Jan. 2006, Vol.21, No.1, pp.95–105 J. Comput. Sci. & Technol.

Communication Between Speech Production and Perception Within the Brain—Observation and Simulation

Jianwu Dang[1,2] (党建武), Masato Akagi[1] (赤木正人), and Kiyoshi Honda[2] (本多清志)

[1] *Japan Advanced Institute of Science and Technology, Ishikawa, Japan*

[2] *ATR Human Information Science Laboratories, Kyoto, Japan*

E-mail: {jdang,akagi}@jaist.ac.jp; honda@atr.jp

Received February 24, 2005; revised August 29, 2005.

Abstract Realization of an intelligent human-machine interface requires us to investigate human mechanisms and learn from them. This study focuses on communication between speech production and perception within human brain and realizing it in an artificial system. A physiological research study based on electromyographic signals (Honda, 1996) suggested that speech communication in human brain might be based on a topological mapping between speech production and perception, according to an analogous topology between motor and sensory representations. Following this hypothesis, this study first investigated the topologies of the vowel system across the motor, kinematic, and acoustic spaces by means of a model simulation, and then examined the linkage between vowel production and perception in terms of a transformed auditory feedback (TAF) experiment. The model simulation indicated that there exists an invariant mapping from muscle activations (motor space) to articulations (kinematic space) via a coordinate consisting of force-dependent equilibrium positions, and the mapping from the motor space to kinematic space is unique. The motor-kinematic-acoustic deduction in the model simulation showed that the topologies were compatible from one space to another. In the TAF experiment, vowel production exhibited a compensatory response for a perturbation in the feedback sound. This implied that vowel production is controlled in reference to perception monitoring.

Keywords speech communication, human mechanism, speech production, speech perception

1 Introduction

Generating and perceiving speech sounds are basic human activities in speech communication. It is a long-standing dream to create a human-like machine with these functions. Speech synthesis and recognition technologies aim to realize such human capabilities using computer systems, and a great deal of success was achieved during the last decade. However, the achievements of both speech recognition and speech synthesis are still far from completion when compared with the human performance.

In terms of speech information processing, a speaker can be considered as an encoder in a speech production system and a listener as a decoder to accomplish speech perception. In this analogy, speech information is exchanged not only among people but also internally, within human brain, because human beings play the role of listeners when they take the part of speakers. Especially, in acquiring a new language, one is both a speaker and a listener. During such a learning process, the loop in human brain between speech production and perception must be closed, as a type of internal "speech chain"[1]. It is natural to believe that excellent human speech processing behaviors are deeply concerned with the closed-loop speech chain in the brain.

For years, a number of experiments have been conducted to investigate the relation between speech production and perception in order to explore the nature of the speech chain. In one of the earliest of these, Lombard (1911) clarified that speech sounds have a higher intensity and fundamental frequency (F0) in a noisy environment than under normal conditions[2]. This phenomenon is known as the *Lombard effect*. Lee (1950) investigated the effects of auditory feedback on speech production using delayed auditory feedback (DAF)[3]. He found that as the delay time increases up to 200ms, the DAF seriously disturbs the normal speech function of the subjects. Kawahara (1994) clarified the influence of deviations in auditory feedback on the ongoing phonation of subjects using transformed auditory feedback (TAF)[4]. The TAF experiments showed that subjects compensate for perturbations in the transmission pathway from their mouth to ears.

On the other hand, several models and theories have been proposed regarding the relation between speech production and perception. One of the most well-known theories is the *motor theory of speech perception*[5,6]. The essential point of this theory is that speech sounds are perceived in terms of articulatory gestures that human beings are innately able to produce. This property endows people with a robust capability for perceiving speech sounds. A number of experiments concerned with this issue were conducted, but as yet no common understanding has been achieved. Savariaux *et al.* suggested that speech is controlled partly by some sort of internal representation of the sound and partly, particularly under perturbed conditions, by monitoring

Regular Paper

This research has been supported in part by the National Institute of Information and Communications Technology and in part by a Grant-in-Aid for Scientific Research of Japan (Grant No. 16300053).

the acoustic signal[7]. Honda *et al.* investigated the compensatory responses of articulators to unexpected perturbation on the shape of the hard palate and found that the auditory feedback accelerated the adaptation of articulators to the unexpected changes and adjusted the fine control of the articulators for generating speech sounds[8]. Their results showed that auditory feedback contributes to control of the articulators in speech production. However, the details of how the auditory transformation affects speech articulation need to be elucidated.

Speech production can be considered as a forward process in the speech chain while speech perception is an inverse process. In these two processes, speech production and perception are interlinked with each other at both the motor planning and acoustic transmission levels. A number of new technologies such as functional Magnetic Resonance Imaging (fMRI), Magnetoencephalography (MEG), and Positron Emission Tomography (PET) were developed in the last two decades, and many investigations using those technologies have been carried out to discover the brain functions in speech production and perception[9,10]. Besides brain imaging approaches, electromyography (EMG) is one of the means to reach the function of speech production at a higher level. Honda (1996) used EMG signals of the tongue muscles to investigate the interdependency between speech production and perception mechanisms[11]. He found a common topological mapping between articulatory and auditory patterns, and further suggested that vowel patterns were compatible in both motor and sensory spaces.

In the following sections, this study first investigates the topologies of the vowel description between motor and sensory spaces via a forward process of speech production, using the physiological articulatory model[12]. The interaction between speech production and perception is then investigated using a TAF experiment by applying a perturbation to a speaker in the transmission pathway. A possible communication approach in the brain is discussed in the final part.

2 Hypothesis of Articulatory-Auditory Linkage

This section briefly reviews the hypothesis of articulatory-auditory linkage proposed by Honda based on EMG measurements[11]. The EMG data were obtained for six tongue muscles during the utterance of nonsense phoneme sequences of /əpVp/, and V denotes English vowels[13]. The basic idea of his work was that the position of the tongue body can be represented by the equilibrium of effective muscle forces, which results in a trajectory in the muscle force space. Since muscle force patterns can be inferred from an integrated EMG waveform, the equilibrium can be estimated via an EMG-based muscle force space. According to this consideration, the trajectory of the equilibrium can be constructed based on the EMG data. Such a trajectory is referred to as a motor trajectory hereafter.

Fig.1(a) shows the motor trajectory derived from the EMG-based equilibrium position. The coordinate with about 45 degrees rotation was consistent with the main directions of the first and second components of the tongue movements[14]. The trajectories for these vowels seem to disperse toward the extreme positions in the muscle force space. As the equilibrium of the muscle forces shifts, the tongue body achieves its equilibrium in the articulatory (geometrical) space. These extreme points can be considered as "virtual" motor targets for the vowels, since the targets are not necessarily attained when the motion is very fast or the tongue body is bounded by the wall of the vocal tract. In Fig.1, the motor targets of the four vowels occupy each of the four quadrants of the muscle force space. Consequently, two front vowels /i, æ/ have relatively straight trajectory towards their targets, and two back vowels /a, u/ show quite looped trajectories.

