A COMPUTATIONAL MODEL FOR 2D ARTICULATION: SPEECH PRODUCTION WITH POTENTIAL USE IN RECOGNITION

Leo J. LEE[1], Jianwu DANG[1,2] and Li DENG[1]
[1]*Department of Electrical & Computer Engineering, University of Waterloo, Waterloo, ON, Canada,*
[2]*ATR Human Information Processing Research Labs, Kyoto, Japan*

ABSTRACT

This paper describes a 2D physiological model of the vocal tract to simulate the dynamic actions of speech production organs including the tongue, mandible, hyoid bone, and the outer wall of the vocal tract. This model has drastically simplified a previous 3D physiological model into an efficient computational form. The movement of the articulators is driven by the phone-dependent, slowly time-varying targets specified in the articulatory space. Appropriate muscle activation signals are automatically generated by comparing the current and target position of the articulators. The smooth trajectories of the articulators thus produced are converted to the time-varying vocal tract area function as the final output of the model. Finally, we discuss the potential use of this simplistic but efficient 2D production model in the state-space or segmental modeling approach to automatic speech recognition, where the articulatory model serves as an underlying hidden speech production state equation with realistic physiological constraints.

1. INTRODUCTION

The current work is based on a previously developed 3D articulatory model [1, 2, 3]. The motivation of this further simplification is not only to achieve real-time natural speech production, but also to use it as an efficient tool in the state-space modeling approach to automatic speech recognition.

2. MODEL CONSTRUCTION

The model consists of the tongue, mandible, hyoid bone, outer wall of the vocal tract and related muscles in two dimensions. The geometric parameters are extracted from volumetric magnetic resonance (MR) images of a Japanese male speaker. All soft tissue and rigid organs are modeled by mass-points and viscoelastic connections, where the stiffness is defined to be extremely large for rigid connections. This uniform modeling of the soft tissue and the rigid organs allows us to put them into the same motion equation system, and effectively shortens the computational time.

The tongue shape is extracted from the MR image of the midsagittal plane. It is divided into six sections with nearly equal intervals along the anterior-posterior direction and ten sections along the tongue surface. This structure roughly replicates the fiber orientation of the genioglossus (GG) muscle. Every node in the mass-spring net of the tongue model connects to all of its adjacent nodes, as shown in Figure 1, while extra connections are added according to the muscular and bony structures. The rigid organ structure is also extracted from the MR images. The upper incisor, hard palate, velum (soft palate), pharyngeal wall and larynx are modeled as the fixed rigid vocal tract wall, while the lower incisor and mandible are modeled as the movable rigid jaw wall. Due to the geometrical complexity of these surface walls, they are modeled by connected segments of broken lines, 39 segments for the tract wall, and 23 for the jaw wall. Among the key speech production organs, the lip and velum are not modeled physiologically at the current stage. They will be taken into account by a separate acoustic model to generate speech sounds.

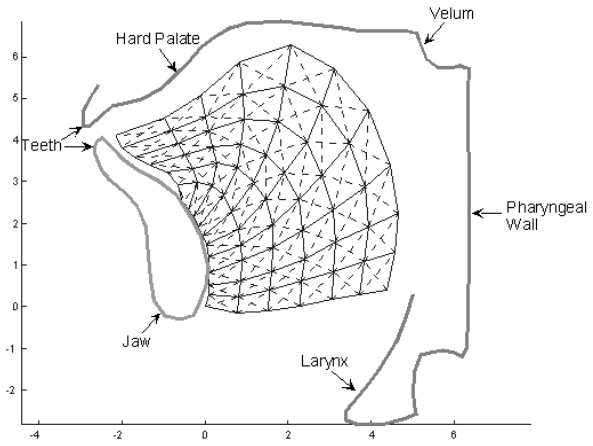

Figure 1. Initial position of the articulators. All dimensions are in cm.

