



Estimation of vocal tract shapes from speech sounds with a physiological articulatory model

Jianwu Dang*

ATR Information Science Division, Kyoto, Japan and Information School, Japan Advanced Information Science and Technology Institute, Ishikawa, Japan

Kiyoshi Honda

ATR Information Science Division, Kyoto, Japan

Received 9th November 2001, and accepted 3rd January 2002

A 3D physiological articulatory model constructed based on volumetric MRI data from a male speaker was used to estimate vocal tract shapes from speech sounds. The advantages of using the model for the inverse estimation are that the model is equipped with the morphological and dynamic constraints that are commonly used for such estimation and possesses the physiological constraints that are involved in human articulation. In this study, a dynamic muscle workspace was introduced to account for temporal variations of the muscle orientation with articulatory movements, and a multipoint control strategy was proposed for flexible control of the tongue tip and tongue dorsum. The control points were used as articulatory parameters, and formants were chosen as acoustic parameters. An articulatory constraint between the F1–F2 difference and tongue dorsum position was introduced in mapping formant patterns to control point positions, where the constraint was obtained based on X-ray microbeam data recorded from the target speaker and five other male speakers. The proposed estimation method was evaluated using vowel-to-vowel sequences. For the target speaker of the model, the average estimation error was 0.16 cm for the vocal tract shapes, and 1.8% for the four lower formants. This implies that our physiological articulatory model can be a valuable tool for the inverse estimation.

© 2002 Academic Press

1. Introduction

The estimation of vocal tract configurations from speech sounds is a longstanding issue in speech research. One of the difficulties in this estimation is that the inverse

*Address correspondence to J. Dang, ATR Information Service Division, Kyoto, Japan. E-mail: jdang@atr.co.jp

transformation faces the problem of *one-to-many nonlinear* mapping (Atal, Chang, Mathews & Tukey, 1978). The one-to-many characteristic arises from the fact that a given segment of speech signal can be generated by an infinite number of vocal tract configurations. The nonlinear characteristic is inherent in the process of speech generation. Its degree of complexity, however, depends on the parameters chosen to represent both articulatory and acoustic spaces. To solve the inverse problem, a number of constraints were introduced in previous studies. The constraints can be roughly classified as morphological (spatial) and dynamic (temporal) constraints. The former is a static constraint, which ensures that the estimated vocal tract configuration is a reasonable vocal tract shape. The latter involves the continuity of vocal tract motion and the effects of coarticulation, such as anticipation and retention.

Over the years, many studies have focused on this issue, and tried to find a better solution for the inverse problem. Schroeder (1967) determined the geometry (log-area) of the human vocal tract by acoustic measurements: formants and acoustic impedance at the lips. Yehia & Itakura (1996) used a similar approach to describe an acoustic-to-geometry mapping. They incorporated a morphological constraint in their estimation while the dynamics of the articulation was not treated. Atal *et al.* (1978) established tables of vocal tract shapes and related them to acoustic representations, in which the morphological constraint was included in the data set. They used the lowest three (log) formant frequencies and the amplitudes of the vocal tract transfer function to obtain log-areas with 20 sections. Schroeter & Sondhi (1994) used an articulatory codebook to estimate vocal tract shapes from speech signals, where the constraint of the vocal tract dynamics was materialized by applying a dynamic programming method to the articulatory codebook. The constraints used in the above methods are based on minimum effort or minimum energy.

In contrast to the above estimations, Shirai & Honda (1978) employed a statistical articulatory model to estimate articulatory parameters from speech waves. They related acoustic parameters to articulatory movements of the speech organs. Okadome, Suzuki & Honda (2000) retrieved articulatory movements from acoustics with phonemic information based on an articulatory-acoustic codebook, which consisted of flesh-point articulatory movements and speech acoustics. Dusan & Deng (1998, 2000) used an analytic method to retrieve vocal tract shapes and compared the estimation with observed data, where they used a statistical articulatory model proposed by Maeda (1990). Among the previous approaches, the method employing articulatory-acoustic codebooks (Okadome *et al.*, 2000) seems advantageous with respect to accuracy because the codebook is equipped with realistic morphological information, although the electromagnetic articulatory (EMA) data offer only partial information, being a constraint to the anterior portion of the vocal tract. To solve the one-to-many problem, minimum energy or minimum acceleration has been employed in the optimal mapping (Schroeter & Sondhi, 1994; Okadome *et al.*, 2000). However, this method lacks flexibility in recovering vocal tract shapes because it is limited by the codebook, which is usually based on data only from a few speakers. Another problem is that this method requires a large database and a great amount of time to search through the database.

Among the articulatory models mentioned above, the one used in the Shirai & Honda study (1978) was a statistical model with five parameters, which was based on an analysis of X-ray data obtained from five Japanese male subjects. Maeda's

model used in the Susan and Deng studies was derived from an X-ray film of a French female subject (Maeda, 1990). The model had eight parameters, and demonstrated more flexibility than the former one because the glottal height, lip protrusion, and velum motions could be controlled. Strictly speaking, however, both of the models work according to a statistical rule but not a physiological law. Therefore, they cannot always reflect physiological constraints correctly, especially for dynamic movements. For a realistic solution of the problem, it is reasonable to consider a physiological articulatory model that faithfully realizes the mechanism of human articulation.

With this in mind, the present study applies a three-dimensional (3D) physiological articulatory model that has been developed for human-mimetic speech synthesis. The geometrical outline and muscular configuration of the articulatory model were based on volumetric MRI data obtained from a male speaker. This model is equipped with physiological constraints resembling human speech organs, and is driven by muscle contraction force according to a target-dependent activation pattern. In this study, the articulatory model is employed to estimate vocal tract shapes from speech sounds. For the readers to understand how the physiological constraints have been equipped with the articulatory model and how the model works, this paper first gives some details about the construction of the articulatory model and the development of the control strategy for the model, and then focuses on the estimation of vocal tract shapes from speech sounds using the model.

2. Construction of articulatory model

A 3D physiological articulatory model has been constructed based on volumetric MRI data obtained from a Japanese male speaker. The earlier version of the model was described in detail in Dang & Honda (1998). The current model consists of physiological parts of the tongue and jaw, kinematical parts of the lips and velum, and a rigid vocal tract wall.

2.1. Configuration of the model

During natural speech, the tongue forms lateral airways by narrowing the tongue blade, or makes a midsagittal conduit by grooving with bilateral tongue–palate contact, as seen in some consonants and the vowel /i/. A realistic model for the tongue should be capable of forming such a midsagittal conduit and side airways, which are essential behaviors for the tongue in speech production. Considering a trade-off between computational cost and model similarity, we have so far constructed a partial 3D model with a 2-cm-thick sagittal layer, instead of a full 3D model.

The tongue model is a partial sagittal representation of a volumetric MR image of the tongue. The sagittal outlines of the tongue were extracted from the midsagittal plane and parasagittal planes 1 cm from the midsagittal plane. The mesh segmentation of the tongue tissue roughly replicates the fiber orientation of the genioglossus muscle, the largest muscle in the tongue. The tongue outline in each plane is divided into 10 radial sections that fan out from the genioglossus'

attachment on the jaw to the tongue surface. In the perpendicular direction, the tongue tissue is divided into six sections concentrically. A 3D mesh model is constructed by connecting the section nodes in the midsagittal plane to the corresponding nodes in the left and right planes. Thus, the model represents the central part of the tongue by a 2-cm-thick layer bounded with three sagittal planes. Fig. 1 shows the initial shape of the tongue model based on the segmentation with the surrounding organs, in which the initial shape replicated the Japanese vowel [e].

