6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

IMPROVEMENT OF A PHYSIOLOGICAL ARTICULATORY MODEL FOR SYNTHESIS OF VOWEL SEQUENCES

Jianwu DANG and Kiyoshi HONDA
ATR Human Information Processing Research Labs, Kyoto, Japan 619-0288
Email: *dan@hip.atr.co.jp* and *honda@hip.atr.co.jp*

ABSTRACT

A 3D physiological articulatory model has been constructed based on volumetric MRI data obtained from a male speaker. The model is driven by muscles according to a target-dependent activation pattern. In this study, we improved dynamic characteristics of the model to produce higher sound quality for vowel sequences. Dynamic characteristics of articulatory organs were investigated using X-ray microbeam data for vowel sequences and vowel-consonant-vowel (VCV) sequences for 11 Japanese speakers. It was found that the velocity of the tongue tip is about 60% faster in transition of vowel-to-consonant than that of vowel-to-vowel, while the velocities of the tongue dorsum and jaw were independent of the sequences. Reaction time, from maximal acceleration to maximal velocity, of the articulators is about 40% shorter in vowel-to-consonant transitions than in vowel-to-vowel transitions. To apply the improved model for speech analysis, articulatory targets were estimated for the vowels in vowel sequences using AbS method, and used to generate the vocal tract shapes for vowel sequences. The vocal tract shapes and synthetic sounds were compared with speech sound and articulatory data from the target speaker. The results showed that our model demonstrates plausible dynamic characteristics of articulatory movement in producing vowel sequences. The simulation error was about 2.5% for the formants, and 0.2 cm for the observation points of the vocal tract.

1. CONSTRUCTION AND IMPROVEMENT OF PHYSIOLOGICAL ARTICULATORY MODEL

A 3-D physiological articulatory model driven by muscle contraction has been developed for human-mimetic speech synthesis. The articulatory model was constructed based on MRI data obtained from a male speaker. The entire model consists of the tongue, jaw and the vocal tract wall.

1.1. Construction of a Physiological Articulatory Model

The tongue shapes were extracted from volumetric MRI data in the midsagittal and parasagittal planes. The basic structure of the tongue tissue model roughly replicates the fiber orientation of the genioglossus muscle. The central part of the tongue including the genioglossus is represented by a 2-cm-thick layer bounded with three sagittal planes. Each plane is divided into six sections with nearly equal intervals in the anterior-posterior direction and ten sections along the tongue surface. The 3D tongue model is constructed by connecting the section nodes in the midsagittal plane to the corresponding ones in the left and the right planes using viscoelastic springs. This model is capable of forming the midsagittal groove and the side airway, which are the essential behaviors of the tongue in producing vowels and consonants. The jaw and hyoid bone are modeled as a rigid body to yield rotation and translation motions. Outlines of the vocal tract wall were also extracted from MRI data in the midsagittal and parasagittal planes (0.7 and 1.4 cm apart from the midsagittal plane on the right side). Assuming that the left and right sides are symmetric, 3D surface shells of the vocal tract wall and the mandibular symphysis wall were constructed using the MRI-derived outlines (see [1,2] for details). 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 velum situation is determined by the opening area of the naso-pharyngeal port. These parameters are taken into accountas acoustic parameters in speech synthesis stage.

1.2 Control Strategy of the Model

The model is driven by twelve muscles for the tongue and eight muscles for the jaw. Muscle activation signals are estimated using target-based control strategy. Three control points are used in the control strategy; 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 0.5 cm inferior to the tip of the mandible incisor. The articulatory target is the coordinate of the final position for each control point during a stable phonation.

The target-based control strategy is to generate muscle activation signals according to a given target for each control point, and feed the activation signals into the muscles to drive the model. For this purpose, a muscle workspace is constructed for each control point to establish a relationship between the target and the activation signals [1]. The muscle workspace consists of muscle force vectors that correspond to a displacement of the control point when the muscles contract. Since orientation of the muscles varies with articulatory movement, the muscle force vector must adjust to the locations of the control points. Therefore, a set of muscle force vectors is calculated for the control points of the tongue in four positions corresponding to the tongue shapes of the rest posture and three extreme vowels of /a/, /i/, and /u/ [2]. For the jaw muscle vector, positions are chosen in the rest position and a wide-open position. The muscle workspace used for generating muscle activation signals is an interpolated workspace that is calculated stepwise based on the set of the muscle force vectors.