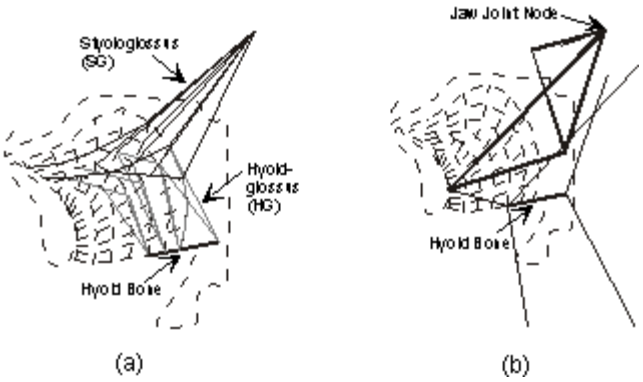

Figure 2. Examples of major muscle groups: (a) hyoglossus (HG) and styloglossus (SG); (b) jaw open muscles (mjawOp).

The anatomical arrangement of the major muscles is based on high-resolution MR images obtained from the same speaker with reference to the literature [4]. All the muscles have been projected to the midsagittal plane so that we can model them in 2D. They are functionally divided into twelve groups, ten of them are responsible for the tongue deformation, while the remaining two are responsible for the jaw movement. Tongue muscles are further divided into a number of branches to realize fast and smooth propagation of the muscle forces throughout the mass-spring net. Some examples of the muscle groups are given in figure 2.

3. COMPUTATIONAL FRAMEWORK

In this section, the computation method of simulating the 2D physiological model is discussed in some detail.

3.1. The Motion Equation

The motion equation governs the movement of the 2D articulatory model according to the physiological constraints. It can be readily derived from the Lagrange's equation in classical mechanics, which is:

$$\frac{\partial}{\partial t}\left(\frac{\partial T}{\partial \dot{q}_a}\right) - \frac{\partial T}{\partial q_a} = Q_a, \quad (1)$$

where T is the kinetic energy, q is the generalized coordinate and Q is the generalized force.

Applying Eq (1) to a single node i at time t and use the usual rectangular coordinate, its motion equation can be shown to be:

$$m_i \ddot{p}_i(t) + \sum_{c=1}^{cn} b_{ic}\dot{p}_{ic}(t) + \sum_{c=1}^{cn} k_{ic} dp_{ic}(t) = f_i(t), \quad (2)$$

where m_i is the mass of node i, $p_i(t) = (x_i, y_i)$ is the position of node i at time t, $p_{ic}(t) = p_i(t) - p_c(t)$ is the length of the connection between node i and node c, $dp_{ic}(t) = p_{ic}(t) - p_{ic}(0)$ is the length increment of the connection with respect to its initial length, cn is the total number of connections with node i, b_{ic} and k_{ic} are the viscous coefficient and stiffness of each connection, and $f_i(t)$ is the active component of the muscle force (see **3.2**) on node i.

In order to carry out the computation on a computer, we have to discretize the above motion equation. Assuming the time interval of discretization is h (*10 ms* in our implementation), the discretized motion equation is:

$$m_i p_i(t) + h\sum_{c=1}^{cn} b_{ic}\left[p_i(t) - p_c(t)\right] + h^2\sum_{c=1}^{cn} k_{ic}\left[p_i(t) - p_c(t)\right] = h^2 f_i(t) + r_i(t), \quad (3)$$

where

$$r_i(t) = m_i\left[2p_i(t-h) - p_i(t-2h)\right] + h\sum_{c=1}^{cn} b_{ic}\left[p_i(t-h) - p_c(t-h)\right] + h^2\sum_{c=1}^{cn} k_{ic}\left[p_i(0) - p_c(0)\right]$$

is the residue term determined by the quantities from the previous time steps and the initial conditions.