To form a 3D model of a deformable tongue, the tongue tissue consists of 120 eight-cornered brick elements (meshes) of elastic bodies. Governing equations were derived based on the finite element method (FEM) to calculate the deformation and movement of the tongue (Dang & Honda, 2001). In the equations, both the damping matrix and stiffness matrix are sparse matrices. An element e_{ij} of the matrices is nonzero only if nodal point i is adjacent to nodal point j . To materialize the matrices in a brief formation, the viscous and stiffness components of the nonzero elements are represented by a viscoelastic spring that connects the adjacent node pair. After connecting all the adjacent nodal point pairs, each mesh consists of 28 viscoelastic springs, in which 12 springs are located at the edges, 12 springs lie on the surfaces, and four springs diagonally connect the opposing vertices inside the mesh. As a result, the FEM-based governing equations of the tongue tissue are represented as a mass-spring network by means of an appropriate simplification

To generate a vocal tract shape, the articulatory model must include the tongue, lips, teeth, hard palate, soft palate (the velum), pharyngeal wall, and larynx. In the present stage, the lips and the velum are not modeled physiologically. The lips are defined by a short tube with a length and cross-sectional area, and the movement of the velum is defined by an opening area of the nasopharyngeal port. They are accounted for in the speech synthesis stage but not in the articulatory movement. Outlines of the vocal tract wall and the mandibular symphysis were extracted from MR images in the midsagittal and parasagittal planes (0.7 and 1.4 cm from the midsagittal plane on the right side). Assuming that the left and right sides are symmetrical, 3D surface models of the vocal tract wall and the mandibular symphysis were constructed using mesh outlines with 0.7-cm intervals in the left-right direction. Fig. 1 shows the model configuration of the vocal tract.

2.2. Arrangement of the tongue muscles

To construct a subject-specific model, the anatomical arrangement of the major tongue muscles were examined based on a set of high-resolution MR images obtained from the target speaker (Dang & Honda, 2001). The boundaries of the muscles were first traced in each slice of the MR images, and then superimposed together so that the major muscles could be identified. Thus, the genioglossus (GG) and the geniohyoid (GH) were identified in the midsagittal plane, and the hyoglossus (HG) and the styloglossus (SG) were mainly found in parasagittal planes. The superior longitudinal (SL) and inferior longitudinal (IL) muscles were seen in both the midsagittal and parasagittal planes. The other intrinsic muscles (transverse and vertical) were not identifiable in the MR images. The orientation of

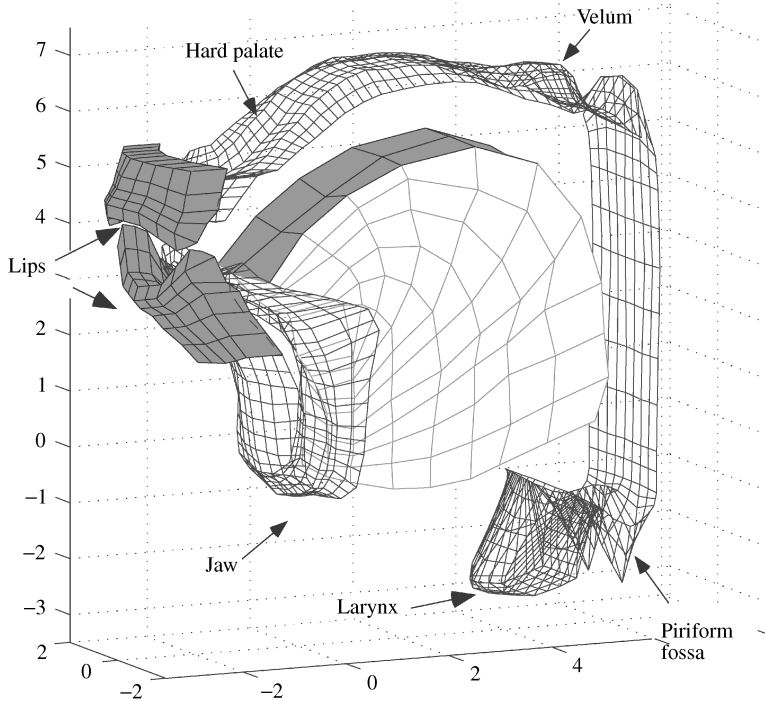


Figure 1. Configuration of the physiological articulatory model.

the tongue muscles was also examined with reference to the literature (Miyawaki, 1974; Warfel, 1993; Takemoto, 2000).

Fig. 2 shows the arrangement of the tongue muscles used in the proposed model. Fig. 2(a) shows the GG, which runs midsagittally in the central part of the tongue. Since the GG is a triangular muscle and different parts of the muscle exert different effects on tongue deformation, it can be functionally separated into three segments: the anterior portion (GGa) indicated by the dashed lines, the middle portion (GGm) shown by the gray lines, and the posterior portion (GGp) indicated by the dark lines. The line thickness represents the approximate size of the muscle units, and the thicker the line, the larger the maximum force generated. Figs 2(b) and (c) show the arrangement of other extrinsic muscles, the HG and SG, in the parasagittal plane, where the thickest line represents the hyoid bone. In addition, two tongue-floor muscles, the geniohyoid and mylohyoid, are also shown in the parasagittal planes. The top points of the mylohyoid muscle bundles are attached to the medial surface of the mandibular body. All the muscles are designed symmetrically on the left and right sides. Fig. 2(d) shows the structure of the vertical muscle in a cross-sectional view sliced at the fifth section from the tongue floor, shown in Fig. 2(e). The left panels of the figure show three intrinsic muscles. The transverse muscle runs in the left-right direction, and its location is plotted in the midsagittal plane with star markers. Altogether, 11 muscles are included in the tongue model.