To generate a movement for the model from a current position to a given target, a displacement vector from the current position toward the target is mapped onto the muscle workspace. Thus, a set of projections is obtained for each muscle vector. The positive projection for a muscle vector implies that the muscle contributes to the movement approaching the target when it contracts, while a negative projection has no contribution for the movement. Therefore, the muscle with a positive projection is excited at this computational step, and the magnitude of activation signal is proportional to the amplitude of the projection. When the activation signals are fed into the muscles, the control point is driven to move to a new position, approaching the target. By iterating the above procedure at each new position, a set of time-varying activation signals is generated, and the control point is driven to approach its target.

1.3 Improvement of the Control Strategy for Sequential Movement Tasks

The above control strategy was designed for the estimation of the activation signals for producing movement towards a stationary target. During speech, however, the target location varies from segment to segment. When articulatory targets are switched from one position to another, desirable movements may be not realized by the above control strategy. When the model aims to reach an extreme target, for example, it results in a large deformation. The strain due to this deformation potentially drives the model to restore the original shape and the control points move toward their initial positions. The motion caused by the stain may strongly interfere with the movement toward the new target. Therefore, the effect of the strain must be considered in the control strategy for sequential tasks to realize quick and smooth movements toward a new target. The following procedure is used to consider the strain problem in the simulation of sequential tasks.

Since the relationship between the strain due to deformation and its effect on the control points is too complicated to be expressed by an analytic description, the first procedure is set to quantify the strain effect on each control point by computing model deformation with no muscle excitation. Because the resultant movement in this computation depends only on the deformation strain, the displacement, referred as to *strain-based displacement*, for each control point reflects the strain effect on the control point. To reduce the interference of the strain with the target-based movement, it requires a force to cancel the strain-based displacement. According to the control strategy described in the previous section, the cancel force can be obtained by a virtual vector that is opposite to the strain-based displacement. The virtual vector is added on the vector from the current position to the next target in estimating muscle activation signals when the sequential task is switched to a new one.

2. INVESTIFGATION OF ARTICULATORY MOVEMENT

To evaluate the performance of the improved model, articulator movements were investigated using the Japanese database obtained by the University of Wisconsin X-ray Microbeam System [7]. The data for eleven Japanese speakers were used and the target speaker for the model (mh19) was included in the database. In the microbeam experiment, four pellets were placed on the tongue surface, named T1 to T4. The pellets were glued onto the tongue surface in the midsagittal plane, and located in about 0.8cm, 2.5cm, 4.5cm, and 6.2cm from the tongue apex. T5 was used for the target speaker as a hanging pellet located in the posterior portion on the dorsum. Other three pellets were used to describe the movement of the upper lip, lower lip, and mandible, respectively. Table 1 shows the utterance list, which consists of vowel sequences with two vowels out of five Japanese vowels, and vowel-consonant-vowel (VCV) sequences. Average duration of the utterance over the eleven subjects was 0.45 seconds for vowel sequences and 0.44 seconds for VCV sequences.

Table 1. 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/

In this study, pellets T1 through T4 on the tongue and the pellet on the mandibular incisor are investigated to evaluate performance the articulatory model. Figure 1 shows the average velocities of each observation point for vowel sequences (left panel) and VCV sequences (right panel). The average velocities are calculated when the pellets transit from the position of the first vowel to the second phoneme, a vowel (in vowel sequences) or a consonant (in VCV sequences). The velocity of the tongue tip (T1) varies with utterances, about 10 cm/s for vowel sequences and 16 cm/s for VCV utterances. It indicates that movement of the tongue tip is about 60% faster in moving to its target for the consonants than that for the vowels. The velocities of the tongue dorsum and jaw were almost independent of the type of the sequences, about 12 cm/s for tongue dorsum and 6 cm/s for the jaw. Comparing the tongue tip with others, it was found that the movement of tongue tip is slower than the tongue dorsum (T3) and the blade (T2) in vowel sequences, although it is faster in moving to the consonantal target.

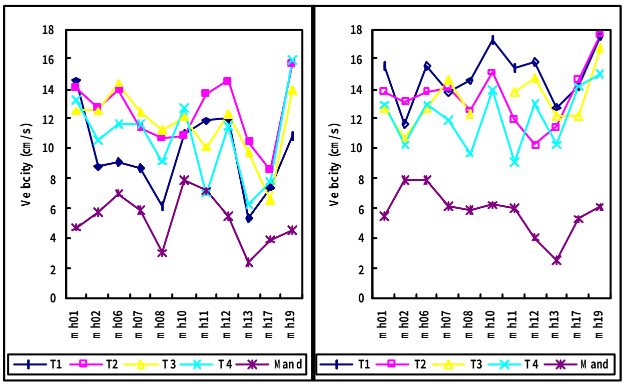

Figure 1. Average velocities in motion from one target to the next for observation points of the tongue and mandible during vowel

sequence (left penal) and VCV sequences (right penal).