Generalizing the above expression to the entire mass-spring model, the final motion equation can be expressed in a matrix form:

$$\left(\boldsymbol{M} + h\boldsymbol{B} + h^2\boldsymbol{K}\right)\boldsymbol{X} = h^2\boldsymbol{F} + \boldsymbol{R}, \quad (4)$$

where $\boldsymbol{X}$ is a vector consisting of the coordinate values (x and y) of all the nodes, $\boldsymbol{M}$ is a diagonal matrix of the masses of all the nodes, $\boldsymbol{F}$ is a vector consisting of the active component of the muscle forces, $\boldsymbol{R}$ is the residue vector, $\boldsymbol{B}$ and $\boldsymbol{K}$ are matrices of viscous coefficient and stiffness with their elements satisfying the following matrix transformation rule:

$$b_{2i+u,2i+v} = \sum_{c=1}^{cn} b_{ic} r_{ic}(u) r_{ic}(v), \quad b_{2i+u,2c+v} = b_{2c+v,2i+u} = b_{ic} r_{ic}(u) r_{ic}(v)$$

$$k_{2i+u,2i+v} = \sum_{c=1}^{cn} k_{ic} r_{ic}(u) r_{ic}(v), \quad k_{2i+u,2c+v} = k_{2c+v,2i+u} = k_{ic} r_{ic}(u) r_{ic}(v)$$

where $u,v=0$ for x direction and $u,v=1$ for y direction, and $r_{ic}(u)$ is the direction cosine from node i to node c.

Although the magnitudes of b_{ic} and k_{ic} are constant, $\boldsymbol{B}$ and $\boldsymbol{K}$ are time-varying matrices because the direction cosines change over time. Therefore the $\boldsymbol{B}$ and $\boldsymbol{K}$ matrices have to be rebuilt at every time step.

3.2 The Muscle Forces

The driving force of this model is induced by muscle contractions. In our work, we adopt a commonly accepted assumption that the muscle force is the sum of passive component (independent of muscle activation) and active component (only present if the muscle is active). Based on the study of Morecki [5], the muscle force is calculated as follows in the current model:

$$\begin{cases} f_1 = k_1\Delta l + b_1\dot{l} \\ f_2 = k_2\Delta l, \quad when \;\; l > l_0, \quad k_2 = 0.002k_1 \\ f_3 = a\left(3.42e^4 + 0.53e^3 - 2.2e^2 + 0.3e + 0.13\right), \end{cases} \quad (5)$$

where $l = l - l_0$, $= l/l_0$. In the above equations, f_1 and f_2 are the passive components of the muscle force, and are taken into account when constructing the $\boldsymbol{K}$, $\boldsymbol{B}$ matrices; f_3 is the active component of the muscle force and is included in the force vector $\boldsymbol{F}$. is the gain coefficient of the muscle, which is determined by control signal.

3.3 Additional Constraints

Besides the motion equation (4), additional constraints have been introduced to make this 2D model more realistic.

3.3.1. Area Constraint for the Tongue Tissue. The soft tongue tissue is modeled as a network of mass-points and viscoelastic springs in the present model. With these connections alone, the incompressible property of the tongue tissue is not adequately represented. Area constraint is

introduced to reflect this property, by which we try to maintain the area of the entire tongue as well as each block (in quadrilateral shape) in the mass-spring net when the tongue deforms. We define the following control error e_A:

$$e_A = a_1 (S - S_0)^2 + a_2 \sum_{j=1}^{bn} (S_j - S_{j0})^2, \quad (6)$$

where S is the entire tongue area at time t, S_0 is the initial tongue area (*27.87 cm²* in our model), bn is the total number of blocks (*60* in our model) in the mass-spring net, S_j is the area of block j at time t, and S_{j0} is the initial area of block j. a_1 and a_2 are the weight factors of the area constraint (see **3.4** for more details). e_A is minimized at every step of the calculation to satisfy the area constraint.

3.3.2. Constraint for the Jaw Joint Node. The movement of the jaw is driven by a torque at the rotational center of the condyle. This rotational center is represented as a jaw joint node in the current 2D model. We assume that the movement of the jaw joint node follows a curved path corresponding to the concave articular groove of the temporomandibular joint [1], which can be approximated by a third-order polynomial:

$$y = y_0 - 2.5(x - x_0)^3 - 4(x - x_0)^2, \quad (7)$$

where (x, y) is the current position of the jaw joint node and (x_0, y_0) is its initial position. Again we define a control error e_J as:

$$e_J = a_3 \left[y - y_0 + 2.5(x - x_0)^3 + 4(x - x_0)^2 \right]^2, \quad (8)$$

and minimize it at every time step. The force at the jaw joint node is also redistributed along the curved path to drive the jaw movement.