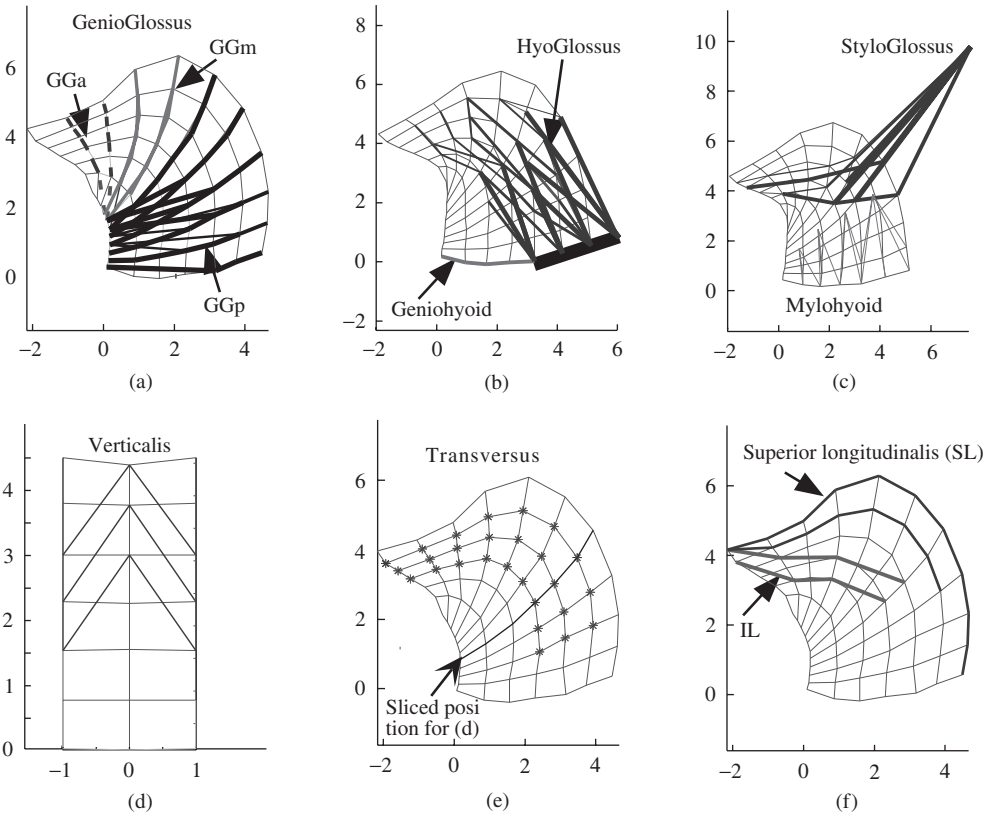


Figure 2. Arrangement of the tongue muscles in the midsagittal and/or parasagittal planes. (b) and (c) show the arrangement in a parasagittal plane and the others for the midsagittal plane (dimensions in cm).

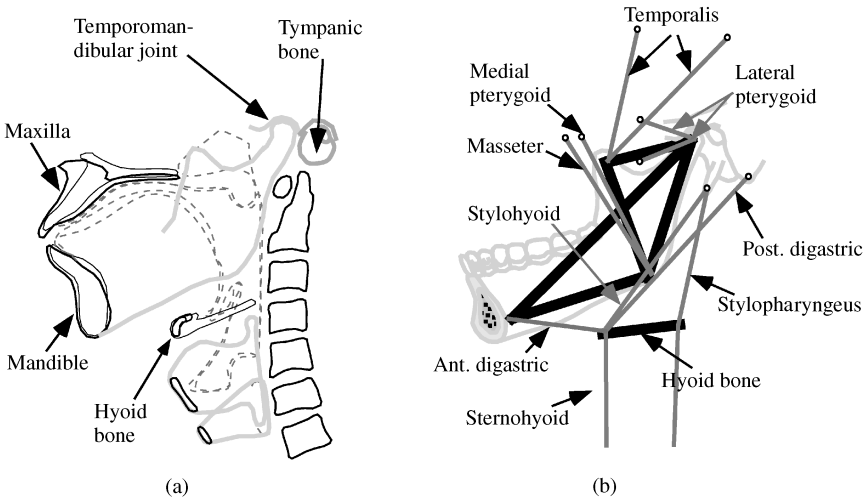


Figure 3. Modeling of the rigid organs based on MR images: (a) extracted framework of bony organs, and (b) model of the mandible and hyoid bone with related muscles.

2.3. Modeling of the rigid organs

The outlines of the rigid organs (the jaw and hyoid bone in the present work) were also traced from the MRI data for the target subject. Although the bony organs were not visualized in the MR images due to the lack of water, the contour of the organs can be identified in the images when soft tissues surround them. Fig. 3(a) shows the bony framework extracted from the midsagittal and parasagittal planes with a 0.7-cm interval. The gray thick lines show the contours of the organs drawn with reference to the anatomical literature, and the dashed lines are the boundaries of the soft tissue. Outlines of the rigid organs traced on the midsagittal plane are shown in the thick dark lines, and the thin lines for those on the parasagittal plane.

Fig. 3(b) shows the model of the jaw–hyoid bone complex. The right half of the mandible is drawn in the background using pale gray lines. The model of the jaw has four mass-points on each side, which are connected by five rigid beams (thick lines) to form two triangles with a shearing beam. These four points, which are similar to those chosen by Laboissière *et al.* (1996), are selected as the attachment points for the jaw muscles. This jaw model is combined with the tongue model at the mandibular symphysis. The model of the hyoid bone has three segments corresponding to the body and bilateral greater horns. Each segment is modeled by two mass-points connected with a rigid beam for each side. Yamazaki (1933) investigated the weight of the cranium and mandible using 92 dry skulls from Japanese specimens. His result showed that the weight of the male jaws was around 90 g. According to this literature, the equivalent mass of the living jaw is roughly estimated to be 150 g including water and the surrounding tissue. To evaluate the mass for the hyoid bone, the structure of the hyoid bone was extracted and measured using volumetric computer topographic data. The volume of the hyoid bone was about 2.5cm^3 for a male subject. Based on this measurement, an equivalent mass was set at 5 g for the hyoid bone. Note that both of the masses are much smaller than those used by Sanguineti *et al.* (1998). The masses are equally distributed in the body nodes. In the present model, rigid beams are also treated as viscoelastic links so that they can be integrated with the soft tissue in the motion equation. Their values are about ten thousand times greater than those used for the soft tissue.

The eight muscles indicated by thin lines in Fig. 3(b) are incorporated in the model of the jaw–hyoid bone complex, where the structure of the muscles is based on the anatomical literature (Warfel, 1993). The small circles indicate the fixed attachment points of the muscles. Since the other rigid organs below the hyoid bone, such as the thyroid and cricoid cartilages, are not included in the present model, two viscoelastic springs are used as the strap muscles. The temporalis and lateral pterygoid are modeled as two units to represent their fan-like fiber orientation. The digastric muscle has two bellies, named the anterior and posterior digastric, and are modeled to connect the hyoid bone at a fixed point. All of these muscles are modeled symmetrically left and right. Jaw movements in the sagittal plane involve a combination of rotation (change in orientation) and translation (change in position). Although there is no one-to-one mapping between muscle actions and kinematical degrees of freedom, the muscles involved in the jaw movements during speech can be roughly separated into two groups: the jaw closer group and opener group.

3. Dynamic control of the model

Development of control strategies is a key issue for a physiological articulatory model to simulate the dynamic physiological processes of speech production. So far, there are two strategies used in controlling a physiological model: the use of electromyographic (EMG) patterns, and equilibrium point hypothesis (EPH). The EPH is basically a plausible approach for controlling the dynamic movement of the tongue and other articulators, since the articulatory system consists of a muscle network and floating rigid bodies. However, the most commonly used version of this theory requires the length parameter and firing information of the muscles, which are difficult to obtain empirically. On the other hand, the observation of EMG signals is limited to only a few large muscles such as the extrinsic tongue muscles. For the above reasons, it is difficult to implement either the EPH approach or EMG signals to control a physiological model to produce an arbitrary articulation.

We have developed a target-based control strategy for our physiological model to simulate an arbitrary articulation (Dang & Honda, 2001). The idea of the target-based control strategy is to find an observable and controllable parameter, and develop an efficient approach to enable mapping between the parameter and muscle activation pattern (MAP). Articulatory targets of the speech organs are chosen as the parameters because they can be obtained and confirmed by experimental approaches such as MRI and X-ray microbeam methods. The MAP and articulatory targets are linked by a muscle workspace for a control point on each articulator. In this study, we further improve the control strategy to achieve a more accurate control.

