A Study On Construction And Control Of A Three-Dimensional Physiological Articulatory Model For Speech Production

by

Qiang FANG

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Jianwu DANG

School of Information Science
Japan Advanced Institute of Science and Technology

March, 2009

Abstract

In the literature, a number of speech scientists tried to reveal the mechanism of speech production based on observed acoustic signals and/or articulatory movements, and proposed a number of theories on the mechanism of speech production. However, most of the speech production theories ignore the motor commands that drive articulators to produce articulatory movements and speech, and the activities of the central neural system that generate motor commands in light of the linguistic representation. This causes a gap between linguistic representation and physical realization of speech in those theories. If we can uncover motor commands and activities of central neural systems in speech production, we may bridge the gap between linguistic representation and physical realization. Here, we follow a bottom-up framework to uncover motor commands and activities of central neural system in speech production. First of all, we need to uncover motor commands from observed articulation; then, we need to uncover activities of central neural system in speech production based on the uncovered motor commands and the linguistic representation of speech. In this thesis, we focus on recovering motor commands involved in speech production, especially in vowel production. Because observed articulations are generated by manipulating speech organs with motor commands, which are difficult to be observed directly, it is almost impossible to know the underlying motor commands if there is no good understanding of the biomechanical properties of speech organs. In this thesis, we uncover the motor commands of vowel production based on a 3D physiological articulatory model that inherently models the biomechanical properties of speech organs. For this purpose, firstly, we constructed 3D jaw and vocal tract wall, and combined them with a 3D tongue model to form a 3D physiological articulatory model, which replicates morphological structure and musculature of the supra-glottal system. To make the proposed model more realistic, the orientation of muscles SG and IL in the 3D tongue are refined according to their anatomical descriptions. In addition, a module named contact handling is incorporated into the physiological articulatory model to deal with the contact between tongue and jaw, and tongue and vocal tract wall. Preliminary evaluation showed

that the model behaves properly when associated muscles are activated. Since the tongue and jaw in the model are driven by associated muscles, it is necessary to understand the detail function of those muscles, especially tongue muscles. Here we quantitative analyze the function of tongue muscles by using the proposed 3D physiological articulatory model to shed light on the general function of individual tongue muscles and the agonist-antagonist properties of tongue muscle pair. The results show that (1) the function of muscle GGa, GGm, GGp, SG, and MH for the movement of both tongue tip and dorsum are consistent with the speculation based on the anatomical orientations; (2) the muscles (T, V, and SL) located in the superficial layer of the tongue contribute most to deformation of the tongue surface; (3) muscle pairs GGm-SL, GGm-HG, GGA-HG act as antagonist muscle pairs for tongue tip, while as agonist of tongue dorsum; (4) muscle pairs GGp-HG, GGp-SL, GGm-SG act as the antagonist muscle pairs for tongue dorsum while act as agonist pairs for tongue tip. After constructing the 3D physiological articulatory model, and evaluate the function of tongue muscles, the 3D physiological articulatory model is implemented to uncover the motor commands of observed isolated vowels. This is done by driving the articulatory model to approximate the observed vowel articulation via an optimization procedure which aims to minimize the 3D difference between simulation results and observations. It shows that the model is able to generate the specific articulations of observed vowel production, and the estimated muscle activations are consistent with those observed from EMG experiments. To account for the vowel articulation with variation and vowels in continuous speech, we elaborated a feedforward control strategy, which maps articulatory target to muscle activations via intrinsic representation of articulatory posture by using General Regression Neural Network. The results show that this method can control the proposed 3D physiological articulatory model with high accuracy. Therefore, it is possible to uncover the articulatory targets of vowels by comparing the model output and observed articulatory movements, and further to uncover the motor commands of vowels in continuous speech by using the feedforward control module.

After constructing the 3D physiological articulatory model, the biomechanical properties of tongue and jaw and the interaction between the tongue and surround structure in speech production are properly modeled. Based on the model, it is easily to uncover the underlying motor commands of vowel production from observed vowel articulation

by using the elaborated feedforward control module for the proposed 3D physiological articulatory model .

Acknowledgments

It would not have been possible to finish this doctoral thesis without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

Fist of all, I would like to give my sincere thanks to my supervisor, Prof. Jianwu DANG. This thesis would not have been possible without his help. His ideas and insightful comments always inspire me to explore new stuffs in my research. His insistent willingness to read my drafts, and to point out ideas and arguments often deeply obscured in my manuscripts had a deep impact on my development as a researcher. My research endeavors would not be successful without him.

I would like to appreciate Professor Masato AKAGI, Associate Professor Isao TOKUDA and Associate Professor Masashi UNOKI of JAIST, and Associate Professor Kohichi OGATA of Kumamoto University. I would like to acknowledge you all for serving as my defence committees. I am really grateful for all of you to take time to reading this thesis and making constructive suggestions.

I would like to give my thanks Assistant Professor Atsuo SUEMITSU for helping me get papers necessary for my research and Previous Assistant Professor Xugang LU of IIPL Lab in JAIST for his frequent discussions and suggestions throughout my research.

I would like to acknowledge the financial support from Professor Takuya Katayama, Japan Advanced Institute of Science and Technology, particularly in the award of a Doctoral Studentship that provided the necessary financial support for this research.

I also will express my thanks to previous and current members in IIPL laboratory for various discussions and suggestions on my research, and to all my friends and staff members in Japan Advanced Institute of Science and Technology (JAIST) for their help, support and friendship. I really appreciate the experiences that I shared with them over the last three years in JAIST.

I also will give my thanks to Prof. Aijun LI, and Prof Zongji WU who led me to the interesting and exciting research field, and offer me a position to continue the work. Es-

pecially, I will give my thanks to Prof. Aijun LI, without her support and encouragement, it is almost impossible to come to JAIST and spend three years to conduct the interesting work here.

Last but not the least, I would like to thank my wife, Ms. Haibo WANG, for her personal support and great patience at all times. I am highly grateful to my wife for her endless love and encouragement, which gave me enormous moral support and helped me go through the difficult times to complete my work. My parents and sister have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Speech, in fact, is one of the few basic abilities that set us apart from other animals and are closely connected with our ability to think abstractly [1]. It makes great influences on the development and functioning of human society and makes development of human culture possible by our ability to share experiences, to exchange ideas, and to transmit knowledge from one generation to another. Hence, the understanding of speech will provide useful insights into the nature and history of human civilization. Moreover, in the modern life of human, people frequently communicate with each other by communication systems and with machines. Therefore, the understanding of speech will help communication engineers developing more efficient communication systems, and help speech engineers provide more friendly interfaces for human-machine communications, so as to improve the qualities of life.

1.1 Background

"When most people consider speech, they think only in terms of moving lips and tongue ... In reality, speech is a far more complex process, involving many more levels of human activity, ..." [1]. In this part, we first have a look on the process of human communication with speech from functional point of view. Then, some of the theories of speech production are reviewed.

1.1.1 Speech chain in human communication

Denes *et al.* [1] described the communication process between two persons with speech from functional point of view. Figure 1.1 gives a schematic description of this process. In the communication process, typically there is a speaker and a listener. The speaker, firstly, arrange his intension, and put that into linguistic representation by selecting proper words and phrases to express its meaning, and by placing these words in the order required by the grammatical rules of the language. This process is associated with activities in the speaker's brain. Then, appropriate motor instructions, in the form of impulses along the motor nerves, are issued from brain to the muscles that activate the vocal organs, e.g. the lungs, the vocal cords, the tongue, and the lips. Therefore, proper sound sources and vocal tract configurations are generated to produce speech. Meanwhile, the speaker pick up auditory as well as somatosensory feedbacks to compare the generated articulatory movements and sounds with those they intended to produce, and make the adjustments necessary to match the results with their intentions. This process is referred to as speech production.

The generated speech sound wave travels through the air between speaker and listener. Pressure variations at the ear activate the listener's hearing mechanism and produce impulses that travel along the sensory nerves to the listener's brain. In the listener's brain, a considerable amount of nerve activities are taking place to translate the nerve impulses arriving from the ear to make the understanding of the speaker's intension. This process is called speech perception.

Speech production and speech perception are two of the most important aspects in the process of speech communication. In the present thesis, we would like to focus on speech production. As shown in Figure 1.1, the speech production process recruits the events occurred in several levels. In the brain, the transmission of a message begins with the selection and ordering of suitable words and sentences. This is attributed to linguistic level. In the physiological level, the central neural system issues neural impulses to activate associated muscles that drive the speech organs. In the physical level, the speech organs are driven by the muscle activations, and consequent speech sound is generated and transmitted to convey the intensions of the speaker to the listener. Therefore, speech production involves activity in at least three levels: linguistic level, physiological level,

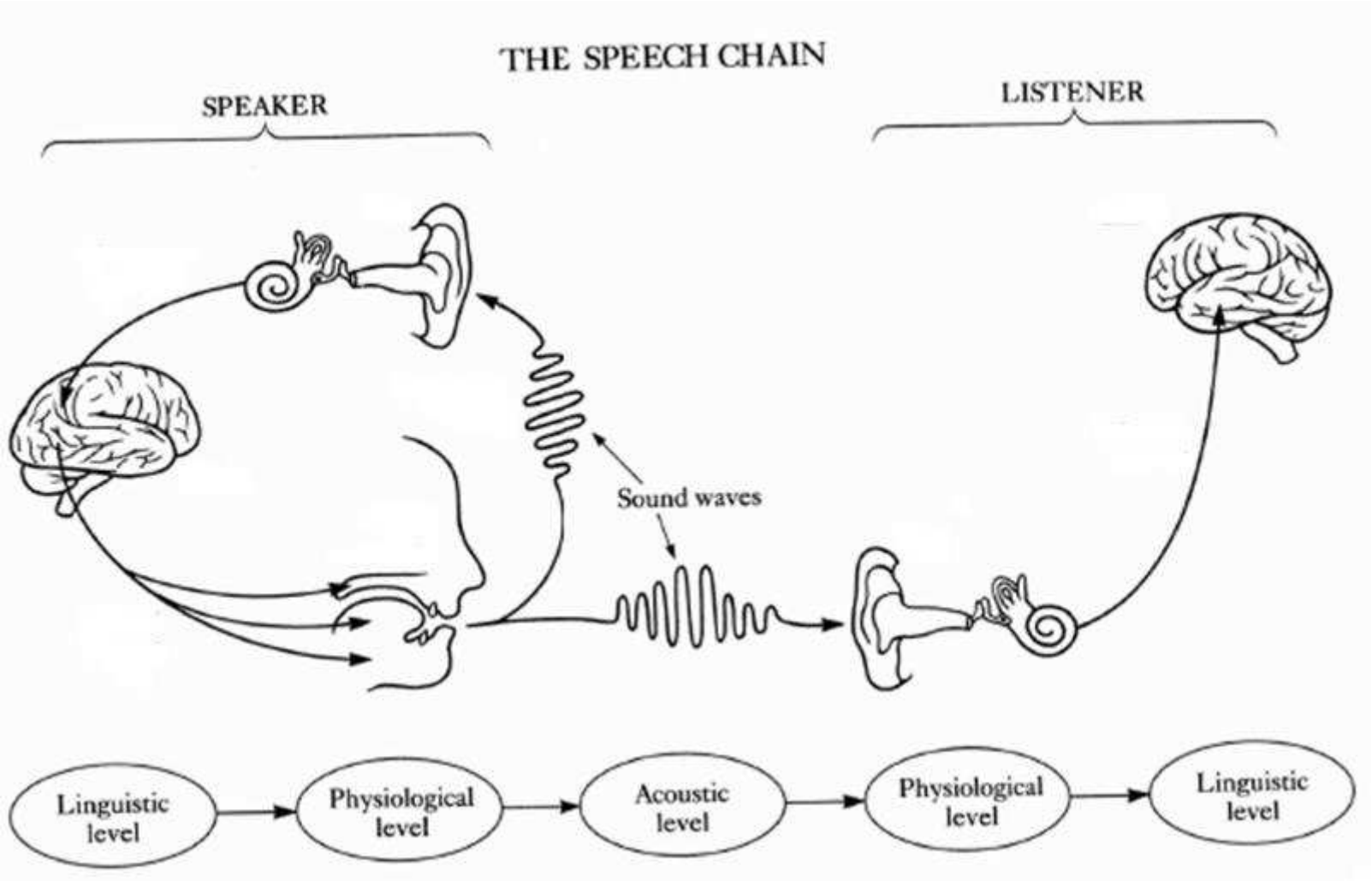

Figure 1.1: Speech chain in human communication with speech (After Denes and Pinson [1]).

and physical level. The comprehension of the activities involved in those three levels will help people to shed light on the detail mechanism of speech production.

1.1.2 Speech production theories

To understand the mechanism of speech production, speech scientists tried to answer some basic questions based on surface articulatory/acoustic observations: (1) What is the underlying targets of speech segments? (2) How the surface articulatory movements and speech sounds are rendered from the underlying targets of the segments that comprise utterances? In the literature, a number of theories have been proposed to explain the mechanism of speech production based on the surface acoustic/articulatory observation. Those theories can be classified into three categories: distinctive feature based theory of speech production, articulatory gesture based theory of speech production, and physical realization theory of speech production.

Distinctive feature based speech production theories

The feature based theories of speech production are the first attempt to explain the mechanism of speech production. In those theories, they took the distinctive features (proposed by Jakoboson *et al.* [2] and Chomsky *et al.* [3]), which are discrete, timeless, and thought to be related with cognition, as the targets for speech segments. In Chomsky and Halle's theory [3], the distinctive features of the segments that comprise an utterance are assimilated and mapping to phonetic representations according to phonological rules. Daniloff and Hammarber [4] adopt the basic idea of Chomsky and Halle. They also specified the targets of a phoneme as the combination of a set of the cognitive oriented distinctive features. The studies of Daniloff and Moll [5] on labial coarticulation and of Moll and Daniloff [6] on velar coarticulation revealed that the rounding of lip and the lowering of the velum can start two, three or even for segments before the influencing one. Those patterns clearly indicated the anticipatory coarticulation which cannot be the product of physical properties of articulatory system. Based on these observations, Daniloff and Hammarber [4] proposed the "feature spreading theory" which accounts for the rendering

of surface observations from distinctive features of individual segments. In the theory, the look-ahead mechanism assigns the coarticulated feature to all preceding unspecified segments and therefore coarticulation extends in time as a function of the number of these segments; the anticipatory coarticulation is blocked by segments specified for a feature that contradicts the spreading feature; unspecified segments acquire the spreading feature. That kind of proposal was able to explain their observed articulatory movement patterns of labial and velar. And the studies of Benguerel and Cowan [7], Lubker [8] on labial coarticulation, fully support the "feature spreading theory". Keating [9] feather developed the distinctive feature based speech production theory by proposing "Window model". She agreed that distinctive feature and phonological rules could not account for the grade nature of the articulatory movements but disagreed that the grade variation should be ascribe to the biomechanical properties of speech organs since they differed among different languages. Hence, she made two basic assumptions: 1). phonological underspecification may persist into the phonetic representation; 2). phonetic underspecification is not categorical but a continuously notion. The output of phonological rules is interpreted in space and time by phonetic implementation rules which provide a continuous representation over time. In the phonetic implementation, the feature values are associated with a range of value - window. For a specified feature, the range is narrow; otherwise the range is wide. And a path connects the windows represent the real articulatory/acoustical trajectory over time. The path is interpolation functions between windows and is constrained by requirement of smoothness and minimal articulatory efforts.

Articulatory gesture based speech production theories

Browman and Goldstein [10] argued that, in the distinctive feature based theories, the description of speech are in two separate domains, which require two distinctive types of representation and relate them to each other with (phonological/phonetic implementation) rules. This makes the phonological and the phonetic levels quite different from each other. Hence, they proposed the Articulatory Phonology theory aiming to bridge the cognitive representation (phonology) and the physical representation (phonetics) of

speech. In articulatory phonology, gestures are characterizations of discrete, physically real events that unfold during the speech production process. And they are treated as basic units of contrast among lexical items as well as units of articulatory action. A gesture in "Articulatory Phonology" is specified using a set of related tract variables, which characterize a dimension of vocal tract constriction (the articulators that contribute to the formation and release of this constriction are organized into a coordinative structure). Tow phonemes will contrast if they differ in gestural composition. This difference can involved the presence or absence of a given gesture, parameter difference among gestures. In the physical level, a gesture is defined by specifying: 1) a dynamic equation (or a set of them); 2) a motion variable or variables; 3) values for the coefficients of the equation; and 4) weightings for individual articulators. And the relations among gestures can be specified abstractly using spatio-temporal phase relations. Utterances are modeled as organized patterns (constellations) of gestures, in which gestural units may overlap in time. The phonological structures defined in this way provide a set of articulatorily based natural classes. Moreover, the patterns of overlapping organization can be used to specify important aspects of the phonological structure of particular languages, and to account, in a coherent and general way, for a variety of different types of phonological variation. Such variation includes allophonic variation and fluent speech alternations, as well as 'coarticulation' and speech errors.

Physical realization theory of speech production

Ohman [11] analyzed Swidish Vowel-Consonant-Vowel (V1-C-V2) utterances based on acoustic and articulatory observations of Swedish, and on acoustic observations of comparable utterances of American English and Russia. He found that the transition of second formant of V1 depended not only on C and V1 but also the identity of V2. Based on the observations, he assumed that the tongue may be regarded as three independently controllable mechanical systems - the apical articulator, the dorsal articulator, and the tongue body articulator. And he argued the "articulatory gestures which are characterized as "dental", "alveolar" or "retroflex" employed the apical articulator; those characterized as "palatal or "velar" employ the dorsal articulator; and those characterized as "back",

"front", "open", "close" etc. employed the tongue body articulators." And the targets of speech segments are specified by the spatial targets of the apical, dorsal and tongue body articulators. To interpret the observed phenomenon that transitions from a fixed consonant to a fixed vowel may differ a lot when different vowels precede the consonant, he proposed a numerical model of lingual coarticulation which is aimed to generate the dynamic movement of the geometrically represented tongue outlines, as shown in Eq. (1.1-1.3).

$$s(x;t) = v(x;t) + k(t)[c(x) - v(x;t)]w_c(x) \tag{1.1}$$

$$v(x) = \alpha a(x) + \beta u(x) + \gamma i(x) \tag{1.2}$$

$$\alpha + \beta + \gamma = 1, 0 \leq \alpha, \beta, \gamma \leq 1 \tag{1.3}$$

where $s(x;t)$ is the resulted tongue shape; $v(x;t)$ is the target shape of vowel; $c(x)$ is the target shape of consonant; $w_c(x)$ is the coarticulation function which denote the amount to which an arbitrary target of vowel is allowed to distort the target of a consonant; $k(t)$ is a time variable parameter which is between zero and one; $a(x)$, $u(x)$, $i(x)$ are the targets of the extreme vowel /a/, /u/, and /i/; and x is the is the distance between the lips and glottis, and t is the time stamp. The effects of the apical and the dorsal articulators on the resulted vocal tract shape are embodied by $w_c(x)$. If the tongue tip in producing consonant allows significant modification by the neighbor vowels, the w_c should show small value in the part corresponding to tongue tip. It is the similar case for modifying the targets for tongue dorsum. And the effects of tongue body articulator are accounted for by the targets of vowel. At any time stamp of the realization of the consonants, it is always based on the dipthong gesture of the trans-consonant vowels. And the gesture of the consonant is realized by perturbing the apical/dorsal gestures of the base dipthongal gesture according to the requirement of the consonant gesture.

Problems of the theories of speech production

Though previous theories of speech production have various detail problems, however, in this part we tried to analyze the general origin of those problems rather than dive into the detail problems of those theories. To address the problems of the current speech production theories clearly, firstly, we make the correspondence between the physical facts in

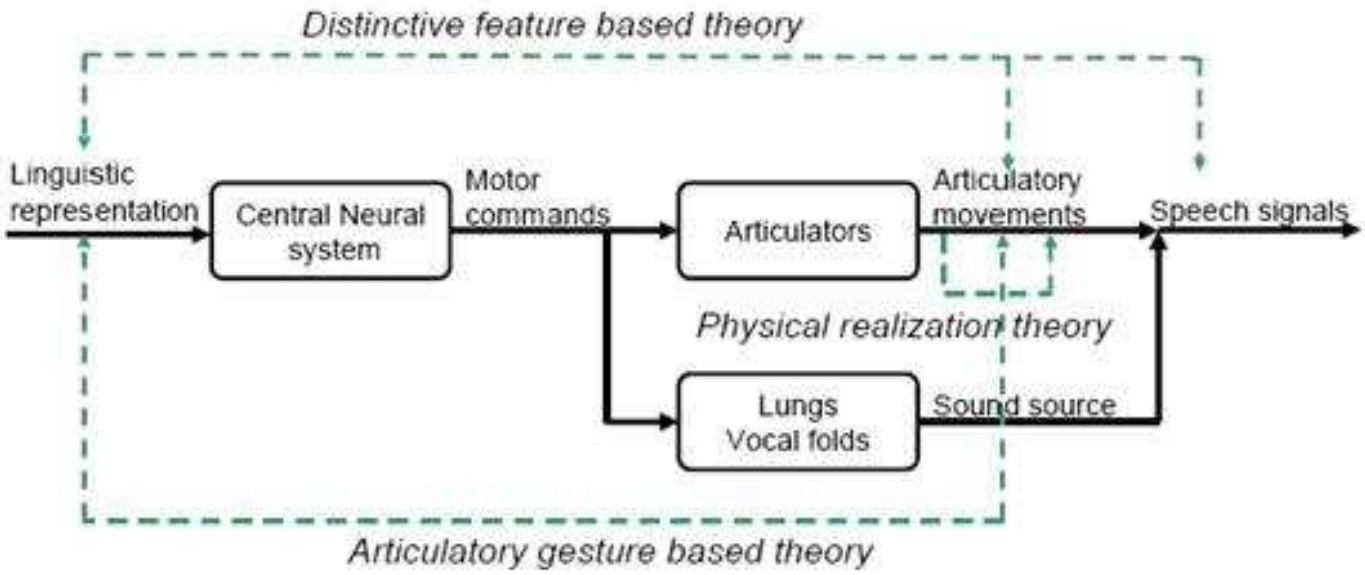

Figure 1.2: The schematic description of the process of speech production and the correspondence between various speech production theories and observations.

speech production and the theories tried to explain those physical facts. Figure 1.2 shows a schematic representation of the process of speech production and the correspondence between the surface observations and the speech production theories.

In Figure 1.2, the blocks represent the modules recruited in speech production. The input and output of each module ride on the solid arrows, and the correspondence between surface (articulatory/acoustic) observations and speech productions are indicated by the dashed arrows. As shown in the figure, the distinctive feature based theories and articulatory gestures based theories, ignore the events in the central neural system that generate motor commands according to the linguistic representation, and the biomechanical effects of the articulatory system; and the Ohman's model [12] only modeling the physical events of articulatory movements in the physical level. All of those theories are only based on partial of the information observed in the process of speech production. The motor commands that embody the control of speech production and the temporal organization of speech segments in utterances, and the activities in the central neural system that transform the linguistic representation into the motor commands, are not taken into account. This may be the reason that there is a gap between the linguistic representation and physical observations of those theories. In addition, those theories differ in the targets for speech segments, and the temporal organization of speech segments in speech flow, and usually give diverse explanation to the same phenomena. Hence, we presume that unless the motor commands and the activities in the central neural system are taken into account, many empirical observations of speech production may only be explained with ambiguity, and the gap between linguistic representation and surface observation cannot be bridged.

Therefore, if we can uncover the motor command and the activities of central neural system in speech production, we may get a better understanding of the detail mechanism of speech production by using the uncovered information and observed articulation. For this purpose, at the first step, we need to uncover the motor commands and the activities of central neural system in speech production; then, based on the uncovered and observed information to elaborate a new theory that bridge the linguistic representation and the physical observations.

1.2 Purpose of this thesis

In the present thesis, we attempt to uncover the motor commands, which reflect the motor control of speech production and temporal organization of speech segments and are difficult to be observed directly, from the observed articulatory posture. This task consists of three parts: 1). uncover the motor commands for isolated vowel production; 2). uncover the motor commands for isolated consonants; 3) uncover the motor commands for continuous speech.

Usually, speech scientists consent that the central control signals are the motor commands [13-18]. In the framework of Equilibrium Position Hypothesis (λ-model) [19], the degree of muscle activation is not specified centrally, but specified by results from the interaction of a central command with the afferent inputs provided by muscle spindles and other proprioceptive afferents. The relationship between muscle length and force is described by a family of length-tension curves. Each is characterized by a different threshold length for motoneuron recruitment, which is centrally controlled. In λ-model, the active muscle force can be formulated according to Eq. (1.4):

$$M = \rho(\exp(c[l-\lambda]^{+}-1),\ [x]^{+} = max(x,0) \tag{1.4}$$

where the threshold length λ is the central control signal, l is the muscle length at current state, ρ is assumed to vary with muscle force generating ability, and may be estimated from each muscle's maximum force capability, c is a form parameter, related to the motoneuron recruitment gradient, and is assumed to be equal for all muscles [20]. Hence the activation of a muscle is determined by the muscle length l and the threshold length λ.

In normal speech, the afferent inputs and the muscle lengths are decided by desired articulatory posture. According to Eq. (1.4), this makes the correspondence between motor commands muscle activations nearly one-to-one. Therefore, in the present thesis, we refer to the muscle activation as motor command.

In the literature, only a few data of muscle activation in speech production are obtain by using Electromyography (EMG) [21-25], which is unable to measure the activation of small intrinsic muscles. And some attempt has conducted to infer the muscle activation from tagged Magnetic Resonant Image (MRI) of articulation [26-29]. But result of which muscle are activated obtained by using tagged-MRI alone are questionable. Furthermore,

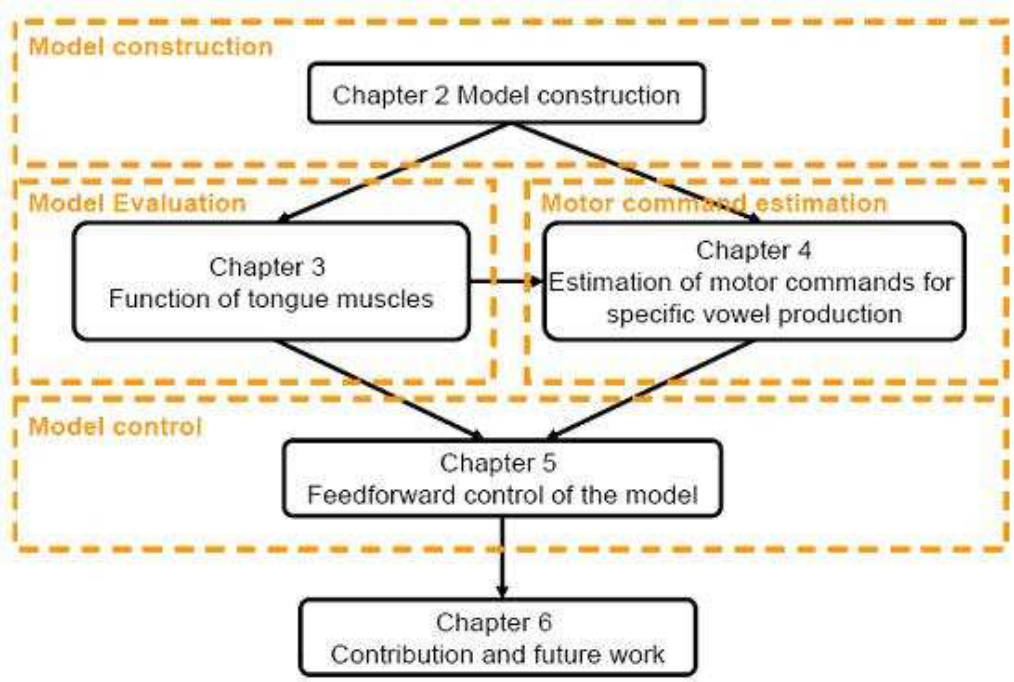

Figure 1.3: Organization of the thesis.

tagged-MRI alone cannot provide the muscle activation level. To uncover motor commands from observed articulation by using a model-based method, we construct a 3D physiological articulatory model, and elaborate a feedforward control strategy for the model.

1.3 Organization of the thesis

In this thesis, we mainly account for uncovering the motor commands in vowel production. Figure 1.3 shows the organization of the present thesis.

In Chapter 2, a 3D physiological articulatory model is constructed by combining a 3D physiological tongue model [30] with 3D jaw and vocal tract wall of the prototype subject. In addition, to make the model more realistic, the anatomical orientations of some muscle

are refined according to the anatomical descriptions [31]. And the contact between tongue and jaw, tongue and palate are correctly handled.

Since the 3D physiological articulatory model is driven by the contraction of associated muscles, it is necessary to evaluate the detail functions of those muscles, especially tongue muscles. In Chapter 3, the function of tongue muscle are quantitative analyzed by using the proposed 3D physiological articulatory model to shed light on the general function of individual tongue muscle and the agonist-antagonist property of tongue muscle pairs, which help to efficiently control the 3D physiological articulatory model.

After constructing 3D physiological articulatory model and evaluating the function of tongue muscles, we proposed a model-based optimization procedure to uncover the motor commands in specific isolated vowel production. This is addressed in Chapter 4.

Usually, the articulation of a same phoneme is with certain variation. The articulation is not exactly the same from time to time. Therefore, the method described in Chapter 4 is expensive for uncovering the motor commands for articulations with variation. In addition, in daily communication with speech, people usually speak to each other with continuous speech rather than isolated phonemes. In continuous speech, the observed articulation is not the same as the desired articulatory targets [32]. To deal with the coarticulated articulations, it requires to uncover desired articulatory targets first, then, to uncover the motor commands for the phonemes of the utterance. All of these need an efficient control strategy to manipulate the physiological articulatory model according to desired target. In Chapter 5, a feedforward control strategy is elaborated by mapping the articulatory targets, specified by articulatory parameters, to muscle activations via the intrinsic representation of vowel production by using General Regression Neural Network. In Chapter 6, we summarize the contribution and future work of present thesis.

Chapter 2

Model construction and preliminary evaluation

The observed articulation is generated by driving articulators with motor commands. It embodies the coeffects of motor command and the biomechanical properties of articulators. Therefore, to uncover motor commands from observed articulation, the biomechanical effects of the articulators should be well understood. A physiological articulatory model is a powerful tool that inherently represents the biomechanical properties of articulators. By using such a model, the effects caused by the biomechanical properties of speech organs in speech production can be reasonably accounted for. Thus, it is expected to help uncover motor command from observed articulation.

The rest of this Chapter is organized in the following way. In Section 2.1, we review the physiological articulatory models appeared in the literature. In Section 2.2, we introduce our 3D physiological articulatory model which approximates the articulators of human both morphologically and biomechanically. In Section 2.3, we incorporate the handling of contact between the tongue and surrounding structures into the model. In Section 2.4, we conduct preliminary evaluation on the behavior of the model by activating muscles recruited in the model individually. In Section 2.5, we discuss the improvements of the 3D model compared with the partial 3D model.

2.1 Introduction

From the 1970's, a number of physiological articulatory models were proposed for various purposes. In this part, we will introduce those works according to the complexity of those models. Firstly, the 2D models [14, 17, 33, 34], including models with fixed jaw-hyoid structures and with jaw-hyoid complex, are introduced. Secondly, the partial 3D model [35], from which our model is extended, is introduced. Thirdly, some of the 3D physiological articulatory models are introduced.

2.1.1 2D physiological articulatory models

Tongue models with fixed jaw and hyoid bone

Perkell [33] devised a 2-D mass-spring physiological articulatory model, which changed its shape only in the sagittal plane, based on X-ray measurements. The model incorporated approximate representations of several basic physiological and biomechanical properties of the tongue: some of its muscular anatomy, conservation of volume, the impenetrability of the vocal-tract walls, rough contractile property of muscle tissue, and passive elasticity and viscous properties of connective tissue. In the model, the mass of the tongue is concentrated in sixteen movable "fleshpoints". The fleshpoints are connected to one another and to bony attachments by 38 active and 47 passive tension-generating elements. The tongue muscles are represented by anatomical groupings of the tension-generating elements which are arranged to enclose 14 quadrilateral areas. Those "volumes" have limited compressibility to allow for the transmission of compression force and conservation of volume. In addition to tensile and volume conservation forces, sliding friction and hard structure impenetrability force act on the fleshpoints in the model to determine their movement.