Dynamic characteristics of the articulators are investigated using time information. The basic assumption is that if a periodical movement is described using the sine curve, the time interval from the maximal acceleration to the maximal velocity is equal to one quarter of its period. This suggests that dynamic characteristics of each observation point can be roughly measured using this time information. Hereafter, the time interval from the maximal acceleration to the maximal velocity is referred as to *reaction time* of the observation point. Figure 2 shows the reaction time for the five pellets, which were obtained from a vowel-to-vowel transition for vowel sequence and vowel-to-consonant transition for VCV sequence. The average reaction time over the eleven subjects was about 50 milliseconds for vowel sequence and 30 milliseconds for VCV sequence. The result shows that actions movements of the articulators are much faster for forming a consonantal posture than for a vowel posture. However, the results did not show any significant difference among the articulators. As also seen in the result of velocity, there is a large individual variation among subjects. The individual variation becomes smaller in VCV sequence. This phenomenon implies that consonant targets constrain the articulators to move to a consistent place because they are more critical for the articulators to achieve.

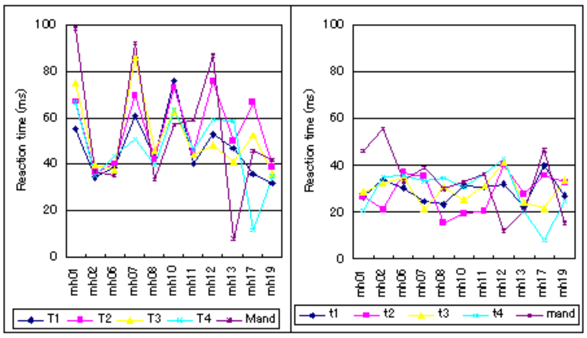

Figure 2. Average time from the maximal acceleration to the maximal velocity for vowel sequence (left panel) and VCV sequence (right panel).

3. SIMULATION OF VOWEL SEQUENCE USING THE AbS METHOD

The improved model is used to estimate the articulatory target for vowel sequence, and further to simulate the vowel sequence using the estimated target. Estimation of the articulatory target is an inverse procedure from acoustic signals to the vocal tract configuration, which is known to have a one-to-many problem. Since our model is equipped with physiological constraints inherent to human speech organs, it is expected to obtain more reliable and accurate result than other methods, in which a number of artificial constraints were incorporated in their estimation [3-5].

3.1 Algorithm of Target Estimation

Both the acoustic and articulatory parameters are used as cues in the estimation of the articulatory targets. The articulatory parameter *X* is a vector consisting of coordinates of the three control points, length and cross-sectional area of the lip tube and the glottis height. The last three parameters are used in the acoustical model for speech synthesis, but did not involved in the computation of the articulatory model. The acoustical parameter is considered as a function *f(X)* of the articulatory parameters *X*. In the estimation, the acoustical vector *f(X)* obtained from the synthetic sound is used to approximate vector f_r which is obtained from the input speech sound. The performance function *J(X)* is defined in the following form.

$$J(X) = \|f_r - f(X)\|_Q^2 + \|X - X_0\|_R^2 + \|X - X_p\|_W^2 \qquad (1)$$

where f_r and *f(X)* are vectors with five elements of the formants. $\|X\|_R^2$ represents the quadratic form of $X^T RX$. *Q*, *R* and *W* are the weight coefficient matrices. The target X_0 of vowel [e] is used as a reference parameter. *X* is the parameter estimated at the previous step, and is used as a constraint in estimating vocal tract shapes of a continuous speech.

