SPEECH PRODUCTION OF VOWEL SEQUENCES USING A PHYSIOLOGICAL ARTICULATORY MODEL

Jianwu DANG and Kiyoshi HONDA
ATR Human Information Processing Research Labs,
2-2 Hikaridai Seikai-cho Souraku-gun, Kyoto, Japan, 619-02

ABSTRACT

This report describes the development of a physiologically-based articulatory model, which consists of the tongue, mandible, hyoid bone and vocal tract wall. These organs are represented in a quasi-3D shape to replicate a midsagittal layer with a thickness of 2 cm for tongue tissue and 3 cm for tract wall. The geometry of these organs and muscles are extracted from volumetric MR images of a male speaker. Both the soft and rigid structures are represented by mass-points and viscoelastic springs for connective tissue, where the springs for bony organs are set to extremely large stiffness. This design is suitable to compute soft tissue deformations and rigid organ displacements simultaneously using a single algorithm, and thus reduces computational complexities of the simulation. A novel control method is developed to produce dynamic actions of the vocal tract, as well as to handle the collision of the tongue to surrounding walls. Area functions are obtained for vowel sequences based on model's vocal tract widths in the midsagittal and parasagittal planes. The proposed model demonstrated plausible dynamic behaviors for human speech articulation.

1. MODEL CONSTRUCTION

To replicate the behaviors of human speech organs, speaker-specific customization of the model was carried out by replicating the anatomical information that was obtained from volumetric MRI data of a male Japanese speaker.

1.1 Design of the Tongue Shape

The tongue tissue model is designed as a thick sagittal layer bounded by three sagittal planes. This design was chosen to form the midsagittal groove of the tongue and the side airway in producing vowels and consonants. The tongue tissue has been modeled commonly using the finite element method [1,2]. Our earlier study aimed at developing an integrated model that combined an FEM model of the tongue and a beam-muscle model of the jaw-larynx system [3]. The computations of movements in this hybrid model were slow because the achievement of an equilibrium between the soft tissue and rigid organs took considerable time. One possible solution to this problem is to model all speech organs using an identical method. To the end, a mass-spring network is used to model both the soft tissue and rigid organs in the current model.

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 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. The tongue tissue model is shown in Fig. 1 with the vocal tract wall. In the tongue model, the mesh lines represent viscoelastic springs, 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 relate a deformation and a stress in the mass-spring network, the mass-points also connect with diagonally adjacent ones by the springs. Thus, the original shape can be restored from a deformation due to the strain forces when external forces are removed. The mass per unit volume is chosen to be 1 g/cm^3 for the tongue tissue, which is the same as that of water.

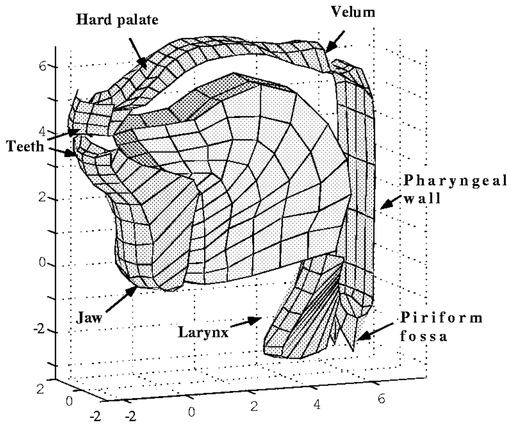


Fig. 1 The oblique view of three-dimensional model of the speech organs. All dimensions are in cm.

The Voigt model was adopted to approximate the properties of the tongue tissue, which consists of a spring parallel to a dashpot. The mechanical parameters for the spring and the dashpot reported in the previous studies deferred widely: the stiffness ranged from 10^4-10^6 dyne/cm^2, and the viscosity from 10^5-10^7 dyne•s/cm^2 [4]. In the present model, parameters were chosen to be 1.54×10^5 dyne/cm^2 for the stiffness and 1.75×10^5 dyne•s/cm^2 for the viscosity.