3.4 Building the Final Computation Equation

In general, it is impossible to satisfy the motion equation and constraints simultaneously unless the articulators remain at their initial positions. A joint minimization scheme is adopted to construct the final computation equation. First we define the control error e_M for the motion equation:

$$e_M = \left| (\boldsymbol{M} + h\boldsymbol{B} + h^2\boldsymbol{K})\boldsymbol{X} - h^2\boldsymbol{F} - \boldsymbol{R} \right|^2. \quad (9)$$

Then we jointly minimize all the control errors by solving the following equation[1]:

$$\frac{\partial e_M}{\partial \boldsymbol{X}} + \frac{\partial e_A}{\partial \boldsymbol{X}} + \frac{\partial e_J}{\partial \boldsymbol{X}} = 0. \quad (10)$$

The first derivative term gives a set of linear terms, while the second and third derivative terms give a set of complicated higher order nonlinear terms. We simplify the calculation by linearize the nonlinear terms with respect to their values at the previous time step. Weight factors a_1, a_2 and a_3 are adjusted so that the total area change is less than *2* percent, the maximum block area change is less than *10* percent and the deviation of the jaw joint node movement from the curved path (e_J/a_3) is less than 10^{-4} at every time step. After the linearization, the final computation equation becomes a set of linear equations (about 140 for the current 2D model), which can be written in the following general form:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}. \quad (11)$$

4. MODEL SIMULATION

In the previous 3D model, the computation equation is solved by a simple iterative method such as Gauss-Seidel, where matrix $\boldsymbol{A}$ can be justified to be diagonal dominant. However, this condition fails for some bone nodes in the 2D model, and experiments show that the Gauss-Seidel algorithm is not applicable any more. Due to the very large stiffness of the bony connections compared with the tissue and muscle connections, the resulting $\boldsymbol{A}$ matrix is quite ill-conditioned (condition number ~ 10^9), so that extra care has to be taken when applying direct solving algorithms. The LU decomposition method with double precision seems to be able to provide a fast and accurate solution to the computation equation so far.

The first step of model simulation is to study the effect of individual muscle groups on the articulators. Figure 3 shows the effect of the contraction of four major tongue muscle groups, which is obtained by exciting each muscle group with unit activation for a fixed duration (*200 ms* in this case) from the initial tongue position.

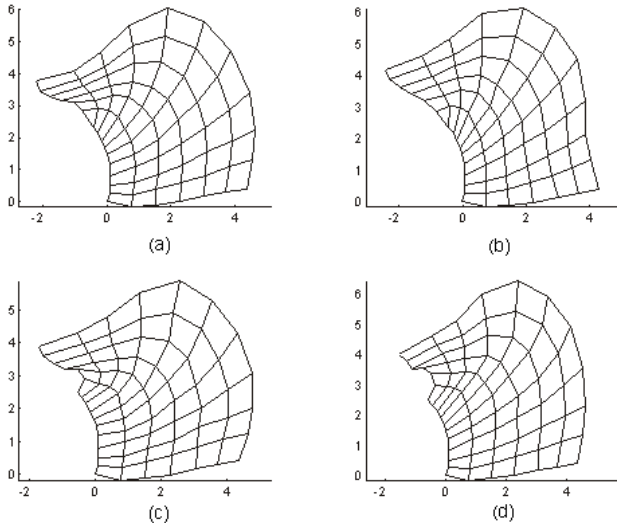

Figure 3. Effects of major tongue muscle contractions: (a) genioglossus anterior (GGa); (b) genioglossus posterior (GGp); (c) SG; (d) HG.

[1] It is also possible to build the computation equation based on the Lagrange multiplier method, but the resulting form will be computationally more expensive.