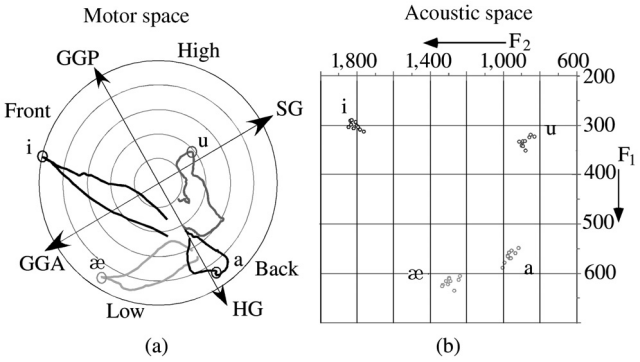

Fig.1. Motor trajectory of the equilibrium position of the tongue in the motor space with reference to the formant patterns. (a) Motor trajectory for the four corner vowels. (b) Distribution of vowel formants in the acoustic space (Modified from [11]).

Fig.1(b) shows F1-F2 patterns for the four vowels measured from the same data using a pitch-synchronous method. The frequency axes are reversed to describe the vowels in the traditional vowel triangle with the high front vowel /i/ in the upper front corner of the plot, and the low back vowel /a/ in the lower back corner of the plot. The interesting finding of this study is that the distribution of the motor targets for the four vowels (Fig.1(a)) roughly coincides with that of vowels in the acoustic space defined by the first and second formant frequencies (Fig.1(b)). This fact is also relevant to the traditional phonetic evidence that the location of the highest point of the tongue dorsum corresponds to the vowel's F1-F2 pattern.

According to the topological consistency in the motor and the acoustic spaces, Honda speculated that the motor-to-acoustic compatibility is not a mere coincidence, but results from the spatial orthogonality observed in both the kinematic and acoustic domains[15]. Adopting this consideration, the curious resemblance of motor and acoustic patterns of the vowels not only pro-

vides physiological evidence that the motor pattern of vowel articulation is compatible with the acoustic pattern, but also suggests an aspect of high-level speech organization in the brain: vowel representations in the motor and auditory spaces are also compatible.

As a general assumption, the high-level motor organization utilizes sensory information as a guide to motor control[6]. This sort of sensorimotor integration typically accounts for visually-aided hand movements, where the object's coordinates in the visual space are mapped into kinematic parameters in the motor space. Although speech articulation does not always undergo the same scheme, there may be a case for the claim that articulatory control is guided with reference to auditory input. The assumption of motor-to-auditory compatibility may not be unreasonable in such a situation. Therefore, at least as far as vowels are concerned, the compatibility of the sensorimotor patterns can be considered as a unique underlying characteristic of speech processes.

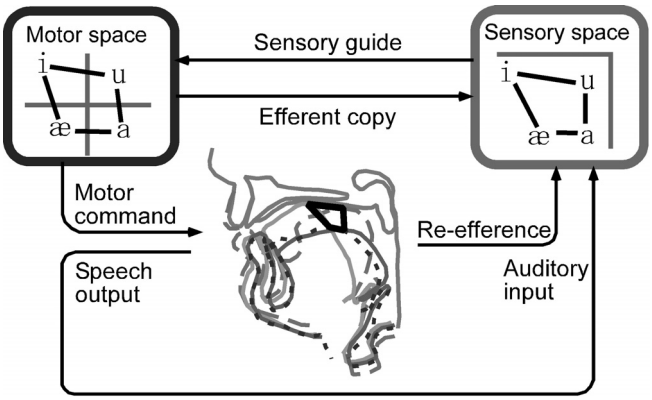

Fig.2. Hypothesis of modular organization of vowels in the brain. Auditory and articulatory representations of vowels show a compatible distribution in both the spaces. (Modified from [16]).

Based on the above analysis, the vowel representations in the motor and auditory spaces arrived at the image shown in Fig.2. The auditory-motor linkage shown by the arrows closes the internal loop of the speech chain, which permits a direct mapping between intended articulatory and auditory patterns. Vowels have analogous motor and sensory representations in the anterior and posterior cortical areas, which connect to each other via a sensorimotor linkage. The backward arrow from the articulatory to auditory spaces is equivalent to the *efferent copy* from motor to sensory areas. In the case of the speech process, the efferent copy is thought to generate articulatorily-induced auditory images, which are evaluated in comparison with somatosensory information as well as with perceived speech sounds. The forward arrow represents the general idea of sensory-guided motor execution, which may be applicable to a situation of auditorily-guided speech production. The looped flow of information between compatible sensory and motor representations further suggests that the two cortical areas for production and perception may work in unison to share the same information. This is what has been argued, in part, by the motor theory of speech perception[6].

3 Simulation of Muscle Activation and Equilibrium Position

The preceding section described the motor-sensory linkage hypothesis on the relation between speech production and perception based on EMG data. EMG measurement succeeds in large articulatory muscles such as the extrinsic tongue muscles, but fails in examining the intrinsic muscles because these muscles are intermingled with the others. Tagged MRI is often used to speculate about the function of the extrinsic and intrinsic muscles of the tongue[17,18], but it is limited because a number of muscles are co-activated, even in executing a simple movement. In contrast, a physiological model of the human speech organs can provide an alternative means to investigate the characteristics of the speech production system. In this section, the relation between the equilibrium position and muscle forces is explored using our physiological articulatory model[12].

3.1 Configuration and Muscular Structure of the Model

A partial 3D physiological articulatory model has been constructed based on volumetric MRI data obtained from a male Japanese speaker. Fig.3 shows the model configuration consisting of the tongue, jaw, hyoid bone, and vocal tract wall. The articulatory model, derived from a previous version[19], achieved a semi-continuum description for the tongue tissue by employing a truss structure of viscoelastic cylinders. The outlines of the tongue body were extracted from the sagittal slices with a 1.0cm interval. Mesh segmentation of the tongue tissue roughly replicates the fiber orientation of the genioglossus muscle. The outline of the tongue body in each plane was divided into ten radial sections that fan out from the genioglossus' attachment on the jaw to the tongue surface. In the perpendicular direction, the tongue tissue was divided into six concentric sections. A 3D mesh model was constructed by connecting the mesh nodes in the midsagittal plane to the corresponding nodes in the left and right planes. Thus, the model represents the principal region of the tongue by a 2cm-thick layer bounded with three sagittal planes. Based on the same MRI data set, the vocal tract wall and the jaw were constructed in 3D with a width of 2.8cm in the left-right dimension.