Perrier *et al.* [14] constructed a 2D physiological tongue model based on finite element method (FEM) by explicitly adapting the geometrical external shape of the model to the tongue contours extracted from an X-ray picture of a specific speaker. The configuration of the tongue is close to the production of schwa. The tongue model consists of 48 elements, which are roughly arranged along the orientation of Genioglossus. The number of

elements is determined through the result of a compromise between anatomical accuracy and low computational costs. Each muscle pair symmetrically located on the left-right side of the tongue is modeled as a unique group of fibers. In addition, muscles, whose effects on tongue shaping in the midsagittal plane are slight, are not modeled. Consequently, the muscles described in the model are the anterior and posterior parts of Genioglossus (GG), Styloglossus (SG), Hyoglossus (HG), Verticalis (V), and superior and inferior parts of the Longitudinals (SL and IL). "the numerous fibers of the transverse muscle have not been represented, " [15]. At the bottom of the model (between the mental spine and the hyoid bone), the Geniohyoid (GH) and Mylohyoid (MH) are functionally modeled as a whole, with a reaction force that is applied to corresponding nodes to limit the amplitude of any downward movement. Finally the Finite Element (FE) structure is inserted in the vocal tract contours measured on the same X-ray picture. "Muscle insertions on bony parts (jaw and hyoid bone) are simulated by imposing "do not move" constraints to the corresponding nodes" [15].

Tongue models with jaw-hyoid complex

In speech production, the jaw-hyoid complex structure is important for manipulating the position of jaw and hyoid bone in speech production. However, in the above two models, the locations of jaw and hyoid are fixed. Attempts have been made to suppress this problem by some speech scientists.

Sanguineti *et al.* [17] constructed a 2D physiological articulatory model based on X-ray measurement using FEM technique. The model geometry is derived from a young female speaker. The position and orientation of the jaw at occlusion are estimated by superimposing a normative model of the jaw on the x-ray data. The tongue surface contour, the hyoid position and orientation, and the larynx height are likewise obtained from the x-ray data set. Jaw motions in the model have two degrees of freedom-orientation in the sagittal plane and translation along the articular surface of the temporal bone. The hyoid has three degrees of freedom, horizontal and vertical position and sagittal plane orientation. The larynx is modeled as a point mass with a single degree of freedom-vertical position. In the tongue model, they included four extrinsic muscles, GG, HG, SG, and MH, and

three intrinsic muscles, SL, IL, and V. Muscles acting on the jaw include a jaw opener, which is responsible for jaw opening, and a jaw closer, which takes charge of jaw closing. This model, virtually, consists of two sub systems. One is the FEM model that governs the movement and deformation of the tongue. The other is the mass-rigid body system, which manipulate the positions of the jaw, hyoid, and larynx.

Honda's articulatory model [34] of speech organs has a two-dimensional configuration which demonstrates a mid-sagittal view of vocal tract organs. The outline of the model is provided by mid-sagittal MRI data from a Japanese speaker. The organs implemented in the model are the tongue, jaw, hyoid bone, laryngeal cartilages, and the vocal folds. Twenty-one muscles are implemented in the model as active components. Two different algorithms are combined in the computation of the effects of muscle contraction: a finite element method (FEM) for tongue deformation and a mass-spring model for displacements of the rigid structures. The FEM tongue model consists of approximately thirty triangular elements, and each element is represented by a plate model which deforms by direct and indirect forces acting on each angle. The mass-spring model for the rigid structures connects the jaw, hyoid bone, thyroid cartilage, and cricoid cartilage by a network of muscles and ligaments. The displacement of these structures is computed by solving motion equations. The shape of the tongue and the position of other articulators are determined by the equilibrium of the muscle forces and ligament stiffness. These biomechanical interactions among the components are expressed in the model as a natural consequence of the overall mass-spring action of the model. In this system, the tongue musculature includes the extrinsic and intrinsic muscles. The extrinsic muscles arise from the surrounding bony structures, inserting into the tongue. To a first approximation, the large deformation of the tongue for vowel articulation is primarily produced by these muscles. The intrinsic muscles run within the tongue in three directions, and they are mainly responsible for tongue blade deformation in consonants. Anatomically, the major extrinsic tongue muscles are the Genioglossus (GG), Hyoglossus (HG), and the Styloglossus (SG). For the purpose of explaining their functional organization based on the EMG data, however, he divided muscle GG into two portions: anterior (GGA) and posterior (GGP), assuming that the middle portion in the model and the real system serves a transitional function between the two parts., The model was computationally slow in achieving equilibrium be-

tween the soft tissue and rigid bodies because these structures are modeled independently.

2.1.2 Partial 3D physiological articulatory model

Dang and Honda [35] constructed a partial 3D physiological articulatory model based on the MRI measurement of a male Japanese subject by using extended-FEM (X-FEM). The outlines of the tongue body are extracted from two sagittal slices: one is the midsagittal plane and the other is a plane 1.0 cm apart from the midsagittal on the left side. The initial shape of the model adopts the tongue shape of a Japanese vowel /e/. Mesh segmentation of the tongue tissue roughly replicates the fiber orientation of the genioglossus muscle. The outline of the tongue body in each plane is divided into ten radial sections that fan out from the attachment of the genioglossus on the jaw to the tongue surface. In the perpendicular direction, the tongue tissue is divided concentrically into six sections. 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, where each mesh is a hexahedron. Thus, the model represents the principal region of the tongue as a 2cm-thick layer bounded by three sagittal planes. The 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 left side). The anatomical arrangement of the major tongue (musclesGG, GH, HG, SG, SL, and IL) are examined based on a set of high-resolution MR images obtained from the prototype speaker. The muscles, which could not be identified form MR images, are arranged according to the anatomical literature [31, 36, 37]. The model of the jaw has four nodes on each side, which are similar to those used by Laboissiere *et al.* [20]. And those four nodes are connected by five rigid beams to form two triangles with a shearing beam. The hyoid bone is modeled as three segments corresponding to the body and bilateral greater horns. Each segment has two nodes connected with a rigid beam. The masses are uniformly distributed over the rigid beams. To provide a uniform computational format, rigid beams are also treated as visco-elastic links with a high Young's modulus so that they can be integrated with the soft tissue in the motion equation.

The essential disadvantages of the 2D and partial 3D models are: (1) it cannot faithfully

represent the morphological structure of the speech organs; (2) the anatomical orientation of related muscle cannot be arranged correctly; and (3) it is difficult to account for the interaction between tongue and surrounding structure correctly.

2.1.3 3D physiological articulatory models

Kiritani *et al.* [38] elaborated an 3D physiological articulatory model for studies of the static characteristics of the tongue as a 3D linear system under simplified geometric boundary constraints based on the anatomical data form Miyawaki [36]. The first version is composed of 14 tetrahedrons, which make that the muscular properties can be approximated by uniformly distributed elastic characteristics within each unit. "The tongue body is surrounded by a hard wall of a ellipsoidal vertical cylinder continued into an ellipsoidal shell simulating the roof of the mouth cavity" [38]. Kakita *et al.* [39] developed it by refining the elements within the model. In the extended model, the tongue consists of 172 tetrahedrons. And the entire set of tetrahedral was subgrouped into 30 functional units, given rise to 33 functional node points. The wall structure housing the tongue were approximated by an elliptic vertical cylinder connected to the upper half of an ellipsoid, representing the side walls and palatal ceiling, respectively. However, due to the coarse representation of the morphological structure of the tongue, it is difficult to make faithful arrangement of the musculature of the tongue, and to represent the tongue configuration with enough accuracy. In addition, in these models, neither the inertial component nor the effects of geometric nonlinearities were represented. To account for the factors, Wilhelms-Tricarico [40] proposed a rigorous method for modeling the soft tissue and then built a 3D tongue model. But computation cost of the proposed model [40] is too high to be implemented to explore nature of speech production.

2.2 The 3D physiological articulatory model

To overcome the problems of the previous physiological articulatory models, we construct a 3D physiological articulatory model by combining the 3D tongue model [30] with sur-

rounding structures, and refined the orientations of SG and IL to make them more realistic. The proposed model is composed of the following components: the tongue body, the jaw and vocal tract wall, associated muscles, and jaw-hyoid complex. In this part, we will introduce them in detail.

2.2.1 Morphological structure of tongue

The 3D tongue within the 3D physiological articulatory model is adopted from Fujita *et al.*'s work [30]. In their work, the initial shape of the tongue is obtained based on the volumetric MR images while producing Japanese vowel /e/. Following the method proposed by Dang [35], the mesh structure of the tongue is based on the anatomic structure of the genioglossus that fans out from the attachment on the mandible. The configuration of the tongue adopts the realistic shape of the MRI-based tongue tissue geometry using five layers in the left-right dimension, which has a maximal width of 5.5 cm in the posterior portion, although the most lateral regions of the tongue root are not included. The mesh structure of the tongue model in the oblique view consists of eleven layers with nearly equal intervals fanning out to the tongue surface from the attachment on the mandible, and seven layers in the perpendicular direction. Totally, the tongue body consists of 240 hexahedrons. Figure 2.1 demonstrates the configuration of the 3D tongue involved in the physiological articulatory model.

2.2.2 Morphological structure of vocal-tract wall and jaw

To constrain the movement of the tongue and generate speech sound, vocal-tract wall and jaw are indispensable. In this model, the contours of the jaw and vocal tract wall are carefully extracted form the MRI images of vowel /e/ superimposed with the lower and upper teeth [41] at the interval of 0.4cm in the left-right dimension. The sagittal images start from the outer surface of the molar on the left side and end at the outer surface of the molar on the right side. Therefore, the vocal-tact wall and jaw can form a complete enclosed space when the jaw clenches with the maxilla. Finally, the distance between the left and right extremes is 5.6cm, and the vocal-tract wall and the jaw consist of 15 layers. In the current model, the vocal-tract wall is composed of upper teeth, hard palate, velum,

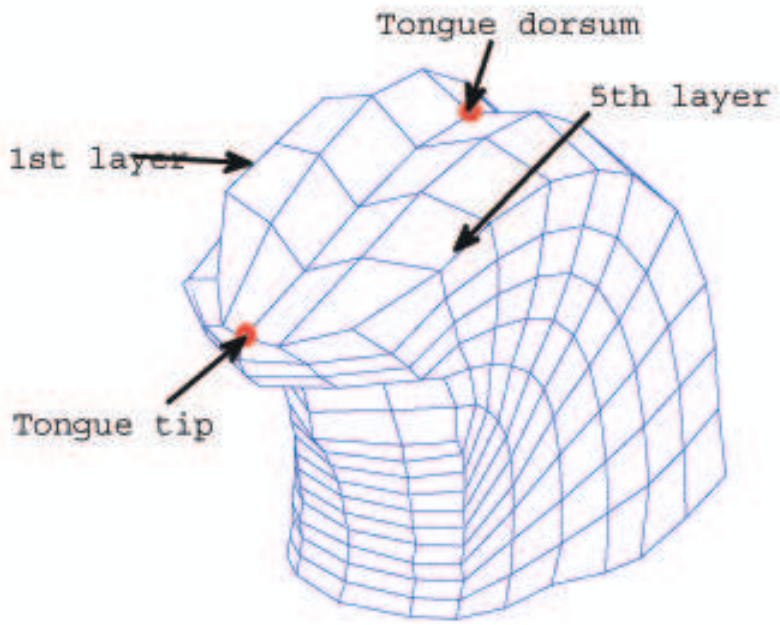

Figure 2.1: The oblique view of the 3D tongue involved in the 3D physiological articulatory model. The first and fifth layers of the tongue in the transversal dimension, and the representative points of tongue tip and dorsum are indicated with arrows.

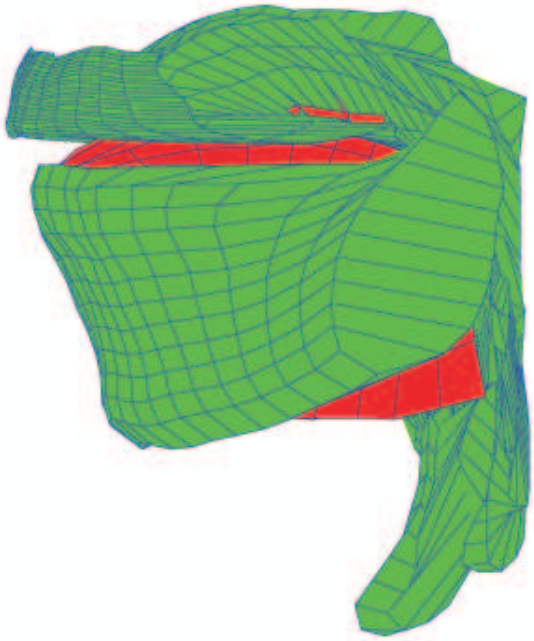

Figure 2.2: The profile of the 3D physiological articulatory model which consists of tongue, jaw and vocal tract wall.

pharyngeal wall, and larynx tube. The position of the velum can be adjusted according to the size of nasopharyngeal port, which is for generating nasal sounds. Other parts of the vocal tract wall are treated as rigid components. Figure 2.2 shows the morphological configuration of the 3D physiological articulatory model.

When the jaw is open, the cheek is considered to form the lateral boundary of the vocal tract immediately outside the lower and upper teeth. Therefore, the area function of the vocal tract can be directly calculated from the cross-sectional planes that intersect with the vocal tract. In this way, the 3D physiological articulatory model can depicts the area function of the vocal tract more accurately for the specific configuration such as for /l/ and for the vocal tract with dynamic variations.

So far, the lips and the velum are not modeled physiologically. They are involved in constructing vocal tract shapes for speech sound generation but not in the generation of articulatory movements. The lips are defined by a short tube with a length and cross-sectional area. And the movement of the velum is described by the opening area of the nasopharyngeal port.

2.2.3 Musculature

To drive the physiological articulatory model according to human mechanism, associated muscles are faithfully arranged in the model based on their anatomical orientations. There are nine muscles included in the tongue model. Three extrinsic muscles, Genioglossus (GG), Styloglossus (SG), and Hyoglossus (HG), are arranged mainly based on the MRI analysis [42]. The intrinsic muscles, Superior Longitudinal (SL), Inferior Longitudinal (IL), Transversus (T), Verticalis (V), are modeled based on the anatomical data from Takemoto [31]. Tongue floor muscles, Mylohyoid (MH) and Geniohyoid (GH), are arranged based on anatomical literature [25]. All the muscles are bilaterally symmetric.

According to the functions of the different parts of GG [21, 23, 24], GG is divided into three portions: GG anterior (GGa), GG middle (GGm), and GG posterior (GGp). Takano [43] found that length of the muscle fibers at anterior part (inside the tongue) and the posterior part (from the styloid process to the insertion) of SG showed significant difference in producing the five Japanese vowels. Hence, SG is divided into 2 parts SGa (the part within the tongue) and SGp (the part from the styloid process to the insertion). And they can be controlled independently. Figure 2.3(a)-(k) illustrates the layout of the major extrinsic, intrinsic muscles in the 3D physiological articulatory model.

Figure 2.3(l) shows the model of the jaw-hyoid complex. The model of the jaw has four nodes on each side, which are connected by five rigid beams (the upper thick lines) to form two triangles with a shearing beam. The jaw combines with the tongue model at the mandibular symphysis. The hyoid bone is modeled as three segments corresponding to the body and bilateral great horns. Each segment has two nodes connected with a rigid beam. Eight muscles indicated by thin gray segments are incorporated into the jaw-hyoid complex, where the structures of the muscles are based on the anatomical literature [37]. Although there is no unique mapping between muscle activations and kinematic degrees of freedom, the muscles involved in the jaw movements during speech can be roughly separated into two groups: the jaw-closer group (jawCl: solid gray segments) and the jaw-opener group (jawOp: dashed gray segments). The force generation of muscle is modeled by functionally relating the length and strain of the muscle length (see Appendix B for detail).

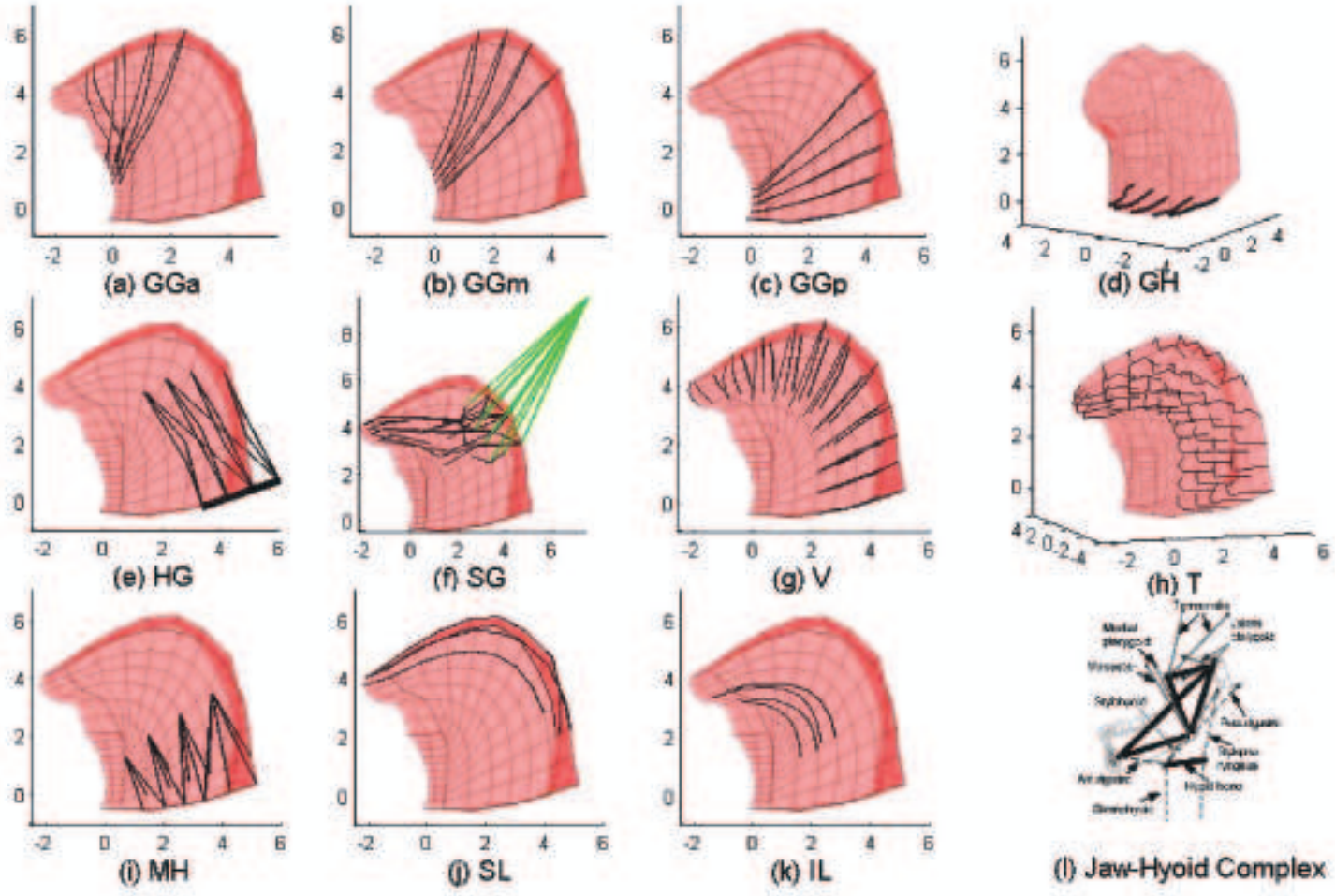

Figure 2.3: The musculature and jaw-hyoid complex of the 3D physiological articulatory model. (a) Genioglossus Anterior (GGa); (b) Genioglossus Middle (GGm); (c) Genioglossus Posterior (GGp); (d) Geniohyoid (GH); (e) Hyoglossus (HG); (f) Styloglossus (SG); (g) Verticalis (V); (h) Transversus (T); (i) Mylohyoid (MH); (j) Superior Longitudinal (SL); (k) Inferior Longitudinal (IL); (l) jaw-hyoid complex (gray solid segments: jaw-closing muscle group; gray dashed segments: jaw-openning muscle group; thick lines: rigid beams).

2.3 Contact handling of physiological articulatory model

In speech articulation, the tongue often contacts the teeth, palate, and jaw as it moves. Dang *et al.* [44] observed that the tongue contacted the hard palate in the lateral area 1.5cm from the midsagittal plane during most of the phonation. Honda *et al.* [45], and Sanguineti *et al.* [16] claimed that it was essential to account for the interactions between soft tissue and bony structures in order to have accurate prediction of vocal tract motion. Therefore, the handling of contact between tongue and surrounding structures is one of the essential tasks for a physiological articulatory model.

Since the shape of the jaw and vocal tract wall are too complex to be described with an analytic function, the contact of the tongue with the vocal tract wall, cannot be combined into the motion equations systematically. As an alternative, we propose a method with two steps to deal with the contact between tongue and surrounding structures: 1). check whether or not the nodes of the tongue cross through the vocal tract wall; 2). calculate the retraction force according to the depth of the penetration of nodes.

If a node is beyond the wall of the vocal tract during articulation, its trajectory must have an intersection with the jaw/vocal tract wall. Since the jaw/vocal tract wall was assembled by triangle planes, we first identify the plane with which the trajectory intersected on the wall and then calculate the collision force of the node when it is bounded on the wall. The following formula is used to estimate the collision force.

$$f_x = \sum_i (k_i \Delta l_{xi} + b_i \Delta l_{xi}/h) \tag{2.1}$$

where i is the index of the cylinders connecting with the node, and h is the computation step. k_i and b_i are the stiffness and viscous coefficient of the cylinder i. Δl_{xi} is the increment of the cylinder i in x-dimension caused by the wall bounding. f_x is the x-component of the resultant bounded force. Using the same approach, the force can be calculated for y-dimension f_y and z-dimension f_z.

Figure 2.4 (a) and (b) gives the comparison of the without and with accounting for contact handling in the physiological articulatory model. Figure 2.4 (a) gives the result of activating muscle GGp by 1N force without accounting for the contact between tongue and jaw. It shows that the tongue tip and tongue dorsum move forward, and make the tongue tip unrealistically thick. Figure 2.4(b) shows the result in the same situation ex-

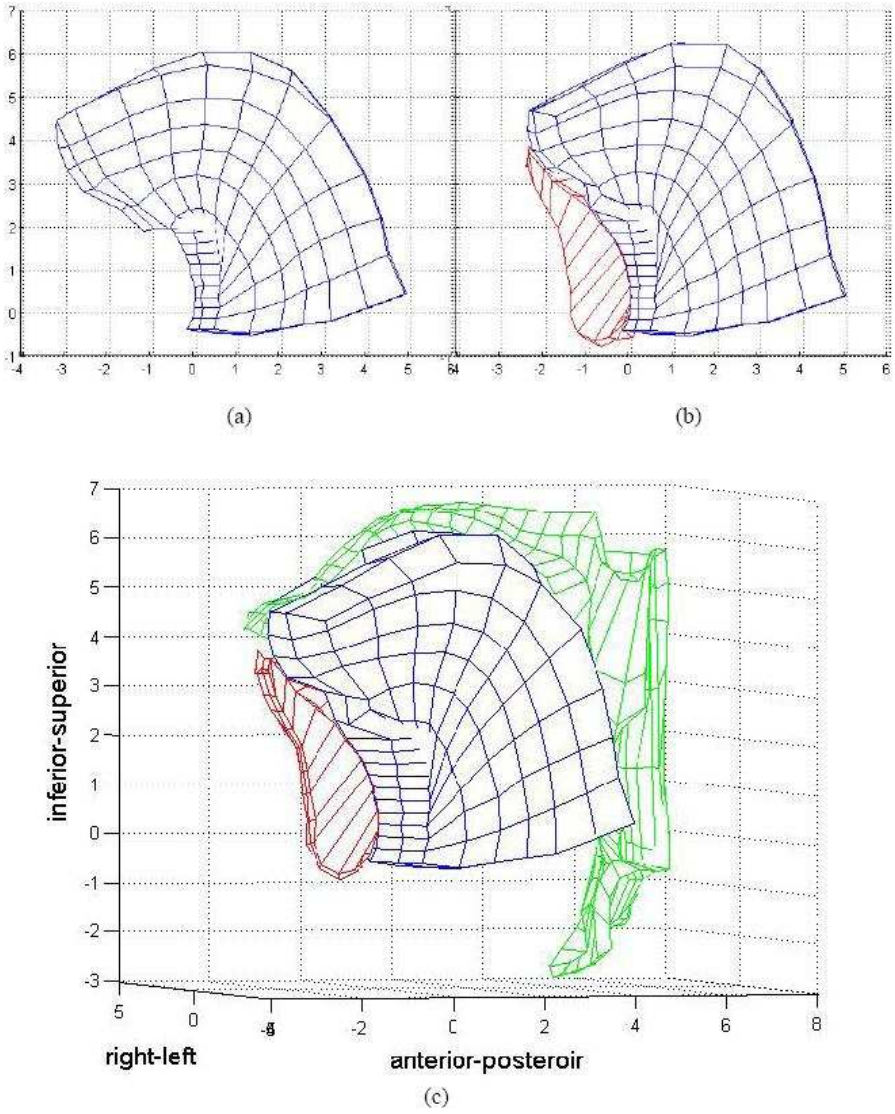

Figure 2.4: Results of incorporating contact handling into the physiological articulatory model: a) activate GGp with 1N without accounting for contact handling; b) activate GGp with 1N with accounting for contact handling; c) contact between tongue and jaw, and tongue and palate.

cept that the contact between jaw and tongue is considered. The tongue tip is stopped by the jaw, and the tongue body goes up. And the shape of the tongue is more realistic. Figure 2.4 (c) shows the result of the contact between tongue and jaw, tongue and palate, which usually happen in speech production. It shows that the contacts between jaw and tongue, and vocal tract and tongue are properly handled.

2.4 Testing of the model

In the literature, the function of the extrinsic and intrinsic muscles were qualitatively speculated based on anatomical descriptions [29, 37, 43]. By referencing those results, we evaluated the behavior of the 3D physiological articulatory model by activating the tongue muscles individually with 1N in 300ms, where the boundary constraints resulted from the jaw and vocal tract wall on the tongue are not taken into account.

Figure 2.5(a) gives the posture of the tongue at its rest position. Figure 2.5(b) demonstrates the result by activating GGa. Comparing with initial shape, the portion of the tongue blade is depressed. The activation of GGm (Figure 2.5(c)) moves the tongue body downward and forward. The activation of GGp (Figure 2.5(d)) makes the tongue body forward and slightly upward. The activation of HG (Figure 2.5(e)) draws the tongue body downward and backward, while the tongue tip goes up with the rotation of the tongue body. Muscle SGa (Figure 2.5(f)) helps to bunch the tongue. Muscle SGp (Figure 2.5(g)) draws the tongue backward and upward. In addition, muscle SL (Figure 2.5(h)) retract the tongue body and elevate the tongue tip. Muscle IL (Figure 2.5(i)) retracts the tongue body and makes the tongue tip slightly downward. The contraction of T narrows the tongue and elevates the tongue dorsum, as shown in Figure 2.5(j), while the contraction of V draws tongue surface to the floor of the mouth, and makes it flatter and wider, as shown in Figure 2.5(k). These results are consistent with muscle function speculated based on their anatomical orientation.

Moreover, the function of jawCl and jawOp are also tested. It shows that the jaw can clench with maxilla when 3N force is applied to jawCl, and form the maximal opening about 2cm between jaw and maxilla in the midsagittal plane when 6N force is applied to jawOp. The results are consistent with the observations reported by Ostry *et al.* [46].

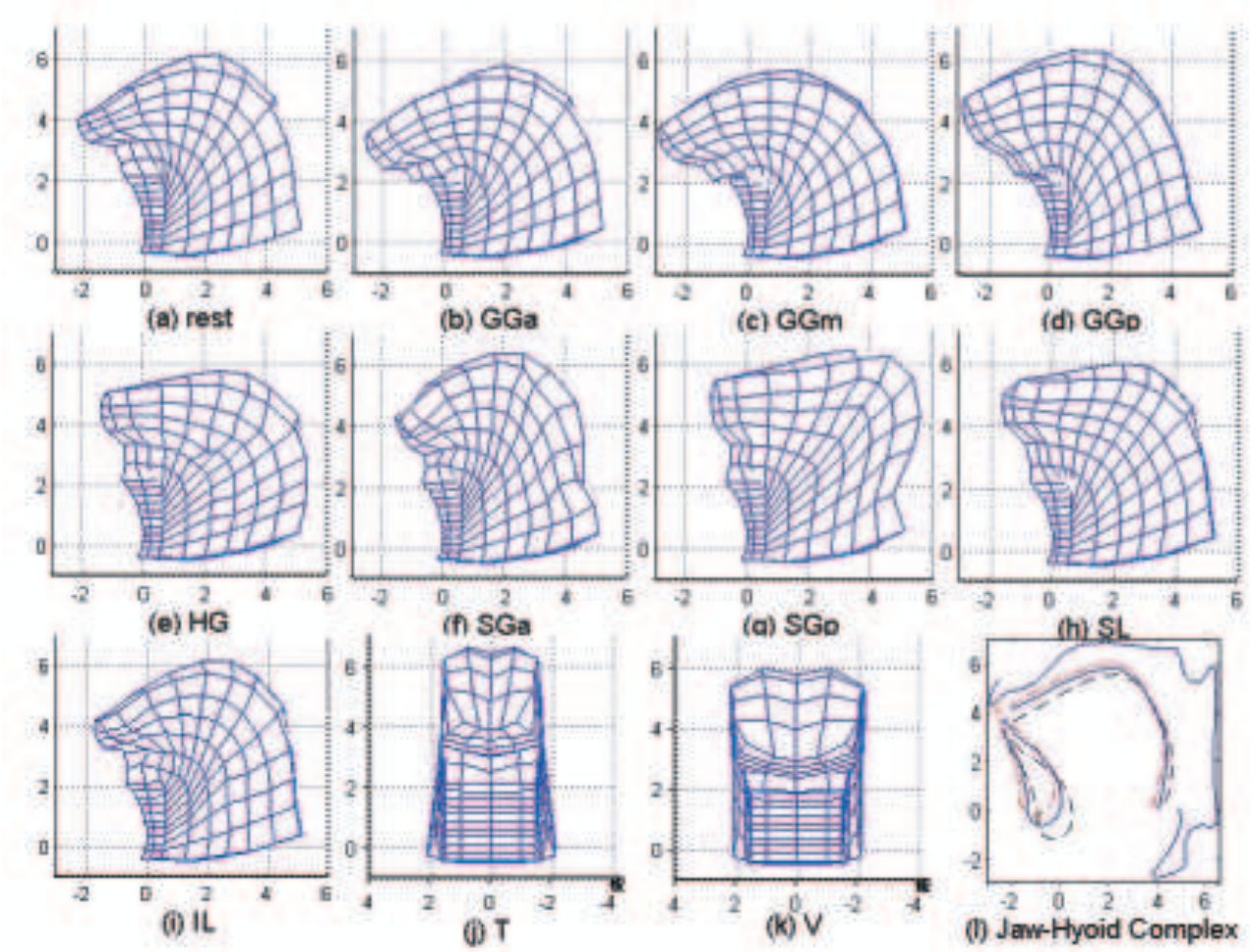

Figure 2.5: Results of activating individual tongue muscle, JawOp, and JawCl. (a) The configuration of the tongue without muscle activation; (b) GGa with 1N force; (c) GGm with 1N force; (d) GGp with 1N force; (e) HG with 1N force; (f) SGa with 1N force; (g) SGp with 1N force; (h) SL with 1N force; (i) IL with 1N force; (j) T with 1N force; (k) V with 1N force; (l) The position of the jaw and the tongue contour in the midsagittal plane under the various activations of muscle JawOp/JawCl (solid curve: JawOp 0N; dashed curve: JawOp 6N; and dotted curve: JawCl 3N).

2.5 Comparison between the 3D model and the partial 3D model

Compared with the partial 3D model, a number of improvements have been made in both morphological structure and musculature of the 3D model. In this part, we will discuss them at length.

2.5.1 Improvement of morphological structure

The most important improvement is that the morphological structure of the tongue model completely replicates the real geometry of the tongue. In the partial 3D model, the tongue model represents the principal region of the tongue as a 2-cm-thick layer bounded by three sagittal planes, which could not describe the tongue properties in the transverse dimension as well as in the volume preserving properties. In the current full 3D model, the tissue of the tongue in the left-right dimension represents realistically by extending the 2cm-think layer to real shape with a maximal width of 5.5cm in the posterior part of the tongue. Consequently, the transverse dimension of the model is not only developed from three layers to five layers, but also the interval between the layers varies with the tongue body dimension instead of the parallel arrangement. Therefore, the new tongue model is more realistic than the partial 3D model.