1.2 Modeling of the Rigid Organs

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 geometries, the mandible is modeled by four mass-points on each side, which form two triangles using five rigid beams including one shearing-beam [5]. 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 offers rotation and translation motions. Each segment of the hyoid bone is modeled by two mass-points connected by a rigid beam. Eight muscles are incorporated in the model of the mandible-hyoid bone complex.

1.3 Construction of the Vocal Tract Wall

To determine a vocal tract shape, it is necessary to incorporate the organs surrounded the tongue in the model. The surrounding organs are the lips, teeth, hard palate, soft palate (the velum), pharyngeal wall, and the laryngeal tube. At this stage, the present model has no lips, and treats the other organs as a single rigid wall. Therefore, the movements of the velum and larynx are not taken into account in the present model. The outlines of the vocal tract wall are extracted from MRI data 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, as shown in Fig. 1. Because of the geometrical complexities, it was not able to derive an analytic function for the surface walls. 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.

1.4 Arrangement of the Tongue Muscles

The anatomical arrangement of the major tongue muscles was determined based on high-resolution MR images obtained from the same target speaker. The genioglossus (GG), geniohyoid (GH), and mylohyoid (MH) were extracted in the midsagittal plane. The superior longitudinal (SL), and inferior longitudinal (IL) were identified in the plane 0.6 cm apart from the midsagittal. The hyoglossus (HG) and styloglossus (SG) were distinguished in the plane 1.5 cm apart from the midsagittal. The orientation of all these tongue muscles was also examined with reference to the literature [6]. Figure 2 shows the location for the extrinsic muscles in the model, (a) for the midsagittal plane, and (b) for the parasagittal plane. The genioglossus (GG), the largest muscle in the tongue, runs midsagittally in the central part of the tongue. Since the triangular muscle GG exerts different effects on tongue deformation in different parts, it can be functionally separated into three muscle bundles: the anterior portion (GGa), middle portion (GGm), and posterior portion (GGp). The thickness of the lines represents the size of the muscle units, the thicker the line, the larger the maximum force produced. The hyoglossus (HG) and styloglossus (SG), shown in the parasagittal plane, are designed to be symmetrical on the left and right sides. Totally, eleven tongue muscles were treated included in the model.

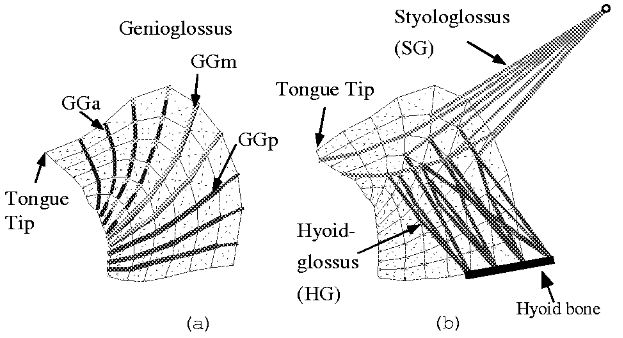


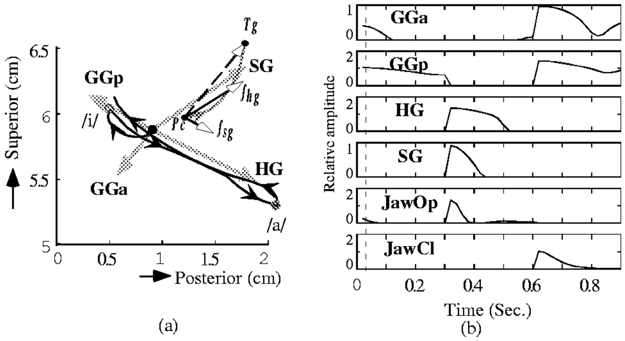
Fig. 2 Structure of extrinsic muscles of tongue model: (a) three bundles of genioglossus muscles (GGa, GGm and GGp) on the midsagittal plane and (b) hyoglossus (HG) and styloglossus (SG) on the parasagittal plane.

