physiologically in this model. They are taken into account in speech synthesis by

incorporating their areas into an acoustic model. The movements of the larynx are not taken into account in this model, which are discussed in another study of the authors [5].

2. MODELING OF THE SPEECH ORGANS

Since the finite element method, which is commonly used in modeling the tongue tissue [1,2], 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. The mesh lines in Fig.1 show the viscoelastic springs in the network of the tongue body, and mass-points are located in the intersections of the mesh lines. The mass-points in the midsagittal plane also connect to the corresponding mass-points in the right and left planes by the springs. To preserve the strain force caused by the deformation of the mass-spring network, each mass-point is also connected with diagonal adjacent ones using the springs in three dimensions. Thus, the original shape can be restored from a deformation when external forces are removed.

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

3. DYNAMIC CONTROL OF THE MODEL

In our previous study, we proposed a target-based control strategy to control the tongue position in producing vowel sequences, in which only extrinsic tongue muscles were involved. To produce consonants, however, it requires to control more than one point of the tongue to form the vocal tract. For this purpose, we propose a multi-point control strategy, which involves the tongue tip, tongue dorsum, and the jaw. The *tongue tip* was defined the first node at the tongue tip in the midsagittal plane of the model. The *tongue dorsum* was represented by the average position of three midsagittal nodes around the highest point in the initial shape. The jaw is controlled using a point at the mandible incisors.

Fig 2 Muscle workspaces for three control points, in which the weight coefficients of the muscles are larger than 0.3.

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, it is necessary to establish a relationship between the target and the activation signals. To do so, each tongue muscle was excited by unit activation for a certain duration. For the given excitation, each control point moves from its 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. Using all the muscle vectors, three vector spaces can be constructed for the control points around their initial positions, respectively, which are referred to as the muscle workspaces. Figure 2 shows the muscle workspaces for the three control

points. The extrinsic tongue muscles and the jaw muscles reign the whole tongue body including the tongue dorsum and tongue tip globally because of their strength. 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.

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

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. Suppose that the tongue dorsum is located in the *current position Pc* and moves backward to a *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 projections 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.

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

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 sum of the signals obtained for all the control points. An ideal control strategy should be able to drive the control points independently to move to the specified articulatory targets. To verify the proposed strategy, we attempt to produce an English phrase “thank you” using the model. This phrase is selected because both the tongue tip and tongue dorsum move from one of the extreme positions to another in articulatory space in producing this phrase. Figure 4 shows the activation patterns of the muscles in producing “thank you”. At the first five steps, the muscle GGm and transversus act to control the tongue tip to the position of “th”. Between steps 10 and 15, HG and Jaw-opener act to move tongue body in the position for “a” sound. The muscle of Longitudinal Inferior 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 “a” position to “k” position. At remaining steps, the muscles act to shape the vocal tract for “you” sound.

To produce dynamic vocal tract shape, the tongue tip, tongue dorsum and jaw were driven by the muscle activation signals to move forward to the given targets. Trajectories of the three control points were plotted in Fig. 5 using the thin lines with various marks. At the first 10 steps shown by the circle mark, 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 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 reached their targets, respectively.

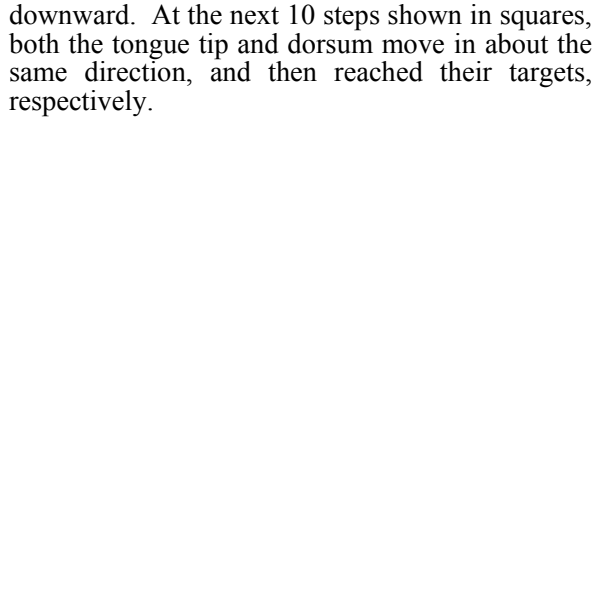

Fig. 4 Muscle activation patterns in producing a phrase "thank you".

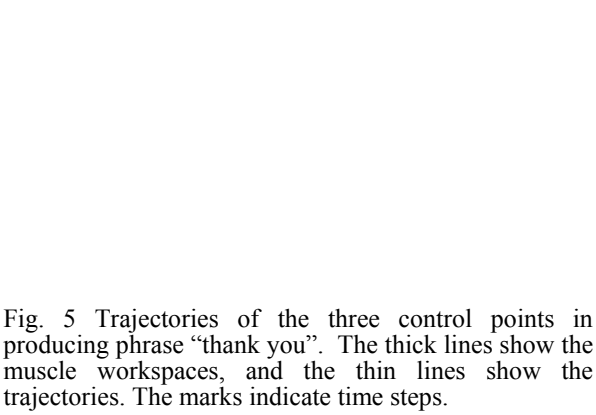

Fig. 5 Trajectories of the three control points in producing phrase "thank you". The thick lines show the muscle workspaces, and the thin lines show the trajectories. The marks indicate time steps.

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 demonstrated a desired performance. This result suggests that this control strategy provides a practical way to drive a physiological model, although it is not perfectly realistic from a physiological point of view.

From the above simulation, 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 (see [3] for more details). A series of area function is obtained in a 20-ms interval from the vocal tract shape. The coupling of the nasal tract to the oral tract is specified by a time-varying nasopharyngeal port. A transmission line model [7] was employed for synthesizing speech sound. According to an informal listening test, the quality of the synthetic sound was quite natural.

5.CONCLUSIONS

A mass-spring network was employed in modeling tongue tissue and rigid organs. In this version, the intrinsic muscles of the tongue were also manipulated in the control strategy. The proposed multi-point control strategy successfully controlled the tongue tip, tongue dorsum, and jaw according to given articulatory targets. This model demonstrated some behaviors characteristic to human speech articulation in producing both consonants and vowels. Natural sound quality was obtained from the dynamic vocal tract shape.

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.

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] Wu, C., Dang, J. and Honda, K. (1999.3). "A design of laryngeal structures for 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.