Comparing with the improvements of the tongue model, the jaw and vocal tract wall are modeled according to the real structure based on volumetric MR images. As a result, the transversal dimension of the jaw and vocal tract wall are extended from 2.8cm to 5.6cm. Due to this improvement, an enclosed cross-sectional area of the vocal tract can be obtained during speech production. Thus, we can obtain more accurate cross-sectional area function of the vocal tract but not estimate it from the vocal tract width via α-β model [47].

2.5.2 Improvement of musculature

Another improvement of the proposed model is that orientations of tongue muscles become more realistic. In the partial 3D tongue model, the arrangements of all the muscles were limited in three parallel layers, which is quite different from the real situation of the muscular structure. In the proposed 3D model, the muscle orientation is arranged as faithfully as possible based on the anatomical knowledge. Accordingly, muscle IL is arranged on the lateral sides of GG in different layers, instead of the previous arrangement, in which they are located in the same layer. For the same reason, muscle HG is symmetrically located in the outside parasagittal planes. The anterior parts of SG is located on the outer lateral layers of GG, and the attach points of the SG to the stylo-process spans about 7.5cm in transversal dimension, close to that of the human anatomic structure. The orientation of muscle T and V are completely replicated in the 3D model by referencing the Takemoto's work [31], which is not able to be realized in the partial 3D model. Altogether, the arrangement of the muscles in the proposed 3D model is more realistic compared with the description of the muscle structures in anatomical literatures [31, 36, 48]. The realistic muscular structure forms the basis for uncovering motor commands more accurately by using model-based numerical experiments.

2.6 Summary

In this chapter, we introduced the construction of a 3D physiological articulatory model. We constructed 3D jaw and vocal tract wall, and combined them with the 3D tongue model [30] into a 3D physiological articulatory model. To drive the model according to human mechanism, the related muscles are functionally formulated in the 3D model. The orientation of the muscle SG, and IL are improved to be more realistic compared with the anatomical descriptions. And the interaction between the tongue and surrounding structures are appropriately accounted for. Preliminary evaluations show that the model behaves properly when the muscles are activated individually. This forms the physical basis for uncovering motor commands in speech production with physiological articulatory model.

Chapter 3

Functions of tongue muscles

The 3D physiological articulatory model introduced in Chapter 2 is driven by contraction of associated muscles. To uncover motor commands from observed articulation by using model-based method, it is necessary to use proper muscle combination and muscle activations to drive the physiological articulatory model to generate the observed articulation. This requires good understanding of the functions of muscles in the physiological articulatory model, especially tongue muscles. In this chapter, we will quantitatively investigate the function of individual tongue muscles on both tongue movement and deformation, and the agonist-antagonist property of tongue muscle pairs.

In Section 3.1, we will introduce previous studies on exploring the muscle function. In Section 3.2, we explored the tongue movements and deformation caused by the contraction of individual tongue muscle. In Section 3.3, we explore the agonist-antagonist property, which offers the function to maintain the position of tongue tip/dorsum, while manipulating the position of the tongue dorsum/tip [35], of muscle pairs with the 3D physiological articulatory model.

3.1 Introduction

In the history, speech scientists tried to make clear the function of individual tongue muscles so as to understand how speech is controlled [29, 48, 49] or facilitate the control of a physiological articulatory model [35]. The function of the tongue muscles are mainly ob-

tained by three kinds of methods: speculations based on muscle orientation, speculations based on tagged-MRI observation, and quantitative analysis based on mode simulation. In the following part, we will introduce those work in detail.

3.1.1 Speculations of muscle function based on anatomical orientation

In the literature, a number of descriptions of the function of individual tongue muscle are speculated based on their anatomical orientation. "The superior longitudinal muscles courses along the length of the tongue, comprising the upper layer of the tongue. By virtue of their courses and insertions, fibers of the superior longitudinal muscle tend to elevate the tip of the tongue." "The inferior longitudinal muscle originates at the root of the tongue and corpus hyoid, with fibers coursing to the apex of the tongue. And the contraction of inferior longitudinal pulls tip of tongue downward, assists the tongue tip retraction." "The fibers of transverse muscles originate at the median fibrous septum and course laterally to insert to the side of the tongue in submucous tissue. The activation of these muscles provides a mechanism of narrowing the tongue." "The vertical muscles of the tongue a run at right angles to the transverse muscles. The fiber of the vertical muscles course from the base of the tongue and insert into the membranous cover. Contraction of the vertical muscles of the tongue will pull the tongue down into the floor of the mouth." "The genioglossus arises from the inner mandibular surface at the symphysis and fans to insert into the tip and dorsum of the tongue, as well as to the corpus of the hyoid bone. .. The contraction of the anterior fibers retracts the tongue, and the contraction of the posterior fibers protrudes the tongue." "The hyoglossus arises from the length of the greater cornu and lateral body of the hyoid bone, coursing upward to insert into the sides of the tongue. The hyoglossus pulls the sides of the tongue down." "The styloglossus originates from the anterior margin of the styloid process of the temporal bone, coursing forward and down to insert to the inferior sides of the tongue. contraction of styloglossus will draw the tongue back and up."[48, 49].

However, the tongue consists of incompressible soft tissue. Contractions of the tongue muscles not only cause tongue deformations along the direction of the muscle orientation,

but in other concerned dimensions, which are difficult to be predicted form the orientations of the muscles.

3.1.2 Estimation of muscle function based on tagged-MRI observation

Kumada *et al.* [26] used tagged-MRI to measure the deformation of the internal structures of the tongue in producing the five Japanese vowels. The muscle fibers of GGa, GGm, GGp, and SL are approximated by connecting the corresponding tagged-points in the tagged-MRI images. By comparing the length of the muscle fibers in producing vowels with those in reference position, they found that: a). the position of the tongue in producing vowel /a/ is lower than the reference position, and GGa is much shorter in producing vowel /a/ than that in reference position. It suggests that the function of GGa is to depress the tongue; b). the position of tongue in producing vowel /i/ is more anterior than that in reference position, GGp and SL are shorter in producing vowel /i/ than that in reference position. It suggests that GGp is to pull the tongue forward, and SL is to shorten the tongue; c). GGp and SL are much shorter in producing vowel /u/ than that in producing vowel /i/, while the tongue is lower than that in producing vowel /i/. This suggests that muscle V is activated to shorten the tongue in vertical axes.

Stone *et al.* [29] claimed that muscle lengthening (expansion) must always be passive, because active muscle contraction causes shortening, however, muscle shortening can be due to active muscle contraction, or to passive compression of the tissue. They utilized a deformable model to extract the principle strain, which reflects the compression and expansion pattern of the local tongue tissue, for small local tongue regions. By inspecting the principle strains, they found that: a). the downward and backward compression seen in the lower third of the tongue is consistent with the contraction of HG; b). compression of the upper tongue was almost uniformly vertical, particularly at midline and right. This is consistent with the contraction of Verticalis. c) the circumference of the tongue surface is shorten. This is consistent with contraction of the Superior Longitudinal.

However, Buchaillard *et al.* [50] used a 3D finite-element biomechanical model of the oral cavity to study the relation between the distribution of the strains observed in the

tongue body and the location of the active tongue muscles. They found that: a). in most cases, the location of the tongue area that underwent high strains and the location of the active muscle show good correlation when single muscle is activated, however; b). for movements involving combined muscle activations, a limited and even no correlation is found. Usually, in speech production, a number of muscles work together to produces a specific articulatory posture [25]. Hence, direct reading of tagged-MRI images would not allow reliable inference of major tongue muscles activated in these cases.

3.1.3 Quantitative analysis based on partial 3D physiological articulatory model

Dang and Honda [35] investigated the functions of tongue muscles on tongue movement based on their partial 3D physiological articulatory model. In their experiments, they moved the control points of the model from the initial position to various start position by activating the four three-muscle combinations (GGp-GGm-GGa, GGp-SG-GGa, HG-GGm-GGa, HG-SG-GGa) with various muscle activations. After the control points arrived at the given start positions, all force were released and the tongue body was driven by specific muscles individually with a given muscle activation (4N). They found that: a). GGa is to lower the tongue tip; b). GGm moves both tongue tip and tongue dorsum downward and forward; c). GGp moves both tongue tip and tongue dorsum upward and forward; d). HG moves the tongue tip backward and slightly upward, while makes the tongue dorsum backward and downward; e). SG drives the tongue tip backward, while makes the tongue dorsum backward and upward; f). IL drives the tongue tip downward and backward, while makes the tongue dorsum upward and backward; g). SL makes the tongue dorsum backward and downward, while moves the tongue tip backward and upward. That was the first study which systematically investigated the muscle functions based on physiological articulatory model. However, due to the limitation of the anatomical orientations of the muscles and the limitation of the reality of the morphological structure implemented in the partial 3D model, the investigations may not give correct results always. Therefore, the functions of the tongue muscle are better to be investigated based on a full 3D model, which is more realistic in both muscle orientation and morpho-

logical structure of the tongue.

3.2 Basic function of individual tongue muscles

Contraction of the tongue muscles will perform movement and deformation in not only the tongue surface but also the part deeply inside the tongue. Both the tongue movement and the tongue deformation are essential for speech production. When the vocal tract is treat as an enclosed acoustic tube in speech production, however, we can mainly focus on the situation of the tongue surface and the surrounding organs. In this part, we investigate the deformation and movement of the tongue surface cause by the contraction of individual tongue muscle.

To investigate the function of individual muscle on the movements and deformation of the tongue surface, we conduct numeric experiments by activating individual tongue muscle with 8 level muscle forces (0.0N, 0.1N, 0.2N, 0.4N, 1.2N, 2.0N, 3.5N, 6.0N) based on the 3D physiological articulatory model, where the boundary constraints of the surrounding structures are not taken into account.

3.2.1 Muscle function on tongue movement

Method

It is widely accepted that the movement of the tongue is nearly symmetric in speech production. Hence, in this study, we use the sagittal movement of two representative points (tongue tip and dorsum) to describe the movement of the tongue with reference to the work of Dang *et al.* [35]. Here, the point for tongue tip is the 11th node along the tongue surface contour in the midsagittal plane, and the point for tongue dorsum is the 7th node along the tongue surface contour in the midsagittal plane (as shown in Figure 2.1).

The tongue muscles are individually activated with monotonically increasing muscle activation from 0N to 6N. The corresponding trajectories of the equilibrium positions of these representative points (tip/dorsum) are extracted from the simulation results to quantitatively evaluate the function of individual muscle.

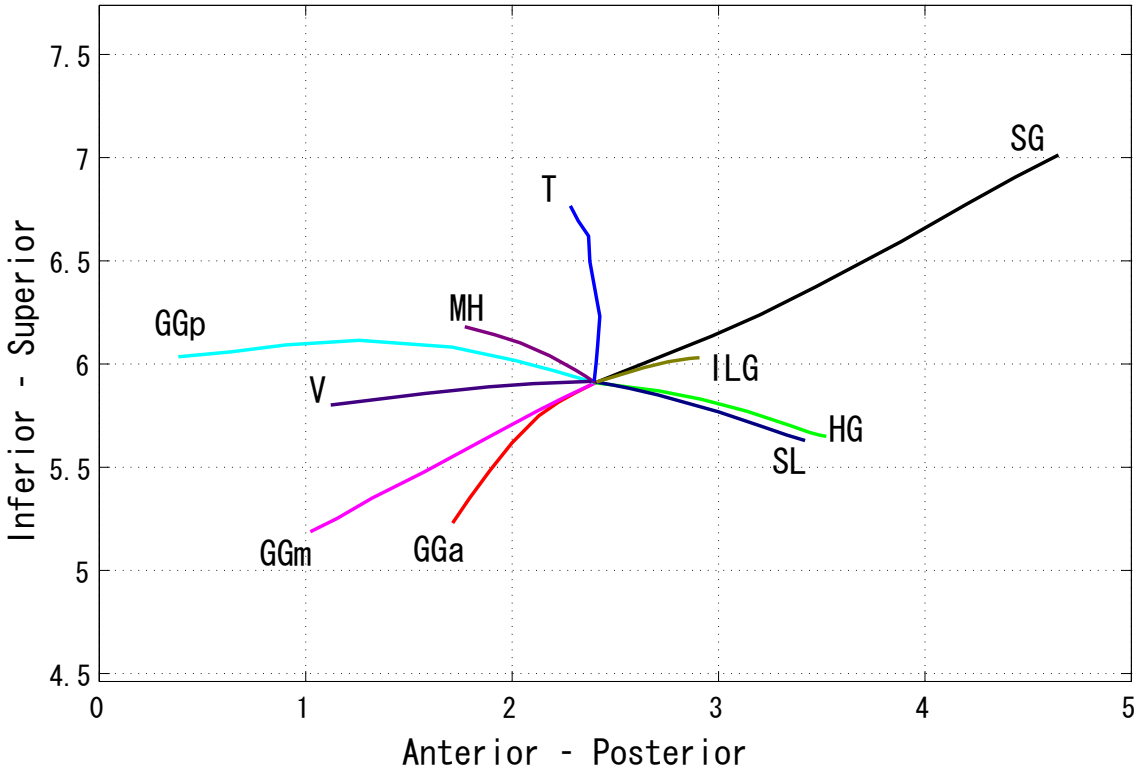

Figure 3.1: Trajectories of the equilibrium positions of the tongue dorsum when corresponding tongue muscle is activated individually. (Unit: cm)

Results

The movements of the representative points of the tongue caused by contraction of individual muscles are shown in Figure 3.1 for tongue dorsum and Figure 3.2 for tongue tip. The centroid points in these two figures are the rest position of the tongue dorsum and tip, respectively. The curves represent the trajectories of equilibrium positions of tongue dorsum/tip when the activation level of the corresponding tongue muscle increases from 0N to 6N monotonically.

As shown in Figure 3.1, the contraction of GGa and GGm pulls the tongue dorsum forward and downward, while shows some difference in the detailed directions. Similar tendencies are also found for tongue tip (shown in Figure 3.2). Muscle V makes the tongue dorsum move forward and slightly downward, and the tongue tip is dropped by the effects

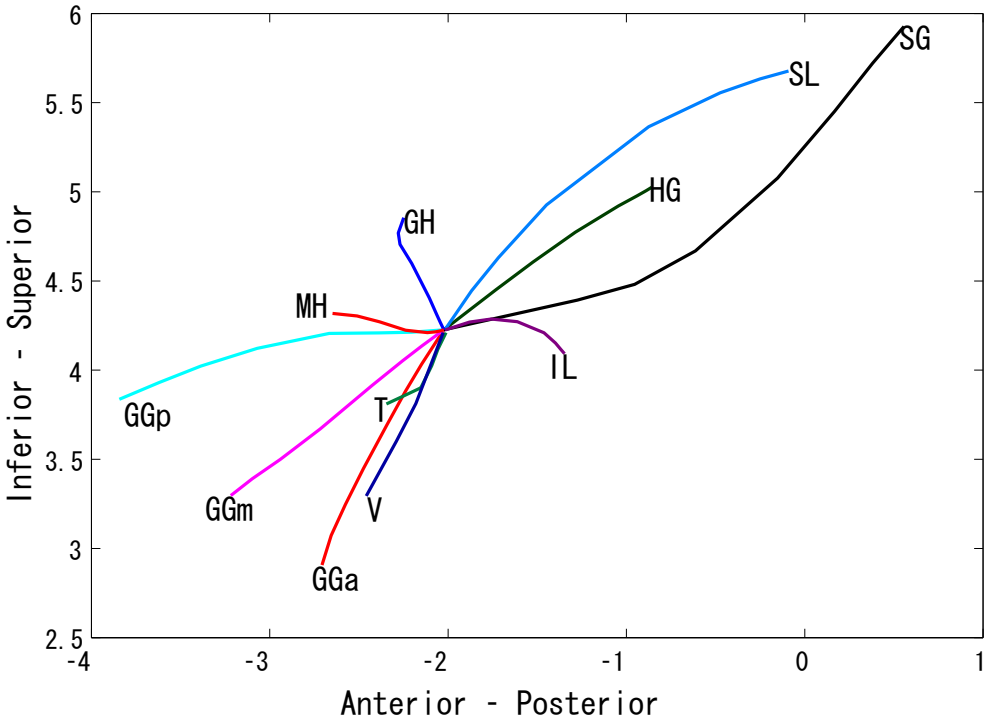

Figure 3.2: Trajectories of the equilibrium positions of the tongue tip when corresponding tongue muscle is activated individually. (Unit: cm)

of extending the tongue length if without the support of jaw. Muscle T moves the tongue dorsum upward, and makes the tongue tip slightly downward by the effects of extending the tongue length if without the support of jaw. The contraction of GGp moves the both tongue tip and dorsum forward without the wall contract. MH contributes to the tongue forward and upward movement. GH makes the tongue tip move upward while does not have clear pattern in the contribution of movement of tongue dorsum. SG draws both the tongue tip and dorsum backward and upward. IL moves the tongue tip backward and slightly downward, while makes the tongue dorsum slightly backward and upward. HG pulls the tongue dorsum backward and downward, while moves the tongue tip backward and upward. SL also pulls the tongue dorsum backward and downward, while moves the tongue tip backward and upward, but shows some difference in the detail direction of the tongue tip movement with that caused by the contraction of HG.

Comparing the trajectories of the equilibrium position of tongue tip and dorsum with speculations based on the corresponding muscle orientation, one can see that, for muscle

GGa, GGm, GGp, SG, and MH, the direction of the movement is consistent with the speculations based on anatomical orientations of these muscles. The contraction of these muscles seems to move the tongue tip and dorsum as an entire object. As for muscle HG, the direction of the movement of the tongue dorsum is consistent with the muscle orientation, while the tongue tip moves to a different direction. As for muscle IL, the direction of the movement of the tongue tip is consistent with the muscle orientation, while the tongue dorsum moves to a different direction.

3.2.2 Muscle function on tongue deformation

Method

To measure the deformation of the tongue surface, we defined the tongue width to represent the average thickness between the 1st and 5th layer (shown in Figure 2.1) of the tongue surface layer in the transversal dimension and the length of tongue contour in the midsagittal plane to express the longitudinal length of the tongue. These two quantities are measured from the simulation results and used in the evaluation. The relative variations in the tongue width and tongue length are calculated according to Eq. (3.1) and Eq. (3.2), respectively.

$$r_w = (w - w_0)/w_0; \tag{3.1}$$

$$r_l = (l - l_0)/l_0; \tag{3.2}$$

where w_0, w are the average tongue widths at the rest state and at the equilibrium state when individual tongue muscle is activated with a certain activation level, respectively; l_0, l are the tongue length in the corresponding situation, respectively.

Results

Using the above method, we measured the relative variation of the tongue width and the tongue length when tongue muscles were activated individually, where the jaw was kept at its rest position. Table 3.1 gives the results of the maximum relative variation of tongue

Table 3.1: The maximal ratios of variations of average tongue width (MW) and the length of tongue contour (ML) in the midsagittal plane.

Muscle	MW	ML
GGa	0.03	0.05
GGm	0.04	0.01
GGp	0.00	0.01
HG	0.03	-0.05
SG	0.02	0.02
SL	0.10	-0.18
IL	0.02	-0.02
T	-0.27	0.22
V	0.16	0.18
GH	0.00	0.06
MH	0.00	0.03

width and tongue length caused by the contraction of individual tongue muscle.

The maximal ratios of the tongue width variation are shown in the second column of Table 3.1. One can see that muscle V and SL show the largest contribution (above 10%) to widening the tongue, and muscle T makes the biggest contribution (27%) to compressing the tongue width. The other muscles contribute to the variation of tongue width less than 5%. A clear pattern can be seen that the muscles located in the superficial layer of the tongue, such as SL, T and V, contribute more to the length change of the tongue, while the muscles, which are oriented deeply inside the tongue, contribute less to the variation of the tongue width.

The maximal ratios of tongue length variation are shown in the third column of Table 3.1. One can see that muscle T and V make the biggest contribution (above 18%) to elongating the length of tongue contour, while muscle SL makes the biggest contribution to shorten the length of tongue contour in 18%. In addition, a clear pattern can be seen that the muscles located in the surface layer of the tongue, such as SL, T and V, contribute more to the length variation of the tongue contour, while the muscles, which are

oriented deeply inside the tongue, contribute less to the length change of the tongue. According to above analysis on the variation of both tongue width and length of tongue contour, the muscles located in the surface layer of the tongue, such as T, SL, and V, make great contribution to tongue surface deformation, while the other muscles show little contribution to tongue surface deformation.

3.3 Agonist-antagonist property of tongue muscle pair

As indicated by EMG experiments [25], in speech production, tongue muscles are co-activated, not activated individually. Therefore, it is necessary to investigate the effects of co-contraction of tongue muscles, especially the agonist-antagonist property which make it possible to maintain the position of tongue tip/dorsum while manipulating the position of tongue dorsum/tip.

Dang *et al.* [35] systematically investigated the agonist-antagonist property of the tongue muscle pairs based on their partial-3D physiological articulatory model. They found that the co-contraction of some muscle pairs show the agonist-antagonist property. However, the results obtained from the partial-3D model cannot give a full image for the co-contraction effects of tongue muscles due to its insufficient representation of morphological structure and musculature of the tongue. In this part, we investigate the agonist-antagonist property of the tongue muscle pairs by numerical experiments based on the proposed 3D physiological articulatory model.

Figure 3.3 shows an example of the agonist-antagonist property of muscle pair GGm-HG. The trajectories of the equilibrium position of muscle GGm and HG for tongue tip lie almost on a straight line, but course to opposite directions (as shown in Figure 3.3(a)). This muscle pair may work as antagonist muscle pair for the tongue tip. However, for the tongue dorsum, these two muscles may work together to depress the tongue (as shown in Figure 3.3(b)). They may work as agonist muscle pairs for tongue dorsum. And we conduct model simulation by appropriately set the activations of these two muscles, and increase them gradually. The results for the tongue tip and tongue dorsum are shown Figure 3.3(a) and Figure 3.3(b), respectively. It says that the tongue dorsum moves to the direction indicated by the brown arrow in Figure 3.3(b) while the tongue tip is kept

in a small area, indicated by the brown ellipsis in Figure 3.3(a). This shows that the muscle pair GGm-HG really work as agonist to depress the tongue dorsum, while work as antagonist to maintain the position of tongue tip.

3.3.1 Method

By inspecting the equilibrium trajectory of tongue muscles for tongue tip and dorsum shown in Figure 3.1 and Figure 3.2, we found that muscle pairs GGm-SG, MH-HG, MH-SL, GGp-HG, GGp-SL may work as agonist for tongue tip while work as antagonist for tongue dorsum; and GGp-SG, GGm-SL, GGa-SL, GGm-HG, GGa-HG may work as agonist for tongue dorsum while work s antagonist for tongue tip.

To quantify agonist-antagonist property of muscle pairs, we introduce the concept of Equilibrium Vector (EV). When a muscle is individually activated with muscle activation *Act*, the vector pointed from the centroid to the equilibrium position is defined as the EV of this muscle at the activation level of *Act*. Then, we define antagonism ability of a muscle pair as its ability to manipulate the tongue tip/dorsum to the direction close to those of the EVs of the muscles recruited in the muscle pair. If the muscle pair has strong antagonism ability, this muscle pair may work as antagonist for tip/dorsum.

As shown in Figure 3.3, the agonist-antagonist property of a muscle pair appears more evidently when both of the two muscles in the muscle pair are activated with higher activation. Hence, we discuss the agonist-antagonist property of the muscle pair during the maximal activation (Act_{max}=6N). For convenience, the muscles involved in a muscle pair are referred to as M_1 and M_2 hereafter, respectively. Accordingly, to explore the agonist-antagonist property of a muscle pair, we measure the antagonism ability of the muscle pair in the above situation.

Figure 3.4 gives a schematic description of the calculation of the antagonism ability of a muscle pair. Here, P_rP_1, P_rP_2 are the EVs of M_1 and M_2 with the muscle activation Act_{max}=6N, respectively. P_{12} is the equilibrium position when both M_1 and M_2 are activated with activation Act_{max}=6N simultaneously. The distance from P_{12} to P_rP_1 and P_rP_2 are denoted as h_1 and h_2, respectively. Then, the antagonism ability of the muscle pair can be indicated by quantities defined in Eq. (3.3-3.4).

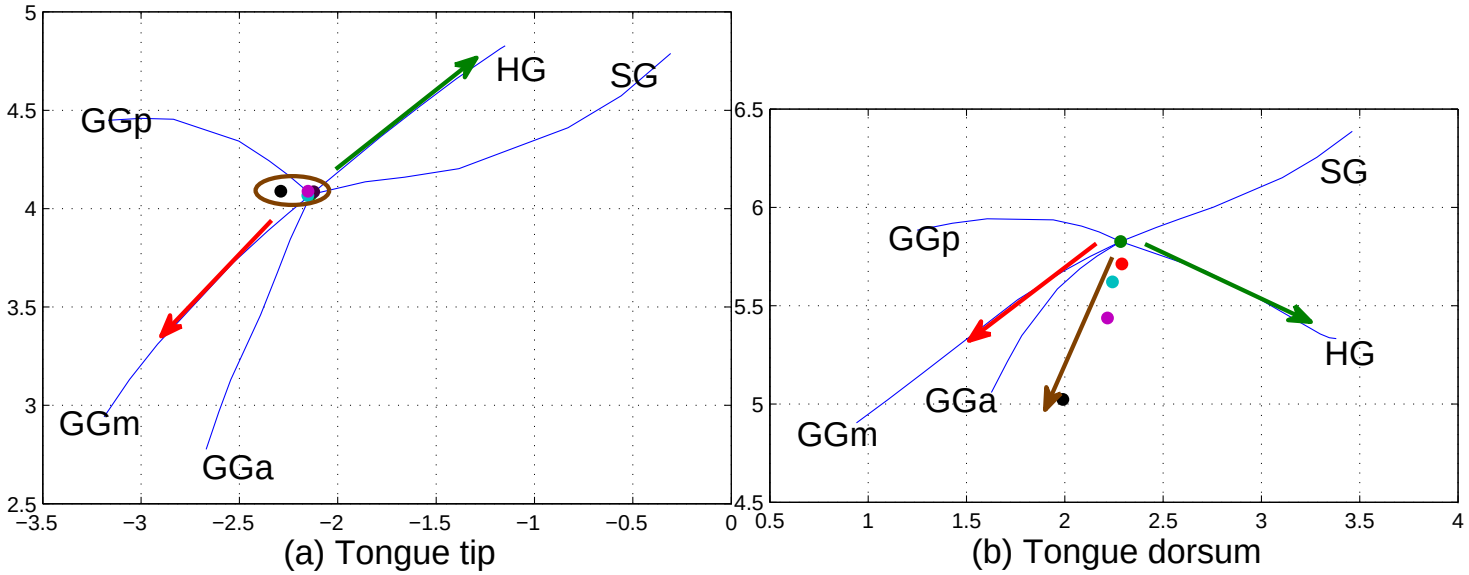

Figure 3.3: An example of agonist-antagonist property of tongue muscle pair.

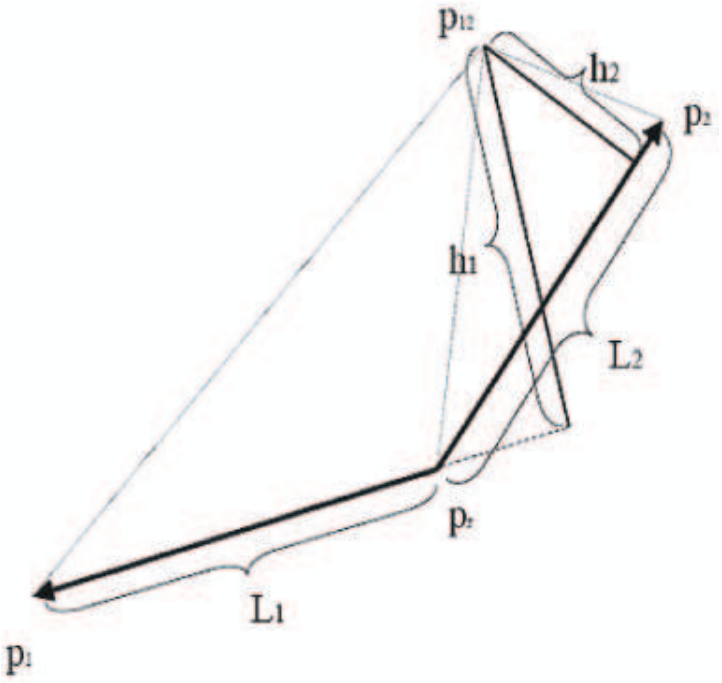

Figure 3.4: A schematic explanation of the meaning of the ratio.

$$r_1 = h_1/L_1; \tag{3.3}$$

$$r_2 = h_2/L_2; \tag{3.4}$$

where L_1 and L_1 are the length of P_rP_1 and P_rP_2, respectively. The smaller the ratios, the stronger antagonism ability the muscle pair has.

3.3.2 Results

Table 3.2 gives ratios of the candidate antagonist muscle pairs for tongue dorsum calculated according to Eq. (3.3-3.4). It shows both r_1 and r_2 of muscle pair GGm-SG, GGp-HG, GGp-SL are less than 17%, while the r_1 for MH-SL and MH-HG are greater than 30%.

Similarly, Table 3.3 gives the ratios of the candidate antagonist muscle pairs for tongue tip calculated according to Eq. (3.3). It shows the r_1 for muscle pair GGp-SG and GGa-SL are greater than 60%, over half of the displacement caused by the contraction of GGa

Table 3.2: The ratios for the candidates of antagonist muscle pairs for tongue dorsum.

	r_1	r_2
GGm-SG	0.07	0.12
MH-HG	0.39	0.25
MH-SL	0.35	0.14
GGp-HG	0.11	0.17
GGp-SL	0.05	0.17

Table 3.3: The ratios for the candidates of antagonist muscle pairs for tongue tip.

	r_1	r_2
GGp-SG	0.88	0.16
GGm-SL	0.20	0.13
GGa-SL	0.68	0.22
GGm-HG	0.16	0.17
GGa-HG	0.10	0.19

and GGp, respectively. For GGm-SL, GGm-HG, and GGa-HG, both r_1 and r_2 are less than 20%.

At the current stage, 30% was chosen as the threshold with reference to the work of Dang *et al.* [35] on the definition of contribution factor. If both r_1 and r_2 of a muscle pair for tongue tip/dorsum is less than 30%, then the muscle pair is considered to be antagonist pairs for tongue tip/dorsum. Consequently, GGm-SL, GGm-HG, GGa-HG are the antagonist muscle pairs for tongue tip, while they act as agonist for tongue dorsum according to this criterion. Similarly, GGp-HG, GGp-SL, GGm-SG are the antagonist muscle pairs for tongue dorsum, while they act as agonist for tongue tip.

3.4 Discussions

So far, for most of the tongue muscles, their functions are mainly estimated from their anatomical orientations [49] or the muscle activity/length with corresponding articulation [25]. In this study, we investigated the function of tongue muscles by using numerical experiment based on the full 3D physiological articulatory model.

3.4.1 Function of GGp

For years, there are several different point of views on the function of muscle GGp. Kumada *et al.* [26] argued that muscle GGp contributes to the elevation of the frontal part of tongue based on MRI observations. Baer *et al.* [25] found that the activation of GGp correlated with the height of the tongue. The higher the tongue was, the stronger activation of GGp was observed by EMG. Nevertheless, Seikel *et al.* [49] argued that the function of GGp was to protrude tongue based on the anatomical orientation of GGp.
As shown in Figure 3.1 and Figure 3.2, the contraction of GGp is to move both tongue tip and tongue dorsum forward if no boundary constraints are taken into account. This result is consistent with the argument [49] based on the anatomical orientation of GGp. Nevertheless, after accounting for the boundary constraints, the horizontal movement of tongue tip is stopped, while the tongue tip and dorsum go up, as shown in Figure 3.5 (a). The results suggest that the function of GGp itself is to move the tongue forward. With the supporting of surrounding structures, the contraction of GGp will elevate the tongue tip and dorsum.

3.4.2 Function of Transversus

The function of muscle T has been speculated from the anatomical structure. "The transversus muscle of the tongue compresses the tongue toward the midsagittal area, effectively narrowing the tongue" [49]. The simulation results show that the contraction of muscle T shows complex behavior. It not only narrows the tongue, but also elongates the tongue. This result supports the speculation of Zemlin [48].