Fig.4 shows the muscular structure of the model (see [19, 20] for the details). Fig.4(a) shows the genioglossus (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) denoted by the dark lines. Figs. 4(b) and 4(c) show the arrangement of the hyoglossus (HG) and styloglossus (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 bundles are attached to the medial surface of the mandibular body. All the muscles are designed symmetrically on the left and right sides. Fig.4(d) shows three intrinsic muscles of the superior longitudinalis (SL), inferior longitudinalis (IL) and transversus. The transversus runs in the left-right direction, and its distribution

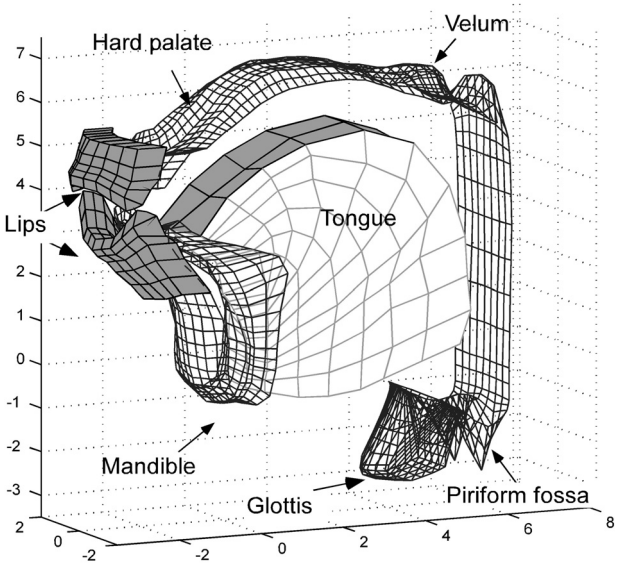

Fig.3. Configuration of the physiological articulatory model.

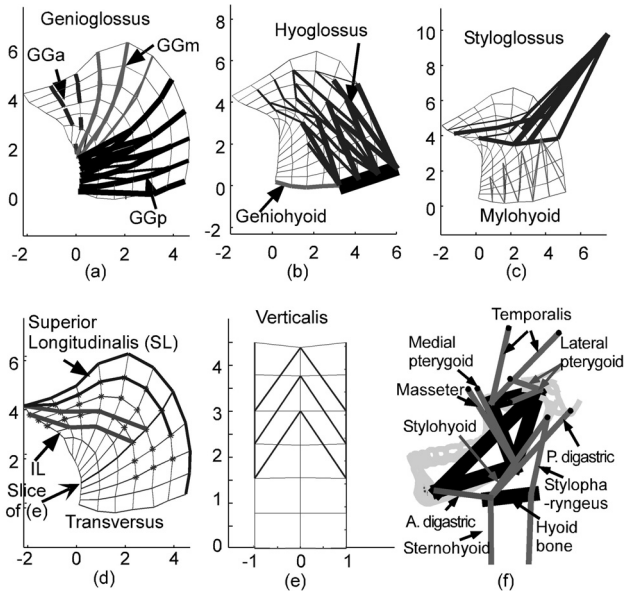

Fig.4. Muscular structure of the model. (a)–(e): Tongue muscles in the midsagittal and parasagittal planes (in cm). (f) Complex of the mandible and hyoid bone.

is plotted in asterisks. Fig.4(e) shows the structure of the verticalis muscle in a cross-sectional view sliced at the 5th section from the tongue floor, denoted in Fig.4(d).

Fig.4(f) shows the model of the jaw-hyoid bone complex. The jaw model consists of four nodes on each side, which are connected by five rigid beams (thick lines) to form two triangles with a shearing-beam. The tongue is attached to the jaw at the mandibular symphysis. The model of the hyoid bone has three segments corresponding to the body and bilateral greater horns. The small circles indicate the fixed attachment points of the muscles. For control purpose, these muscles are separated into two groups: the jaw closer group and the opener group.

3.2 Muscle Forces and Equilibrium Positions

In general, the property of vowels mainly depends on the position of the tongue dorsum, while most of the consonants are shaped by the tongue tip. Accordingly, the tongue tip and tongue dorsum are used to represent equilibrium positions of the tongue. The previous study[12] shows that the equilibrium position (EP) for each muscle shifts monotonically as the activation force increases, where the equilibrium generally achieves in about 300ms after a force is applied. The monotonically shifting EP shapes a trajectory for each muscle.

Fig.5 shows the EP trajectory of the tongue muscles for the tongue tip and tongue dorsum, when each muscle is activated individually by a variable force from zero to six Newtons (N) with eight levels. The HG moves the tongue tip backward and slightly upward as the activation level increases. The SG drives the tongue tip backward horizontally while the GGa draws the tongue tip downward. The tongue tip is driven forward-upward by the GGp and forward-downward by the GGm. The intrinsic muscles SL and IL move the tongue tip largely upward-backward and downward-backward, respectively. These muscle effects indicate that the EP depends on both the muscle and its activation level, but is independent of the previous situation of the tongue. The trajectory built on the EPs can be considered as a "vector" that spreads out from the initial position as the activation level increases. The EP vectors form a stationary coordinate that bridges the motor space and articulatory space. Thus, the EP-based coordinate provides an invariant mapping function between the motor space and articulatory space. The mapping is unique from the motor space to articulatory space, but inverse mapping faces the one-to-many problem.

The right panel shows a number of larger EP vectors for the dorsum. Differing from the tongue tip, the extrinsic muscles (except the GGa) have definitely larger EP vectors than the others. The EP vectors of the GGp and HG show completely opposite directions. This indicates that the GGp and HG work antagonistically in governing the tongue dorsum. Similarly, the GGm and

SG are another antagonistic muscle pair. Therefore, there are two major antagonistic muscle pairs for controlling the dorsum. This simulation supports that basic consideration of tongue muscle function[11] in which the four extrinsic muscles share four quadrants in the rectangular coordinate. The approximate orthogonal relation of the tongue muscles provides another inference of a straightforward mapping from muscle activation to articulation. Since the GGm has the same function as the GGa but is stronger than the GGa for controlling the dorsum, the muscle GGm is used in the latter model simulation instead of the GGa.