2. DYNAMIC CONTROL OF THE MODEL

A new control method for the articulatory model is developed based on an assumption that muscle activation patterns depend on the geometric distance between the current position and the articulatory target. To produce vowel sequences, this study employed four extrinsic muscles of GGp, HG, SG, and GGa for the tongue, and the jaw opener and closer muscles for the jaw in the current stage.

The key issue in developing a target-based control strategy is to determine a tongue position at an arbitrary moment in the geometrical space. To do so, the *tongue position* was defined by the average position of five midsagittal nodes on the tongue surface from the tip to the dorsum. When exciting a single tongue muscle by a unit activation for a certain duration, the tongue position moves from the initial position to a new position, and thus it forms a vector in the geometric space, which is referred to as a muscle vector for the muscle. Figure 3 (a) shows the tongue muscle vectors by thick gray arrows for four extrinsic muscles. These muscle vectors form a space, referred to as a muscle workspace, which can be directly mapped on the geometrical space. Suppose that the tongue is located in the current position *Pc* and moves forward to a target *Tg*, the dashed line from *Pc* to *Tg* forms a vector, named 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's length. At the *current computational step* shown in Fig. 3 (a), SG and HG are the active muscles. When the activation signals are computed at each computational step and drive the tongue to move to a new position, a trajectory of the tongue position, indicated by the thin gray arrow, and time-varying muscle activation patterns are obtained.

The dark line with v-arrows shows the trajectory of the tongue position in /iai/ sequence. Note that the tongue trajectory is a resultant path of the tongue movement superimposed on the jaw movement. Figure 3 (b) shows the activation signal patterns for the four extrinsic muscles and for the opener and closer groups of the jaw muscles. It is interesting to find that the muscle activation signals resemble the EMG signals observed in the physiological experiments [7,8]. The fact that the antagonistic muscles show a reciprocal pattern suggests that the proposed method based on articulatory targets can be used in place of the method using EMG signals. If the co-contraction between the agonist and antagonist muscles is not taken into account, the muscle activation signals can be uniquely obtained by a given target sequence. Therefore, this control strategy offers a practical way to drive a physiological model, even though it is not perfectly realistic from a physiological point of view. The activation signals for the opener and closer muscle groups are generated in the same way using a muscle workspace of the jaw muscles.



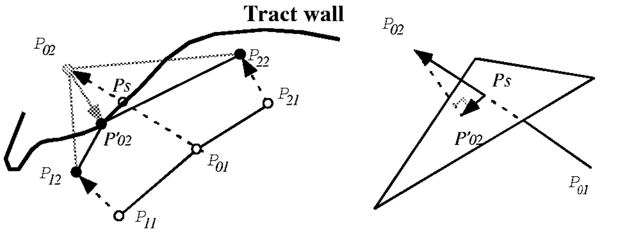
(a) (b)

Fig 3. Model control using an articulatory target sequence: (a) tongue trajectory in tongue muscle workspace, and (b) generated muscle activation signals. The signals for the jaw opener (JawOp) and closer (JawCl) groups were obtained in jaw muscle workspace.

3. CONTACT OF THE TONGUE AND THE TRACT WALL

In speech articulation, the tongue makes contact with the teeth, hard palate, and mandible. Therefore, the contact between the tongue and the outer wall of the vocal tract is one of the critical factors in achieving accurate and stable control of the tongue. Since the outer wall is too complex in shape to be represented by an analytic function, the collision of the tongue on the wall cannot be combined with the motion equations of the model. Alternatively, a method is proposed to compute tongue deformation when a rigid wall is introduced in the movement path of the nodes. Figure 4 shows a diagram for explaining this method. Suppose that the points P_{01}, P_{11}, and P_{21} represent the positions of three nodes m_0, m_1, and m_2 on the tongue surface at time 1. When a certain force is applied, these three nodes would move to P_{02}, P_{12}, and P_{22} at time 2 if there were no vocal tract wall. The dashed lines with an arrow show the pathways of the nodes. When the tract wall is introduced the pathway of the nodes, node m_0 hits the wall at Ps and receives a reaction force. Then, the node should arrive at an equilibrium position P'_{02}, where Ps is the intersection of the trajectory of m_0 and the wall.