5. DISCUSSIONS

5.1. Control Strategy

The control strategy of this 2D physiological model is based on the targets specified in the articulatory space, and similar to the one used in the 3D model [3]. In order to simulate an arbitrary phoneme, we have to control some key points of the articulators to form a specific vocal tract shape. We propose a multi-point control strategy, which includes the tongue tip, tongue dorsum and the jaw, to fulfill this requirement. The *tongue tip* is defined to be the front-most tongue node when the tongue is at its initial position, the *tongue dorsum* is represented as a weighted average of three tongue nodes around the highest point of the tongue model, and the *jaw* is defined to be the position of the mandible incisor.

First, we build a muscle workspace on each control point by examining the effect of each muscle group on it, as shown in figure 3. In order to drive the control points towards the target positions, their current positions are compared with the target positions and the difference vectors are projected on the muscle workspace to activate muscle groups with a positive projection at every time step. It remains as a main challenge to drive the control points to move towards some crucial positions quickly and accurately in order to form some critical constrictions. We will solve this problem along the following two paths: (1) Calculate the momentum of the articulators at every time step and add an appropriate control to offset their inertia when they reach the critical targets; (2) Increase the viscous component significantly when the articulators are at or close enough to the targets so that they can stay there in balance. These two approaches will be further justified by comprehensive model simulations.

5.2. Potential Application in Speech Recognition

Here we briefly discuss the potential use of the present computationally efficient speech production model in automatic speech recognition. The current speech recognition technology is mainly based on accounting for acoustic variability of speech via the use of a large number of model parameters. This has intrinsic disadvantages because the nonlinear process from articulatory domain to acoustic domain turns a low-pass signal (articulator's motion represented by Eq(4)) to a wide frequency-band one, and the temporal smoothness in the articulatory domain no longer exists in the acoustic domain. This makes it difficult to directly and efficiently model speech coarticulation in the acoustic domain. We advocate the use of the temporal smoothness property of the articulatory process as the internal hidden process of the speech model in speech recognizers. Since the smoothness as informed by the articulator motion equation (4) applies not only to segments within-phones but also across phones, it permits economical and effective descriptions of speech coarticulation that is impossible to do by using the acoustic-domain information only [6].

The key implementation issue includes development of a mathematical framework in which the articulatory process is modeled as the parametric hidden (unobservable) process that provides natural information flow from one phoneme to another. In the state-space modeling framework we are currently developing, the 2D articulatory model described in this paper is used as the dynamic state equation whose system matrix is initialized from the physiological parameters permitting realistic synthesis of speech sounds. In the deployment of the recognizer, these parameters are on-line adapted according to Bayesian methodology using a small amount of adaptation data from the acoustic domain only.

6. CONCLUSIONS

A computationally efficient 2D physiological model has been developed, which is capable of generating almost all the phones in the acoustic domain. Its potential application in automatic speech recognition has been briefly discussed. Comprehensive simulations of this model to generate realistic speech are underway, and algorithms are under development to use this model in automatic speech recognition.

ACKNOWLEDGMENTS

The authors would like to thank NSERC Canada for the funding of this project. Leo Lee would like to thank the Ontario graduate scholarship (OGS) for providing the financial support of his graduate study at the University of Waterloo.

REFERENCES

[1] J. Dang and K. Honda, "A Physiological Model of a Dynamic Vocal Tract for Speech Production," *Tech. Report of ATR*, TR-H-247, 1998.

[2] J. Dang and K. Honda, "Speech Production of Vowel Sequences Using a Physiological Articulatory Model," *Proc. ICSLP98*, vol. 5, pp. 1767-1770, 1998.

[3] J. Dang and K. Honda, "Speech Synthesis of VCV Sequences Using a Physiological Articulatory Model," *137th ASA Meeting* (Berlin, Germany), March 1999.

[4] K. Miyawaki, "A Study of the Musculature of the human Tongue," *Ann. Bull. Res. Inst. Logoped. Phoniatrics*, Univ. Tokyo, vol. 8, pp. 23-50, 1974.

[5] A. Morecki, *Biomechanics of Engineering – Modeling, Simulation, Control*, Spring-Verlag, New York, 1987.

[6] L. Deng, "A Dynamic, Feature-Based Approach to the Interface between Phonology and Phonetics for Speech Modeling and Recognition," *Speech Communication*, vol. 24, no. 4, pp. 299-323, 1998.