Suppose $\hat{X}_k$ is *k'th* approximation of *X*, function *f(X)* can be linearized around $\hat{X}_k$. The $(k+1)$*'th* approximation of $\hat{x}_{k+1}$ can be obtained from the following equation,

$$\hat{X}_{k+1} = \hat{X}_k + l\left\{A^T QA + R + W\right\}^{-1}\left\{A^T Q(f_r - f(\hat{X}_k)) + R(X_0 - \hat{X}_k) + W(X_p - \hat{X}_k)\right\} \qquad (2)$$

where $A = \partial f(\hat{X}_k)/\partial X$, and coefficient is calculated to meet the condition of $f(\hat{X}_{k+1}) \le f(\hat{X}_k)$. It is difficult to find an analytical solution for the partial derivative *A* because of the complicated relation between the acoustical parameter and articulatory parameter. This study employs the method proposed by Shirai and Honda [5] to find an approximate solution. Giving a small variation of x around $\hat{X}_k$, we can obtain a value of *df(X)* corresponding to the variation. The ratio of *df(X)* to x is one of the approximate solutions for the partial derivative (see [6] for details).

3.2 Target Estimation and Model Simulation

The analysis-by-synthesis method is used for target estimation. Acoustic signal of vowel sequences is analyzed using LPC-cepstrum, and phoneme boundary of vowels is determined by delta cepstrum. The acoustic data were recorded with the articulatory data simultaneously in the X-ray microbeam experiment. The typical formant patterns of the vowels are used as templates to classify the input vowel. The initial target for the first vowel is determined by a set of articulatory targets of typical vocal tract shapes for five Japanese vowels. Muscle activation signals are derived stepwise based on the articulatory targets, and fed into the muscles to drive the model to produce a vocal tract shape. Synthetic sounds are generated using an area function obtained from the vocal tract shape. The synthetic sound is processed using the same procedure as that used in the input speech sound. The new articulatory target is estimated to minimize the distance between the acoustical parameters of the synthetic sound and recorded speech. A physiologically plausible vocal tract shape and articulatory target are gradually achieved by iterating above procedures (see [6] for details).

The vocal tract configuration of the articulatory model is calculated according to a given target and the previous configurations. For this reason, the vocal tract configuration of the second vowel cannot be independent of the first vowel completely. In estimation of the second vowel, therefore, model simulation is also started from the first vowel. In this stage, the target for the first vowel is the one estimated in the above processing, and the initial target for the second vowel is determined based on its formants as done for the first vowel. When the targets are switched from the first one to the second, the force to cancel out the strain is calculated, and taken into account in estimating the muscle activation signals. Then, the procedure described above is iterated to reach a plausible articulatory target for the second vowel.

After the above procedure, articulatory targets of the vowels were obtained for a vowel sequence. Vowel duration is extracted from the recorded sounds and equipped with the targets. Using the target and vowel length, a series of vocal tract shapes can be obtained for the vowel sequence by using the articulatory model. Synthetic sounds are produced based on the vocal tract shapes.

4. RESULTS

In this study, model simulation was carried out for the vowel sequences alone. The performance of the model is evaluated by comparing the model simulation with the articulatory data from the target speaker.

4.1 Evaluation of Dynamic Characteristics of the Model

Dynamic characteristics of the articulators are measured using maximal velocity and reaction time. To compare the model simulation with articulatory data, observation points of the model are defined corresponding to the positions of the five pellets used in the X-ray microbeam experiment for the target speaker, and are labeled using the same name. Figure 3 shows the maximal velocity (left penal) and the reaction time (right penal) obtained from the model and the articulatory data. The average difference between the model and the articulatory data is about 1.56cm/s for the velocity, and 3 milliseconds for the reaction time. The model simulation is consistent with the articulatory data within 15% for the velocity, and 8% for the reaction time. This result shows that the model approximately demonstrates dynamic characteristics of the articulators of the target speaker. The target for the second vowel can be reached more accurately by using the improved control strategy than the unimproved one. The time interval from the target alteration to the maximal velocity were reduced about 20%, though the result did not show any significant difference for the velocity and the reaction time.

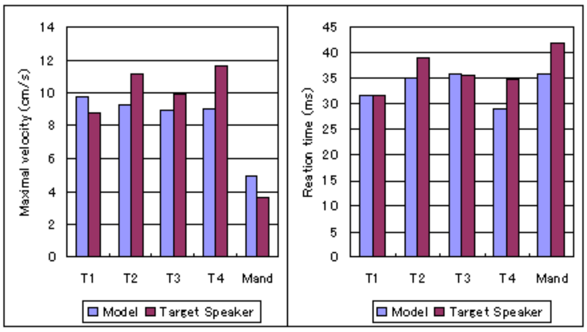

Figure 3. Comparison of the model simulation and the articulatory data from the target speaker for the maximal velocity and the reaction time.