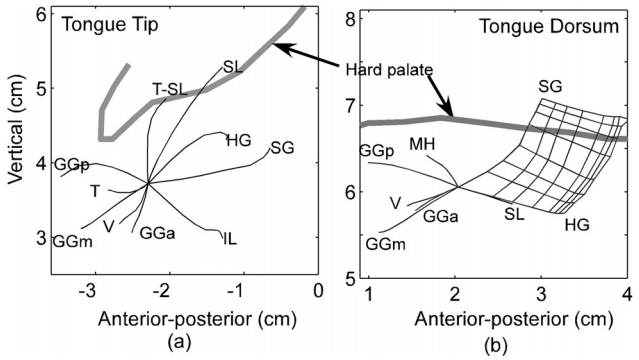

Fig.5. Trajectory of the equilibrium position for each muscle, where the activation force varying from zero to six N. (a) Tongue tip. (b) Tongue dorsum. (The network shown in the right panel consists of the contour lines corresponding to the force levels.)

According to the contribution of the muscles to articulatory movements and to control freedom, five major muscles are taken into account in controlling the tongue dorsum, and nine muscles and one muscle group are used for the tongue tip. A mapping between the motor space and articulatory space is obtained based on the selected EP vectors. Fig.5 shows an example of the map consisting of the EPs of SG and HG in the right panel, where the contour lines of the meshwork correspond to the eight force levels. This map is named the *equilibrium position map* (EP-map). With this EP-map, any arbitrary position inside the region of the map can be reached using the forces interpolated from the contour lines.

3.3 Equilibrium Positions of Co-Contraction

The simplest way to form a tongue shape is to activate two agonist muscles simultaneously, such as use of the EP-map shown in Fig.5. During speech articulation, however, the situation is much more complex because more than two agonist muscles possibly work together to reach a target and/or an agonist-antagonist co-contraction possibly takes place. To clarify the relation between the planning level and the articulatory level, it is necessary to investigate the mechanism of co-contraction. For this purpose, we designed eight three-muscle groups, consisting of one master muscle and a subordinate muscle pair, to generate various possible co-contractions during speech articulation[12].

Fig.6 shows an example of the co-contractions of the muscle groups. The thick dark lines show a part of the EP vectors in the coordinates. The thin dark lines and pale lines illustrate the equilibrium trajectories of the two subordinate muscles in the group. The attachment of the thin lines corresponds to activation levels of the master muscle in the same muscle group. The combination of the SG and the muscle pair of the GGp and SL can move the tongue tip to an apical target by the muscle pair and control the dorsal target by the SG. It is interesting to find that the muscle pair works as a synergistic pair for the tongue tip but as an antagonist pair for the tongue dorsum. If a proper force ratio is kept for the GGp and SL, the tongue tip position can be manipulated while the tongue dorsum remains unchanged. With the mechanism, the dorsum position depends on the activation level of the master muscle and the force ratio of the two subordinate muscles, no matter how strong the force of the subordinate muscles is. This mechanism is capable of realizing a compatible target set for both the tongue tip and dorsum, and of maintaining the stability of a kinematic system.

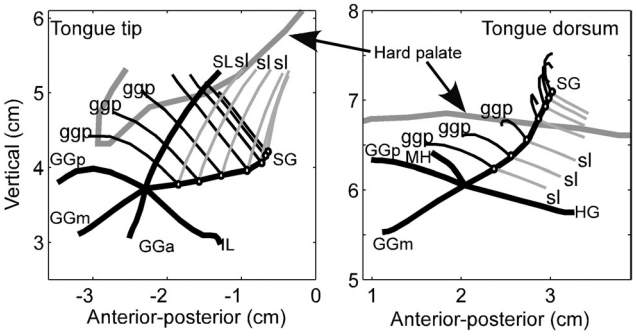

Fig.6. Co-contractions within a muscle group: a master muscle (SG) and a subordinate muscle pair (GGp and SL).

Co-contraction of the agonist and antagonist muscles is the basic muscle activation in speech articulation. The above results show that the co-contraction within such muscle groups enlarges the control flexibility and the degree of freedom of the model while increasing model stability. This suggests that such muscle groups are the basic control unit for the motor commands during speech.

4 Comparisons Among the Motor, Articulatory, and Acoustic Spaces

In reference to the relation between muscle activation and equilibrium position of the articulators discovered above, this section examines the relationship among motor commands, kinematic movements, and the acoustic properties of vowels.

4.1 From Articulatory Target to Motor Command

It is necessary to estimate motor commands from a sequence of articulatory targets in order to realize speech production with the articulatory model. However, such estimation faces the one-to-many problem from articulatory targets to muscle contraction patterns. To solve the problem, we adopt the minimal energy and optimality principle in estimation of motor commands from articulatory targets based on the proposed co-contraction mechanism.

Fig.7 shows an example of estimating motor commands from a given articulatory target using the co-contraction mechanism. The target in this example is the consonantal target /t/ in /ta/, which consists of two features. One feature is crucial for determining the target configuration of the consonant /t/, shown by the filled circle on the tongue tip. The other feature is on the dorsum, shown by the open circle on the right panel, which reflects the coarticulation required by the following vowel /a/, but is not decisive for the consonant. In Fig.7, one can see that the HG is the master muscle that drives the tongue dorsum toward the dorsal target, while the subordinate muscle pair of GGp and SL realizes the apical target. There are three combinations—ggp_1-sl_1, ggp_2-sl_2, and ggp_3-sl_3—which are capable of moving the tongue tip to reach the apical target. The difference among these combinations is that they have different co-contraction levels with the HG, where the activation force of HG is 0.0, 0.1, and 0.2N, respectively. Since all three combinations can guarantee the crucial feature, the decision finally depends on the behavior of the combinations in realizing the indecisive feature of the dorsum. The circled numbers in the dorsal coordinate are the predicted locations for these three combinations. The location of circled 3 is the best of the three expected positions since it is closest to the given dorsal target, shown by the open circle. Therefore, the combination of ggp_3-sl_3 and HG with 0.2N is optimal for the given target, and thus the motor command is determined.

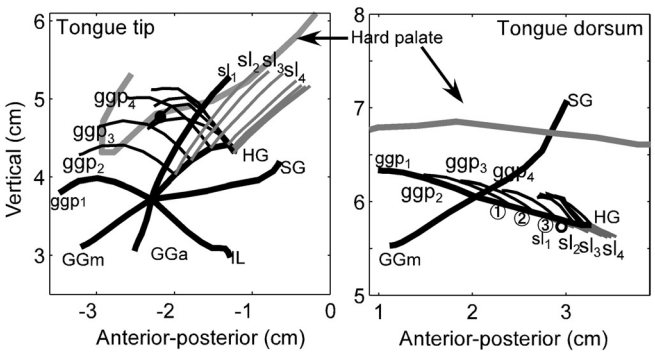

Fig.7. Example to generate motor commands based on muscle co-contraction.