3.1. Construction of a dynamic muscle workspace

During speech, the tongue tip usually has more freedom than other parts of the tongue and jaw. To simulate this function, therefore, it requires controlling different parts of the tongue and the jaw with different degrees of freedom. In this study, three control points are used to produce articulatory movements: the tongue tip, tongue dorsum, and jaw. The control point for the tongue tip is the apex of the tongue in the midsagittal plane. The control point for the dorsum is the weighted average position of the highest three points in the initial configuration in the midsagittal plane. The control point for the jaw is 0.5cm inferior to the tip of the mandible incisor.

In the multipoint control strategy, a muscle workspace is constructed for each control point. Each muscle vector in the muscle workspace corresponds to a displacement of the control point when the muscle contracts. During speech, both muscle orientations and the original position of the muscle workspaces vary with the jaw and tongue movements. Therefore, the muscle force vectors in the workspace should be adjusted according to changes in the muscle orientation. To account for the variation in the muscle orientation, this study constructs a set of muscle workspaces, referred to as *typical muscle workspaces*, whose origin is chosen to cover the articulatory space. Thus, a dynamic muscle workspace corresponding to an arbitrary position can be interpolated based on the typical workspaces for a control point.

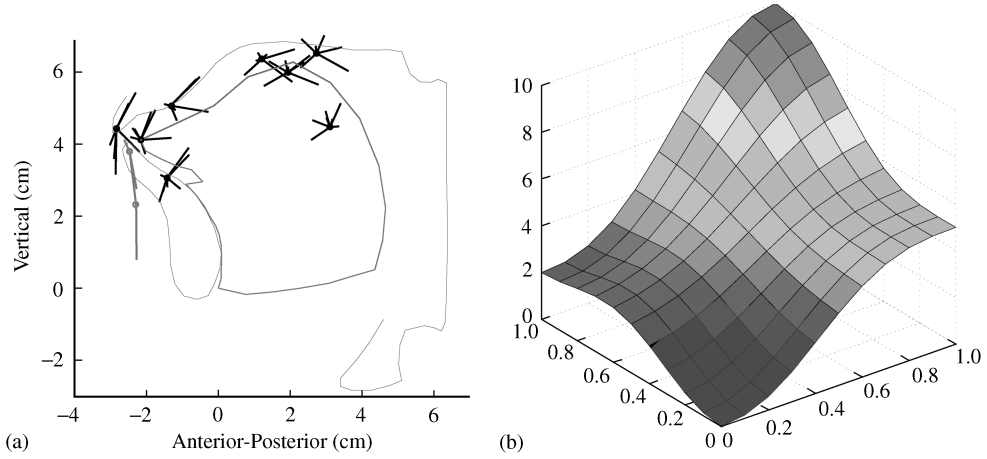


Figure 4. Typical muscle workspaces for three control points (a). Four muscle workspaces were built for tongue tip (dark lines) and tongue dorsum (light lines), and two for the jaw (light lines). (b) An example of the interpolation surface using four points.

Fig. 4(a) shows the muscle workspaces for the three control points. Four typical workspaces are constructed for the control points of the tongue tip and tongue dorsum, respectively, and two workspaces for the jaw. The dark lines indicate the workspaces for the tongue tip, and light lines around the tongue dorsum indicate the workspaces for the dorsum control point. The central typical workspace of the tongue tip and dorsum corresponds to the rest position, while the others roughly correspond to the positions of the three extreme vowels /a/, /i/, and /u/. The distribution of the typical workspaces is designed to cover the articulatory space of both vowels and consonants. For the jaw, the origin of the typical workspaces is chosen in the rest position and a wide-open position. To construct a typical workspace for a control point, we first input a certain activation pattern into the muscles to drive the control points to a specific position, and then excite each muscle using unit activation with a fixed duration individually. For the given excitation, each control point moves from the specific position to a new position. This displacement forms a vector in the geometric space, referred to as the muscle vector for a control point. Iterating the procedure for all of the specific positions, the typical muscle workspaces are established for each control point. For an arbitrary position of a control point, a dynamic muscle workspace is interpolated for the control point based on the typical workspaces. This interpolation is carried out using the following formulas

$$V = \frac{\sum_{i=1}^n L_i v_i}{\sum_{i=1}^n L_i}; \quad L_i = \prod_{\substack{j=1 \\ j \neq i}}^n l_j^2 \quad (1)$$

where V denotes the muscle vector of the dynamic muscle workspace, v_i is the muscle vector in the basic workspace i , and l_j is the distance from the current

position to the origin of typical workspace j . The number of typical workspaces n is four for the tongue, and two for the jaw. Fig. 4(b) shows an example of the interpolated surface from four points using Equation (1). It demonstrates that the interpolation has a quadratic surface with a relatively flat characteristic surrounding the reference points. Because of the flat property around each original point, interpolation is acceptable in cases where some targets are a little bit out of the region covered by the workspaces.

3.2. Generation of muscle activation signals

Since the muscle workspace is compatible with the geometrical space, the mapping of the control point between the geometrical space and the muscle workspace is straightforward. If a control point moves in an arbitrary direction, its displacement can be decomposed into several components parallel to the muscle force vectors. The amplitude of the vector component reflects how much the contraction of the muscle contributes to the displacement of the control point. Thus, an articulatory movement is related to a set of muscle contraction forces. Therefore, muscle activation signals can be obtained for any arbitrary movement using this approach. Fig. 5 shows an example of generating muscle activation signals according to a given target in a simplified muscle workspace of the tongue dorsum control point. The muscle workspace in Fig. 5(a) consists of muscle vectors of the four extrinsic tongue muscles, shown by the thick dark arrows. P_c indicates the current position of the control point and T_g is the target position. The dashed line from P_c to T_g 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.

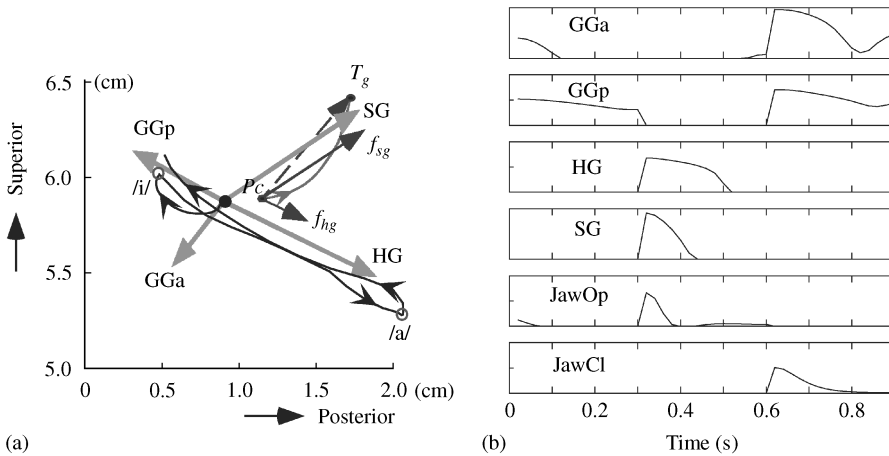


Figure 5. Generation of muscle activation signals using the target-based control strategy. (a) An example of the target-based control procedure in a simplified muscle workspace of the tongue dorsum, and (b) generated muscle activation signals for the vowel sequence /iai/, where the signals of jaw opener and closer groups were obtained similarly from the jaw muscle workspace.