The equilibrium position of node m_0 contacting the wall can be estimated by the forces on node m_0 at the collision with the outer wall. The total energy of m_0 at the collision can be approximately represented by the potential energy from Ps to P_{02}. The component of the momentum ($m_0 v_0$) of m_0 in a given direction is proportional to that of the vector from Ps to P_{02} in the same direction. When the node hits the outer wall, the reaction force of the wall on m_0 is opposite of the direction to the velocity v_0 and proportional to its amplitude. The tangential component of the force on m_0 can be reasonably represented by the projection of the vector Ps-P_{02} on the wall surface of a triangular plane that node m_0 hit. Assuming that the plane has homogeneous properties in all directions, displacement of P_{02} on the triangular plane in each direction is proportional to the component of the momentum, *i.e.*, the vector Ps-P_{02}. Therefore, the equilibrium position can be approximated using the projection of P_{02} on the triangular plane, shown by P'_{02} in Fig. 4 (b). This deformation changes the length of the springs connected with the node. For instance, the spring between P_{12} and P_{02} is shortened to the one between P_{12} and P'_{02} due to the collision. This effect induces additional forces for the adjacent nodes. These forces are calculated according to the length increments of the concerned springs in comparison with their length in the case without the outer wall, and are taken into account as an input for the next computation step.



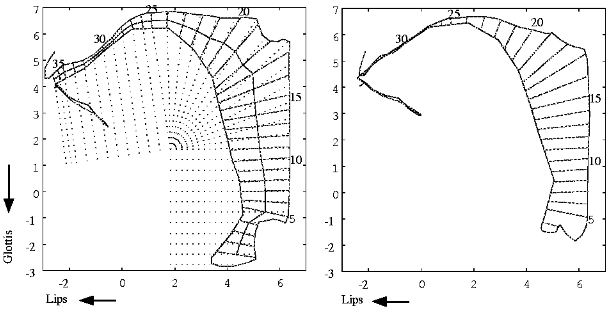
(a) (b)

Fig. 4 Collision of the surface nodes of the tongue and vocal tract wall: (a) deformation during collision and (b) an equilibrium position of the node on the wall.

4. SYNTHESIS OF VOWEL SEQUENCE

Based on the above method, muscle activation signals are generated according to a given articulatory target sequence, and the model is driven by the activation signals to produce a dynamic change of vocal tract shape. Since the present model does not provide a full 3D model of the vocal tract, realistic area function of the vocal tract has to be estimated using the information of the partial vocal tract of the model.

This study obtains vocal tract area function by two steps: determination of vocal tract width of the model and computation of the area function from the width information. Figure 5 shows an example of estimating area function for vowel /i/. A grid line system [9] shown by doted lines is implemented on the midsagittal plane to determine the vocal tract width. The central line of the vocal tract is defined by a curve passing through the mid-points of the grid line intersections with the outer and inner boundaries of the tract (see Fig. 5 (a)). The orientation of each cross-sectional plane is decided based on the central line by optimizing the angle of the cross-sectional planes and the central line so that there are no abrupt angle changes between the adjacent cross-sectional planes. The solid lines show the position and orientation of the obtained cross-sectional planes of the vocal tract, and the line length defines vocal tract width. The vocal tract width on the parasagittal planes is measured using the same cross-sectional planes. The result is shown in Fig. 5 (b). As shown in this figure, two sides of the tongue contact with the hard palate in the anterior portion for /i/, and the airway is constructed by the groove of the tongue in the midsagittal plane.