The above process describes how to estimate motor commands for one muscle group. To find the optimal motor command for a given target, above estimation is carried out for all muscle groups in the EP maps, and the square summation of the muscle forces is calculated for each group. According to the minimal energy principle, the muscle group with the smallest force summation is judged as the optimal one for the given target. Note that although this process of estimating motor commands is feasible, the computational cost seems too high for a human to realize it in real time. For this reason, it is likely that human beings may use a more efficient strategy to determine motor commands for a given target based on a "knowledge database" which has been filled with trained motor commands. However, the above approach may be adopted in building the knowledge database for acquiring a new language.

4.2 Motor Command Estimation for Dynamic Components

The motor command estimated by the above approach is dependent only on the given target, but independent of either the current positions or the past trajectory. Since the command does not vary until the articulation is completed, such a command-dependent force is referred to as a *static force*. In a dynamic system, it is known that muscles are not always activated according to the minimal energy principle. It is conceivable that all of the muscles with positive contributions are activated at the same time to achieve a given target quickly. Thus, muscle activation is also concerned with both the target and the current situation of the kinematic system. In addition to the static force, therefore, a *dynamic force* must be taken into account of model control. For this purpose, we have proposed a so-called *muscle workspace*, and reduce the distance from an arbitrary position to its target stepwise via the muscle workspace[19].

Fig.8 shows a simplified muscle workspace of the dorsal control point. This muscle workspace consists of muscle vectors of the four extrinsic tongue muscles, shown by the thick dark arrows, where the muscle vectors correspond to the displacement which resulted from exciting each muscle individually using a unit activation with a fixed duration. Since the muscle workspace is compatible with the geometrical space, the mapping 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 components 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 (motor commands) can be obtained for any arbitrary movement using this approach.

In Fig.8, *Pc* denotes the current position of the con-

trol point and *Tg* is the target position. The dashed line from *Pc* to *Tg* forms a vector, referred to as an *articulatory vector*. When the articulatory vector is mapped onto the muscle workspace, a set of projections is obtained for the muscle vectors. Based on the penalty function, the result arrives at that only the positive projections should be taken into account in the force generation, while the negative ones are to be ignored. From a physiological point of view, this result implies that a muscle will be activated only when the muscle contraction contributes positively to the movement towards the target (see [19] for details).

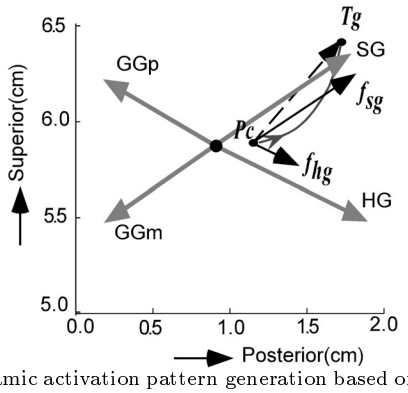

Fig.8. Dynamic activation pattern generation based on the muscle workspace.

4.3 Comparisons Between Motor, Articulatory, and Acoustic Spaces

This section investigates the linkages among the motor, articulatory, and acoustic spaces via the forward process of speech production from an articulatory target to acoustic output using the physiological articulatory model. Vowel sequences of /aui/ and /aiu/ are used in this simulation.

According to the given target of the vowel sequences, static forces are estimated based on the EP-map, and dynamic ones are obtained via the muscle workspace. Fig.9 shows the muscle force trajectory for utterances /aui/ and /aiu/, where the muscle forces were normalized individually. As mentioned in the previous section, we use the muscle GGm instead of GGa, because GGm is more powerful than GGa in governing the tongue dorsum. Although the EP vectors of GGp-HG pair and GGm-SG pair were not exactly orthogonal, for simplification these four muscle vectors were plotted as an orthogonal coordinate. In the simulation, the trajectory for /u/ shows a more complicated shape than that of the others. This tendency is the same as that seen in the EMG data shown in Fig.1(a). Fig.10 shows articulatory trajectories that were generated using the forces shown in Fig.9. In the case of /aui/, a quite looped motor trajectory is seen in simulation and in the EMG data, but is not seen in the articulatory trajectory.

To produce synthetic sounds, cross-sectional area functions of the vocal tract were derived from the articulatory movement and implemented in a transmission line model[19]. Formants of the synthetic sounds of /aui/ and /aiu/ were calculated for each 10ms. Fig.11 shows the acoustic patterns consisting of F1 and F2 for vowels /a/, /i/ and /u/ out of the utterances /aui/ and /aiu/. Some differences in the formants are seen in the transition between the vowels.

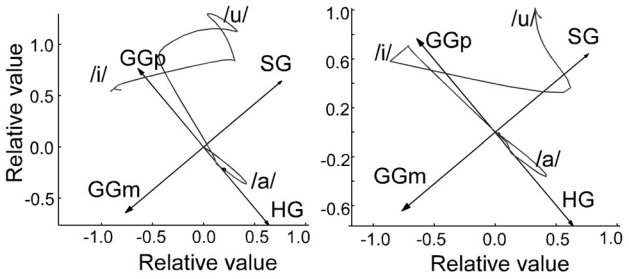

Fig.9. Motor trajectories for utterances /aui/ and /aiu/, where the forces were normalized individually.

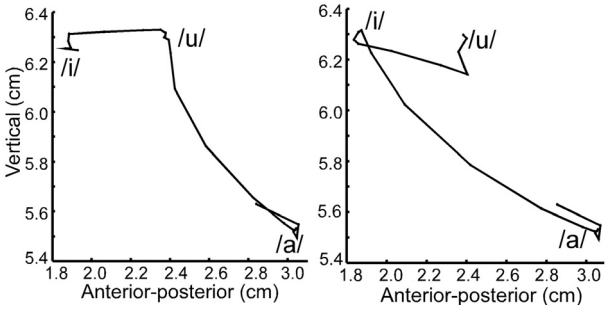

Fig.10. Articulatory trajectories of the dorsum for utterances /aui/ and /aiu/, where the dimension is cm.

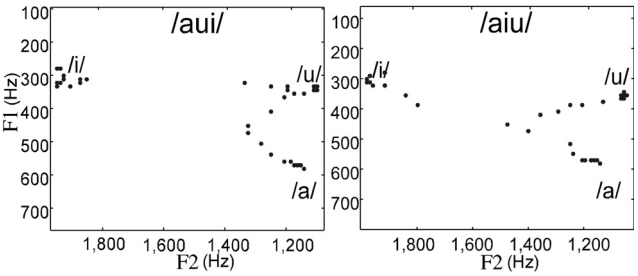

Fig.11. Acoustic patterns of the first formant (F1) and second formant (F2) for utterances /aui/ and /aiu/.