Supposing that the projection of the articulatory vector for muscle vector v_i is $\alpha_i v_i$ and the projection of the optimal vector for the control point to move towards the target is $\beta_i v_i$, a performance function is defined as the summation of the squared difference of the vector components between the articulatory and optimal vectors,

$$g(\beta) = \sum_i (\beta_i - \alpha_i)^2 \quad (2)$$

Here, we introduce a penalty function

$$P(\beta) = \begin{cases} 0, & \beta \geq 0 \\ \beta^2, & \beta < 0 \end{cases} \quad (3)$$

and construct a cost function

$$F(\beta, A) = g(\beta) + A \sum_i P(\beta_i) \quad (4)$$

where A is an extremely large positive number. The component β_i of the optimal vector is solved by minimizing the cost function. As a result, the activation signal for the optimal vector is $\beta_i \approx u(\alpha_i) \alpha_i$ for muscle i , where $u(\alpha_i)$ is the unit step function, 1 for $\alpha_i > 0$, and 0 for the else. This means that the positive projections alone contribute to the optimal movement, and the negative ones can be ignored. From a physiological point of view, the result implies that a muscle will be excited only when the displacement caused by the muscle contraction has a direction consistent with that of the movement towards the target.

Accordingly, the SG and HG are the active muscles at the current computational step shown in Fig. 5(a). As the activation signals are computed at each step and fed to the muscles, the control point is driven to approach its target. The thin gray path shows the resultant trajectory moving towards the target, where the gray arrow indicates the optimal vector at the current step. Fig. 5(a) also demonstrates the trajectory of the tongue dorsum in simulating the vowel sequence /iai/, where the arrows indicate the direction of the trajectory. The corresponding activation signals are shown in Fig. 5(b) for the four tongue muscles and two jaw muscle groups. The activation signals for the jaw opener and closer muscle groups were generated using the same approach in the muscle workspace of the jaw. It is not surprising to find that the patterns of the muscle activation signals are similar to those of the EMG signals observed in physiological experiments (Baer *et al.*, 1988; Dang & Honda, 1997). That is, the GGp and GGa are active during the production of /i/, while the HG and SG are the major active muscles for /a/. The fact that the antagonistic muscles show a reciprocal pattern suggests that the proposed control method can be reasonably used in place of the method using EMG signals.

It is known that the extrinsic tongue muscles and jaw muscles reign the whole tongue body including the tongue dorsum and tongue tip. To insure more freedom for the tongue tip and tongue dorsum in the model, a weight coefficient was defined for each muscle at a specific control point. Large coefficients were defined for the extrinsic muscles to control the tongue dorsum while the intrinsic muscles had large coefficients to control the tongue tip. The weight coefficients were determined by model simulation. Table I shows the weight coefficients of the muscles for each control point.

TABLE I. 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
Vertical	1	0	0
Transverse	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

4. Method for estimating vocal tract shapes from vowel sounds

There were two major approaches used in estimating vocal tract shapes in the previous studies: estimating area functions using an idealized acoustic tube, and deriving articulatory parameters using a statistical articulatory model. To solve the one-to-many problem, morphological and dynamic constraints have been incorporated in this estimation. However, the reliability and plausibility of the estimation are in question because the above approaches did not account for the physiological constraints of human speech organs. To obtain a more reliable and accurate estimation, we propose an estimation method that employs the physiological articulatory model.

4.1. Parameter selection for the estimation

Nine articulatory parameters were selected for the estimation of vocal tract shapes from vowel sounds. They are represented as a vector,

$$X = (J_x, J_y, T_x, T_y, D_x, D_y, L_a, L_l, G_h)^T \quad (5)$$

The first three element pairs in the vector are the x - y values of the jaw, tongue tip, and tongue dorsum, respectively. They correspond to the three control points of the model described in the preceding section. These variables are involved in the articulatory model to calculate articulatory movements. The last three parameters are the cross-sectional area and length of the lip tube, and the height of the glottis. These three parameters are used to complete a vocal tract shape for generating speech sounds, but not involved in the calculation of articulatory movements. In this study, the velum was not used as an articulatory parameter because the nasal sounds were not dealt with.

For speech sounds, several kinds of parameters can serve to describe their acoustic characteristics. Among them, the formant pattern of the first formant (F1)

and second formant (F2), namely the vowel space, is known to be compatible with the articulatory space (Stevens & House, 1955; Moore, 1992; Stevens, 1999). According to this relation, it is easy to find a mapping between formant patterns and the control points of the articulatory model. Therefore, formant frequencies of speech sounds were chosen as acoustic parameters in this estimation, which were obtained using the LPC-Cepstrum method. The acoustic vector consists of five parameters: the four lower formants, and the difference between F1 and F2. Although the F1–F2 difference is not completely independent of the other parameters, as will be seen in a latter section, the difference can be used to derive an articulatory constraint for model control directly.

4.2. Acoustical model for speech synthesis

The transmission line model is used in this study to generate speech sounds from vocal tract area functions of the articulatory model. Since the present model does not provide a full 3D model of the vocal tract, a realistic area function of the vocal tract has to be derived from the partial vocal tract of the model. To do so, a grid line system (Tiede & Yehia, 1996) is implemented on the midsagittal plane of the model, and thus a series of line segments bounded by the outer and inner boundaries of the tract are obtained. The central line of the vocal tract is defined as a curve passing through the midpoints of the line segments. Assuming that the sound propagates in a plane wave along the central line, the angle of each cross-sectional plane is adjusted to be as perpendicular to the central line as possible. Slicing the vocal tract using the cross-sectional planes, a set of vocal tract widths is obtained for the midsagittal and parasagittal planes, respectively. The area function of the vocal tract is calculated using an improved $\alpha\beta$ model, which has the following shape:

$$A(x) = [\alpha(x)W(x)^{\beta(x)}]^2 + \gamma(x)g_t \quad (6)$$

where x is the distance from the glottis. $A(x)$ is the calculated area at x , and $W(x)$ is the average value of the vocal tract widths in the midsagittal and parasagittal planes. $\alpha(x)$ and $\beta(x)$ are the proportional and exponential coefficients of the width, respectively. The first term transforms the vocal tract width to an area except for the space between the upper and lower teeth. The second term is used to compensate for the discontinuity of the vocal tract area function caused by the bilateral space between the upper and lower dental arches. The compensation area is the product of the interdental gap g_t and the width $\gamma(x)$ from the inner edge of the teeth to the cheek wall. The gap g_t is determined by the degree of jaw opening, and the width $\gamma(x)$ is supposed to have a value up to 3 cm to account for the left and right sides together. The coefficients of $\alpha(x)$, $\beta(x)$ and $\gamma(x)$ were determined by minimizing the distance between the calculated and MRI-based area functions for five Japanese vowels obtained from the target speaker.

A series of area functions are obtained from the time-varying vocal tract shapes with a 16-ms interval. An algorithm for the transmission line model was developed based on that proposed by Maeda (1996). A glottal area function served as the sound source, where the interaction between the sound source and the vocal tract was taken into account. The effects of viscous components and wall vibration of the

vocal tract were considered in this model. The piriform fossa and paranasal cavities were included in this model as side branches.