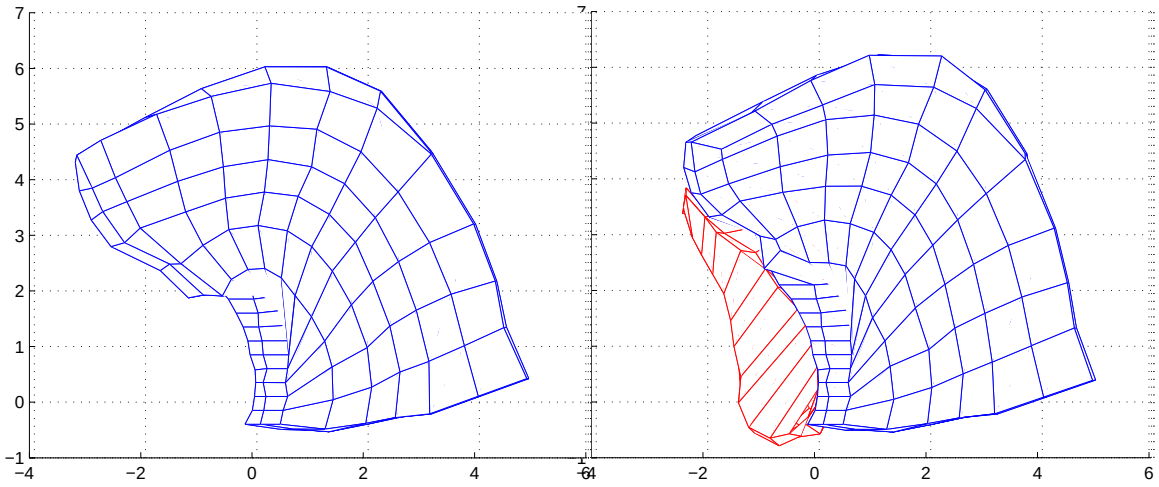

Figure 3.5: The simulation results which activating muscle GGp with 1N force in two situations: a) with boundary constraints of the jaw and vocal tract wall; b) without boundary constraints of the jaw and vocal tract wall.

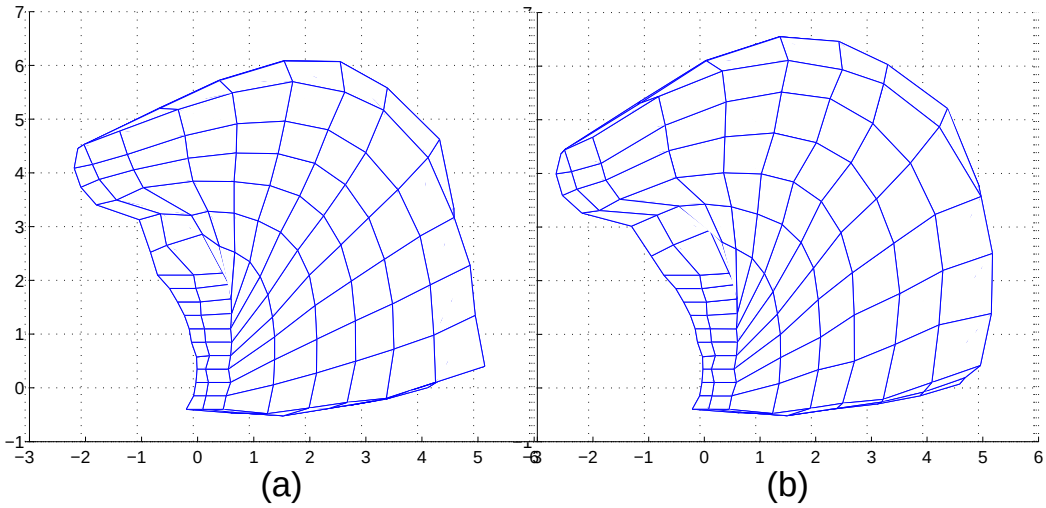

Figure 3.6: The lateral view of the effects by contraction of muscle T by taking boundary constraints into account. (a) Muscle T is activated by 0N force; (b) muscle T is activated by 2N force when boundary of surrounding structure is taken into account.

In addition, the contraction of T contributes to bunch the tongue. Figure 3.6 shows the effects of the contraction of T from the lateral view. The left panel shows the configuration at rest state, and the right panel shows the configuration of the tongue when T contracts with 2N force. It is clear that the dorsum in the right panel is higher than that in the left panel, while the height of the tongue tip almost does not change.

3.4.3 Properties of co-contraction of tongue muscles

Similar to the findings of Dang *et al.* based on the partial 3D physiological articulatory model [35], co-contraction of several tongue muscles pairs show agonist-antagonist properties. It shows that GGm-SL, GGm-HG, GGa-HG act as antagonist muscle pairs for

tongue tip, while as agonist of tongue dorsum; GGp-HG, GGp-SL, GGm-SG work as the antagonist for tongue dorsum while act as agonist for tongue tip. This property makes it possible to maintain the position of tongue tip/dorsum, meanwhile, manipulate the position of tongue dorsum/tip.

This property may help us to understand coarticulation in continuous speech, in which the crucial articulatory configuration should be guaranteed while the others show large variance, from the perspective of motor control. For example, when produce a sequence of "tatatata", the tongue dorsum almost kept at the same position, while the tongue tip moves upward and downward. This may be realized by activating muscle GGp, SL, and HG. Among the three muscles, HG act as an independent muscle to manipulate the position of tongue dorsum, while GGp and SL may work as agonist to moves the tongue upward and downward, and act as antagonist to maintain the position of tongue dorsum. In addition, the agonist-antagonist property of tongue muscles implies an efficient way to organize model simulations so as to construct a feedforward control strategy to manipulate the 3D physiological articulatory model.

3.5 Summary

In this Chapter, we investigated the function of tongue muscles on both tongue movement and tongue deformation, and the agonist-antagonist property of tongue muscle pairs by using numeric experiments based on the 3D physiological articulatory model.

The analysis of the tongue movements, by activating individual tongue muscle with monotonically increasing muscle activations, show that the functions of muscles GGa, GGm, GGp, SG, and MH are consistent with the speculations based on their anatomical orientations. When muscle HG contracts, the direction of the movements of tongue dorsum is consistent with the muscle orientation, while the tongue tip moves in the direction different from the muscle orientation of HG. When muscle IL contracts, the direction of the movements of tongue tip is consistent with the muscle orientation, while tongue dorsum moves in the direction different from the muscle orientation of IL.

The analysis of the tongue deformation illustrates that the muscles located in the superficial layer of the tongue, such as T, SL, and V, make great contribution to tongue

surface deformation, while the other muscles show little contribution to tongue surface deformation.

In addition, the agonist-antagonist property of tongue muscle pairs is investigated with numerical analysis. It is found that the muscle pairs GGm-SL, GGm-HG, GGA-HG act as antagonist for tongue tip, while as agonist for tongue dorsum; GGp-HG, GGp-SL, GGm-SG act as the antagonist for tongue dorsum while act as agonist for tongue tip. This property may help us to understand coarticulation form physiological point of view, and implies an efficient way to organize simulations for model control.

So far, we have constructed a 3D physiological articulatory model and evaluate the functions of tongue muscle by using the 3D physiological articulatory model. In next chapter, we will uncover motor command for vowel production by using the proposed 3D physiological articulatory model and the knowledge of functions of tongue muscles.

Chapter 4

Estimation of motor commands for vowel production

In Chapter 2 and Chapter 3, we introduced the 3D physiological articulatory model and the functions of the tongue muscles involved in the model. In this chapter, we will uncover the motor command for vowel production by using the proposed 3D physiological articulatory model and the knowledge of functions of tongue muscles.

The organization of the rest part of this chapter is in the following way. In Section 4.1, we review the history of estimating activity of tongue muscles. In Section 4.2, the target articulations obtained by MRI are introduced. In Section 4.3, we come up with the cost function that measures the difference between observation and model-based simulation. In Section 4.4, the procedure that estimates the muscle activations of the corresponding articulation is introduced. In Section 4.5, we present an example by using the proposed model-based method to estimate the muscle activation in producing vowel /a/. In Section 4.6, the morphological difference between observation and simulation, and corresponding estimated muscle activations are presented.

4.1 Introduction

During speech production, speech organs are driven by coordinated muscle activations to manipulate the vocal tract shape and provide proper sound source. Tongue is the most

important speech organ that forms vocal tract shape for producing most of the vowels and consonants, and is driven by the contraction of tongue muscles. Since most of the tongue muscles locate deep inside the tongue body, it is difficult to measure the activations of the those muscles directly. For years, speech scientists have developed several methods to explore the activations of tongue muscles during speech production. Among them, tagged-MRI [29, 38, 51] and EMG [21, 23-25] are the most popular approaches.

Tagged-MRI is an approach that helps to trace the internal motion of a deformable body. It is widely used to analyze the cardiac motion and function [52, 53]. Since the 1990's, it has been implemented to estimate activations of tongue muscles by comparing the measured length of muscle fibers [26, 28, 54] or principal strain of the tongue body [29, 55] with a reference configuration. However, the results of obtained by tagged-MRI alone are questionable, because the tongue muscle may be activated even if it is elongated/expanded, and a limited and even no correlation is found between shortening/compression pattern of the tongue tissue while the tongue is driven by the combined muscle activation.

EMG is a technique that detects the electrical potential generated by muscle cells when the muscle is activated. The amplitude of the electrical signal measured by EMG depends on the location that the electrodes attach to [56] as well as the activation level of the muscles. EMG is used in many types of researches, including those concerned with biomechanics, motor control, neuromuscular physiology, movement disorders, postural control, and physical therapy. Baer *et al.* [25] explored the activation of most of the major extrinsic muscles, e.g., Genioglossus posterior, Genioglossus anterior, Styloglossus, Hyoglossus, Geniohyiod, and Mylohyiod, by inserting hooked wire electrodes into those muscles during producing English vowels in the context of /əpvp/. Their data were widely referenced. However, it is difficult to measure the activations of small intrinsic muscles with EMG, such as Transversus (T) and Verticalis (V), which may also play important roles in speech production.

As pointed out above, the previous methods have a variety of flaws in exploring activations of tongue muscles. To suppress those problems, we proposed a physiological articulatory model based method to estimate muscle activations in vowel production. It is assumed that if a faithful physiological articulatory model generates the same articulatory posture as the observed articulation, the muscle activations applied in the model are the possible

ones that the speaker used in the observed articulation. In this way, the activations of major extrinsic muscles as well as small intrinsic muscles are possible to be estimated without suffering uncertain relation between the compression pattern of tongue and muscle activations.

Since the proposed 3D model is more realistic in morphological structure of the tongue and surrounding organs, and the musculature from the anatomical perspective, it can be expected that the results about the muscle activations during speech production estimated by the proposed 3D model are more reliable than those by 2D and the partial 3D models. For this reason, we estimate the muscle activation by using the proposed 3D physiological articulatory model via an optimization procedure, which minimizes the difference between model simulations and corresponding MRI observations.

4.2 MRI-based observation of vowel articulation

In this study, we used the ATR MRI database as the target articulations of vowel production. Figure 4.1(a)-(e) shows the configurations of the vocal tract in producing the five Japanese vowels. And Figure 4.1(f) shows the contours of the tongue in producing the five Japanese vowels. The contours of the tongue for the five vowels show two characteristic patterns of Japanese vowel production. First, while the tongue dorsum is higher in high vowels (/i/ and / u/) and lower in a low vowel (/a/), the highest point of the tongue dorsum for /u/ is not in the rearmost but is between /e/ and /o/. Second, the tongue tip tends to remain in front even in back vowels (/a/ and /o/) and the tongue blade shows a downward indentation. The length of the tongue contours in MRI observation varies from 11.64 (vowel /e/) to 12.96cm (vowel /a/).

Since the tongue is a 3D object, its movement and deformation are carried out in three dimensions [43]. Accordingly, the information of movements and deformations of tongue in both sagittal and transversal dimensions are indispensable. In this study, the tongue width is used as the variable to represent the transversal (left-right) variation of the tongue. When producing the five Japanese vowels, however, lateral parts of the tongue often contact the surrounding structures. It is difficult to measure the tongue width from the lateral surface directly. For this reason, we follow Takano *et al.* [43] and use the sharp

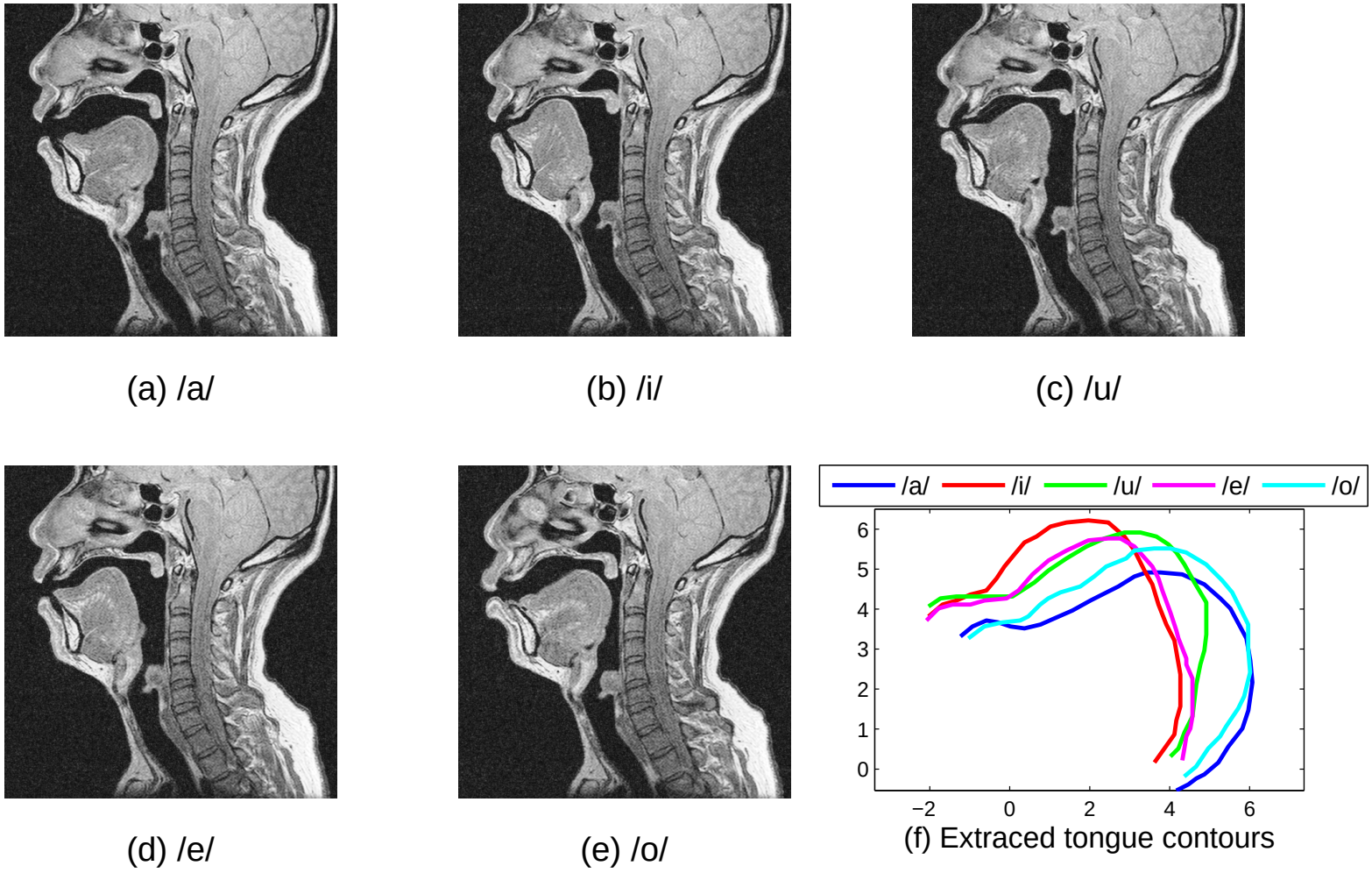

Figure 4.1: The observations of the MR images in the midsagittal plane for the five Japanese vowels: (a) /a/; (b) /i/; (c) /u/; (d) /e/; (e) /o/; and (f) the extracted tongue contours in the midsagittal plane (unit:cm).

bending points of the deep branches of the lingual arteries as the landmarks, which are identified on each side of the tongue by using the coronal images of the subject while producing the five Japanese vowels (see Takano *et al.* [43] for details). The transversal distance between the right and left landmarks was measured to serve as the measurement of the tongue width.

4.3 Cost function for minimizing differences between observation and simulation

To estimate the underlying muscle activation in vowel production, the 3D physiological articulatory model is driven to approach the postures of the observed articulations by adjusting the activation of associated muscles via an optimization procedure. The cost function of Eq. (4.1) is designed to minimize the difference between the simulated and observed postures in 3D. The right side of Eq.(4.1) is the summation of two terms: the first term accounts for the difference of the tongue contours in the midsagittal plane, while the second term is related to the difference in the transversal dimension. Using this equation, the 3D difference of the tongue shapes between the simulation and observation can be accounted for to some extent.

$$D = \frac{(\boldsymbol{s} - \boldsymbol{s}_0)^T(\boldsymbol{s} - \boldsymbol{s}_0)}{N_{\boldsymbol{s}}} + \frac{(\boldsymbol{w} - \boldsymbol{w}_0)^T(\boldsymbol{w} - \boldsymbol{w}_0)}{N_{\boldsymbol{w}}} \tag{4.1}$$

where $\boldsymbol{s}_0$ and $\boldsymbol{s}$ are the vectors that represent the midsagittal contours of the tongue in the observation and simulation, respectively; $\boldsymbol{w}_0$ and $\boldsymbol{w}$ are the tongue width vectors obtained from observation and simulation, respectively; $N_{\boldsymbol{s}}$ and $N_{\boldsymbol{w}}$ are the number of dimensions of vector $\boldsymbol{s}$ and $\boldsymbol{w}$, respectively. Both $\boldsymbol{s}$ and $\boldsymbol{w}$ are functions with regard to muscle activation $\boldsymbol{f}$.

To calculate the difference between $\boldsymbol{s}$ and $\boldsymbol{s}_0$, we define a semi-polar coordinate system (shown in Figure 4.2)which avoids matching corresponding nodes between observation and simulation and approximately represents the cross-sectional dimension of the vocal tract. The contour of the tongue surface in the midsagittal plane is represented by the lengths of an ordered sequence of line segments that are nearly perpendicular to the tongue surface

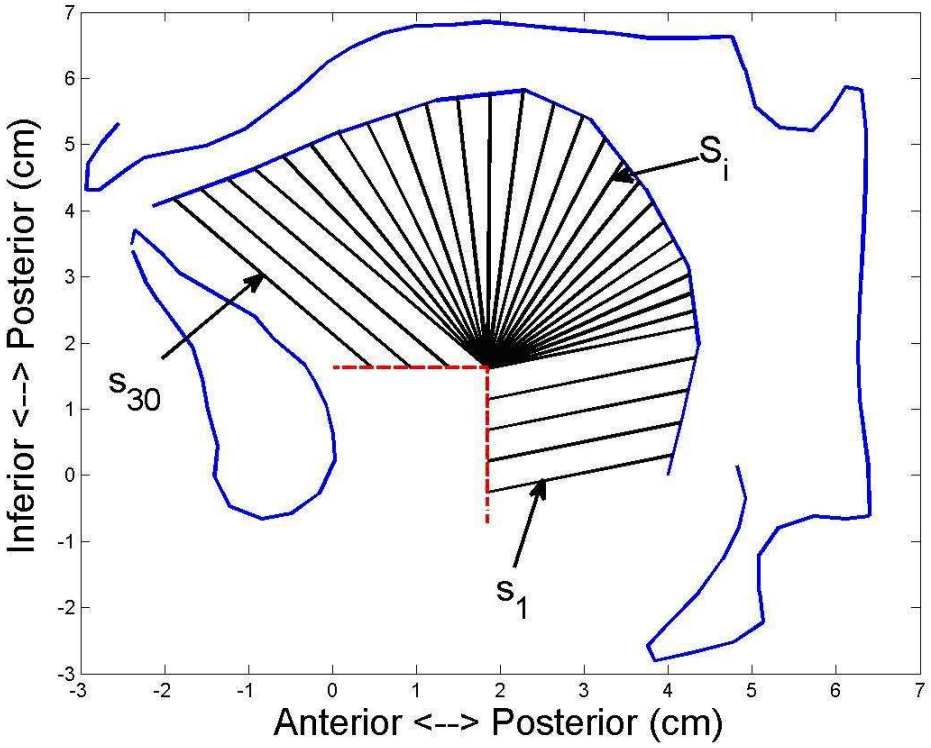

Figure 4.2: The semi polar system for evaluating the difference between the real tongue shape and simulation result.

(s_1, s_2, ..., s_{30}). If the tongue contour intersects with a grid line, then the value of this dimension determined by the grid line is the length between the origin of the grid line and the intersection, otherwise, the corresponding length is set to be zero. To measure the difference in the transversal dimension, the tongue width in the simulation is defined as the average distance between the 2nd and 4th layers (shown in Figure 2.1) in the left-right dimension, which approximates the distance between the anatomical landmarks mentioned in Section 4.2.

Therefore, estimating the muscle activation arrives at looking for the activation vector $\boldsymbol{f}$ of the tongue muscles that minimize the difference between observation and simulation. Nonetheless, it is difficult to explicitly define an analytic relationship between $\boldsymbol{s}$ and $\boldsymbol{f}$, and $\boldsymbol{w}$ and $\boldsymbol{f}$ due to the complicated interaction between the tongue and surrounding structures. Hence, the gradient descent algorithm, which is widely used in various optimization tasks, is implemented to estimate the underlying muscle activation, as defined in Eq.(4.2).

$$\boldsymbol{f}_n = \boldsymbol{f}_{n-1} - \frac{0.05}{\lambda}\nabla_{\boldsymbol{f}} D; \lambda = max\{|d_i||i = 1, ..., N_f\} \tag{4.2}$$

where $\boldsymbol{f}_n$ is the force vector at the current step, $\boldsymbol{f}_{n-1}$ is the force vector at the pervious step, d_i is the *ith* element of vector$\nabla_{\boldsymbol{f}}$D, and $N_{\boldsymbol{f}}$ is the number of elements in vector $\nabla_{\boldsymbol{f}}$D. To calculate the gradient of D with regard to $\boldsymbol{f}_n$, we adopt the method proposed by Shirai and Honda [57] to approximate the partial derivative vector. That is, for a given small variation of $\Delta\boldsymbol{f}$ around $\boldsymbol{f}_{n-1}$, we obtain the corresponding variation Δ[D($\boldsymbol{f}$)] by model simulation.

4.4 Procedure of estimating muscle activation

To make the simulated tongue shapes as close to the observations as possible and guarantee the rationality of the estimated activation patterns of tongue muscles, in this study, we start our estimation with reference to the observed EMG signals. Then, additional muscles or unnecessary muscles are activated or deactivated to reduce the difference between model simulations and objective observations based on the knowledge on the functions of tongue muscles. Comparing the anatomical orientation of the muscles in the proposed

Table 4.1: The muscles used in the initial conditions (the second column) which are derived from the experiment of [25].

Vowels	Muscle combinations
/a/	SG, HG
/i/	GGm, GGp, MH
/u/	GGp, HG, SG
/e/	GGm, HG, MH, SG
/o/	SG, HG

3D model and the description on the position of the electrodes attached to in Baer's experiment [25], we notice that the muscle GGa in Baer's experiment corresponds to the muscle GGm in our model, whilst the others have consistent arrangements. Therefore, we can decide the initial setting with reference to their measurements. When we inspect the EMG observation presented by Baer *et al.*, the activation of muscle GGm (GGa in Baer *et al.*'s experiments) in producing vowel /a/, /u/, and /o/, SG in producing vowel /i/, and GH in producing all the vowels are less than 1/8 of the maximum observed muscle activation in the experiments. Accordingly, those muscles are treated as inactive in the initial conditions.

We put forward a procedure shown in Figure 4.3 to estimate the muscle activation. At first, the activation of JawOp/JawCl is estimated by approximating the jaw in the model to the MRI observation. Then, we choose the muscle combinations for producing the vowels according to the EMG observations in Baer's experiments [25] as the initial configurations (see Table 4.1 for details), and set their initial activation levels of the muscles with reference to the relative amplitude of the EMG signal. After that, the optimization method (see Section 4.3) is implemented to obtain the optimal muscle activations that minimize the difference between simulation and target MRI observation with current muscle combination. If the difference is greater than a predefined threshold, we refine the muscle combinations by manually activating additional muscle if necessary and/or deactivating the muscle with little contribution, then go back to the optimization step. Otherwise, the optimal muscle activations and the corresponding tongue shape are out-

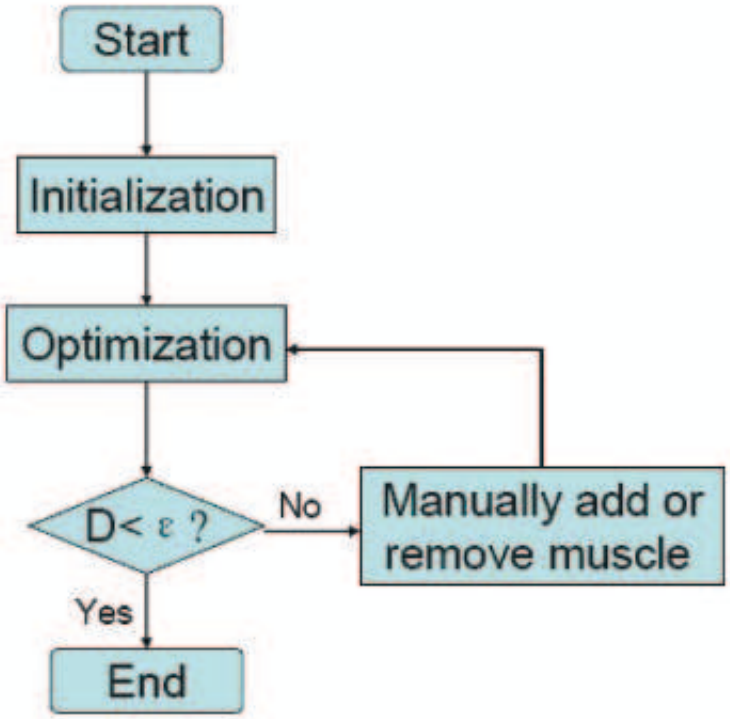

Figure 4.3: The procedure for estimating muscle activations during vowel production.

put.

4.5 An illustration of implementing the proposed model-based method

Figure 4.4 shows an example of estimating the muscle activations for producing vowel /a/ according to the above procedure, where the observed tongue contour is indicated by the curve with square markers. When activating the muscles observed in EMG experiment alone, we obtained the simulation shown by the curve with open circles. One can see that there is an obvious difference in the anterior portion although in the posterior portion the simulated contour of the tongue is consistent with the observation. Especially, the length of the contour of the simulation in the anterior portion is much shorter than that of the observation. Accordingly, to elongate the anterior part of the tongue, muscle T and V are

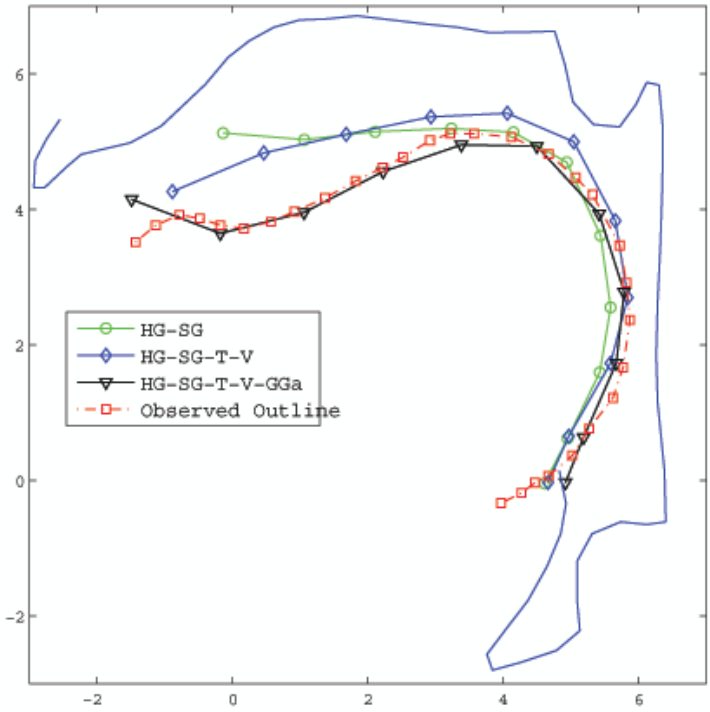

Figure 4.4: The simulation that is most similar to the corresponding observation. The curve consists of the squares is the observed outline of the tongue in the midsagittal plane. The curve consists of circles is the outline of the simulated tongue in the midsagittal plane by considering HG and SG only. The curve consists of diamonds is the outline the simulated tongue by considering HG, SG, T and V. The curve consists of triangles is the outline of the simulated tongue by considering HG, SG, T, V, and GGa.

selected with reference to the anatomical literature [48] and numerical experiments [58]. After taking T and V into account, the tongue contour becomes closer to the observation, indicated by the curve with diamond makers. However, there are still differences between the simulation and the corresponding observation at tongue blade. This suggests that additional muscles are necessary to depress this part. It is known that the GGa is an appropriate candidate with reference to the anatomical structure [31] and experiment data [28, 29, 59]. After taking GGa into account, the contour with inverse triangles obtained by simulation becomes coincide with the observation.

4.6 Results

In this part, we present the results of morphological difference between the observations and corresponding simulations, and the estimated muscle activations by using the proposed model-based method.

4.6.1 Differences of tongue shape in the midsagittal plane and transversal dimension

The model based method provides good approximation to the MRI observation both with contour of the tongue in the midsagittal plane and the tongue width in the left-right dimension. The comparisons of differences in the midsagittal plane and transversal dimension between the simulation and observation are shown in Figure 4.5 for the five vowels. The final configuration in the midsagittal plane for the simulations (solid curves) and target MRI observations (dashed curves) of the five vowels /a/, /i/, /u/, /e/, and /o/ are shown in Figure 4.5(a)-(e), respectively. One can see that the simulations coincide with the target MRI observations when the model is driven by the estimated muscle activations. The average differences (err_A) between the simulations and observations along the contour are 0.10, 0.11, 0.12, 0.15, and 0.08cm for vowel /a/, /i/, /u/, /e/ and /o/, respectively. The difference at the constrictions (err_C), which are crucial for vowel production and denoted by ellipsis, are 0.09, 0.09, 0.03, and 0.05cm for /a/, /i/, /u/, and /o/, respectively. Figure 4.5(f) gives the tongue width in producing the five vowels /e/, /a/, /i/, /u/, and /o/ obtained by using model simulation and (gray bars) obtained by MRI (black bars). The quantitative analysis indicates that the differences of the tongue width between the simulations and observations for each vowel are less than 0.07cm.