By means of the three figures, we illustrate an image of the speech production process in the motor space, articulatory space and acoustic space. In the EMG data and model simulation, the motor trajectory for /u/ is quite looped. In the acoustic space shown in Fig.11, one can see that the formant trajectory is looped for the /aui/ sequence, while this is not seen in the articulatory space. Accordingly, the motor trajectory in the motor space and the formant trajectory in the acoustic space show a higher similarity in the topologies than the others. This might suggest that the mapping between the motor area and auditory area in the brain be more

likely straightforward rather than via the articulatory space.

5 Interaction Between Speech Production and Perception

In the above sections, the linkages among motor and sensory spaces were examined using a forward process from the motor planning level to the acoustic level via articulation level. This section verifies the linkage by a feedback process using a transformed auditory feedback (TAF) approach.

It is commonly understood that in acquiring a new language, especially before an infant has any practice producing speech sounds, the primary mode of control should be a feedback-based strategy through a loop between speech production and perception. The *motor theory of speech perception* suggests that the acoustic signal is perceived in terms of articulatory gestures[6], while the other study emphasizes that the feedforward system becomes practiced over time after the speech skill is acquired, and that the need for feedback-based control is finally eliminated[21]. So far, a number of experimental studies have been conducted to verify this issue. However, the majority of them focus on a long-term learning effect [cf. 20, 22]. To our knowledge, there has been no report on the function of simultaneously monitoring of speech perception. This experiment is an attempt to clarify whether or not speech production is controlled with reference to the perception monitoring.

5.1 Configuration of the TAF Experiment

To obtain concrete evidence regarding interactions between speech production and perception, EMG signal and video media as well as acoustic recordings were employed in the experiment to investigate different aspects of the interaction. Fig.12 shows the configuration of the experimental setup. The purpose of using the EMG and video media is to discover whether or not subjects have compensated for the TAF, but for some reason that compensation does not appear in the acoustic output. To clarify this situation, the Chinese vowels /i/ and /ü/ ([i] and [y] in IPA) were chosen as the speech material, because the articulation difference in these vowels, with or without lip protrusion, can be easily observed. Physiologically, there are three major muscles, the depressor labii inferioris, zygomatic major, and buccinator, involved in such articulations. Activations of these three muscles can be observed using a surface EMG method. To that end, the above muscles were measured using EMG signals with five channels.

Except for the lip protrusion in producing [y], these two vowels have about the same vocal tract shape. Acoustically, the first formant (F1) is consistent with these two vowels within 6%. The second formant (F2) for [y] is about 15% lower than that for [i], while the third formant (F3) of [y] is 25% lower than [i]. One Chinese subject participated in this experiment, who was born in eastern China and has no history of any speech or auditory disease. In this experiment, the subject was asked to maintain a stable pronunciation of the vowel [i] for 5s. In 70% of the trials, the feedback sound was randomly transformed from [i] to [y] at the time of 2.2s from the beginning of the trials.

5.2 Experimental Results

Fig.13 shows examples of EMG measurements with and without TAF in the right and left panels, respectively. The upper panels show the EMG signals and the lower panels are their envelopes. The EMG signals were obtained from the depressor labii inferioris, which shows strong activity in producing Chinese [i]. In the case without TAF, the EMG signals show large amplitude at the beginning and then gradually decrease. In the other words, the depressor labii inferioris is activated to form the lip shape for /i/, and then the activation is

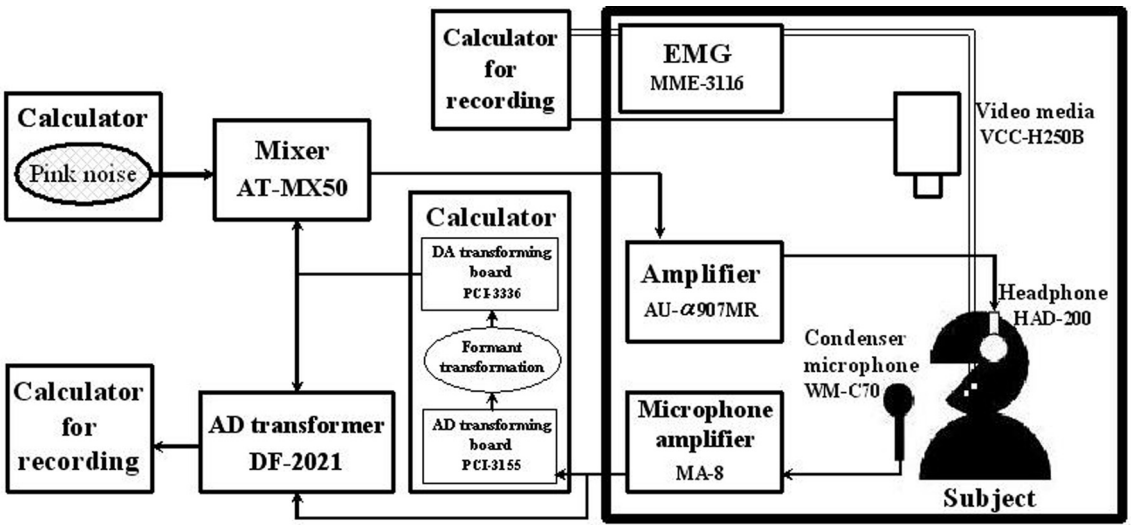

Fig.12. Configuration of the transformed auditory feedback experiment system.

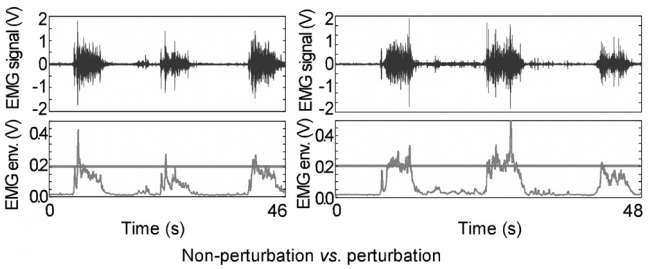

Fig.13. Measurements of the EMG signals without TAF (left) and with TAF (right).

reduced gradually for maintaining the configuration alone. In the case with TAF, the perturbation was applied at 2.2s from the beginning of the trial. The activation of the muscle is strong at the beginning, gradually weakens, and then gets stronger again after the TAF is applied. It seems that when the feedback sound deviates from [i] due to the TAF, the subject attempted to maintain an [i]-like configuration by increasing the activation of the muscle to compensate for the perturbation. This suggests that speech production be controlled with reference to perception monitoring. This interaction was seen in 71% of the trials with the transformed auditory feedback.