4.3. Algorithm of acoustic-to-articulatory mapping

The procedures of the model simulation from articulatory targets to speech sounds are calculation of a vocal tract shape using the physiological articulatory model according to the given articulatory parameters, and generation of speech sounds from the vocal tract shape using the above acoustic model. The inverse processing of the above procedures, therefore, can be represented as an estimation of articulatory parameters from speech sounds. Generally speaking, acoustical parameters can be considered as a (nonlinear) function of the articulatory parameters, denoted by X . As shown in earlier sections, physiological constraints of human articulation are incorporated in the proposed articulatory model. The processing from an articulatory target to a synthetic sound can be described in the following state-space formulas:

$$V(k+1) = \Phi V(k) + \Psi X + AC \quad (7)$$

$$Y(k) = H(V(k), S(k)) \quad (8)$$

Formula (7) describes that vocal tract shape V at time $k+1$ depends on the vocal tract shape at time k , articulatory target X and constraint C , where Φ , Ψ and A are transform matrices for the corresponding components. Φ and Ψ are fully determined by the physiological properties of the articulatory model. Formula (8) means that the synthetic sound Y is generated based on the vocal tract shape and sound source, where H is the acoustical model described above. Although the function between Y and X is difficult to express as an analytic function, it can be described in a general form, $Y=f(X)$. Using the simplified description, the inverse estimation can be realized by reducing the distance between the acoustic vectors obtained from the speech sound and the synthetic sound. Employing the analysis-by-synthesis (AbS) method, the “true” articulatory parameters are gradually approached by modifying the articulatory vector X . A performance function is defined and minimized to realize the above processing, which is given in the following:

$$J(X) = \|y - f(X)\|_Q^2 + \|X - X_0\|_R^2 + \|X - X_p\|_W^2 \quad (9)$$

where y and $f(X)$ are acoustic vectors with five elements of speech and synthetic sounds, respectively. Vector X has nine elements. $\|X\|_R^2$ represents the quadratic form of $X^T R X$. The weight matrices of $Q(5 \times 5)$, $R(9 \times 9)$, and $W(9 \times 9)$ are simplified as diagonal matrices in this study. The parameter X_0 , the articulatory target of vowel [e], is used as a reference. X_p is the parameter estimated at the previous step, and is used as a constraint in estimating vocal tract shapes for continuous speech.

Supposing that $\hat{x}$ is the vector giving a minimal value of performance function $J(X)$ and its k th approximation is $\hat{x}_k$, function $f(X)$ can be linearized around $\hat{x}_k$,

$$f(X) \cong f(\hat{X}_k) + \left. \frac{\partial f(X)}{\partial X} \right|_{X=\hat{X}_k} \bullet (X - \hat{X}_k) \quad (10)$$

Substituting Equation (10) into Equation (9), and setting the partial derivative of $J(X)$ to be zero, the $(k+1)$ th approximation of $\hat{x}_{k+1}$ can be obtained from the following equation,

$$\hat{X}_{k+1} = \hat{X}_k + \lambda \{A^T Q A + R + W\}^{-1} \{A^T Q (y - f(\hat{X}_k)) + R(X_0 - \hat{X}_k) + W(X_p - \hat{X}_k)\} \quad (11)$$

where $A = \partial f(\hat{X}_k)/\partial X$ is the Jacobian matrix, and coefficient λ is determined to meet the condition of $f(\hat{X}_{k+1}) \leq f(\hat{X}_k)$.

The vector obtained from Equation (11) is the new target, and used to drive the articulatory model. The weight matrix R is calculated by the following formula:

$$R_k = (R_0 - R)\gamma^k + R \quad (12)$$

where $\gamma=0.8$. R_k approaches R as k increases.

It is difficult to find an analytical solution for the partial derivative matrix, $\partial f(\hat{X}_k)/\partial X$, because of the complicated nonlinear relationship between the acoustical parameters and articulatory parameters as shown in formulas (7) and (8). For this reason, this study employs the method proposed by Shirai & Honda (1978) to approximate the partial derivative matrix. That is, given a small variation of δx around $\hat{X}_k$ in articulatory space, increment $df(X)$ can be obtained in acoustic space corresponding to the variation. The partial derivative is computed by the difference method according to the increments and the variation. Physiological constraints of the articulatory model are involved in the relationship between the δx and $df(X)$. In this estimation, $R=0.05\mathbf{I}$, $R_0=0.1\mathbf{I}$, and $Q_d=\{2.0, 1.3, 1.0, 0.8, 1.2\}$, where $\mathbf{I}$ is the unit matrix, and Q_d represents the diagonal elements of Q .

4.4. Procedure of the estimation

Fig. 6 shows the procedure used in the estimation. An input speech sound is digitized at a sampling rate of 16 kHz, and is preemphasized to reduce the radiation effects. After flattening the spectral characteristics using a second-order adaptive filter, the formant frequencies of the input sounds are obtained via the LPC-Cepstrum method. To determine a target for the articulatory model in the initial stage, articulatory target sets were prepared for the typical vocal tract shapes of five Japanese vowels. The initial target is determined by interpolating the two target sets whose formant patterns are closer to that of the input sound. Based on the articulatory target, muscle activation patterns are derived stepwise via the muscle workspaces, and fed into the muscles to drive the model to form a time-varying vocal tract shape. A synthetic sound is generated using area functions obtained from the vocal tract shape. The acoustic vector of the synthetic sound is obtained using the same processing as that used for the input sound. To find the minimum for the performance function, Jacobian matrix is calculated using the method described above. The calculating procedure is indicated by the loop with dashed lines, which combines the model physiological constraints into the inverse estimation. A new articulatory target is estimated to reduce the distance between the acoustical parameters of the input and synthetic sounds. The physiologically plausible vocal tract shape is gradually reached by iterating the target renewing steps, which are indicated by the frame with a dashed line in the figure.

TABLE II. A comparison of estimated targets with known targets

Vowels	/a/	/i/	/u/	/e/	/o/
Target error (cm)	0.15	0.06	0.30	0.11	0.04
Formant (%)	1.18	3.15	3.69	2.36	3.34

extent in obtaining the formant pattern of /u/. This phenomenon indicates that the proposed method still faces the one-to-many problem. In synthesizing the above sounds, the glottal height was fixed. As a test, however, the glottal height was treated as a variable, which was modified in the estimation. In the result, the length of the laryngeal tube in the estimation varied within 2% among the five vowels. This indicates that the method can offer a reliable estimation.

The estimation method was also examined using acoustic parameters. The estimation error was defined as the average value of the difference between input sound and synthetic sound for the four lower formants. As shown in Table II, the estimation errors were less than 4% for all of the vowels. The results for articulatory and acoustic simulations demonstrated the validity of the articulatory model for the inverse estimation.

5.2. Articulatory constraint for model-based estimation

As seen above, the compensation between lip protrusion and tongue position in /u/ causes a deterioration in the estimation accuracy. This problem might be solved using formant patterns to constrain the tongue position because the tongue position and formant pattern have a straightforward relationship (Stevens, 1999). In addition, our model has some advantages for realizing this constraint. One of the advantages is that the tongue dorsum control point of the model is compatible with the tongue position so that the constraint can be directly implemented in controlling the model. The other is that the effects of the tongue position movements on the vocal tract shape are easily quantified using the model simulation, while such effects are difficult to describe with other methods such as the one using area functions as articulatory parameters.