(a) Midsagittal plane

(b) Parasagittal plane

Fig. 5 Cross-sectional shapes of the vocal tract in the midsagittal plane (a), and the parasagittal planes (b). (The dimensions are in cm)

Area functions of the vocal tract are estimated using an improved α-β model [3] with vocal tract widths in the midsagittal and parasagittal planes.

$$A = \alpha_1 w_m^{\beta_1} + \gamma_1 w_m + \alpha_2 w_p^{\beta_2} + \gamma_2 w_p$$

where w_m is the vocal tract width in the midsagittal plane and w_p is the average value of the widths in two parasagittal planes. $\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2$ are a function of the distance from the glottis, and are determined by minimizing the difference between the estimated and MRI-based area functions for 5 Japanese vowels. Unlike the conventional methods, this estimation not only uses the width information of the midsagittal plane, but also employs the information from parasagittal planes. Estimated area functions are shown in Fig. 6 for vowel sequence /iai/. The thick lines show the area functions in stationary segments for vowels /i/ and /a/, while the thin lines illustrate area functions during the transition. Since the rate of movement of the tongue and jaw from one articulatory target to another is constrained by the physiological property of the model, the targets in a sequence can be reached for a long vowel duration, but they may not be if the duration is short. This behavior is similar to movement smoothing, or coarticulation, in human speech articulation. This feature of the model allows us to manipulate speech rate of the synthetic sound only by adjusting vowel duration without affecting sound quality in the transitional segment.

5.CONCLUSIONS

In this study, mass-points and viscoelastic springs were employed in modeling tongue tissue and rigid organs. This model demonstrates two major advantages over other physiological models based on the finite element method (FEM). First, the system consisting of mass-points and viscoelastic springs is stable at fast generating a large deformation of a soft tissue continuum. Second, the soft tissue and rigid organs can be integrated in the same motion equation system using mass-points and viscoelastic springs. This design reduces computation time greatly. The computing time of our model is about 50 times of real time using the Sun Workstation Ultra-30. The model shows some behaviors characteristic to human speech articulation, which are reflected by natural quality of the synthesized sounds in a vowel-to-vowel transition.

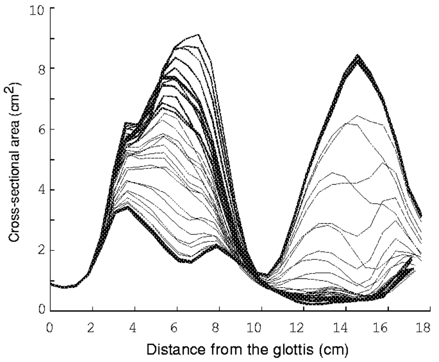


Fig. 6 Vocal tract area functions calculated from the physiological model for vowel sequence /iai/. Thin lines show the area functions during the transition between /i/ and /a/.

6.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] Hirai, H., Dang, J., and Honda K. (1995)" A physiological model of speech organs incorporating tongue-larynx interaction," J. Acoust. Soc. Jpn, 52, 12, 918-928. (in Japanese)

[4] Sakamoto, T. and Saito, Y. (1980). *Bionics and ME - From the basic to measurement control*, Tokyo Electric Machinery University.

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

[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] Baer, T., Alfonso, J., and Honda, K. (1988). “Electromyography of the tongue muscle during vowels in /epvp/ environment,” Ann. Bull. R. I. L. P., Univ. Tokyo, 7, 7-18.

[8] Dang, J. and Honda, K. (1997). "Correspondence between three-dimensional deformation and EMG signals of the tongue," Proc. of ASJ spring meeting, 241-242.

[9] Heinz J. and Steven K.•‘On the derivation of area functions and acoustic spectra from cineradiographic film of speech,” J. Acoust. Soc. Am. 36, 1037 (1964).