Acoustical analysis was also carried out in this experiment. The formant difference between the trials with and without TAF was calculated for the first three formants and is shown in Fig.14. As seen in this figure, there is a large difference in the second formant. In the transformation process from [i] to [y], the frequency is reduced by 220Hz for F2 and 750Hz for F3. When the transformed sound is fed back to the speaker, F2 increased by about 70Hz in his speech sound. This increment clearly shows that the speaker endeavored to maintain the "monitored" sound correctly by compensating for the formant reduction. However, F3 does not show any significant compensation for [y] concerned with the transformation.

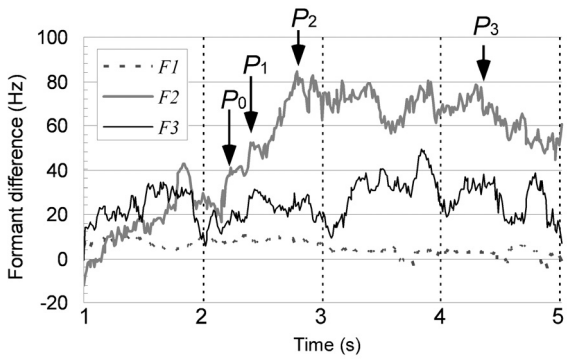

Fig.14. Formant difference caused by the TAF. P_0 is the time applying the TAF. P_1 and P_3 are the start and end points of the compensation. P_2 is the point of the maximal compensation.

In Fig.14, P_0 indicates the timing, 2.20s, at which the TAF was applied. Focusing on the changes in F2, P_1 at 2.35s seems to be the start point of the compensation for the TAF. It took about 150ms from application of the TAF to observing an obvious change in F2. This reaction time is reasonable considering the auditory processing time. The maximal compensation of F2 occurred at about 2.74s. It was about 290ms from the start to the maximum compensation. This time is about the same as the latent time for speech articulation, which is about 300ms [cf. 23]. When the TAF was terminated at 4.4s shown by P_3, the formant was restored monotonically. This implies that speech production is controlled ordinarily by monitoring the acoustic signal.

6 Discussion and Summary

In this paper, we start from an EMG-based hypothesis in which speech communication in human brain is carried out efficiently via a topological mapping between the motor and sensory spaces[11]. Descriptions of speech production were demonstrated in the motor and acoustic levels using a model-based forward process. Auditory monitoring of speech production was examined by a transformed auditory feedback experiment. The communication between speech production and perception within the brain is discussed based on that investigation.

6.1 Summary of the Model Simulation

The physiological articulatory model used in this simulation is constructed based on volumetric MRI data and anatomic data, which can realize the speech production process from articulatory target to acoustic output[12]. The simulation demonstrates that each muscle drives the articulators (the tongue and jaw) toward an equilibrium position (EP) corresponding to the magnitude of muscle activation, and the EP monotonically shifts as muscle activation increases under the condition that the other muscles' activation is unchanged. The EP depends only on the level of muscle activation but is independent of past articulatory movements and deformations. This suggests that there exist an invariant map between the motor space and articulatory space. It can be reasonably considered that the invariant map of motor space to articulation is the basis for the communication between speech production and perception inside the brain.

To realize the above mapping in a realistic articulation task, co-contraction of the agonist and antagonist muscles must be considered in speech production. So far, it is not clear how human beings control a complicated co-contraction in a muscle-woven mass such as the tongue body. This study demonstrates an efficient way to realize the muscle co-contraction, and also demonstrates plausible articulatory activities when implementing the co-contraction in the model. Using the proposed co-contraction mechanism, the tongue tip and tongue dorsum can be controlled independently to some extent, and the flexibility of the model and the degree of freedom for model control are enlarged. Accordingly,

it might be reasonable to speculate that human beings adopt such an efficient approach in speech articulation.

To estimate motor commands for articulatory targets, all possible combinations of muscle activations are investigated in the EP maps according to the given articulatory target, and the optimal one is identified based on the minimal energy principle. This process for estimating motor commands is feasible but the computation may be too heavy for human beings to accomplish in real time. It is likely that for a given target human beings may use a more efficient strategy to determine a motor command based on their "knowledge database", while the proposed approach is possibly adopted in building the knowledge database for acquiring a new language.

This simulation demonstrates a speech production process from the given articulatory target through the motor space, kinematic space, to the acoustic space. It is verified that the three spaces have compatible topologies in describing speech sounds, at least for vowels. Among the three topologies, the motor space and acoustic space show higher similarity. This might suggest that the mapping between the motor area and auditory area in the brain be most likely straightforward but not always via the articulatory space. This result basically supports the hypothesis that speech communication in human brain is carried out in an efficient way via a topological mapping between the motor and sensory areas[11].

6.2 Summary of the TAF Experiment

In the TAF experiment, we employed electromyographic (EMG) signals, image information, and acoustic analysis to acquire more evidence for an instantaneous linkage between speech production and perception. Chinese vowels [i] and [y] were chosen as the speech materials since their articulatory differences can be observed from outside. The surface EMG was used to capture muscle activations in forming the major articulatory difference in the lip gesture. The speaker was asked to maintain vowel [i] for 5s, while feedback sound was randomly transformed from [i] to [y]. The EMG signals showed that the speaker attempted to emphasize [i] when the feedback sound was transformed. This emphatic effort was seen in 71% of the trials with the auditory perturbation.

Acoustic analysis also shows the compensation of the speech production side for the auditory perturbation. In the transformation from [i] to [y], the frequency was reduced by 220Hz for F2 and 750Hz for F3. When this TAF was applied in the transmission pathway, F2 of [i] increased by about 70Hz in the acoustic output. This increment clearly shows that the speaker maintains the vowel property when the F2 of perceived sound is lower than that of the expected one. In other words, the speaker attempts to compensate for the "error" in speech production immediately with reference to the auditory monitor. It implies that auditory monitoring normally takes place during speech production.

As shown in the TAF experiment, speech production is controlled with perception monitoring. This implies that human speech processing is a closed-loop control in which articulatorily-induced auditory images and sensory-guided motor execution communicates in the brain during speech. In contrast, the processing in both speech synthesis and recognition systems is an open-loop control, the disadvantage of which is a lack of sensitivity to the dynamics of the system. With reference to the human mechanism, it is essential to construct a speech processing system with a closed-loop control. To achieve this goal, it is necessary to develop efficient descriptions of speech features and a powerful matching method, similar to that proposed in this study, so that both speech synthesis and recognition systems can be formed by the same processing mechanism. In this study, however, we could not devise a concrete image for an efficient communication approach. Discovering and implementing the communication approach between speech production and perception in human brain is still a challenging topic remaining for future studies.