To quantify the acoustic-to-articulatory relation, six male Japanese speakers were selected from a Japanese articulatory database produced using the X-ray microbeam system at the University of Wisconsin (Hashi, Westbury & Honda, 1998). Table III shows the speech materials used in this investigation, which consists of vowel sequences and vowel-consonant-vowel sequences. To build a relation between acoustics and geometry, the horizontal and vertical positions of the pellets on the tongue and jaw were selected as the articulatory parameters, and F1 and F2 as the acoustic parameters. The result revealed that the horizontal position of the tongue dorsum has a high correlation with the frequency difference between F1 and F2. Fig. 7(a) shows the correlation between the horizontal displacement of the tongue dorsum (5.7cm back from the tongue apex) and the F1–F2 difference for the target speaker, where the F1–F2 difference is shown in a log scale. The correlation coefficient was -0.956 for the target speaker, and the residual standard variance was 0.026. This result implies that vowel formant and articulatory movement can be

TABLE III. Speech material used in this investigation

Vowel sequence	/ae/, /ai/, /au/, /ea/, /ei/, /ia/, /ie/, /iu/, /ou/, /ua/, /ui/, /uo/
VCV sequence	/aka/, /ata/, /asha/, /apa/, /ara/, /aza/, /aba/, /ada/, /ama/, /aha/, /awa/, /aya/, /acha/, /ana/, /aga/, /asa/

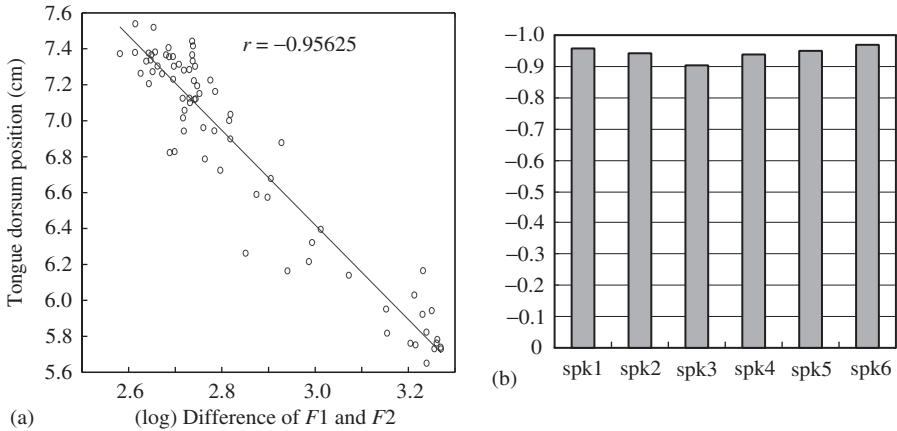


Figure 7. The horizontal position of the tongue dorsum and the F1–F2 difference for the target speaker (a), and correlation coefficients for six male speakers (b).

mapped one to other in a straightforward manner. Fig. 7(b) shows the correlation coefficients for the six male speakers. The correlation ranged between -0.91 and -0.96 . The average correlation coefficient was -0.942 , and the average residual standard variance was about 0.03 . Using the F1–F2 difference to predict the tongue dorsum position, the average prediction error was 0.181 cm, and the maximal error was 0.244 cm among the six speakers. The small prediction error and residual variance are evidence that formant-to-tongue position mapping can be used as an articulatory constraint for the inverse estimation.

The relationship between jaw movement and formant pattern was also investigated. The result showed that jaw movement also had a high correlation with the first formant, whose average value was 0.66 for the six speakers. This correlation between jaw position and vowel formant, however, is not high enough to be a constraint for the model.

5.3. Examination of the inverse estimation using articulatory data

The articulatory data were recorded with acoustic data simultaneously in the database. The speech materials used in the model-based acoustic-to-articulatory mapping were the vowel sequences shown in the upper part of Table III. The procedure for estimating the vocal tract shape for the first vowel is exactly the same as that described in Section 4.4, while the estimation for the second vowel always

begins from the first vowel to account for preservatory coarticulation. The inverse estimation was carried out for the target speaker using the articulatory model with and without the articulatory constraint described above.

Both the vocal tract shape and the corresponding acoustic parameters were used in the examination for each vowel, which were extracted from the stable parts of vowel sequences of the observation and the simulation. Four tongue pellets that were glued at 0.8, 2.6, 4.7 and 6.8 cm from the tongue apex, named as T1–T4, and the mandibular incisor pellet were used as a measure for the estimated vocal tract shapes. The pellets were projected on the midsagittal plane of the model for evaluation of estimation errors. The distances from the tongue pellets to the model tongue surface and from the jaw pellet to the jaw control point were measured. Fig. 8 shows the average value over the distances measured for all of the pellets during vowel sequences. The left panel shows the results without the articulatory constraint. In this case, the average error over all of the utterances was about 0.2 cm. Among the five pellets, T4 and T1 showed larger errors than the other pellets. The right panel shows the estimation error when the articulatory constraint was implemented. Compared with the case without the constraint, the estimation accuracy is generally improved. The average distance for the vowel sequences of /iu/ and /ou/, which had the largest error in the case without the constraint, was improved about 0.1 cm. The estimation error was reduced more remarkably for the vowel /u/ in the vowel sequences. This is probably because, in generating the formants of the vowel /u/, the lip configuration and tongue position can compensate for one another. After introducing the articulatory constraint, a reasonable region of the tongue position was guaranteed, and thus the estimation error is decreased.

Fig. 9 shows the average error for the four lower formants of the synthetic and recorded vowel sequences. The left panel is the result without the constraint, and the right panel is the result with the constraint. In the left panel, the average error over all of the utterances was about 2.5%, ranging from about 1 to 4%. When the articulatory constraint was employed, an obvious improvement can be seen in the right panel, especially for the vowels with a large error. Comparing the two conditions, the largest error that is seen in the /a/ of /ai/ and the /u/ of /iu/ decreased from about 4.2 to 3% and the average error dropped to 1.8%. The results

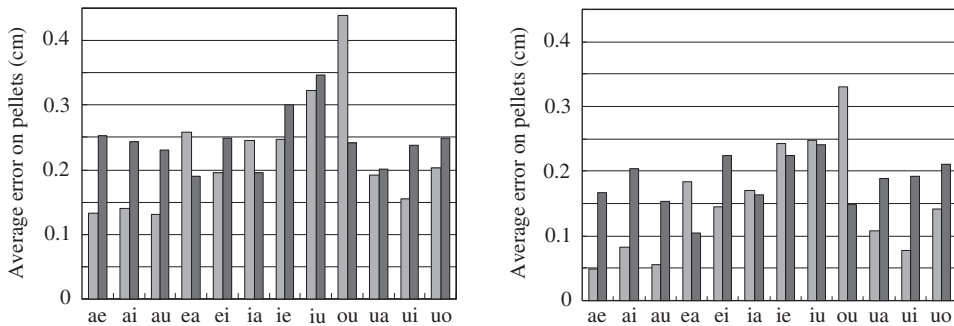


Figure 8. Averaged values of the distances from the tongue pellets to the model tongue surface and from the jaw pellet to the jaw control point for the vowels in the vowel sequences from the target speaker: (a) without the articulatory constraint and (b) with the constraint.