4.6.2 Estimated muscle activations

The muscle activations corresponding to the above optimal simulations are also obtained by the propose method (as shown in Figure 4.6). The black bars denote the estimated muscle activation in producing the five vowels, and the gray bars denote the corresponding normalized EMG observations. In the figure, the muscle is activated if its activation level

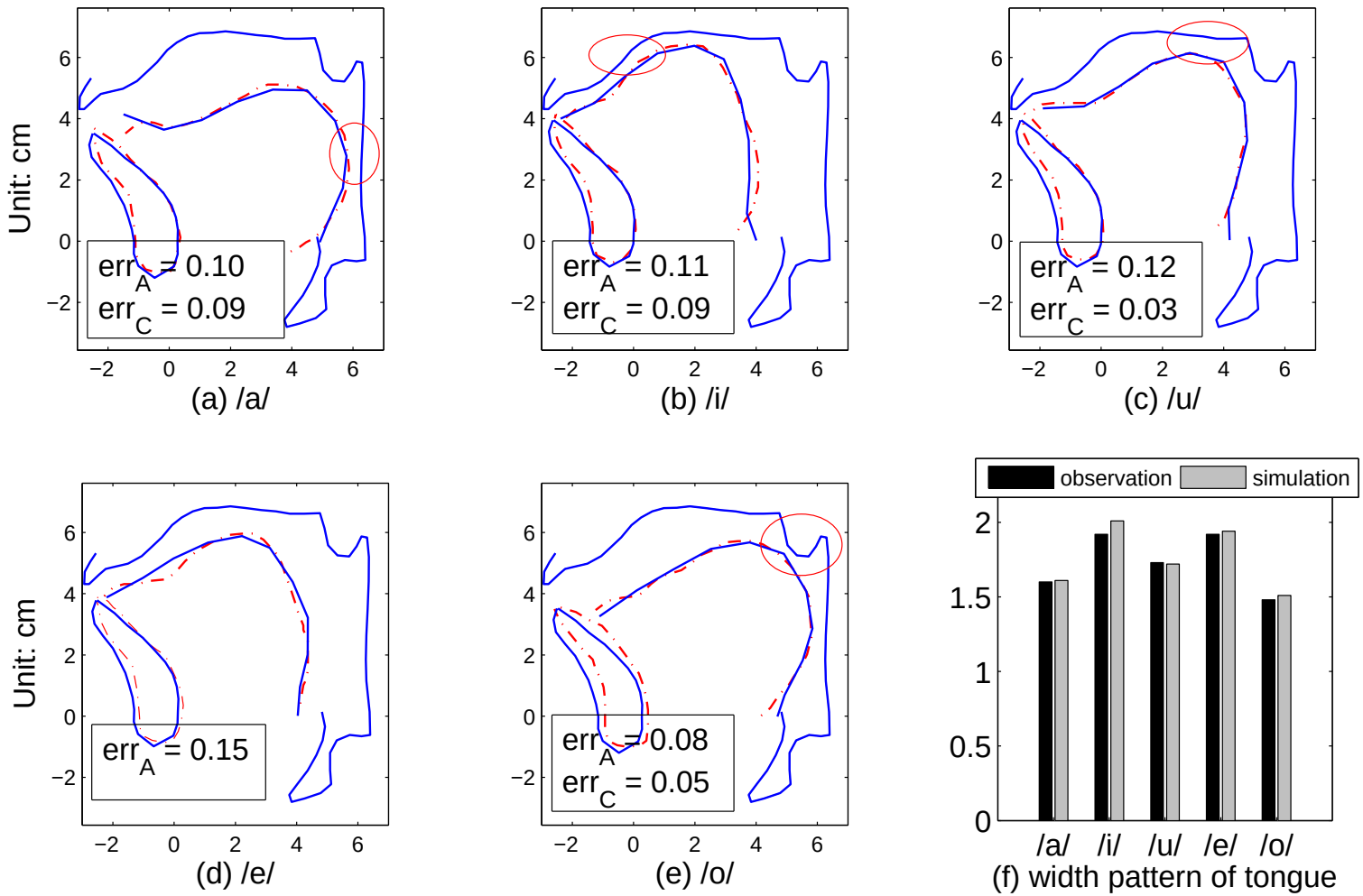

Figure 4.5: The result obtained by optimization. Panel a, b, c, d, and e are the outline of tongue in the midsagittal plane for the observation (dash curves) and simulation (solid curves) of vowel /a/, /e/, /u/, /e/ and /o/, respectively. Panel f is the relative tongue width pattern of vowel /e/, /a/, /i/, /u/, and /o/, respectively.

is non-zero. For producing vowel /a/, muscle GGa, HG, SGp, T, and V are activated; for producing vowel /i/, muscle GGa, GGm, GGp, SGa, SGp, T, V and MH are activated; for vowel /u/, muscle GGa, GGp, HG, SGa, SGp, T and V are activated; for vowel /e/, muscle GGa, GGm, GGp, SGa, SGp, T, V, and MH are activated, for vowel /o/, muscle GGa, HG, SGa, SGp, T, V are activated.

Muscle GGa, T, and V are found to be activated in the production of all the five vowels, though the levels of activation show some differences. In addition, the tongue muscles show relative stronger activation in producing four peripheral vowels (/a/, /i/, /u/, and /o/) than that in producing relative neutral vowel /e/. HG and SGp show stronger activations in producing back vowels /a/, /o/, and /u/. GGa and MH are activated with relative higher activation level for vowel /i/ and /e/ than for other vowels. GGp showed strong activations in high vowels /i/ and /u/.

4.7 Discussions

4.7.1 Activations of tongue muscles in vowel production

Kakita *et al.* [39] estimated the activation of tongue muscles with an 3D physiological articulatory model, which is governed by a set of quasi-static equations. However, in their study, they did not evaluate differences between the simulated tongue shapes and real articulations, which are one of the important factors to evaluate the reliability of the estimated muscle activation. So, it is difficult to compare our results with their results directly.

Baer *et al.* [25] measured the activations of major extrinsic tongue muscles when producing /əpvp/ utterances. They found that 1) GGa (corresponding to GGm in our 3D model) and MH were more active for front vowel /i/ and /e/ than for back vowels; 2) GGp was the most active in producing high vowel /i/ and /u/; 3) HG and SG showed stronger activations in producing back vowels /a/, /o/, and /u/; and 4) GH was more active for the front vowels than that for the back vowels. The results obtained with the proposed 3D model-based method are consistent with most of the observations by EMG, except that GH seem have no activation on the production of vowels. This suggests that

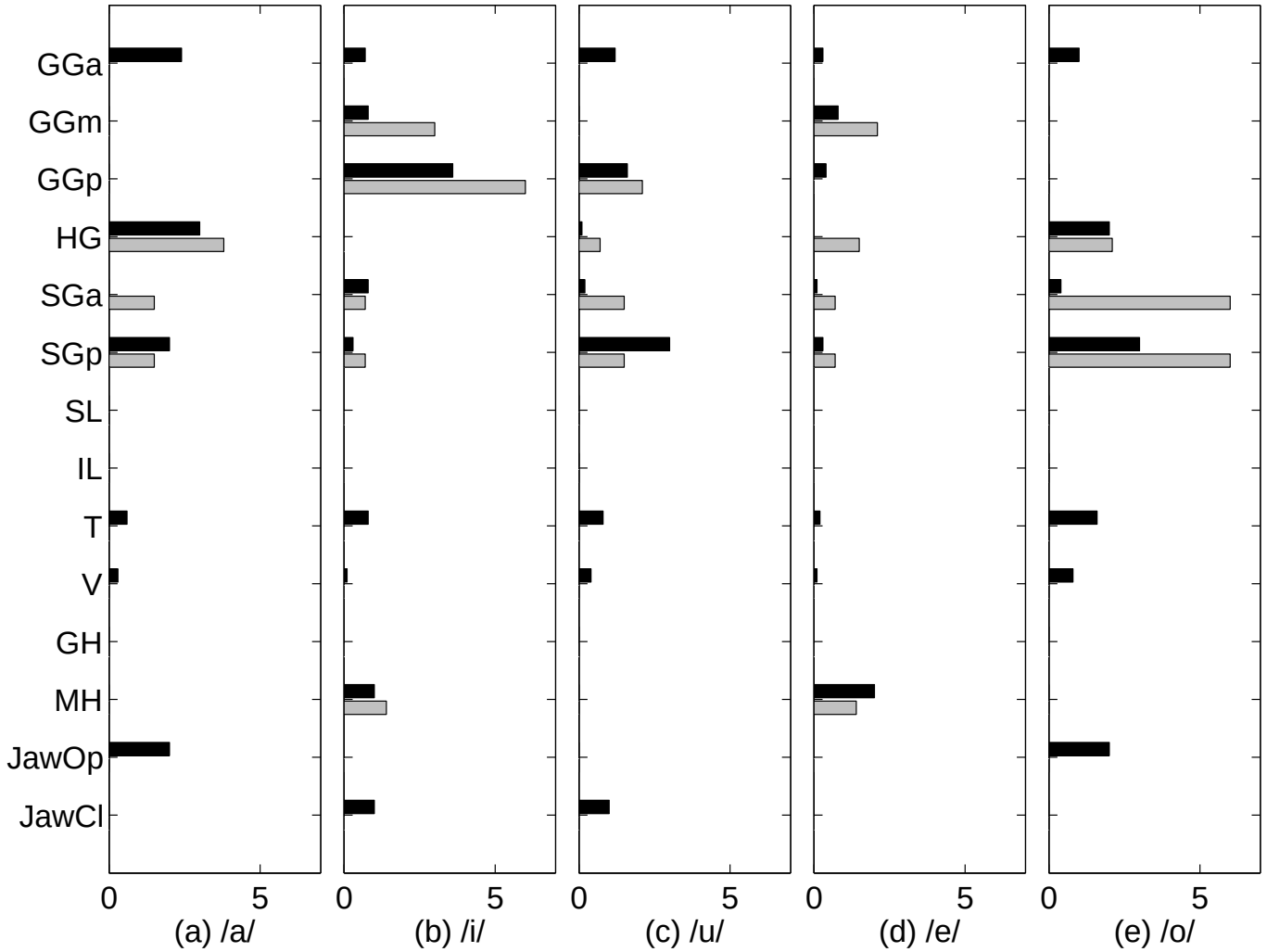

Figure 4.6: The activation of tongue muscles for the five Japanese vowels obtained by the 3D physiological articulatory model-based method. The black bars are the results obtained by proposed method, and the gray bars are the corresponding normalized EMG measurements. (Unit: Newton).

the proposed method is able to reproduce reasonable muscle combinations in producing the five Japanese vowels.

4.7.2 Function of GGa in vowel production

In this study, the GGa is defined as a subdivision of the GG that runs vertically in the anterior part of the tongue. The function of GGa has been widely investigated in vowel production. Stone *et al.*[29], and Niimi *et al.* [28] found that the GGa was associated with local depressions of the anterior tongue for vowel /a/. Takano *et al.* [43] found that GGa was shorter in back vowels than in front vowels. They suggested that GGa was activated in producing /a/ and /o/. In our study, we found that GGa was activated in producing all the vowels. It is widely accepted that the tongue is pushed forward and upward by activating the muscle GGp in producing vowel /i/ [34, 43, 60, 61], and the constriction formed by the anterior portion of the tongue and the corresponding part of the palate should be precisely controlled [62-64]. In our simulation, when we activated the muscles for producing the vowel /i/ as shown in Figure 4.6, while deactivating GGa, the anterior portion of the tongue collapsed fully on the palate, and no airway is formed in this portion. Activating GGa forms a small duct in the anterior part by pulling down the midsagittal area of the tongue. This simulation supports the speculation about the function of GGa in producing high vowel /i/. These results are consistent with those found by Kakita *et al.* [39] in their model-based work.

4.7.3 Function of T and V in vowel production

So far, most previous studies focused on the function of the extrinsic muscles in vowel production. Only few of them tried to shed light on the functions of the intrinsic muscles T and V in vowel production. Takano *et al.* [60] found that T played an important role in elevating the tongue blade by comparing motion patterns obtained from the tagged-MRI observations and the simulation based on a four-block tongue model. Stone *et al.* [29] found that the V contributed to the backing and lowering of the tongue surface in producing vowel /a/. However, nobody mentioned the co-contraction of muscles T and V, and

its function in vowel production. In this study, we found that the length of the tongue in MRI observation varies from 11.64 (vowel /e/) to 12.96cm (vowel /a/), especially the anterior part is elongated. According to the anatomical structure of the tongue muscles, no muscle is directly responsible for extending the tongue in longitudinal direction. And the co-contraction of the major extrinsic muscles manipulates the whole tongue body, especial the posterior portion of the tongue. In contrast, the intrinsic muscles contribute to the local deformations. Among the intrinsic muscles, the co-contraction of muscles T and V can shrink the cross-sectional area of the tongue and extend the tongue in the longitudinal direction due to the incompressibility of the soft tissue. After taking the co-contraction of T and V into account, the tongue length varies from 11.74 to 12.48cm in producing vowels, which approximates the range of the tongue length variation in MRI observations. Meanwhile, the variations of the tongue width are consistent with the MRI observation, which is concerned with the activation of muscle T. The consistency in 3D reveals that the co-contraction of intrinsic muscles T and V play an important role in controlling the elongation of the tongue in vowel production.

4.7.4 Control units of SG

The muscle SG is known to play an important role in producing back vowels /u/ and /o/ [25]. However, there is little information on the control unit of muscle SG in speech production. Most of the physiological articulatory models [15, 33-35] treat the whole SG as one control unit. In contrast, Takano *et al.* [43] found that the three parts of SG (SGa, SGm ,and SGp see [43] for the details) showed different behaviors in producing Japanese vowels. The different behaviors may be resulted from separate controls on the different parts of SG. we divide the SG into two parts, the SGa (inside the tongue) and the SGp (from the styloid process to the insertion), and control them independently in the model based estimation process. The result demonstrated that the activation ratio of SGa to SGp varies between 0 and 2.7 among the production of the Japanese vowels, rather than always the same amount of activation. This suggests that our two part treatment for muscle SG is appropriate.

4.8 Summary

In this chapter, we estimated muscle activations for vowel production by driving the articulatory model to approximate the observed vowel articulation via an optimization procedure which aims to minimize the 3D difference between simulation results and observations. It shows that the model can generate the specific articulations of observed vowel production. And the estimated muscle activations are consistent with those observed from EMG experiments [25].

In addition, the model based estimation method revealed that muscle T, V, and GGa are activated in producing all the five Japanese vowels. T and V play an important role in manipulating the length of the tongue surface in the longitudinal direction, and GGa shows important function in the active control of the tongue blade.

In order to provide more freedom to the proposed 3D model, we separated the SG into two portions in control our model: anterior portion (inside the tongue), and posterior portion (from the styloid process to the insertion). The estimation demonstrates that the two-part treatment of SG is appropriate and gives optimal matching between the model simulations and the MRI observations.

In this study, we used relative simple task, vowel production, to testify our proposed estimation method. The results implied that the 3D physiological articulatory model based estimation method is available and powerful to uncover the muscle activations of the intrinsic muscles as well as extrinsic muscles in vowel production.

In this chapter, we estimated the motor commands (muscle activation) from observed specific articulation. However, usually, in continuous speech, the observed articulation is not the same as the desired articulatory targets [32]. To deal with coarticulated articulations, it requires uncovering desired articulatory targets first, then, uncovering the motor commands for the phonemes of the utterance. Therefore it is necessary to elaborate an efficient control strategy to manipulate the physiological articulatory model according to desired articulatory target. This problem will be dealt with in the next chapter.

Chapter 5

Feedforward control of the 3D physiological articulatory model

Usually, the articulation of a same phoneme is with certain variation. The articulation is not exactly the same from time to time. The method described in Chapter 4 is expensive for uncovering the motor commands for articulations with variation because it needs an explicit optimization for each articulation. In addition, in daily communication with speech, people usually speak to each other with continuous speech rather than isolated phonemes. In continuous speech, the observed articulation is not the same as the desired articulatory targets [32]. To deal with the coarticulated articulations, it requires uncovering desired articulatory targets first, then, uncovering the motor commands for the phonemes of the utterance. All of these need a feedforward control strategy to manipulate the physiological articulatory model according to desired target.

Figure 5.1 shows the flowchart of feedforward control of physiological articulatory model. It consists of "Motor Command Generator", which generator motor command according to input target, and "Physiological articulator model", which execute the motor command. Within the flow of feedforward control, the essential part is how to generate motor commands according to input target.

In the present Chapter, we attempt to elaborate a motor command generator, which can be represented by mappings from articulatory targets to motor commands, for the physiological articulatory model. The articulatory target, here, is a kind of representation in physical level. In the "Motor Command Generator", an articulatory target specified by

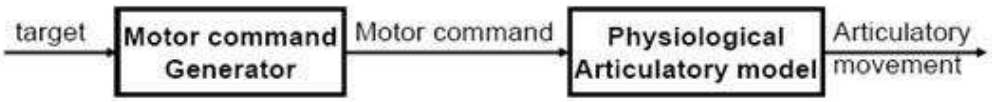

Figure 5.1: The flowchart of feedforward control of physiological articulatory model.

articulatory parameters is transformed to corresponding intrinsic representation of the target articulation, then, a proper mapping function is chosen to output appropriate motor command according to the intrinsic representation. These will be address at length in this chapter.

To elaborate the "Motor Command Generator" module for feedforward control of the physiological articulatory model, firstly, we conduct model simulation to obtain the paired articulatory posture-muscle activation data. This is addressed in Section 5.2. Then, a set of articulatory parameters, which are readily to be interpreted from articulatory point of view, are extracted from acoustically and articulatorily constrained simulations to describe the articulatory targets. This is handled in Section 5.3. At last, the functional relationship between articulatory posture and muscle activations is elaborated by using neural network. And this is described in detail in Section 5.4.

5.1 Introduction

In the literature, several methods for generating motor commands from input targets have been proposed for feedforward physiological articulatory models. They can be categorized into two groups: the mapping from target to λ commands, and the mapping from target to muscle activation.

5.1.1 Target to λ commands

Sanguineti et al. [16] inferred the λ commands by driving a 2D physiological articulatory model to approximate observed tongue shapes, which were acquired by means of low-intensity X-rays at a sampling rate of 50Hz. Then, they used factor analysis to extract the components of articulatory postures and λ commands, respectively. And the articulatory components and λ command components are associated by a linear mapping. However, in their paper, they evaluated neither the articulatory accuracy nor the acoustic accuracy which were achieved by their control strategy.

Perrier et al. [18] mapped the required acoustic target to a set of λ commands to control their 2D model to generate vowel-consonant-vowel sequences. In their study, the conducted 8800 different simulations to describe a large variety of tongue shapes; including both vowels and consonants, by the randomly sampling λ space. The area functions of the generated vocal tracts were computed with an enhanced area function model [65]. Since their biomechanical model does not include any description of the lips, the cross-sectional area of the acoustic tube surrounded by lips is approximated with average lip areas for spread lips (3 cm^2) and for rounded lips (0.5 cm^2), respectively. For each obtained articulation, the two area functions are obtained by combining with the two types of area of lip tube. Corresponding formant values were computed from the final tongue shape of each simulation by a transmission line model [66] based on the calculated area function. Finally, a functional model of the relations between motor control variables (λ-command) and formant patterns (F1, F2, F3) was elaborated by using Radial Basis Function neural network. Using the elaborated control module, the λ commands can be generated according to the desired acoustic target, and the differences between the obtained formant frequencies and desired formant frequencies are less than 3%.

The λ-model formulated in Eq. (1.1) was first derived for skeletal muscles. The length of skeletal muscle is easily to be measured. So, there is only one unknown variable in Eq. (1.1). This makes it easy to evaluate the λ commands. However, it is almost impossible to measure the length of fibers of the tongue muscles because they highly interweave with each other. So, there are two unknown variables in Eq. (1.1). This makes it difficult to evaluate the λ commands estimated from the observed articulation.

5.1.2 Target to muscle activation

Though λ-model is theoretically plausible, the λ commands are difficult to be observed and to be quantitatively measured. Dang and Honda [35, 47] proposed their control strategies from functional point of view. In those strategies, the muscle activation is manipulated directly.

Muscle workspace

Dang and Honda [47] proposed a method, named muscle workspace, to generate muscle activation according to the difference between the current positions of the control points of the model and the desired targets. In that method, three control points (the tongue tip, tongue dorsum, and jaw) are used to describe the sagittal movements of their articulatory model. 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, and the control point for the jaw is 0.5 cm inferior to the tip of the mandible incisor. In their multipoint control strategy, muscle workspaces are constructed for each control point. Each muscle vector in a muscle workspace corresponds to a displacement of the control point when the corresponding muscle contracts. The muscle force vectors in the workspace are adjusted according to the changes of the muscle orientation caused by the movements of jaw and tongue. This is realized by constructing a set of typical muscle workspaces, the distribution of which is designed to cover the articulatory space of both vowels and consonants. Four typical workspaces are constructed for the control points of the tongue tip and tongue dorsum, respectively, and two workspaces for the jaw (As shown in Figure 5.2). And the muscle workspace at arbitrary position is obtained by the interpolation of the typical muscle workspaces.

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 the direction towards the target, 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

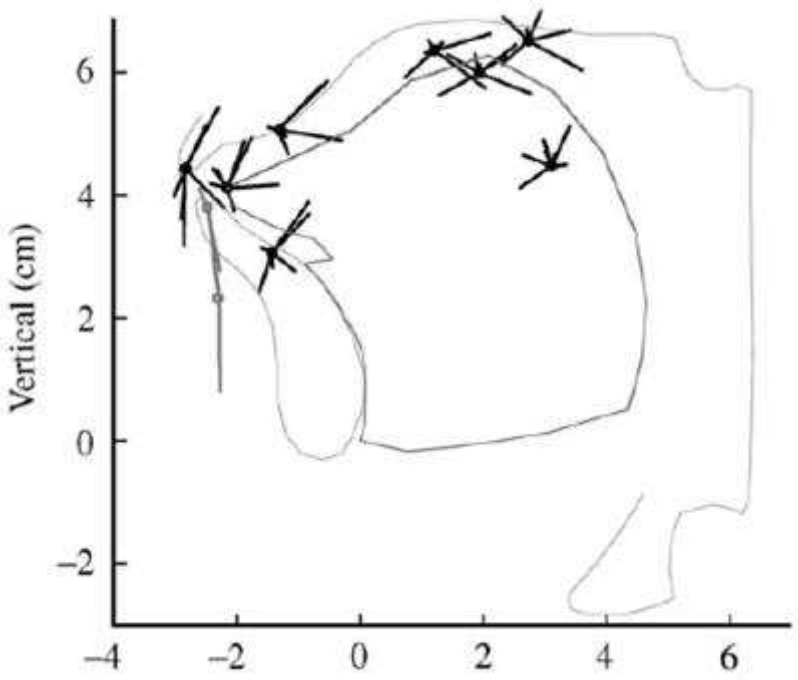

Figure 5.2: Typical muscle workspaces for three control points. Four muscle workspaces were built for tongue tip (dark lines surround the tongue tip) and tongue dorsum (light lines surround the dorsum), and two for the jaw (light lines). (After Dang and Honda [47])

the displacement of the control point. The muscle activation signals can be obtained for any arbitrary movement using this approach.

Ep-map

Dang and Honda [35] found that the control points, which start from extreme positions, converge to sufficient small regions when the activation duration is sufficiently prolonged, and the relationship between a muscle force and an equilibrium position is unique based on model simulation. This relation provides a connection between a muscle activation and a spatial point in the articulatory space which is invariant for a given muscle structure.

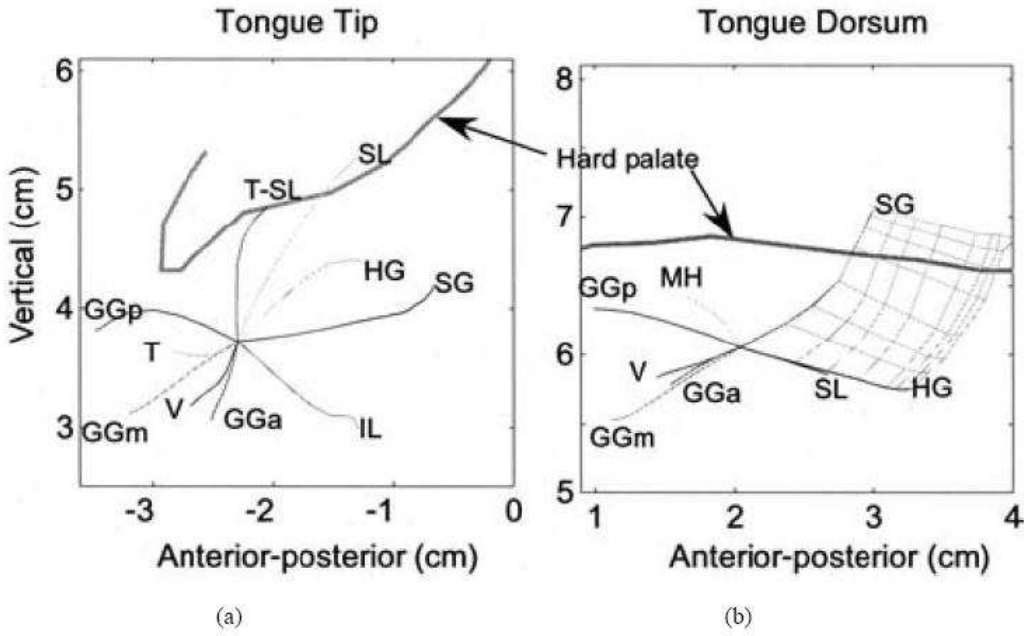

Figure 5.3: Coordinates consisting of the equilibrium positions corresponding to the activation forces ranged between 0 and 6 N. The net in the right panel consists of the contour lines of the EPs of SG and HG. (After Dang and Honda [35])

.

Using such a connection, a unique mapping can be obtained from a muscle force to a spatial position.

Based on those findings, they got the Equilibrium Position (EP) vector for each muscle by activating the tongue muscle with eight level muscle activations (0.0, 0.1, 0.2, 0.4, 1.0, 2.5, 4.0, and 6.0 N), and elaborated a coordinate based on the EPs for each control point. Figure 5.3 (a) and Figure 5.3(b) shows the coordinates for the tongue tip and tongue dorsum, respectively. Since the EPs shift monotonically, the equilibrium position can be expected to move along the path consisting of the EPs as the muscle force varies continuously, as long as the forces of other muscles remain unchanged. Thus, the mapping between the spatial points and the muscle forces can be obtained based on the selected EP vectors. An example is shown in the right panel of Figure 5.3(b) by a contour net, which consists of the EPs of the SG and HG. The contour lines correspond to the six

force levels. Such a net of contour lines is named the equilibrium position map (EP-map). With the EP-map, any arbitrary point inside the region of the map can be reached using the activations interpolated from the contour lines.

The muscle workspace and EP-map are essentially multipoint control strategies, the control points of the model are controlled independently. However, as a coordinated structure of articulation, those control points are physically related with each other. The independent control of those control points may cause contradiction in realizing their targets. Moreover, the muscle workspace and EP-map are essentially multipoint control strategies only take the sagittal movement of the tongue into account.

Since the tongue is a 3D soft object, its behavior under the activation of tongue muscles are too complicated to be handled by consider its sagittal movements only. In this chapter, we attempt to control the whole shape of tongue not only the positions of the control points.

5.2 Model simulation

To construct a feedforward control module for the proposed 3D physiological articulatory model, it is necessary to elaborating mappings from input target, either acoustic or articulatory, to the corresponding muscle activation. This needs a large number of paired target-muscle activation data. So far, there are no empirical data that can be used to infer the functional relation between muscle activations and the articulatory posture. Hence, first of all, we conduct model simulations to obtain the data of articulatory posture-muscle activation pairs.

5.2.1 From muscle activations to articulatory posture

In Chapter 3 and Chapter 4, we have evaluated the functions of tongue muscles and the evaluated the ability of the model in realizing specific articulations, which provides the information of how many muscles are involved in the vowel articulation and the function of tongue muscles in vowel production. Therefore, here, we organize model simulations by

accounting for the function of individual tongue muscles and the agonist-antagonist property of tongue muscle pair (see Chapter 3 for details), and the number of muscles involved in vowel production (see Chapter 4 for details). Totally, 30 6-muscle combinations (details are shown in Appendix C) are used in model simulation. Within each 6-muscle combination, Transversus together with Verticalis manipulate the width and length of the tongue. And among the other 4 muscles, two of them are antagonist for tongue tip/dorsum, and the other two muscles are chose to enlarge the region that the tongue covers by activating the selected antagonist. Eight-level muscle activations (0.0, 0.1, 0.2, 0.4, 1.0, 2.0, 4.0, and 6.0N) are assigned to each tongue muscle; three-level activations (0.5, 1.5, 3.0N) are assigned to jawCl; and six-level activations (0.0, 0.5, 1.2, 2.4, 4.0, 6.0N) are assigned to jawOp.

Figure 5.4 shows the distribution of the 11 nodes along the tongue surface in the mid-sagittal plane of simulation results. The area with different colors corresponds to the dispersion of individual tongue nodes. One can see that the simulation covers large variations of the tongue shapes. Moreover, it shows that the tongue can form constrictions and full closure at the alveolar and the palate-velar part, which are important for apical and dorsal consonants, respectively.

5.2.2 From articulatory posture to formants

To account for feedforward control of the model for vowel production, we need to constrain the simulation results by using perception criterions. Therefore, it is necessary to calculate the acoustic response of the vocal tracts based on the area function of the vocal tract.

To calculate the cross-sectional area function of a vocal tract, which is enclosed by the tongue, jaw, vocal tract wall, and cheek, the vocal tract is extracted and segmented into several short tubes by a grid-plane system (as shown in Figure 5.5(a)). Then, the cross-sectional area function and length of the vocal tract are evaluated by concatenating the cross-sectional area and length of the short tubes along the vocal tract, respectively. As shown in Figure 5.5(b), the cross-sectional planes usually are irregular polygons. To calculate the area of the irregular polygon, we decompose it into several regular components -

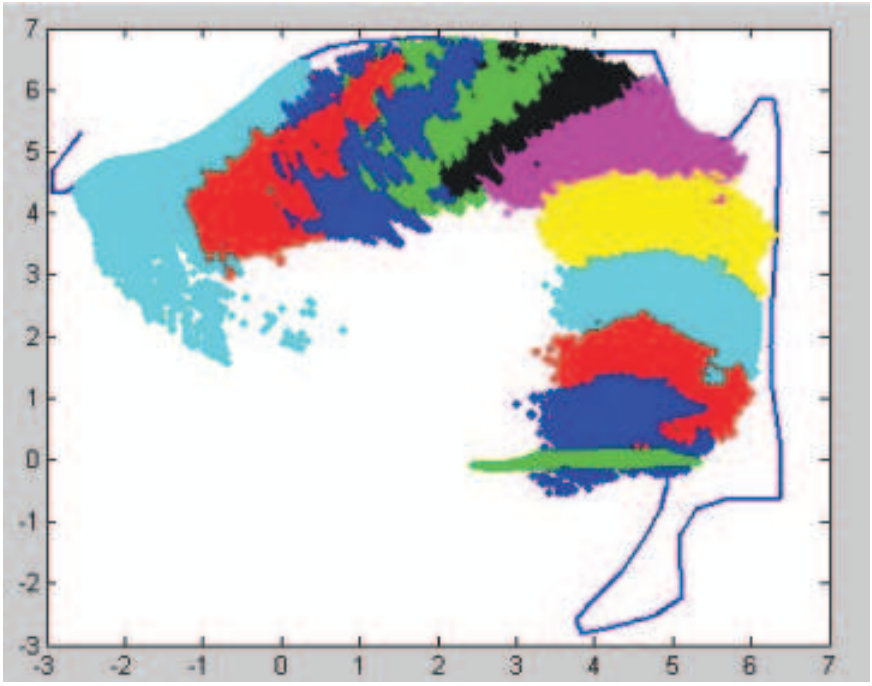

Figure 5.4: The distribution of the 11 nodes along the tongue surface in the midsagittal plane of parts of the simulation results.

trapeziums. The area of the irregular polygon is represented by the sum of the area of the component trapeziums. At the current moment, the lips are not physiologically modeled in the model. So, for the part surrounded by lips, we use a short tube to approximate the tube configured by lips. The corresponding cross-sectional area and length of the lip tubes for the five Japanese vowels are adopted from the MRI measurements [67].

The acoustic response of the vocal tract is calculated by using the frequency domain transmission-line model [68]. In that model, the loss from viscous effect, the yielding wall and the thermal exchange are taken into account. And the first two resonant peaks are picked up by using peak picking technique.

Figure 5.6 shows the dispersion of the first two resonance peaks of the vocal tracts obtained by model simulation. The area covered by blue spots is the vowels space generated by the physiological model. It is consistent with the distributions of the recorded vowels [69] and that obtained by Maeda [70] with geometrical articulatory model. The ellipses, in Figure 5.6, are the regions of the acoustic targets of 5 Japanese vowels of the prototype subject within the limen of 5% for F1 and 10% for F2 [71, 72], respectively. It shows that the simulations can cover the sustained vowels of the prototype subject. The articulatory

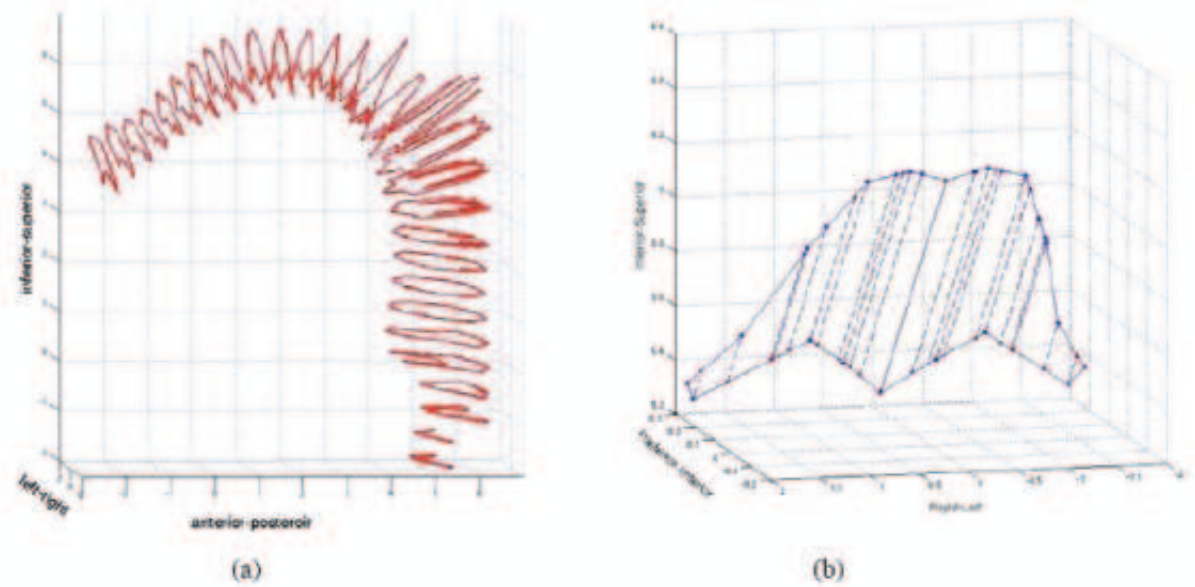

Figure 5.5: a) The cross-sectional shapes of the vocal tract obtained using the grid-plane system; b) Polygon segmented into trapeziums.

postures, whose acoustic responds located in the ellipses in Figure 5.5 are chosen for further processing.