Acknowledgements The authors would like to thank Rie Matsuoka and Xu-Gang Lu for their contribution to the TAF experiment, and also thank Emi Murano for her instruction in the EMG experiment. The authors very appreciate the valuable comments from Donna Erickson.

References

[1] Denes P, Pinson E. The Speech Chain. 2nd Edition, New York: W.H. Freeman and Co. 1993.

[2] Lombard E. Le signe de l'elevation de la voix. Annales Maladies Oreilles Larynx Nez Pharynx, 1911, 37: 101–119.

[3] Lee B S. Effects of delayed speech feedback. *J. Acoust. Soc. Ame.*, 1950, 22: 824–826.

[4] Kawahara H. Interactions between speech production and perception under auditory feedback perturbations on fundamental frequencies. *J. Acoust. Soc. Jpn.*, 1994, 15(3): 201–202.

[5] Liberman A M, Cooper F S, Shankweiler D P, Studdert-Kennedy M. Perception of the speech code. *Psych. Rev.*, 1967, 74(6): 431–461.

[6] Liberman A M, Mattingly I G. The motor theory of speech perception revised. *Cognition*, 1985, 21: 1–36.

[7] Savariaux C, Perrier P, Orliaguet J. Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *J. Acoust. Soc. Ame.*, 1995, 98(5): 2428–2442.

[8] Honda M, Fujino A, Kaburagi T. Compensatory responses of articulators to unexpected perturbation of the palate shape. *J. Phonetics,* 2002, 30: 281–302.

[9] Nota Y, Honda K. Brain regions involved in control of speech. *Acoust. Sci. & Tech.*, 2004, 25(4): 286–289.

[10] Sakai K L, Homae F, Hashimoto R, Suzuki K. Functional imaging of the human temporal cortex during auditory sentence processing. *Am. Lab.*, 2002, 34: 34–40.

[11] Honda K. Organization of tongue articulation for vowels, *J. Phonetics*, 1996, 24: 39–52.

[12] Dang J, Honda K. Construction and control of a physiological articulatory model. *J. Acoust. Soc. Ame.*, 2004, 115(2): 853–870.

[13] Baer T, Alfonso J, Honda K. Electromyography of the tongue muscle during vowels in /əpvp/environment. *Ann. Bull. R. I. L. P., Univ. Tokyo,* 1988, 7: 7–18.

[14] Maeda S. Compensatory Articulation During Speech: Evidence from the Analysis of Vocal Tract Shapes Using an Articulatory Model. Hardcastle, Marchal Speech Production and Speech Modeling, Dordrecht: Kluwer Academic Publishers, 1990, pp.131–149.

[15] Carré R, Mrayati M. Articulatory-Acoustic-Phonetic Relations and Modeling, Regions and Modes. Speech Production and Speech Modeling, Hardcastle W, Marchal A (eds.), Netherland: Kluwer Academic Publishers, 1990, pp.211–240.

[16] Honda K, Kusakawa N. Compatibility between auditory and articulatory representations of vowels. *Acta Otolaryngol. (Stockh)*, Suppl., 532: 103–105.

[17] Niimi S, Kumada M, Niitsu M. Functions of tongue-related muscles during production of the five Japanese vowels. *Ann. Bull. R. I. L. P., Univ. Tokyo,* 1994, 28: 33–40.

[18] Stone M, Davis E, Douglas A, Ness Aiver M, Gullapalli R, Levine W, Lundberg A. Modeling motion of the internal tongue from tagged cine—MRI images. *J. Acoust. Soc. Am.,* 2001, 109(6): 2974–2982.

[19] Dang J, Honda K. Estimation of vocal tract shape from sounds via a physiological articulatory model. *J. Phonetics,* 2002, 30: 511–532.

[20] Houde J, Jordan M. Sensorimotor adaptation in speech production. *Science,* 1998, 279(5354): 1213–1216.

[21] Callan E, Kent D, Guenther H, Vorperian K. An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system. *Journal of Speech, Language, and Hearing Research,* 2000, 43: 721–736.

[22] Purcell D, Johnsrude I, Munhall K. Perception of altered formant feedback influences speech production. In *Proc. ISCA Workshop on Plasticity in Speech Perception*, London, UK, 2005, pp.15–17.

[23] Masaki S, Honda K. Estimation of temporal processing unit of speech motor programming for Japanese words based on the measurement of reaction time. In *Proc. ICSLP 94*, Yokohama Japan, 1994, pp.663–666.

Jianwu Dang received his B.E. and M.S. degrees from Tsinghua Univ., China, in 1982 and 1984, respectively. He worked for Tianjin University as a lecturer from 1984 to 1988. He was awarded the Ph.D. Eng. from Shizuoka Univ., Japan in 1992. Dr. Dang worked for ATR Human Information Processing Lab., Japan from 1992 to 2001. He joined the University of Waterloo, Canada, as a visiting scholar for one year in 1998. He has been with the Japan Advanced Institute of Science and Technology (JAIST) since 2001, where he is a professor. He joined the Institute of Communication Parlee, Center of National Research Scientific (CNRS), France, as a research scientist the first class for one year in 2002. His research interests are in all of the fields of speech science, especially in speech production. He is a member of the Acoustic Societies of America and Japan, and also a member of the Institute of Electronics, Information and Communication Engineers.

Masato Akagi received the B.E. degree in electronic engineering from Nagoya Institute of Technology in 1979, and the M.E. and the Ph.D. Eng. degrees in computer science from Tokyo Institute of Technology in 1981 and 1984. In 1984, he joined the Electrical Communication Laboratory, Nippon Telegraph and Telephone Corporation (NTT). From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with the School of Information Science, JAIST, where he is currently a professor. His research interests include speech perception mechanisms of humans, and speech signal processing. Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, and the Sato Prize for Outstanding Paper from the ASJ in 1998.

Kiyoshi Honda graduated from Nara Medical University in 1976 and joined the Faculty of Medicine at the University of Tokyo to work in the voice clinic and conduct speech research. He was also a visiting scholar at Haskins Laboratory, New Haven, for three years from 1981. He was awarded a Ph.D. degree in medical science in 1985. Dr. Honda moved to Kanazawa Institute of Technology in 1986 and continued speech research as an associate professor. He then moved to ATR in 1991 to be the supervisor of the Auditory and Visual Processing Research Laboratories and Human Information Processing Research Laboratories. He was also a senior scientist in the field at the University of Wisconsin for three years from 1995. Currently he is the head of Department of Biophysical Imaging (speech production group) at ATR Human Information Science Laboratories. His research work focuses on speech science and physiological experimental phonetics using MRI to investigate the form and function of the speech organs.