4.2 Evaluation Using Acoustic and Articulatory Data

From the above simulation, a series of vocal tract shapes were obtained corresponding to the vowel sequence. Since the present model does not provide a full 3D vocal tract shape, the vocal tract area function has to be estimated using the partial vocal tract provided by the model. We first determine the vocal tract widths in the midsagittal and parasagittal planes of the model, and then estimate vocal tract area functions using the width information (see [1] for details). A series of area functions are obtained from the vocal tract shapes in a 20-ms interval. A transmission line model is employed to synthesize speech sound. Voice source is glottal area functions, where the interaction between the sound source and vocal tract is taken into account. The effects of viscous components and wall vibration of the vocal tract are considered in this synthesis. Figure 4 shows the simulation error averaged over the first four formants of the synthetic sound and real sound for vowel sequences. The average error over all the utterances is smaller than 2.5%, ranged from about 1% to 4%.

Since there is no unique relationship between sound signals and vocal tract configurations, the use of the synthetic sound alone is not sufficient to evaluate the model behaviors. For this reason, we also compared the vocal tract shape of the model simulation with the articulatory data. To do so, the vocal tract shape of each vowel is extracted from the stable parts of vowel sequences for the model simulation and the articulatory data. The pellets of the jaw and T1 though T4 are mapped on the extracted vocal tract from model simulation. Distances between the jaw pellet and the jaw control point is used to evaluate the simulation error for the jaw. The minimal distance from each of the pellets T1 through T4 to the tongue surface of the model is used to evaluate the estimation error in the tongue shape. The average error over all the utterances is about 0.2 cm. Among the five pellets, T1 and T4 showed a larger error than the others.

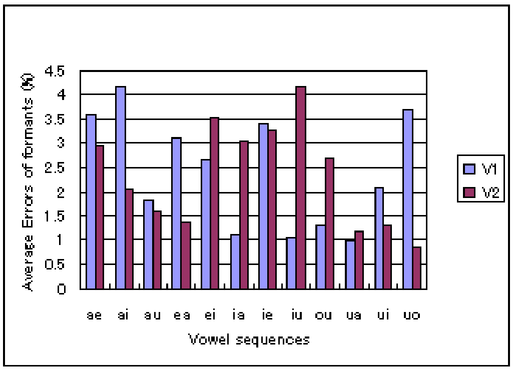

Figure 4. Simulation error averaged over the first four formants for the first and second vowels in the vowel sequences, respectively.

5. CONCLUSION

In this study, control strategy of the model was improved by taking the strain of the deformation into account in estimating muscle activation signals when the targets are switched from one to another. Model performance was evaluated using articulatory data and acoustic data for the target speaker, which was from X-ray microbeam data. The average error of the simulation was about 2.5% for the formants, and about 0.2 cm on the pellets of the vocal tract. The result showed that the improved model demonstrated good dynamic characteristics in simulating vowel sequences. Comparing our estimation with the others studies [3-5], the estimation error in this study showed slightly larger for acoustic parameters. Unlike the other studies, however, this study provided not only the acoustic evaluation but also an articulatory evaluation. An articulatory evaluation is very important because acoustic signal alone is difficult to measure the accuracy of the evaluation for the one-to-many problem.

ACKNOWLEDGMENTS: This work has been supported in part by CREST of JST (Japan Science and Technology).

6. REFERENCES

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

[2] Dang, J., Sun, J., Deng, L. and Honda, K. (1999). "Speech synthesis using a physiological articulatory model with feature-based rules," Proc. ICPhS-99, 2267-2270 (San Francisco, USA).

[3] Shroeder, M. R. (1967). “Determination of the geometry of the human vocal tract by acoustic measurement, ” J. Acoust. Soc. Am., 4, p. 1002.

[4] Yehia, H. and Itakura, F. (1996). ”A method to combine acoustic and morphological constraints in the speech production inverse problem,” Speech Comm. 18, 151-174.

[5] Shirai, K. and Honda, M. (1978). “Estimation of articulatory parameters from speech sound, “ Trans. IECE, 61, 409-416 (1978).

[6] Dang, J. and Honda, K. (2000). “Estimation of vocal tract shape from speech sounds via a physiological articulatory model”, 5th Speech Production Seminar, (Munich, Germany).

[7] Hashi, M., Westbury, J. and Honda, K. (1998) “Vowel posture normalization,” J. Acout. Soc. Ame., 104, 2426-2437.