5.2.3 Articulatory constraints on acoustically constrained articulatory postures

In the history, a number of literature [73-75] noticed that speaker compensate for the situations when some of the speech apparatus is disrupted. When an unanticipated mechanical perturbation is inflicted on the lower lip during the labial explosive consonant /p/ or /b/, the closure between the lower and upper lips is accomplished by a down shift of the upper lip [76-78]. Similar compensatory movements were observed in bilabial phonemic tasks by applying an electrical perturbation to the lower lip [79] and in bilabial or lingua-dental phoneme tasks by applying a mechanical perturbation to the jaw. Similar phenomenon are also observed in the experiment by inserting Bite-block between teeth of speakers [73, 80-83]. Atal et al. [84] also noticed this phenomenon that similar acoustic sounds can be generated from divers vocal tract configurations in their simulations based on a geometric articulatory model. Hence, the simulation data should be constrained by

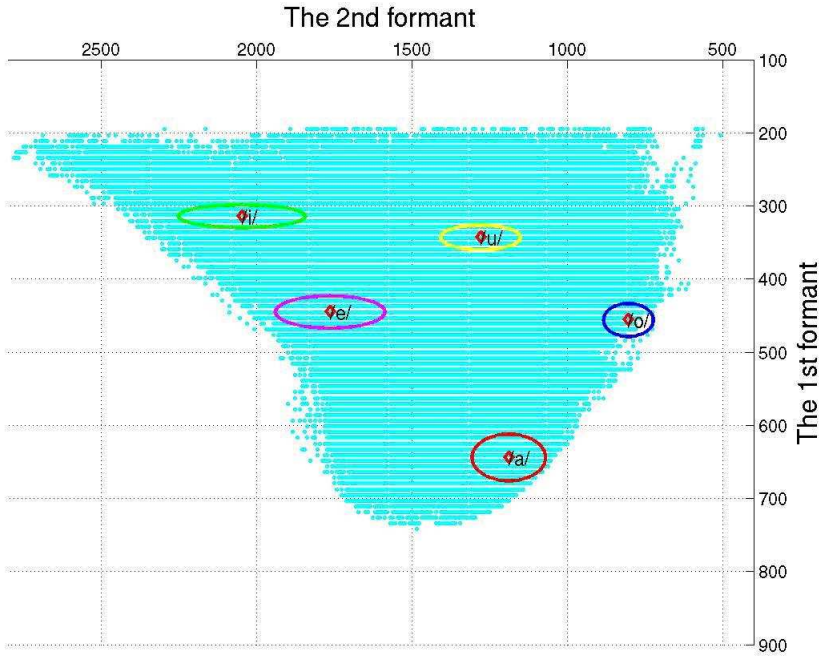

Figure 5.6: The distribution of the acoustic responds in the F1-F2 plane.

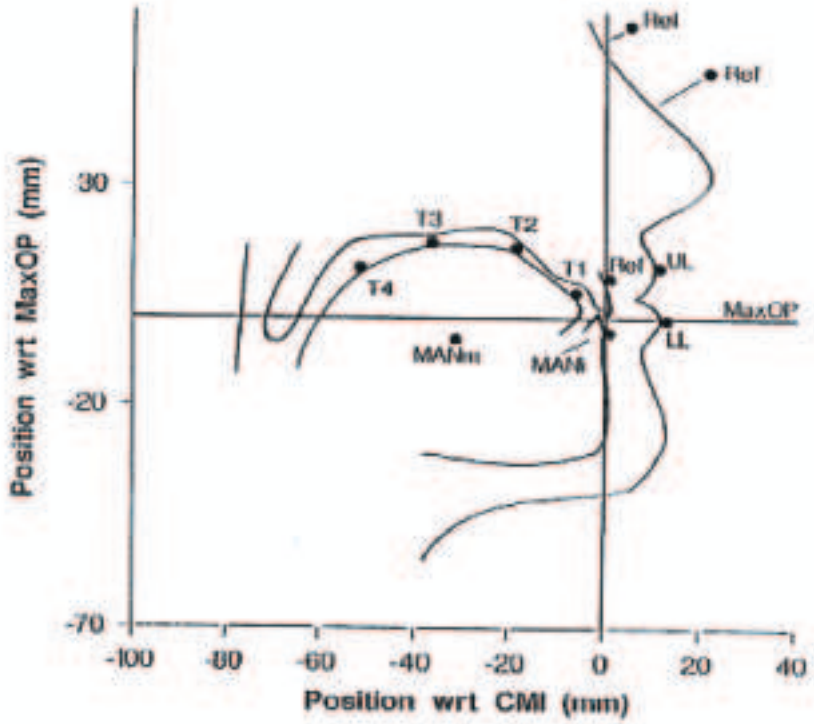

Figure 5.7: The approximate pellet placement locations

the articulatory data corresponding to normal articulation.

In addition, Qin et al. [85] found that the majority of normal speech is produced with a unique vocal tract shape based on the analysis on large scale empirical data. So, the articulatory record of normal speech will be a good template for constraining the simulated articulatory postures. In this study, the simulation articulatory postures are constrained by the x-ray microbeam data of the prototype subject which are selected from the ATR x-ray microbeam database.

The usual pellet constellation in the X-ray microbeam experiment includes eleven pellets, placed generally according to the scheme shown in Figure 5.7. Two of the reference pellets, highest on the nose bridge and at the incisors, were used to establish a floating two-dimensional, cranial-based coordinate system within which the positions of other pellets could be described. Defining the coordinate system in this way allowed complete removal of head motion from the motions of the remaining "articulator" pellets, as long as head motion was simple, involving only pitching rotation (about axes normal to the midsagittal plane), and/or translation relative to axes lying in the midsagittal plane. Two pellets were routinely attached to the mandible: one (MANi) glued to the buccal surface

of the central incisors, analogously to the maxillary incisal pellet, in the pocket formed by the central diastema and the enamel-gingival border; and, a second (MANm) glued in the vicinity of the juncture between the first and second mandibular molars. Four pellets were attached along the longitudinal sulcus of speaker's tongue. The most ventral of these (T1) was typically placed roughly 0.7 mm posterior to the apex of the extended tongue. The most dorsal (T4) was placed as rear as the speaker would tolerate without gagging. Two medial pellets (T2, T3) were placed so that the distance between the front and rear-most pellets was divided into three roughly equal segments. One pellet each was attached to the upper (UL) and lower lip (LL).

As mentioned above, only the positions of T1 and MANi are exactly known. At first, we found the corresponding points $T1_{ar}$ and $MANi_{ar}$ on the physiological articulatory model for T1 and MANi. When constrain the simulation results by the observed articulatory movement data, $T1_{ar}$ and $MANi_{ar}$ should locate in the areas covered by 2 times standard deviation of T1 and MANi, respectively. Meanwhile, the tongue contour should pass through the area covered 2 times standard deviation of T2, T3, and T4. By using these articulatory constraints, finally, 364, 312, 573, 437, and 423 samples are obtained for vowel /a/, /i/, /u/, /e/ and /o/, respectively.

5.3 Extraction of articulatory parameters

So far, we obtain paired articulatory posture-muscle activation data corresponding to normal vowel articulations. Then, in the next step, we need construct a mapping from articulatory postures to muscle activation. However, the original articulatory posture vector consists of the (x, y, z) coordinates of the nodes on the tongue surface, which is difficult to interpret the articulatory posture from articulatory point of view directly, and brings difficulties to map articulatory postures to muscle activations due to it high dimension and relative small training set. Therefore, at first, we need to extract a set of articulatory parameters to describe the articulatory posture, then, map the articulatory parameters to muscle activations. In this part, we extract the articulatory parameters based on the constrained simulation results by using Linear Component Analysis (LCA) method.

5.3.1 Linear component analysis of articulatory posture

In LCA, it is assumed that the movement and deformation of articulators in speech production can be described by the combination of several linear components, and each component associates with one degree of freedom of the articulators. The advantage of using LCA is that every extracted control parameter has a well-defined articulatory interpretation.

As shown in Eq. (5.1), $\boldsymbol{s}$ is the articulatory posture vector, $\boldsymbol{s}_m$ is the average articulatory posture vector of the data for analysis, $\boldsymbol{M}$ is the transformation matrix, each column of $\boldsymbol{M}$ corresponds to one basic articulatory posture vector, and $\boldsymbol{a}$ is a vector of loadings factors.

$$\boldsymbol{s} = \boldsymbol{s}_m + \boldsymbol{M}\boldsymbol{a}; \tag{5.1}$$

In contrast to Principle Component Analysis (PCA), in LCA, the correlation between the columns of $\boldsymbol{M}$ is allowed.

The loading factors of the basic articulatory posture vectors are determined in the light of the following procedure: a). the data that describe the jaw movement are feed to PCA. Then the corresponding loading factor is subject to linear regression to extract the first basic articulatory posture vector; b). the residue are obtained by subtracting the influence of jaw, and are feed to PCA to extract the other component of tongue movements; and c). The corresponding articulatory posture vectors, the rest column vectors of $\boldsymbol{M}$, are determined by linear regression on the extracted factor loading.

5.3.2 Articulatory parameters

By using LCA, six linear components are found. They are Jaw height (JH), Tongue body advancing (TB), Tongue body arching (TD), Tongue tip elevation (TT), Tongue width (TW), and Tongue blade indention (TI). Figure 5.8 shows the effects by manipulate the value of individual factors. It shows that: a). With the increasing of JH, the jaw moves upward, and cause some tongue deformation at the tongue tip and rear part of the tongue body; b). With the increasing of TB, the tongue moves forward and upward; c). With the decreasing of TD, the tongue body moves upward and backward, while the tongue

Table 5.1: Articulatory parameters (first column) extracted by LCA procedure, the minimum (the second column) and max value (the third column) of each parameter, and their contributions (the fourth column).

Articulatory Parameters	Min. of Parameters	Max. of Parameters	Explained Variance
JH	-0.51	0.58	36%
TA	-5.03	3.57	42%
TD	2.77	3.87	12%
TT	1.80	2.13	4%
TW	-2.25	1.85	2%
TI	-1.12	1.94	1%

tip moves backward and downward; d). With the decreasing of TT, the tongue tip moves upward; e). With the increasing of TW, the tongue body is narrowed; and f). With the increasing of TI, the indention pattern at the tongue blade is more and more clear.

Table 5.1 shows the minimum and maximum of each component and the variance explained by each component for each component. Among the extracted six components, JH and TB show the largest contribution to explain the variance of the tongue shapes in vowel production, 36% and 42%, respectively. The extracted six components explain about 97% of the total variance.

Comparing with the components extracted by similar method based on X-ray [70] and MRI measurement [86], similar components (JH, TB, TD, TT) are founded although the source of data and the shape vectors for analysis show some difference. Besides, additional components, TI and TW, are found to have clear patterns from articulatory point of view. The average differences between the positions of original tongue surface nodes and the reconstructed ones are 0.07 cm, and the standard deviation is 0.03 cm. This suggests that the articulatory posture can be represented and reconstructed by these six linear components with high accuracy.

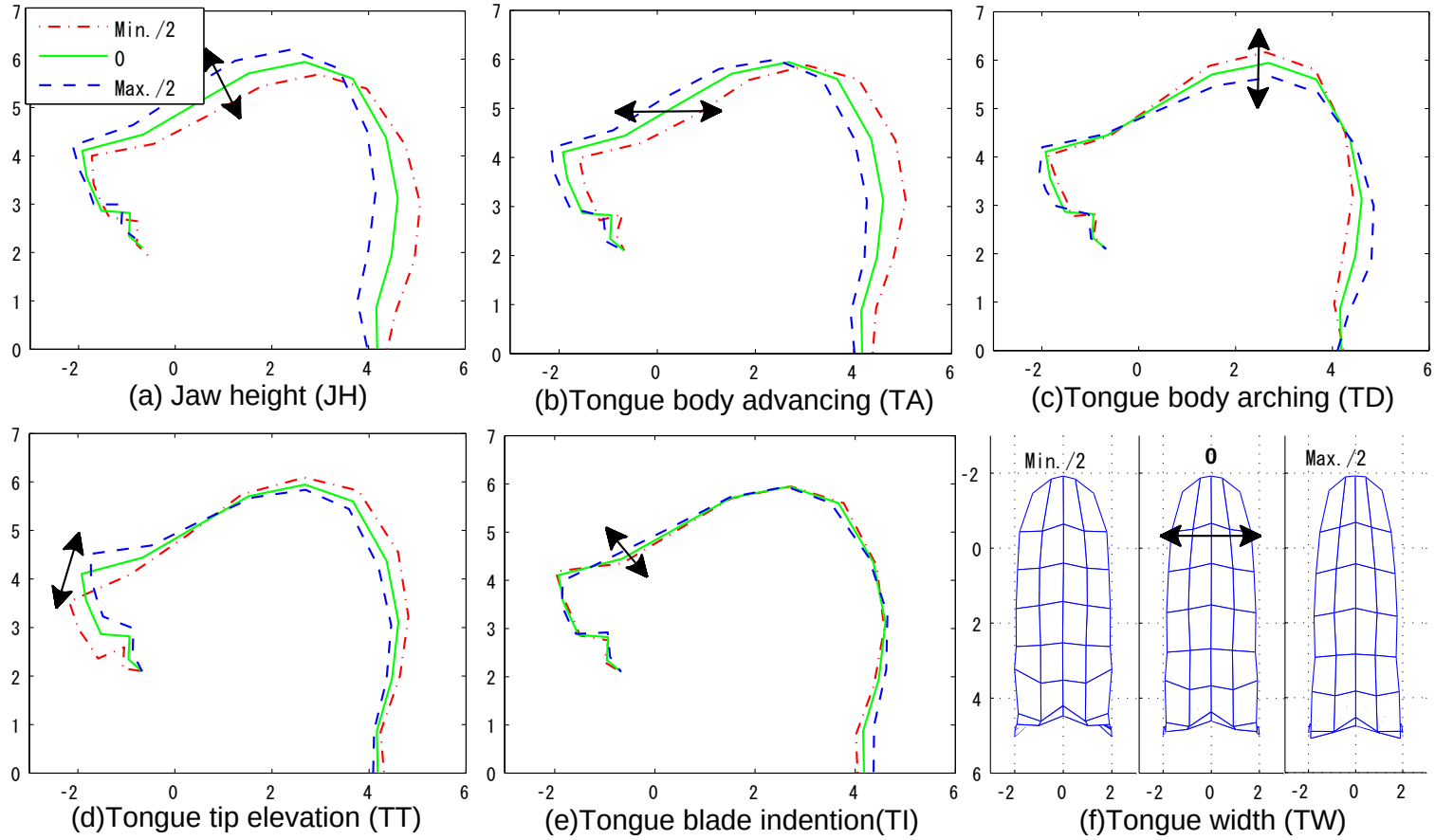

Figure 5.8: The parameters that describe the tongue shape from articulatory point of view: (a) Jaw height; (b) Tongue body advance; (c) Tongue body arching; (d) Tongue tip elevation; (e) Tongue blade indention; (f) Tongue width. In panels (a)-(e), the curves in green are the average tongue shape, the curves in red are those when corresponding component take half of the minimum value (Min./2), and the curves in blue are those when corresponding component take half of the maximum value (Max./2). The minimum and maximum values of each parameter are shown in Table 5.1.

5.4 Motor command generation

In vowel production, the tongue contacts with different parts of the vocal tract wall according to the position of the required constrictions. This makes the boundary constraints different among vowels. For this reason, it is necessary to cluster the articulatory posture and construct function from articulatory targets to muscle activation for each cluster. Moreover, in the feedforward control process, an appropriate function should be correctly chosen to generate muscle activation according to the input articulatory targets. This requires define an appropriate measurement that reflects the similarity relationship between the original articulatory postures in terms of the extracted articulatory parameters (see Section 5.3).

Our experiments show that Euclidian distance is an appropriate measurement of the dissimilarity between original articulatory postures. And the original articulatory posture can be represented by articulatory parameters with Eq. (5.1). Accordingly, the dissimilarity of the original articulatory postures can be derived in terms of the extracted articulatory parameters, as shown in Eq. (5.2).

$$d_{ij} = ||\boldsymbol{s}_i - \boldsymbol{s}_j|| = (\boldsymbol{a}_i - \boldsymbol{a}_j)^T \boldsymbol{M}^T \boldsymbol{M} (\boldsymbol{a}_i - \boldsymbol{a}_j); \tag{5.2}$$

where $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are original shape vectors, $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ are corresponding articulatory parameter, $\boldsymbol{M}$ is the same as that in Eq. (5.1).

Because the columns in M are not orthogonal to each other, the matrix $\boldsymbol{M}^T\boldsymbol{M}$ is not diagonal. This makes it difficult to define a simple similarity measurement that keeps the similarity property of the original articulatory postures in terms of the articulatory parameters directly. For this reason, we first look for a low dimensional space in which the similarity of the original articulatory postures can be kept by using the Euclidean distance between the representations of the articulatory postures in the low dimensional space. Then, the articulatory parameters are transformed into the low dimension representation, and the mapping from the low dimensional representation to muscle activation is elaborated. In this way, the feedforward control strategy is elaborated from articulatory parameters to muscle activation via the low dimensional representations. In the rest part of this chapter, we will introduce them in detail.

5.4.1 From articulatory posture to intrinsic representation

Multidimensional scaling (MDS) is a technique that has been widely used as a method of exploratory data analysis. It is used to find representations of samples in a low dimensional space, given a matrix of dissimilarities or distance between the samples. In the obtained low dimensional space, the similarity between objects in the original space is kept. Here we use the method proposed by Webb [87], in which every original articulatory posture can be transformed into a lower dimensional space, which kept the dissimilarity relation between the articulatory posture with other articulatory postures, with a combination of a set of radial basis function by using Eq. (5.3).

$$\boldsymbol{f(s)} = \boldsymbol{W\phi(s)} \tag{5.3}$$

where $\boldsymbol{W}$ is the weighting coefficient matrix, $\boldsymbol{\phi(s)}$ is a vector with *jth* element $\phi_j(\boldsymbol{s})$, $\phi_j(\boldsymbol{s})$ is a radial basis function of , $\boldsymbol{s}$ is the original articulatory posture. The coefficients in $\boldsymbol{W}$ can be obtained by minimizing the cost function defined in Eq. (5.4).

$$\boldsymbol{W} = \underset{\boldsymbol{W}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=1}^{N}(q_{ij}(\boldsymbol{W}) - d_{ij})^2,\ q_{ij} = ||f(s_i) - f(s_j)|| \tag{5.4}$$

where d_{ij} is the Euclidian distance between the *ith* and *jth* original articulatory postures, $q_{ij}(\boldsymbol{W})$ is the Euclidian distance between the representation of *ith* and *jth* articulatory postures in the transformed low dimensional space.

In this way, the similarity structure of data in the original space can be kept by in the transformed low dimensional space by using the measurement of Euclidean distance. Figure 5.9 shows the dispersion transformed low dimensional space. It demonstrates that these vowels are well separated in the transformed low dimensional space. And the dispersion of the vowels are quite similar to the results obtained by Dang and Lu [88], which they referred as the intrinsic dimension of vowels. Hereafter, we adopt the words "intrinsic presentation" as the representation of the articulatory postures in the low dimensional space obtained by the proposed nonlinear transformation.

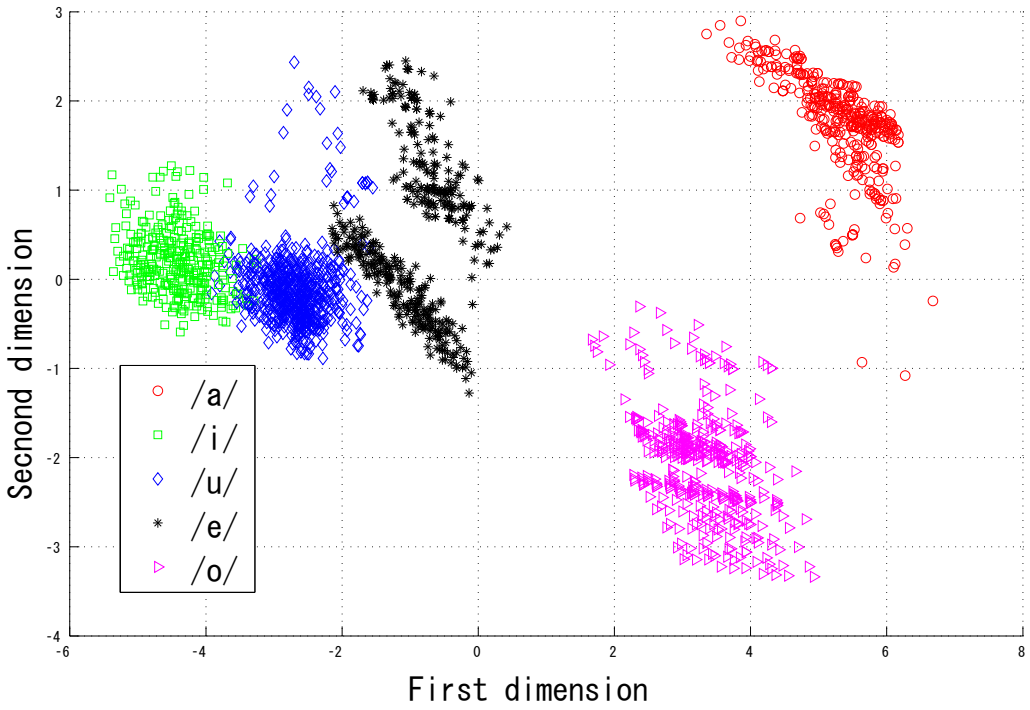

Figure 5.9: The dispersion of the vowels in low dimensional space obtained by MDS analysis.

5.4.2 From articulatory parameter to intrinsic representation

As demonstrated in Section 5.4.1, the obtained intrinsic representation preserves the dissimilarity relation between the articulatory postures with the measurement of Euclidean distance. Therefore, the mapping from articulatory postures to muscle activations is divided into two steps: 1). the articulatory parameters are transformed into intrinsic representation by nonlinear functions; and 2). a direct mapping is elaborated form the intrinsic representation to muscle activations. In this section, we will deal with the issue of how to transform the articulatory parameters into corresponding intrinsic representations. And in the next section, the mapping from intrinsic representation to muscle activation is elaborated.

As illustrated in Eq. (5.1), the original articulatory posture s can be represented by the weighted sum basic shape vectors - the columns in matrix $\boldsymbol{M}$, a and the average shape vector $\boldsymbol{s}_m$; moreover, the intrinsic representations can be obtained form an nonlinear function, as shown by Eq. (5.3). Therefore, by substituted Eq. (5.1) into Eq. (5.3), we get:

$$\boldsymbol{f(a)} = \boldsymbol{W\phi(Ma+s_m)} \tag{5.5}$$

where matrix $\boldsymbol{M}$, the average shape vector $\boldsymbol{s}_m$ are already obtained in Section 5.3. And the radial basis function $\boldsymbol{\phi}(\boldsymbol{s})$, and the weighting matrix $\boldsymbol{W}$ are obtained in Section 5.4.2. In this way, the intrinsic representation of original articulatory postures can be obtained directly from the articulatory parameters by using Eq. (5.5).

5.4.3 Mapping intrinsic representation to muscle activation

Articulatory posture clustering

Since the intrinsic representation keep the similarity structure of the data in the original shape space, we construct the mapping from the intrinsic presentation to the muscle force combination so as to control the physiological articulatory model. To account for the categorical differences of the data caused by different boundary constraints resulted from the interaction between tongue and surrounding structures, the k-means method is applied to clustering the articulatory posture based on the measurement of Euclidian distance. And

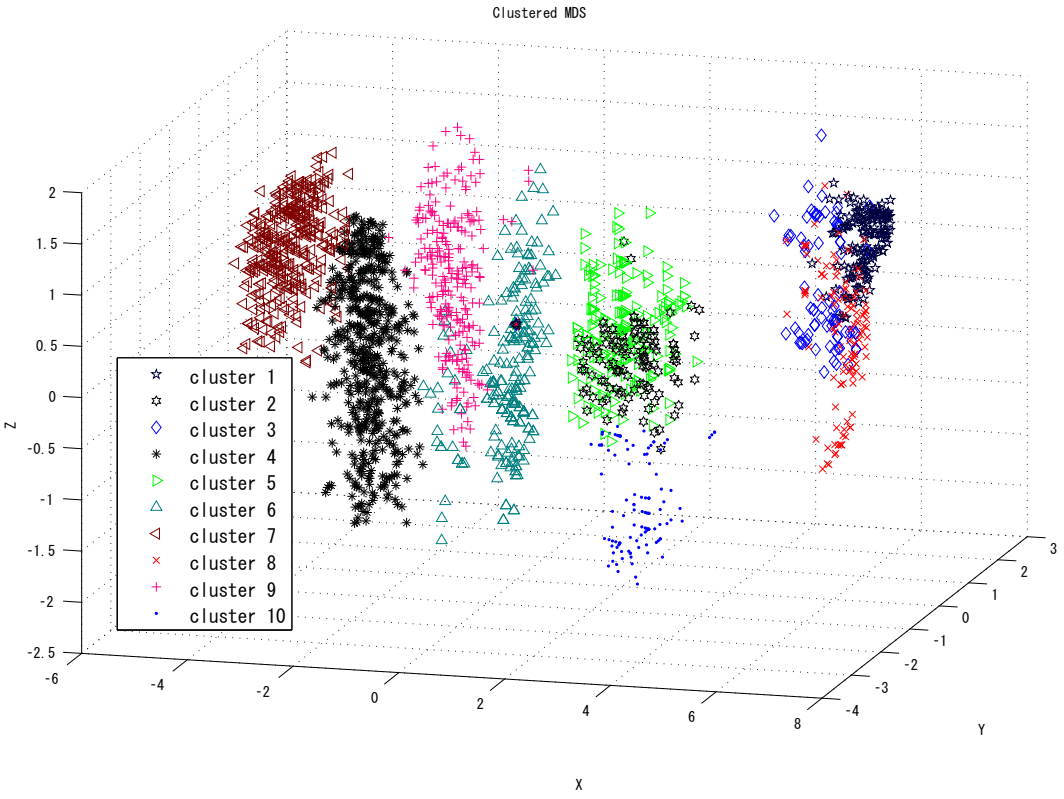

Figure 5.10: Clustering results of the intrinsic representation of vowel production.

the 'elbow' criterion is used to determine the number of clusters in the data.

As shown in Figure 5.10, both vowel /i/ and /u/ are mainly represented by one cluster, respectively; vowel /e/ is represented by two clusters; and /a/ and /o/ are represented by three clusters, respectively. The details of the number of samples in each cluster are shown in Appendix D.

General Regression Neural Network

For each cluster, the mapping from the intrinsic representation to muscle activations is elaborated. This mapping, actually, is to construct a functional relationship between articulatory input and muscle activation output. This is called regression in the field of machine learning, or system identification in the field modern control theories. In current study, we use the General Regression Neural Network (GRNN) for this task.

The regression of a dependent variable, $\boldsymbol{y}$, on an independent variable, $\boldsymbol{x}$, is the compu-

tation of the most probable value of $\boldsymbol{y}$ for each value of $\boldsymbol{x}$ based on a finite number of possibly noise measurement of $\boldsymbol{x}$ and the associated values of $\boldsymbol{y}$. Usually, it is necessary to assume some functional form with unknown parameters. The values of those parameters are chosen to make the best fit the observed data. The GRNN allows the appropriate form to be expressed as a probability density function that is determined from the observed data using Parzen window estimation [89].

Assume the $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ represents the known joint probability density function of random vectors $\boldsymbol{x}$, $\boldsymbol{y}$. And let $\mathbf{x}$ be a particular measured value of random vector $\boldsymbol{x}$. The conditional mean of $\boldsymbol{y}$ given x is given by Eq. (5.6).

$$\mathbf{E}[\boldsymbol{y}|\boldsymbol{x}] = \int_{-\infty}^{\infty} \boldsymbol{y}\boldsymbol{f}(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} \tag{5.6}$$

When the density $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ is not known, it must be estimated from samples of observations of $\boldsymbol{x}$ and $\boldsymbol{y}$. For nonparametric estimation of $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$, the Pazen window method is implemented. And the probability estimator $\hat{\boldsymbol{f}}(\boldsymbol{x},\boldsymbol{y})$ is based on the samples $\mathbf{x}_i$ and $\mathbf{y}_i$ of random vector x and y. When Gussian window is used, $\hat{\boldsymbol{f}}(\boldsymbol{x},\boldsymbol{y})$is represented as shown in Eq. (5.7),

$$\hat{f}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{N}\frac{1}{(2\pi)^{(p+d)/2}\sigma^{(p+d)}}\sum_{i=1}^{N}\exp\left(-\frac{(\boldsymbol{x}\text{-}\boldsymbol{x}_i)^T(\boldsymbol{x}\text{-}\boldsymbol{x}_i) + (\boldsymbol{y}\text{-}\boldsymbol{y}_i)^T(\boldsymbol{y}\text{-}\boldsymbol{y}_i)}{2\sigma^2}\right) \tag{5.7}$$

where N is the number of samples, p is the number of dimensions of $\boldsymbol{x}$, d is the number of dimension of $\boldsymbol{y}$, and σ is the width of the Gaussian window. Substituting Eq. (5.7) into Eq. (5.6), the desired conditional mean of $\boldsymbol{y}$ given $\boldsymbol{x}$ is expressed by Eq. (5.8)

$$\hat{y}(\boldsymbol{x}) = \frac{\sum_{i=1}^{N}\boldsymbol{y}_i\exp\left(-\frac{(\boldsymbol{x}\text{-}\boldsymbol{x}_i)^T(\boldsymbol{x}\text{-}\boldsymbol{x}_i)}{2\sigma^2}\right)}{\sum_{i=1}^{N}\exp\left(-\frac{(\boldsymbol{x}\text{-}\boldsymbol{x}_i)^T(\boldsymbol{x}\text{-}\boldsymbol{x}_i)}{2\sigma^2}\right)} \tag{5.8}$$

To elaborate the mapping from the MDS space to corresponding articulatory parameters, The Generalized Regression Neural Network (GRNN) is implemented. For each cluster, 80% of the data are chosen as the training set, and the left 20% as testing set. The General Regression Neural Network (GRNN) is applied to approximate the underlying function of each cluster.

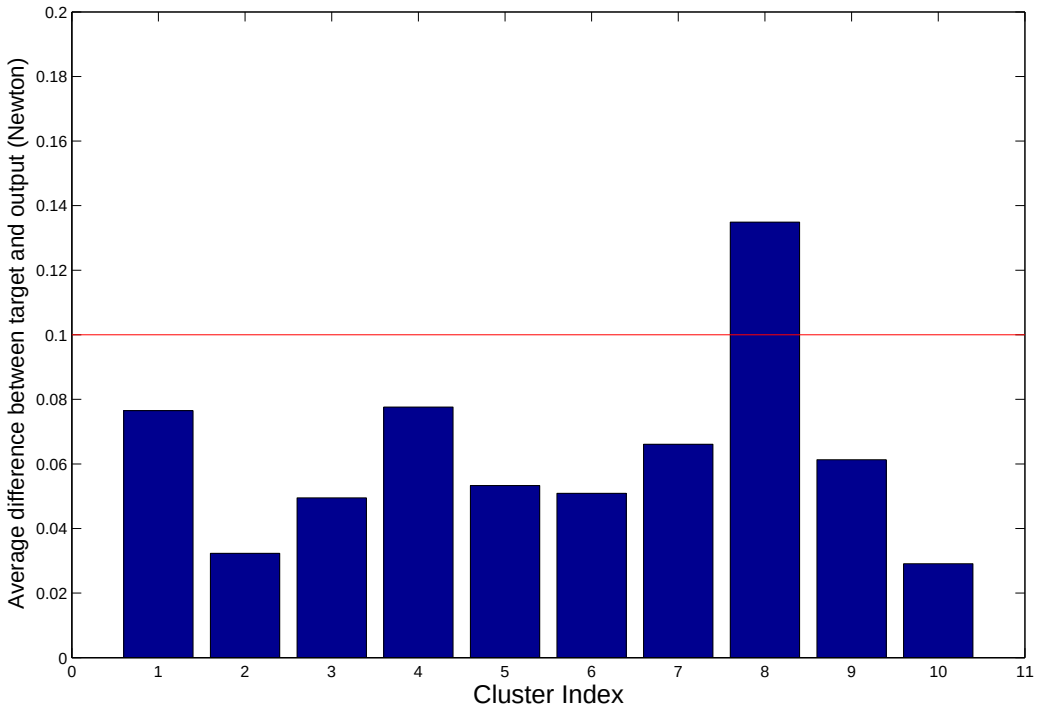

Figure 5.11: The average difference between the target muscle activation and that estimated by the GRNN in each cluster.