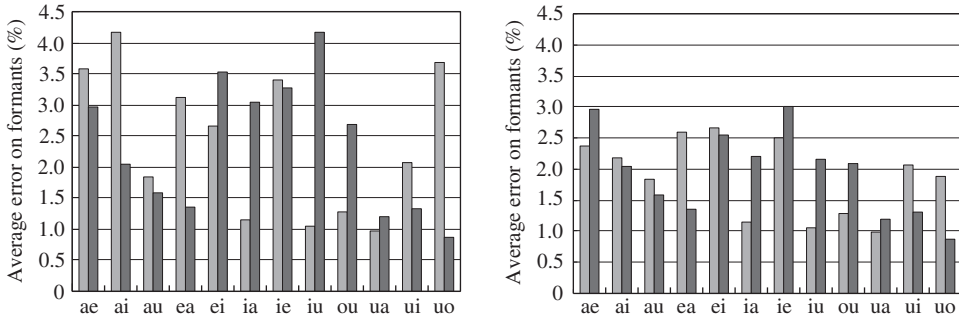


Figure 9. Simulation errors averaged over the four lower formants for the vowels in the vowel sequences from the target speaker: (a) without the articulatory constraint and (b) with the constraint.

indicate that the model-based estimation method demonstrated a good performance for acoustic-to-articulatory mapping.

6. Conclusions

This study described an improvement of the control method for our physiological articulatory model and an application of the model to acoustic-to-articulatory mapping. A multipoint control strategy employing dynamic muscle workspaces provided a higher flexibility in controlling the tongue tip, tongue dorsum, and jaw for producing vowels and consonants. The physiological articulatory model demonstrated a number of advantages in the inverse estimation because of the physiological constraints inherent in the model.

The relation of tongue positions and vowel formants were investigated using X-ray microbeam data for six male speakers. Because they have a high correlation and lower residual variance, formant-to-tongue position mapping was used as an articulatory constraint in the model control. Such a one-point constraint can be used for the present estimation method due to the fact that the model can easily account for the effects of tongue position movement on the vocal tract shape. As shown in the results, the use of the articulatory constraint improved the accuracy of acoustic-to-articulatory mapping. Note that the lip/tongue compensation for /u/ does not only affect the formant frequencies but also other acoustic characteristics such as spectral tilt. For this reason, more accurate constraints are required for further studies.

The estimation accuracy of vocal tract shapes was evaluated using articulatory data obtained from the target speaker for vowel sequences. The average value of the minimal distance from the tongue pellets to the model tongue surface and the distance from the jaw pellet to the jaw control point were measured. For the target speaker, the average distance was 0.16cm for the vocal tract shape, and the estimation error was 1.8% for the four lower formants.

In the estimation, the estimation error was a cumulative result of all of the errors that occurred in calculating the articulatory movement and generating the synthetic

sound. One of the errors was seen in calculating area functions from the vocal tract widths. In this study, the proposed method demonstrated a good performance in estimating vocal tract shapes for vowel sequences. To develop the method as a practical tool, estimation for speech sounds with consonants is also necessary. These two issues remain for future studies.

This work was supported in part by CREST of JST (Japan Science and Technology).

References

- Atal, S., Chang, J., Mathews, J. & Tukey, W. (1978) Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique, *Journal of the Acoustical Society of America*, **63**, 1535–1555.
- Dang, J. & Honda, K. (1997) Correspondence between three-dimensional deformation and EMG signals of the tongue. In *Proceedings of the Spring Meeting of Acoustical Society Japan*, pp. 241–242. (in Japanese)
- Dang, J. & Honda, K. (1998) Speech production of vowel sequences using a physiological articulatory model. In *Proceedings of the International Conference of Spoken Language Processing*, Vol. **5**, pp. 1767–1770.
- Dang, J. & Honda, K. (2001) A physiological articulatory model for simulating speech production process. *Journal of the Acoustical Society of Japan (E)*, in press.
- Dang, J., Sun, J., Deng, L. & Honda, K. (1999) Speech synthesis using a physiological articulatory model with feature-based rules. In *Proceedings of the ICPhS-99*, pp. 2267–2270.
- Dusan, S. & Deng, L. (1998) Recovering vocal tract shapes from MFCC parameters. In *Proceedings of the International Conference of Spoken Language Processing*, Vol. **7**, 3087–3090.
- Dusan, S. & Deng, L. (2000) Acoustic-to-articulatory inversion using dynamic and phonological constraints. In *the 5th Speech Production Seminar*, Munich, Germany, pp. 237–240.
- Hashi, M., Westbury, J. & Honda, K. (1998) Vowel posture normalization, *Journal of the Acoustical Society of America*, **104**, 2426–2437.
- Maeda, S. (1990) Compensatory articulation during speech: evidence from the analysis of vocal tract shapes using an articulatory model. In *Hardcastle, Marchal speech production and speech modeling*, pp. 131–149. Dordrecht: Kluwer Academic Publishers.
- Maeda, S. (1996) Phonemes as concatenable units: VCV synthesis using a vocal-tract synthesizer. In *AIPUK 31*, (A. Simpson & M. Patzod, editors), pp. 127–232.
- Miyawaki, K. (1974) A study of the muscular of the human tongue, *Ann. Bull. Res. Inst. Logoped. Phoniatrics Univ. Tokyo*, **8**, 23–50.
- Moore, C. (1992) The Correspondence of the vocal tract resonance with volumes obtained from magnetic resonance image, *Journal of Speech and Hearing Research*, **35**, 1009–1023.
- Okadome, T., Suzuki, S. & Honda, M. (2000) Recovering of articulatory movements from acoustic with phonemic information. In *The 5th Speech Production Seminar*, Munich, Germany, pp. 229–232.
- Schroeder, M. R. (1967) Determination of the geometry of the human vocal tract by acoustic measurement, *Journal of the Acoustical Society of America*, **4**, 1002–1010.
- Schroeter, J. & Sondhi, M. (1994) Techniques for estimating vocal tract shapes from speech signal, *IEEE Transactions on Speech Audio Processing*, **2**, 133–150.
- Shirai, K. & Honda, M. (1978) Estimation of articulatory parameters from speech sound, *Transactions on IECE*, **61**, 409–416 (in Japanese).
- Stevens, K. (1999) *Acoustic phonetics*. Cambridge, MA: MIT Press.
- Stevens, K. & House, A. (1955). Development of quantitative description of vowel articulation,” *Journal of Acoustics*, **24**, 175–184.
- Takemoto, H. (2000) Morphological analyses of the human tongue musculature for three-dimensional modeling, *Journal SLHR*, 95–107.
- Tiede, M. K. & Yehia, H. (1996) , In *Proceedings of the ASA-ASJ 3rd Joint Meeting*, pp. 861–866.
- Warfel, J. (1993) *The head, neck, and trunk*. Philadelphia and London: Led & Febiger.
- Yamazaki, K. (1933) The weight of the cranium and mandible with comparison of the dental and bony regions of the mandible, *Japanese Journal of Dentistry*, **26**, 769–796 (in Japanese).
- Yehia, H. & Itakura, F. (1996) A method to combine acoustic and morphological constraints in the speech production inverse problem, *Speech Communication* **18**, 151–174.

AUTHOR QUERY FORM

**HARCOURT
PUBLISHERS**

JOURNAL TITLE: DATE: 12/2/2002
ARTICLE NO. : 20020167

Queries and / or remarks

Manuscript Page/line	Details required	Author's response
1	Pl. check if the correspond- ing author is indicated cor- rectly	
24	update ref. Dang & Honda (2001)	
25	Provide vol. for ref. Take- moto, 2000 & journal title in full for the same.	
24	provide journal title in full for ref. Miyawaki, 1974.	
8/15	Ref. Laboissiere et al. (1996) not listed in ref-list. pl.check	
8/26	Ref. Sanguineti et al. (1998) not listed in ref. list. pl.check	
13/16	Ref. Baer et al., 1988 not listed in ref. list. pl.check	
24	Ref Dang et al., 1999 not cited in text. pl. check	