5.4.4 Results

The difference between network output and target muscle activations are shown in Figure 5.11, where the abscissa represents the categories obtained by clustering analysis conducted in Section 5.4.3.1, and the ordinate denotes the average difference between the target muscle force and the muscle force estimated by the trained GRNN for the test set of each cluster. The horizontal line denotes the minimal none-zero muscle activations for each muscle in simulation. It shows that, for most of the clusters, the difference between the target and estimation is less than half of the minimal none-zero muscle activation for each muscle in our simulation.

In addition we use the estimated muscle activation to drive the proposed 3D physiological articulatory model to evaluate the difference between the target articulatory posture and that obtained by the driven the model with estimated muscle activations, and the difference between their corresponding acoustic consequences. The results are shown in

Table 5.2: The average and standard deviation of difference between the postures of the target articulation and those obtained by activating model with muscle activation estimated by GRNN (in cm), and average and standard deviation of between the corresponding acoustic consequences (F1, and F2 in Hz).

	Shape Dev.	Shape Std.	Acoustic Dev. (F1/F2)	Acoustic Std.(F1/F2)
Cluster 1	0.03	0.06	6.9/14.6	6.2/14.5
Cluster 2	0.13	0.09	9.5/14.5	5.1/12.1
Cluster 3	0.07	0.07	20.8/24.3	6.4/21.1
Cluster 4	0.07	0.06	5.2/26.2	7.4/23.1
Cluster 5	0.06	0.07	24.4/10.4	9.6/7.9
Cluster 6	0.08	0.08	68.0/65.0	28.7/25.4
Cluster 7	0.04	0.05	9.8/55.0	9.5/49.8
Cluster 8	0.02	0.05	9.8/20.5	7.9/20.8
Cluster 9	0.03	0.06	9.1/20.7	33.2/15.7
Cluster 10	0.03	0.06	18.7/9.5	5.0/9.5

Table 5.2. It shows that in most cases, the articulatory posture and corresponding acoustic consequences obtained by activating the proposed physiological articulatory model are close to the required articulatory and acoustic target.

5.5 Discussions

5.5.1 Intrinsic representation of vowel production

The intrinsic representation of vowel production in this study is obtained by using the criterion: keep the topology structure of the original data. We use metric multidimenstional scaling technique for this task, where each original data can be transformed into its intrinsic representation with an explicitly defined nonlinear function. As shown in Figure 5.9, vowel /i/, /a/, and /o/ occupy the vertices of the extracted triangular structure, while

/u/ and /o/ locate between /i/ and /o/ and between /i/ and /a/, respectively. Moreover, the vowel categories are clearly separated, although we did not require the discrimination between the samples for the vowel categories.

Dang and Lu [88] used manifold learning technique to reveal the intrinsic degree of vowel production and perception. Figure 5.12 shows the results. Although the detail algorithm differs from MDS, the basic criterion is the same as MDS: preserve the topology structure of original data in the transformed space. The obtained results for both vowel production and vowel perception are similar to what is shown in Figure 5.8.

Kroger et al. [90] elaborated the phonetics map which consist of formants (F1, F2, and F3) and abstract articulatory parameters (high-low, front-back, rounded-unrounded) by using self organizing map to train the proto-vocalic vectors. The phonetic map plays the central role in their neurocomputational model of speech production and perception. It connects the motor plan, sensory (both somatosensory and auditory) feedback with, and linguistic representation (phonemic map). The obtained phonetic map also show similar structures as what we obtained by using multidimensional scaling.

Nowadays, more and more brain imaging experiments results showed that the same brain areas responded both during motor execution and when listened to the sound of an action made by the same effector [91-93]. It suggests that the speech production and perception may share a common encoding of neural activities in human brain. The intrinsic representations obtained by the articulatory measurements (EMA measurement [88] or articulatory posture), by auditory representations (MFCC coefficients [88]), and both of them showed similar dispersion. Comparing the intrinsic presentation and the findings of brain imaging experiments, it is possible that the intrinsic representation may correspond to the common encoding of the neural activities in human brain.

5.5.2 Relation with feedforward control of speech production

The purpose of this chapter is not to construct feedforward control for speech production, but to construct feedforward control of the physiological articulatory model. Hence, there are several differences between elaborated feedforward control of the physiological articulatory model and feedforward control of speech production. Firstly, the inputs to

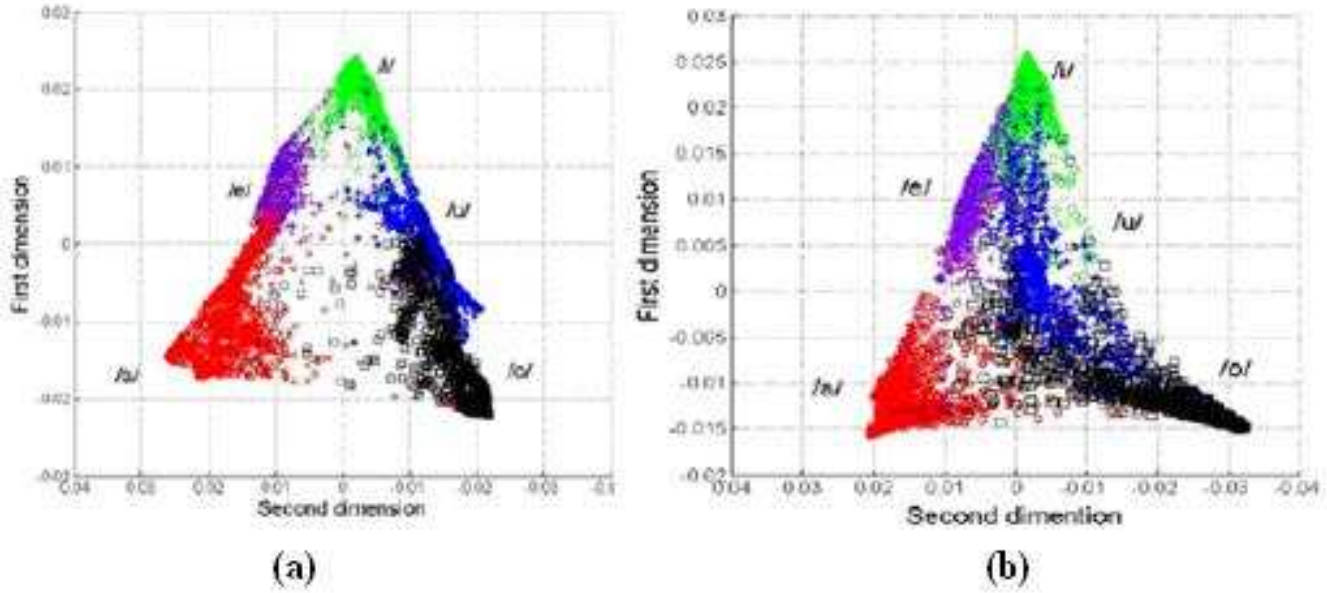

Figure 5.12: The intrinsic structure of vowel production: (a) result obtained from articulatory data; (b) result obtained from acoustic data.(After Dang and Lu [88]).

these two feedforward control modules are different. For speech production, the input is the abstract linguist representation, while for the model control the input is physical articulatory target. Secondly, the mapping functions for the model control are just made to map intrinsic representation to muscle activation, in which the detail mechanism of the real control process of speech production are not taken into account. For example, we clustered the articulatory postures in the intrinsic representation space, and built the mapping function for each cluster. In real situation, human may use the same mechanism to produce all the vowels.

Compare the feedforward flowchart for the physiological articulatory (shown in Figure 5.1) model with the neurocomputational model of speech production and perception, shown in Figure 5.13, the target articulatory postures specified by articulatory parameters, intrinsic representation, and the mapping from intrinsic representation to the motor commands are analogue to "somatosensory map", "phonetic map", the process from "phonetic map" to "primary motor map" (indicated by dashed frames in Figure 5.13), respectively. If the neurocomputation model reveals the real process of speech production and perception, the input of the feedforward control for the physiological articulatory model is analogue to the somatosensory feedback, while the mapping from intrinsic representation to the motor commands is analogue to part of the feedforward control of speech production.

5.6 Summary

In the present Chapter, we elaborated a "Motor Command Generator" module, which is the essential part of feedforward control of the 3D physiological articulatory model, by constructing mapping from articulatory targets to corresponding muscle activations. In the control command generating process, the articulatory postures specified by articulatory parameters are transformed to intrinsic representation of the articulation; then, proper mapping function are chosen to output corresponding muscle activations according to the intrinsic representations of the articulatory target.

To this end, firstly, model simulations are conducted based on the obtained knowledge of the functions of tongue muscles. And the simulation results are constrained with acous-

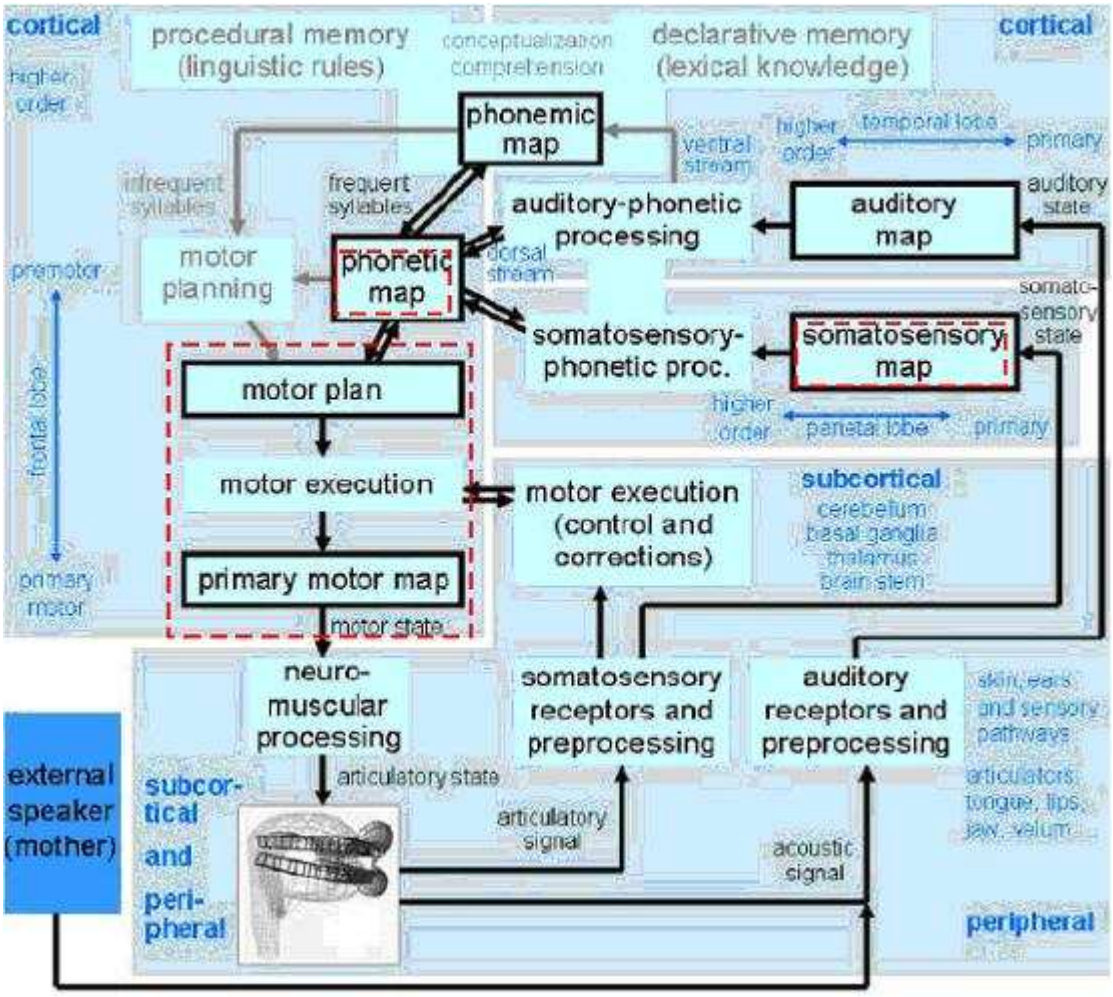

Figure 5.13: The structure of neurocomputational model of speech production and perception. (After Kroger et al. [90])

tical and articulatory constraints to extract the normal articulation for the five Japanese vowels. Secondly, articulatory parameters that describe the articulatory postures are obtained by using the Linear Component Analysis method. It shows that the extracted 6 parameters (JH, TB, TD, TT, TW, and TI) are able to depict the articulatory posture with high accuracy. The average difference between the original tongue surface nodes position and the reconstructed ones is 0.07 cm, and the standard deviation is 0.03 cm. Thirdly, the motor command generation module is elaborated by establishing mapping from articulatory targets to muscle activations via intrinsic representation of vowel articulation with General Regression Neural Network. The results show that this method can control the proposed 3D physiological articulatory model with high accuracy.

Chapter 6

Summary and Future work

6.1 Summary of this thesis

As mentioned in Chapter 1, the previous speech production theories are either jump directly from surface observations to linguistic representations or only modeling the surface observation directly due to the poor understanding of the effects of the biomechanical articulatory system and the activities of the central neural system in speech production. Accordingly, usually there are ambiguous interpretations on the same surface observations. To understand the speech production in a natural way, it is necessary to shed light on the details of what really happens when the speech is generated from its linguistic representations. This requires, first of all, to get the observation of motor commands the activities of central neural system; then use the uncovered information together with the surface observations to obtain an appropriate speech production theory that bridge the gap between the linguistic representation and surface observations naturally. However, motor commands and activities of central neural system are difficult to be observed directly. In the present thesis, we attempt to uncover the motor commands involved in speech production based on observed articulations. As demonstrated in Figure 1.1, the observed articulation is the output of the articulatory system with the input of motor commands. If we can model the biomechanical characteristics of articulatory system, it is possible to uncover motor commands from the observed articulation.

Therefore, firstly, we constructed 3D jaw and vocal tract wall and combined them with a 3D tongue to form physiological articulatory model which models the biomechanical

characteristics of articulators and the musculatures that drive the model according to human. The orientation of some tongue muscles are refined to be more realistic compared with their anatomical descriptions. And muscle SG is divided into two independently controlled parts, SGa and SGp, to provide more degrees of freedom to the proposed model. In addition, the contact between tongue and surrounding structures are correctly handled. Preliminary evaluation shows the proposed model behaves properly when the muscles involved in the model are activated individually.

After modeling the biomechanical characteristics of the tongue and surrounding structures, it is necessary to develop a method that utilize the physiological articulatory model to uncover the motor commands from observed articulation. Since, in speech production, the articulators are manipulated by activating associated muscles, and traditional technique, e.g. EMG, only can provide information of activations of partial muscles, the understanding of the detail function of related muscles, especially tongue muscles, may help to uncover motor commands from observed articulations. To this end, the general function of individual tongue muscle and the agonist-antagonist properties of tongue muscle pair were analyzed by using model simulation. It was found that: 1). the function of muscle GGa, GGm, GGp, SG, and MH for the movement of both tongue tip and dorsum were consistent with the speculation based on the anatomical orientations; 2). the muscles (T, V, and SL) located in the superficial layer of the tongue contributed most to length and width deformation of tongue surface; 3). muscle pairs GGm-SL, GGm-HG, GGA-HG worked as antagonist for tongue tip, while as agonist for tongue dorsum; 4). muscle pairs GGp-HG, GGp-SL, GGm-SG acted as the antagonist for tongue dorsum while act as agonist for tongue tip.

To utilize the physiological articulatory model to uncover motor commands in vowel production, at first, it is necessary to evaluate whether the model is able to produce the observed specific articulations and whether the muscle activation used in the model is consistent with the observed EMG signals. Therefore, we uncovered the motor commands of specific articulation of isolated vowels based on knowledge of the functions of tongue muscles and the observed activation of part of tongue muscles in EMG experiment by using the proposed 3D physiological articulatory model. Here, it was assumed that if the physiological articulatory model generates the same articulatory posture as

the observed one, the uncovered motor commands were the possible ones that the speaker used in the observed articulation. In light of this assumption, we used an optimization procedure to decrease the 3D dimensional difference between the model simulation and target articulation by gradually adjusting the activation of associated muscles. To avoid the local minimum which may be apart from the true value in the optimization procedure, the initial muscle activations are set according to observation obtained by EMG measurement in producing vowels. It is found that the 3D model is able to generated produce the observed articulations of isolated vowels, and the corresponding muscle activations are reasonable compared with the EMG observations. And theoretically, the proposed model-based optimization procedure is easily extended to uncover the motor commands (muscle activations) for isolated consonants according to the same assumption.

Usually, the articulation of a same phoneme with certain variation, the articulation is not exactly the same from time to time. Therefore, the method described in Chapter 4 is expensive for uncovering the motor commands for articulations with variations. In addition, in daily communication with speech, people usually speak to each other continuous speech rather than isolated phones. In continuous speech, the observed articulation is not the same as the desired articulatory targets [32]. To deal with the coarticulated articulations, it requires uncovering the desired articulatory targets first, then, uncovering the motor commands for the ingredients of the utterance. All of these need an efficient control strategy to manipulate the physiological articulatory model according to desired target. To tackle the problem, we constructed a motor command generation module to generate motor commands according to articulatory targets of vowels to motor commands (muscle activation) for the feedforward control of the physiological articulatory model. The articulatory targets specified by articulatory parameters are transformed to their intrinsic representation; then, proper mapping function is chosen to generate corresponding motor commands. Hence, first of all, we conducted model simulations to provide paired articulatory posture-muscle activation data for elaborating mapping from articulatory posture to muscle activation. Based on the simulation results, we extract a set of articulatory parameters, which are readily to be interpreted from articulatory perspectives, to specify the articulatory targets. Then, a set of mappings are elaborated from articulatory targets to muscle activation via the intrinsic representation. The results show this control method

can realize both the observed articulation and corresponding acoustic consequence with high accuracy in most of the cases of vowel production. In the case of isolated vowel production, the observed articulatory posture can be thought of as the articulatory target for isolated vowel. Hence, the motor commands for isolated vowel production is able to be uncovered by using the feedforward control flowchart. As for the motor commands for the production of vowels in continuous speech, a extra procedure, similar to Wei *et al.*'s work [32], is need to estimate the articulatory targets for the observed articulatory movement, then, uncover the motor commands for those vowels based on the estimated articulatory targets. This is left to future work.

By conducting the above work, the first aim, to uncover the motor commands for isolate vowels, is realized. For a given observed articulation of isolated vowels, the corresponding motor command is able to be estimated by using the proposed feedforward flowchart. As for the second and third aims, to uncover the motor commands for isolated consonants and phonemes in continuous speech, they are left to future work.

6.2 Contribution of this thesis

Due to the work addressed in Chapter 2-5, a number of contributions are made for uncovering the motor commands involved in speech production, which may help to realize our final goal - bridge the gap between linguistic representation and surface observation. Usually, articulation is easier to be observed than corresponding motor command. However, motor commands are impossible to be correctly inferred from observed articulation without the understanding of the biomechanical properties of speech organs. In this thesis, we proposed a physiological articulatory model-based method to uncover the underlying motor commands from observed articulation for isolated vowels.

Firstly, we constructed a 3D jaw and vocal tract wall, and combined them with a 3D tongue model [30] to form a 3D physiological articulatory model. The orientations of some of the muscles involved in the 3D tongue model were refined, e.g. the attachments of muscle SG spanned over an interval of 7.5cm rather than the unrealistic parallel arrangement of the posterior portion of SG, which is an improvement from Fujtia's model [30]. In addition, contact handling was incorporated into the model, where the inter-

action between tongue and surrounding structures are appropriate handled. With these improvements, the morphological structure and musculatures are more realistic than the 3D tongue model [30], and partial-3D model [35], and the physical characteristics of interaction between tongue and surrounding structures are more reasonable. With the correct modeling of the biomechanical properties of speech organs and the correct modeling of musculatures, it makes an appropriately model that is able to account for biomechanical characteristics of articulators, which are required for uncovering motor commands from observed articulation.

Secondly, to utilize the proposed 3D physiological articulatory model to uncover motor commands underlying observed articulation, it is necessary to shed light on the detail functions of the muscles that drive the physiological articulatory model, which are difficult to be obtained by using the traditional techniques, e.g. tagged-MRI. In the present thesis, we proposed a model-based method to provide an efficient way to explore the detail function of tongue muscles. By using the physiological articulatory model, the functions of individual tongue muscles and tongue muscle pairs of interest can be independently explored, which is difficult to be inspect from observed data because any articulatory movements are generated by activating a set of associated muscles simultaneously. The results confirmed several speculations on the function of tongue muscles based on their orientations (the function of GGa, GGm, GGp, SG, and MH), and revealed the function of T, V, and SL on tongue width and length deformation. In addition, through quantitative analysis, it is found that there are agonist-antagonist muscle pairs in manipulation of the tongue.

Thirdly, the thesis presented an optimization procedure, which minimized the 3D morphological difference between observation and simulation obtained by driving the model with muscle activations, to uncover the motor commands for isolated vowel production. By comparing the final output of the optimization procedure with the observed articulation and EMG signals, it was found that the uncovered motor commands were consistent with observed EMG signals.

Fourthly, thanks to the 3D reality of the tongue and the sounding structures, the area function of the vocal tract is able to be calculated directly instead of by using α-β model [47] which maps the midsagittal dimension of the vocal tract into area function. By using

the calculated area function and a frequency domain transmission line model of vocal-tract, we are able to incorporate perceptual constraints on the simulation results, which are used in elaborating the feedforward flow for the physiological articulation model.

Fifthly, a motor command generation module is elaborated based on the constrained simulation results for feedforward control of the proposed 3D physiological articulatory model. The result shows that the feedforward control flow can control the proposed 3D physiological articulatory model with high accuracy both acoustically and articulatorily. By using the proposed feedforward control flow, the motor commands for isolated vowel can be readily uncovered from observed articulation. In addition, the feedforward control provides an efficient method to manipulate the physiological articulatory model, which is required to uncover the articulatory targets for the phonemes in continuous speech and further to uncover the underlying motor commands for the phonemes that comprise the continuous speech.

6.3 Future work

In the present thesis, a 3D physiological articulatory model has been elaborated, and a preliminary attempt on feedforward control strategy of vowel production has been presented. To further investigate the mechanism of speech production, the following work should be done.

Firstly, more work are needed to refine the work of feedforward control of vowel production. In Chapter 5, we propose a framework for the feedforward control of the vowel production in normal situation. For a given articulatory target of the vowel production, the muscle activation is determined by a General Regression Neural Network via the intrinsic representation of articulatory targets. More work should be done to refine the mapping form intrinsic representation to motor commands consistent with the mechanism of speech production in the motor commands generation level, in which people thought that maybe the production of the vowels just use the same strategy to generate corresponding motor commands.

Secondly, we need to extend the current control strategy for the control of consonants. In the current step, we mainly concerned with uncovering the motor commands for vowel

production. However, speech utterance consists of not only vowels but also consonants. To uncover the motor commands in speech communication, it is necessary to extend the current control strategy for consonants in the future.

Thirdly, it is necessary to incorporate a framework of coarticulation for uncovering the motor commands of continuous speech. The segments in speech flow influence each other in both planning level and physical level. Based on the assumption that the effects from the physical level is complete taken into account by using the physiological articulatory model, the coarticulation is mainly concerned with the interaction between segments in the planning level. This will be reflected by the motor commands issue to associated muscle in speech production. However, most of the models that deal with coarticulation are from linguistic/cognitive point of view, and do not concerned with how the segments are arranged in temporal order from motor control point of view. Therefore, it is necessary to elaborate a framework of coarticulation from the perspective of motor control, which may promote the understanding of speech from a more realistic perspective, and understand of higher level activities in the brain while producing speech.

Fourthly, the feedback control branch needs to be incorporated into the control loop. As shown in Figure 1.1, in the speech production process, the speaker pick up auditory as well as somatosensory feedbacks to compare the generated articulatory movement and sounds with those they intended to produce, and make necessary adjustments to match the results with their intentions. By incorporating the feedback control into the feedforward control strategy, it is possible to investigate more complicated speech acts, such articulatory compensation under unexpected perturbation.

Appendix A

Physical modeling of the 3D tongue

A.1 Truss structure of the elementary mesh

In our present study, we adopt the displacement-based FEM as the basis of the modeling effort, which is referred to as an eXtended Finite Element Method (the X-FEM). The principal advantages of the X-FEM are that the finite element framework (sparsity and symmetry of the stiffness matrix) is retained and a single-field variational principle is used [94].

In X-FEM, 3D deformation of a continuum is described by the displacements of nodes. By taking advantage of this, the hexahedron mesh mentioned above is reconsidered based on the common assumption that tongue tissue consists of isotropic material. For an isotropic material, the relation between the nodes can be easily represented by using an elastic solid to connect the nodes in all directions within the hexahedron. Therefore, we use Hookean elastic bodies [95], hereafter referred to as cylinders, to connect all the nodes in the mesh. In the hexahedral mesh, there are 12 edges, corresponding to 12 cylinders. On each of the six surface planes, there are two cylinders connecting the diagonal nodes. Inside the hexahedron, there are four cylinders transversely connecting the diagonal vertices. Altogether, each hexahedron mesh is constructed from 28 visco-elastic cylinders. These cylinders in a basic mesh are shown in Figure A.1.

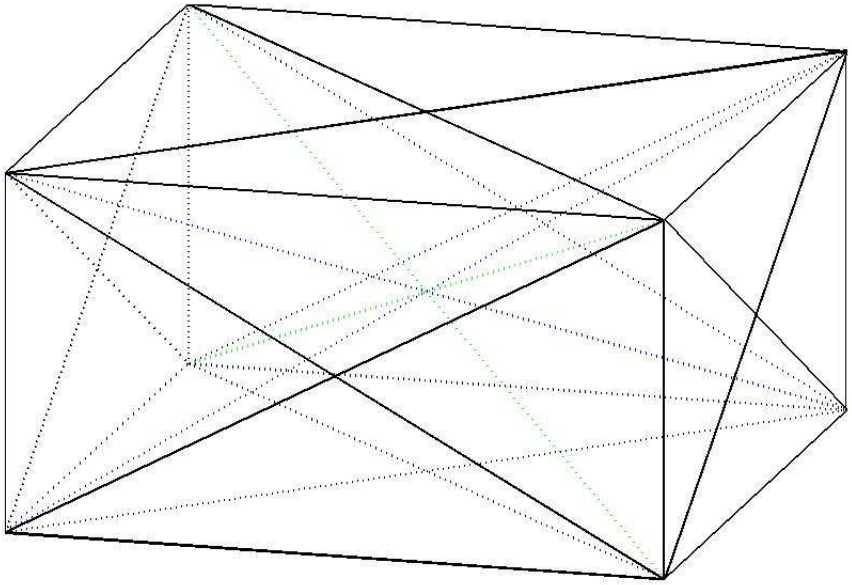

Figure A.1: The truss structure of each mesh in the tongue model.

A.2 Motion equation

For any finite element system, the linear dynamic response of it can be governed by the motion equation Eq. (A.1).

$$\boldsymbol{M}\ddot{\boldsymbol{x}} + \boldsymbol{B}\dot{\boldsymbol{x}} + \boldsymbol{K}\boldsymbol{x} = \boldsymbol{f} \tag{A.1}$$

Where $\ddot{\boldsymbol{x}}$, $\dot{\boldsymbol{x}}$ and $\boldsymbol{x}$ are the acceleration, velocity, and displacement vectors of the finite element assemblage; $\boldsymbol{M}$, $\boldsymbol{B}$, and $\boldsymbol{K}$ are the mass, damping, and stiffness matrices, respectively; and $\boldsymbol{f}$ is the vector of externally loads.

Here we use a simple element truss to demonstrate how to obtain the matrices $\boldsymbol{M}$, $\boldsymbol{B}$ and $\boldsymbol{K}$. As shown in Figure A.2, such a truss frame has one truss element and 2 end nodes which are labeled 1, 2. They correspond to one truss element and it nodes in the tongue model, respectively. The original length, Young's Modulus, and cross-sectional area of the truss element are denoted by l0, **E**, **A**, respectively. Both **E** and **A** are assumed to be constant along the truss element.

To calculate the matrices $\boldsymbol{M}$, $\boldsymbol{B}$ and $\boldsymbol{K}$, the corresponding matrices $\boldsymbol{M}^e$, $\boldsymbol{B}^e$ and $\boldsymbol{K}^e$ of the truss element should be calculated first. Then the $\boldsymbol{M}^e$, $\boldsymbol{B}^e$ and $\boldsymbol{K}^e$ are assembled into $\boldsymbol{M}$, $\boldsymbol{B}$ and $\boldsymbol{K}$ according to the connecting relation between the nodes of the tongue model.

A.2.1 Element stiffness matrix $\boldsymbol{K}^e$

In an idealized truss, externally applied forces as well as reactions can act only at the nodes. All member axial forces can be characterized by the x, y and z components of these forces, denoted by f_x , f_y and f_z, respectively. The components at node i will be identified as f_{xi}, f_{yi} and f_{zi}, respectively. The set of forces at all nodes can be arranged as a 6-component column vector called $\boldsymbol{f}$. Classical structural mechanics tells us that the displacements of the truss are completely defined by the displacements of the nodes. The x, y and z displacement components are denoted by u_x, u_y and u_z, respectively. The values of u_x, u_y and u_z of node i are called u_{xi}, u_{yi} and u_{zi}. Like node forces, they are arranged into a 6-component vector called $\boldsymbol{u}$. Thus, the force cause by the compression/elongation of the cylinder subject to Eq. (A.2).

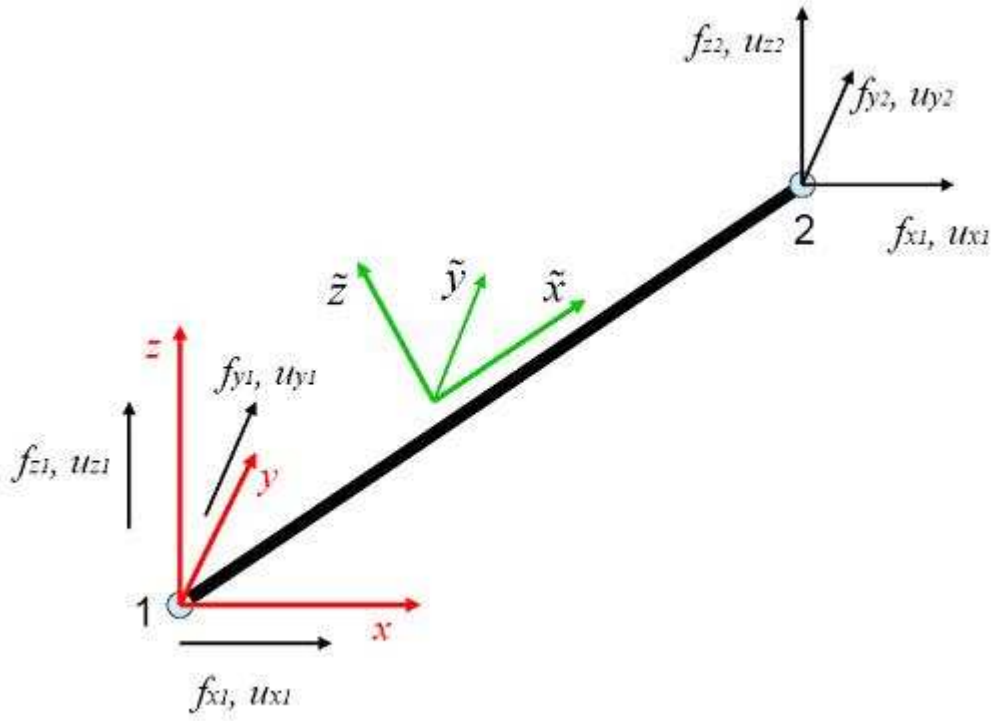

Figure A.2: One truss element within the hexahedra mesh of the tongue model

$$\boldsymbol{f_K} = \boldsymbol{K}^e \boldsymbol{u} \tag{A.2}$$

To calculate the stiffness matrix $\boldsymbol{K}$, first, we assign a local Cartesian system ($\tilde{x}$, $\tilde{y}$, $\tilde{z}$), in which axis is aligned along the axis of the element. The positive direction of runs from node 1 to node 2. The origin is arbitrary and may be placed at the member midpoint or at one of the ends for convenience. Therefore, the force in the local coordinates systems is calculated by Eq. (A.3)

$$\tilde{\boldsymbol{f}}_{\boldsymbol{K}} = \tilde{\boldsymbol{K}}^e \tilde{\boldsymbol{u}} = \begin{pmatrix} 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{y}_1 \\ \tilde{z}_1 \\ \tilde{x}_2 \\ \tilde{y}_2 \\ \tilde{z}_2 \end{pmatrix} \tag{A.3}$$

Then, the coordinates in the local coordinates system ($\tilde{x}$, $\tilde{y}$, $\tilde{z}$) are transformed into those in the global coordinate system (x, y, z) by Eq. (A.4).

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos(x,\tilde{x}) & \cos(x,\tilde{y}) & \cos(x,\tilde{y}) \\ \cos(y,\tilde{x}) & \cos(y,\tilde{y}) & \cos(y,\tilde{y}) \\ \cos(z,\tilde{x}) & \cos(z,\tilde{y}) & \cos(z,\tilde{y}) \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} \tag{A.4}$$

Finally, the force vector in the global system can be calculated according to Eq. (A.5)

$$\boldsymbol{f_K} = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix} \tilde{\boldsymbol{f}}_{\boldsymbol{K}} == \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix} \tilde{\boldsymbol{K}}^e \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix}^{-1} \boldsymbol{u} \tag{A.5}$$

And the stiffness matrix of the truss element can be evaluated by Eq. (A.6)

$$\boldsymbol{K}^e = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix} \tilde{\boldsymbol{K}}^e \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix}^{-1} \tag{A.6}$$

A.2.2 Element consistent mass matrix $\boldsymbol{M}^e$

In this study, we use the consistent mass matrix in represent the mass matrix in Eq (A.1). This method can preserve the linear momentum as well as the angular momentum of truss element. This is done by taking the kinetic energy as part of the governing function. The

kinetic energy of an element with density ρ, which occupies the domain Ω_e and moves with velocity field $\boldsymbol{v}_e$, is $\mathbf{T}^e$(as shown in Eq (A.7)).

$$\mathbf{T}^e = \frac{1}{2}\int_{\Omega_e} \rho(\boldsymbol{v}^e)^T \boldsymbol{v}^e \, d\Omega_e \tag{A.7}$$

Following the FEM philosophy, the element velocity field is interpolated by shape functions, as shown in Eq.(A.8):

$$\boldsymbol{v}^e = \boldsymbol{N}_{\boldsymbol{v}}^e \dot{\boldsymbol{u}}^e \tag{A.8}$$

where $\dot{\boldsymbol{u}}^e$ is the vector of node velocity, and $\boldsymbol{N}_v{}^e$ is the shape function of the truss element. Therefore,

$$\mathbf{T}^e = \frac{1}{2}(\dot{\boldsymbol{u}}^e)^T \int_{\Omega_e} \rho(\boldsymbol{N}_{\boldsymbol{v}}^e)^T \boldsymbol{N}_{\boldsymbol{v}}^e \, d\Omega_e \dot{\boldsymbol{u}}^e \tag{A.9}$$

where the element consistent mass matrix follows as the Hessian of $\mathbf{T}^e$:

$$\tilde{\boldsymbol{M}}^e = \frac{\partial^2 \mathbf{T}^e}{\partial(\dot{\boldsymbol{u}}^e)^2} = \int_{\Omega_e} \rho(\boldsymbol{N}_{\boldsymbol{v}}^e)^T \boldsymbol{N}_{\boldsymbol{v}}^e \, d\Omega_e \tag{A.10}$$

In our case, the shape function matrix of a 2-end truss element in 3D space is:

$$\boldsymbol{N}_{\boldsymbol{v}}^e = \begin{pmatrix} 1-\varepsilon & 0 & 0 & \varepsilon & 0 & 0 \\ 0 & 1-\varepsilon & 0 & 0 & \varepsilon & 0 \\ 0 & 0 & 1-\varepsilon & 0 & 0 & \varepsilon \end{pmatrix}, \ \varepsilon = \frac{\tilde{x} - \tilde{x}_1}{\tilde{x}_2 - \tilde{x}_1} \tag{A.11}$$

Therefore, the element consistent mass matrix in the local coordinate system is:

$$\tilde{\boldsymbol{M}}^e = \frac{\rho A_0 l_0}{6} \begin{pmatrix} 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix} \tag{A.12}$$

where A_0 is the original cross-sectional area of the truss element. At last, the element consistent mass matrix in the global coordinate system is calculated according to Eq. (A.13)

$$\boldsymbol{M}^e = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix} \tilde{\boldsymbol{M}}^e \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix}^{-1} \tag{A.13}$$

A.2.3 Element damping matrix $\boldsymbol{B}^e$

To account for the dynamic response of a soft tissue, the dissipation caused by velocity-dependent damping must be taken into account. Here, we treat the truss element as a visco-elastic body. According to Fung [95], there are three types of models for representing a visco-elastic material: the Voigt model, the Maxwell model, and the Kelvin model. The Voigt model is good for describing a solid body. When a force is applied on the Voigt model, a deformation gradually builds up as the spring shares the load. After the force is released, the dashpot displacement relaxes exponentially, and the original length is restored from the deformation. In present study, we adopt the Voigt model to consider the viscosity of the soft tissue. The procedure for calculating element damping matrix is the same as that for calculating the element consistent mass matrix.

$$\boldsymbol{B}^e = \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix} \tilde{\boldsymbol{B}}^e \begin{pmatrix} \boldsymbol{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{T} \end{pmatrix}^{-1}, \; \tilde{\boldsymbol{B}}^e = \frac{bA_0 l_0}{6} \begin{pmatrix} 2 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{pmatrix} \tag{A.14}$$

where b is viscosity coefficient of the soft tissue.

Finally, $\boldsymbol{M}^e$, $\boldsymbol{B}^e$, and $\boldsymbol{K}_e$, are assembled into the $\boldsymbol{M}$, $\boldsymbol{B}$, and $\boldsymbol{K}$ of Eq. (A.1). After determining the matrices $\boldsymbol{M}$, $\boldsymbol{B}$, and $\boldsymbol{K}$, with the incorporation of volume conservation and collision handling, appropriate dynamic response of the physiological articulatory model can be obtained when external forces are applied to it according to Eq. (A.1).

A.3 Volume conservation

The tongue body is commonly considered to consist of incompressible soft tissue. However, in the above tongue model, the hexahedron itself lacks volume preserving properties.

It is necessary to incorporate a constraint to maintain the volume of the tongue tissue. Totally, the tongue consists of 240 hexahedrons. And each hexahedral mesh consists of 28 cylinders. The volume of each hexahedron is the weighted sum of the virtue volume these cylinders. Here, first, we introduce the calculation of the volume of the truss mesh base on the virtue volume of the cylinders. Then, a method is proposed to consider the property of volume conservation of the tongue tissue.

A.3.1 Volume of the truss mesh

Supposing that the cylinder in the hexahedra mesh is with a uniform radius r, its cross-sectional area varies uniformly when a force is loaded in the axial direction (x-direction) alone. Thus, the cross-sectional area of the cylinder can be calculated according to Eq. (A.15)

$$A = \pi r^2,\ r = r_0(1 - \nu\frac{l - l_0}{l_0}) \tag{A.15}$$

where l is the lengths of the cylinder after applying a force, r_0 and r are the cylinder's radii corresponding to l_0 and l, respectively, and ν is the Poisson ratio of the material. Thus, the volume of the cylinder is:

$$V = \pi r^2[\frac{\nu^2}{l_0^2}l^3 - \frac{2\nu(\nu+1)}{l_0}l^2 + (1+\nu)^2 l] \tag{A.16}$$

In the assembled meshes, a cylinder can be shared by several adjacent meshes. In this case, the volume of the cylinder is assumed to distribute equally over the shared meshes. Hence, a set of weighting coefficients is set according to the reciprocal of the number of the meshed which share the cylinder. Among the 28 cylinders, there are four cylinders inside the hexahedron, which connects the four diagonal nodes of a hexahedron. Hence, the corresponding weighting coefficient is 1. In addition, there are 12 face cylinders which are shared by two hexahedrons. The corresponding weight is 0.5. Moreover, there are 12 edge cylinders which are shared by 4 hexahedrons. Therefore, the corresponding weighting coefficient is 0.25. The volume of the hexahedra is the weighted sum of the volume of the cylinders.

A.3.2 Volume conservation of the tongue

To account for the volume conservation properties of the tongue, we think the position of the nodes with in the tongue should satisfy Eq. (A.1) as well as minimize the difference between the current volume and the original volume. The Lagrange multiplier first comes to mind. However, the constraint of the volume constancy introduced by the Lagrange multiplier did not always work well; it sometimes interfered with tongue movement. One of the resulting phenomena, for example, was that on occasion the tongue could not move in response to a small change in force. It might be that the vectors for retaining constant volume are not distributed continuously over the multi-dimensional space consisting of the nodes' coordinates. For this reason, a procedure for minimizing volume changes is introduced to reproduce tissue incompressibility.

In the truss-structure model, the total volume of the tongue equals the summation of the volumes of all the cylinders. Therefore, minimizing the changes in volume for all cylinders can achieve a volume constraint for the tongue body. When a force is applied on the cylinder i in the axial direction, the change in the cylinder's volume is

$$\triangle V_i(\boldsymbol{x}) = \pi r_0^2 l_0 - \pi r_0^2 (1 - \nu \frac{\triangle l(\boldsymbol{x})}{l_0})^2 (l_0 + \triangle l(\boldsymbol{x})) \tag{A.17}$$

where the variation of the radius is represented by the length increment $\triangle l$ and the Poisson ratio ν. Using the Houbolt integration method [96], the motion Eq. (A.1) is rewritten in the finite difference expansions of $\boldsymbol{Dx} = \boldsymbol{b}$, where D denotes the resultant matrix on the left side of Eq. (A.1) and b is the vector consisting of known terms one the right side of Eq. (A.1). The constraint is combined with the motion equation system by adding the volume difference, and then minimizing the total error. Thus, the final system equations are derived from the following formula,

$$\frac{\partial}{\partial x}[(\boldsymbol{Dx} - \boldsymbol{b})^T (\boldsymbol{Dx} - \boldsymbol{b}) + \alpha \sum_i \triangle V_i(\boldsymbol{x})] = 0 \tag{A.18}$$

where α is the coefficient to adjust the tolerance of the volume changes in the tongue body.

Table A.1: Physical parameters used in the physiological articulatory model.

Parameter	Value
Density of soft tissue	1.0 g/cm^3
Young's modulus	30 kPa
Viscosity	3 kPa
Poisson ratio	0.49
Gravity	0 m/s^2

A.4 Physical parameters

In the model, the visco-elastic properties of the soft tissues are represented using cylinder elements that connect adjacent nodes in the truss structure. Each cylinder has a virtual volume to maintain the volume conservation property of the soft tissue. The Young's modulus and the viscosity coefficient of the cylinders are modified based on numerical experiments based on the model with reference to those used in the partial-3D physiological articulatory model [35]. The density of soft tissue and Poisson ratio are adopted from previous studies. The major parameters are shown in Table 2.1. The muscle model for generating muscle forces is also adopted from partial 3D model [35].

Appendix B

Muscle force generation model

In this study, we adopts the muscle force generation model in the partial 3D model [35]. In that model, to formulate a generalized model of the muscle, the authors accepted a commonly assumption: a force depending on muscle length is the sum of the passive component (independent of muscle activation) and the active component (dependent on muscle activation).

Figure B.1(a) shows a diagram of the rheological model for a muscle sarcomere [97], which is an extended from Hill's model [98]. The muscle model consists of three parts that describes the nonlinear property, the dynamic (force-velocity) property, and the force-length property. The properties of the muscle sarcomere can be described by a set of differential equations:

$$\sigma_1 = k_1\varepsilon \tag{B.1}$$

$$\frac{\dot{\sigma}_2}{k_2} + \frac{\sigma_2}{k_2} = \dot{\varepsilon} \tag{B.2}$$

$$(\sigma_m + \sigma_3)(k_3 + E) + \frac{k_3\, d(\sigma_m + \sigma_3)}{dt} = b_3 E\dot{\varepsilon} + k_3 E\varepsilon \tag{B.3}$$

$$\sigma = \sum_{i=1}^{3} \sigma_3 \tag{B.4}$$

$$\sigma_m = E\varepsilon^2 \tag{B.5}$$

$$\varepsilon = \frac{l - l_0}{l_0} \tag{B.6}$$

where σ_1, σ_2, σ_3 are the stresses of each part, and σ is the total stress of the sarcomere; l is the current length of the muscle sarcomere, and l_0 is the original length of the muscle

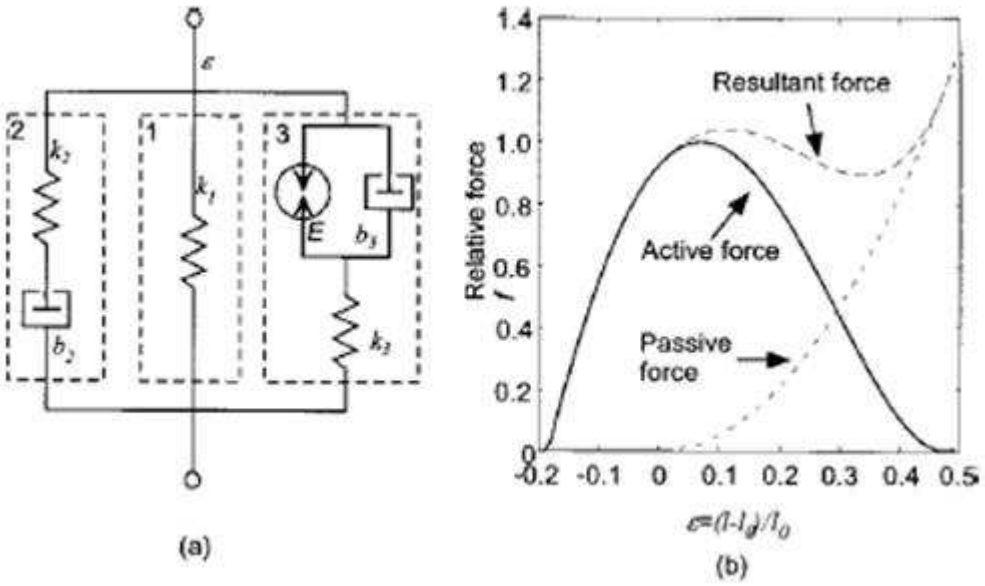

Figure B.1: Muscle modeling: (a) a general model of muscle unit: k and b are stiffness and viscosity, E is the contractile element; (b) generated force varies with stretch ration ε (After Dang and Honda [35]).

sarcomere at rest state.

The first three equations in Eq. (B.1) describe parts 1, 2, and 3 of the muscle model. Part 1 is a nonlinear spring k_1, which is involved in generating force only when the current length of the muscle sarcomere is longer than its original length. The value of k_1 is selected as k_1=$0.05k_0\varepsilon$, where ε>0 and k_0 is the stiffness of the tongue tissue. Part 2 consists of a Maxwell body and is always involved in force generation. According to the second equation of Eq. (2.20), the force generated by this part is determined by two factors: the velocity of the muscle length and the previous force of this branch. As shown in the literature [20, 40, 99], the force-velocity characteristic of the muscle is treated as independent of the previous force. To emphasize the effect of the velocity of the muscle length, a relatively larger stiffness and a smaller viscous component are used in this part. The values of k_2 was set to be twice that of the tongue tissue, while b2 was on the order of one tenth of that used in the tongue body. Part 3 of the muscle sarcomere corresponds to the active component of the muscle force, which is the Hill's model consisting of a contractile element parallel to a dashpot and then cascaded with a spring. This part generates force as a muscle is activated; its characteristics are described by the third equation. In model computations, however, we use a force-length function of the muscle tissue instead of the third equation. The force-length function was derived by matching the simulation and empirical data using the least square method [97]. The function arrived at a fourth-order polynomial of the stretch ratio of the muscles.

$$\sigma_3 = 22.5\varepsilon^4 + 3.498\varepsilon^3 + 4.718\varepsilon^2 + 1.98\varepsilon + 0.858 \tag{B.7}$$

which has a similar shape to that used by Wilhelms-Tricarico [40]. This empirical formula is valid for -0.185<ε<0.49. The active force is assumed to be zero if ε is out of the given range. Figure B.1(b) shows the relationship between the stretch ratio of the muscle sarcomere and the generated force including the passive force. This figure demonstrates the force-length characteristic of the muscle model. Since a muscle consists of a number of muscular fibers with various lengths and thickness, the general lumped rheological parameters of the muscle tissue are not sufficient for determining the muscle-generated force. For this reason, we introduced a parameter, the "thickness" of the muscle fiber, into the force generation. The thickness works as a coefficient for all the three parts of the

muscle sarcomere, which ranges from 0.1 to 4. The value for a given muscle is determined by making the maximum force of the muscles consistent with empirical data [17, 20].

Appendix C

Muscle combinations used in model simulation

GGa-HG-GH-MH + T, V + jawOp/jawCl

GGa-HG-IL-SG

GGa-HG-IL-SL

GGa-HG-MH-GGp

GGa-HG-SG-GH

GGa-HG-SG-SL

GGa-HG-SL-GH

GGa-HG-SL-MH

GGm-HG-GH-GGa

GGm-HG-GH-GGp

GGm-HG-GH-MH

GGm-HG-IL-GH

GGm-SG-GGa-HG

GGm-SG-GGa-SL

GGm-SG-IL-GGp

GGm-SG-IL-HG

GGm-SG-IL-MH

GGm-SG-IL-SL

GGm-SG-MH-GGa

GGm-SL-GH-GGa

GGm-SL-GH-GGp

GGm-SL-GH-MH

GGm-SL-HG-GH

GGm-SL-HG-MH

GGm-SL-IL-GH

GGm-SL-IL-HG

GGm-SL-SG-GH

GGm-SL-SG-HG

GGp-HG-GGa-SL

GGp-HG-GGm-GGa

Appendix D

The number of samples in each cluster

Table D.1: The number of samples in each cluster.

	/a/	/i/	/u/	/e/	/o/	Total
Cluster 1	159	0	0	0	0	159
Cluster 2	0	0	0	0	113	113
Cluster 3	0	0	0	0	0	86
Cluster 4	0	1	523	0	0	524
Cluster 5	0	0	0	0	197	197
Cluster 6	0	0	27	203	0	230
Cluster 7	0	311	19	0	0	330
Cluster 8	119	0	0	0	0	119
Cluster 9	0	0	4	234	0	238
Cluster 10	0	0	0	0	113	113

Bibliography

[1] Denes, P. B. and Pinson, E. N., The Speech Chain: the physics and biology of spoken language. 1993, New York: W. H. Freeman and Company.

[2] Jakoboson, R., *et al.*, Preiminaries to speech analysis. 1963, Cambridge, MA: MIT Press.

[3] Chomsky, N. and Halle, M., The sound pattern of English. 1968, New York: Harper and Row.

[4] Daniloff, R. and Hammarberg, R., On defining coarticulation. Journal of Phonetics, 1973. 1: p. 239-248.

[5] Daniloff, R. and Moll, K., Coarticulation of lip rounding. J. Speech. Hear. Res., 1968. 11: p. 707-721.

[6] Moll, K. and Daniloff, R., Investigation of the timing of velar movements during speech. J. Acoust. Soc. Am., 1971. 50: p. 678-684.

[7] Benguerel, A. P. and Cowan, H., Coarticulation of upper lip protrusion in French. Phonetica, 1974. 30: p. 41-55.

[8] Lubker, J. F. and Gay, T., Anticipatory labial coarticulation: Experimental, biological, and linguistc variables. J. Acoust. Soc. Am., 1982. 71(2): p. 437-448.

[9] Keating, P. A., The window model of coarticulation: articulatory evidence, in Papers in Laboratory Phonetics I: Between the Gammar and Physics of Speech J. Kington and M.E. Beckman, Editors. 1990, Cambridge University Press. p. 451-470.

[10] Browman, C. P. and Goldstein, L. M., Towards an articulatory phonology. Phonology Yearbook, 1986. 3: p. 219-252.

[11] Ohman, S., Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am., 1966. 39: p. 151-168.

[12] Ohman, S., Numerical model of coarticulation. J. Acoust. Soc. Am, 1967. 41(2): p. 310-320.

[13] Perrier, P., *et al.*, The Equilibrium Point Hypothesis and Its Application to Speech Motor Control. J Speech Hear Res., 1996. 39: p. 365-378.

[14] Perrier, P., *et al.*, Control of tongue movments in speech: the Equilibrium Point Hypohtesi perspective. Journal of Phonetics, 1996. 24: p. 53-75.

[15] Payan, Y. and Perrier, P., Synthesis of V-V sequences with a 2D biomechanical tongue model controled by the Equilibrium Point Hypothesis. Speech Communication, 1997. 22(2-3): p. 185-205.

[16] Sanguineti, V., *et al.*, A control model of human tongue movements in speech. Bio. Cybem., 1997. 77: p. 11-22.

[17] Sanguineti, V., *et al.*, A dymamic biomechanical model for neural control of speech production. J. Acoust. Soc. Am., 1998. 103(3): p. 1615-1627.

[18] Perrier, P., *et al.* Modeling the production of VCV sequences via the inversion if a biomechanical model of the tongue. in INTERSPEECH 2005. 2005. Lisbon, Portugal.

[19] Feldman, A.G., Once more on the Equilibrium Point Hypothesis (-model) for model control. Journal of Motor Behavior, 1986. 18(1): p. 17-54.

[20] Laboissiere, R., et al., The control of multimuscle system: Human jaw and hyoid movement. Biol. Cybern., 1996. 74: p. 373-384.

[21] MacNeilage, P. and Sholes, G., An electromyographic study of tthe tongue during vowel production. J. Speech Hear. Res., 1964. 7: p. 209-232.

[22] Hirose, H., Electromyography of the articylatory muscles: current instrumentation and technique. Haskins Laboratories Status Report, 1971. SR-25/26: p. 73-86.

[23] Smith, T., A phonetic study of the functions of the extrinsic tongue muscles. Working Papers in Phonetics, UCLA, 1971. 18.

[24] Miyawaki, K., A preliminary report on the electromyographic study of the activity of lingual muscles. Ann. Bull. RILP, 1975. 9: p. 91-106.

[25] Baer, T., *et al.*, Electromyography of the tongue muscle during vowels in /epvp/ environment. Ann. Bull. RILP., Univ Tokyo, 1988. 7: p. 7-18.

[26] Kumada, M., *et al.*, A study on the inner structure of the tongue in the production of the 5 Japanese vowels by tagging snapshort MRI. Ann. Bull. RILP, 1992. 26: p. 1-13.

[27] Kumada, M., *et al.*, A Study on the Inner Structure of the Tongue for Production of the 5 Japanese Vowels by Tagging Snapshot MRJ; a Second Report. Ann. Bull. RILP, 1993. 27: p. 1-12.

[28] Niimi, S., *et al.*, Functions of tongue related muscles during production of the five Japanese vowels. Ann. Bull. RILP, 1994. 28: p. 33-40.

[29] Stone, M., Modeling the motion of the internal tongue from tagged cine-MRI images. J. Acoust. Soc. Am., 2001. 109(6): p. 2974-2982.

[30] Fujita, S. and Dang, J., A computational tongue model and its clinical application. Oral Science International, 2007. 4(2): p. 97-109.

[31] Takemoto, H., Morphological analysis of the tongue musculature for three dimensional modeling. Journal of Speech and Hearing Research, 2001. 44: p. 95-107.

[32] Wei, J., *et al.*, A model-based learning process for modeling coarticulation of human. IEICE Transcations on Information and Systems, 2007. E90-D(10): p. 1582-1591.

[33] Perkell, J. S., A physiologically-oriented model of t tongue activity in speech production. 1974, MIT.

[34] Honda, K., Orgnization of tongue articulation for vowels. Journal of Phonetics, 1996. 24: p. 39-52.

[35] Dang, J., Honda, K., Construction and control of a physiological articulatory model. J. Acoust. Soc. Am., 2004. 115(2): p. 853-870.

[36] Miyawaki, K., A study on the musculartur of the human tongue. Ann. Bull. RILP, 1974. 8: p. 23-49.

[37] Warfel, J., The Head, Neck, and Trunk. 1993, Philadelphia and London: Led & Febiger.

[38] Kiritani, S., *et al.*, A computational model of the tongue. Ann. Bull. RILP, 1976. 10: p. 243-251.

[39] Kakita, Y., *et al.*, Compuation of mapping from muscular contraction patterns to formant patterns in vowel space., in Phonetic linguistic, V.A. Fromkin, Editor. 1985, Academic Press: New York. p. 133-144.

[40] Wilhelms-Tricarico, R., Physiological modeling of speech production: Methods for modeling soft-tissue articulators. J. Acoust. Soc. Am., 1995. 97(5): p. 3085-3098.

[41] Takemoto, H., *et al.*, A method of tooth superimposition on MRI data for accurate measurement of vocal tract shape and dimensions. Acoustical Science and Technology, 2004. 25(6): p. 468-474.

[42] Dang, J. and Honda, K., A physiological model of a dynamic vocal tract for speech production. Acoustical Science and Technology, 2001. 22: p. 415-425.

[43] Takano, S. and Honda, K., An MRI analysis of the extrinsic tongue muslces during vowel production. Speech Communication, 2007. 49(1): p. 49-58.

[44] Dang, J., *et al.* 3D observation of the tongue articulatory movement for Chinese vowels. in Technical Report of IEICE. 1997.

[45] Honda, K., *et al.* A physiological model of speech production and the implication of tongue-larynx interaction. in ICSLP1994. 1994. Yokohama.

[46] Ostry, D. J. and Munhall, K. G., Control of jaw orientation and position in mastication and speech. Journal of Neurophysiology, 1994. 71: p. 1528-1545.

[47] Dang, J. and Honda, K., Estimation of vocal tract shape from sounds via a physiological articulatory model. Journal of Phonetics, 2002. 30: p. 511-532

[48] Zemlin, W.R., Speech and Hearing Science: Anatomy and Physiology. The fourth Edition ed. 1998: Allyn & Bacon.

[49] Seikel, J.A., et al., Anatomy & Physiology for Speech, Language, and Hearing 3th Edition ed. 2005: Thomson Delmar Learning.

[50] Buchaillard, S., et al. To what extent does Tagged-MRI technique allow to infer tongue muscles' activation pattern? A modelling study. in InterSpeech2008. 2008. Brisbane, Australia.

[51] Hashimoto, K. and Sasaki, K., On the relationship between the shape and position of the tongue for vowels. Journal of Phonetics, 1982. 10: p. 291-299.

[52] Park, J., et al. Model-based Analysis of Cardiac Motion from Tagged MRI Data. in Proceedings of the IEEE Seventh Symposium Computer-Based Medical Systems. 1994: 40-45.

[53] Axel, L., Tagged MRI-Based Studies of Cardiac Function, in Functional Imaging and Modeling of the Heart. 2003, Springer Berlin / Heidelberg.

[54] Niitsu, M., *et al.*, Tongue movement during phonation: a rapid quatitative visualization using tagging snapshot MRI imaging. Ann. Bull. RILP, 1992. 26: p. 149-156.

[55] Dang, J., *et al.* Observation and simulation of Large-scale deformation of tongue. in ISSP06. 2006. Brazil.

[56] Stone, M., Laboratory techniques for investigating speech articulation. The handbook of Phonetic Sciences. 1997: Blackwell Publishers.

[57] Shirai, K. and Honda, M., Estimation of articulatory parameters from speech sound. Trans. IECE, 1978. 61: p. 409-416.

[58] Fang, Q., et al., Investigation of functions of tongue muscles for model control. Chinese Journal of Phonetics (in press), 2008.

[59] Davis, E., *et al.* A continuum mechanics representation of tongue motion in speech. in ICSLP1996. 1996. Philadelphia, USA.

[60] Takano, S., *et al.* Investigation of the intrinsic tongue muscles for production of /i/ using tagged cine-MRI and four cube FEM model. in The 2nd International Symposium on Biomechanics, Healthcare and Information Science. 2008. Kanazawa, Japan.

[61] Perkell, J. S., Properties of the tongue help to define vowel categories: hypotheses based on physiological-orineted modeling. Journal of Phonetics, 1996. 24: p. 3-22.

[62] Fant, G., Acoustic Theory of Speech Production. 1960: Moution & Co.

[63] Stevens, K. N., The quantal nature of speech: evidence frome articulaotry-acoustic data, in Human communication: A Unified view, P.B.a.D. Dunes, E.E, Editor. 1972, McGraw-Hill: New York.

[64] Stevens, K. N., On the quantal nature of speech. Journal of Phonetics, 1989. 17: p. 3-45.

[65] Perrier, P., *et al.*, Vocal Tract Area Function Estimation From Midsagittal Dimensions With CT Scans and a Vocal Tract Cast: Modeling the Transition With Two Sets of Coefficients. Journal of Speech and Hearing Research, 1992. 35: p. 53-67.

[66] Badin, P. and Fant, G., Notes on vocal tract computations. STL QPSR, 1984. 2-3: p. 53-108.

[67] Takemoto, H., Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. J. Acoust. Soc. Am., 2006. 119(2): p. 1037-1049.

[68] Adachi, S. and Yamada, M., An acoustical study of sound production in biphonic singing, Xoomij. J. Acoust. Soc. Am. , 1999. 105(5): p. 2920–2932.

[69] Peterson, G.E. and Barney, H.L., Control methods used in a study of the vowels. J. Acoust. Soc. Am., 1952. 24(2): p. 175-194.

[70] Maeda, S., Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. Speech production and modeling. 1990: Kluwer Academic Publishers.

[71] Flanagan, J.L., Speech analysis synthesis and perception New York/Berlin: Springer-verlag. 1972.

[72] Nakagwa, T., *et al.*, Tonal difference limens for second format frequencies of synthesized Japanesed vowels. Ann. Bull. RILP., 1982 16: p. 81-88.

[73] Flege, J.E., *et al.*, Compasating for a bite block in /s/ and /t/ production: Palatographic, acoustic, and perceptual data. J. Acoust. Soc. Am., 1988. 83(1): p. 212-228.

[74] Honda, M., *et al.*, Compensatory responses of articulators to unexpected perturbation of the palate shape. Journal of Phonetics, 2002. 30: p. 281-302.

[75] Gomi, H., *et al.*, Compensatory articulation during bilabial fricative production by regulating muscle stiffness. Journal of Phonetics, 2002. 30: p. 261-279.

[76] Gracco, V.L. and Abbs, J.H., Dynamic control of the perioral system during speech: kinematic analyses of autogenic and nonautogenic sensorimotor process. Journal of Neurophysiology, 1985. 54(2): p. 418-432.

[77] Abbs, J. H. and Gracco, V. L., Sensorumotor actions in the control of multi-movement speech gestures Trends in Neurosciences, 1983. 69: p. 391-395.

[78] Abbs, J. H. and Gracco, V. L., Contral of complex motor gestures: orofacial muscle responses to load perturbations of lip during speech. Journal of Neurophysiology, 1984. 51(4): p. 705-723.

[79] Folkins, J.W. and Zimmermann, G.M., Lip and jaw interaction during speech: responses to perturbtion of lowe-lip movment prior to bilabial closure. J. Acoust. Soc. Am., 1982. 71(5): p. 1225-1233.

[80] Fowler, C. A. and Turvey, M. T., Immediate compensation in bite-block speech. Phonetica, 1980. 37: p. 306-326.

[81] Lindblom, B., *et al.*, Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. Journal of Phonetics, 1979. 7: p. 147-161.

[82] McFarland, D. H. and Baum, S. R., Incomplete compensation to articulatory perturbation. J. Acoust. Soc. Am., 1995. 97: p. 1865-1873.

[83] Baum, S. R., *et al.*, Compensation to articulatory perturbation: perceptual data. J. Acoust. Soc. Am., 1996. 99: p. 3791-3794.

[84] Atal, B. S., *et al.*, Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Acoust. Soc. Am., 1978. 63(5): p. 1535-1555.

[85] Qin, C and Carreira-Perpinan, M.A. An Empirical Investigation of the Nonuniqueness in the Acoustic-to-Articulatory Mapping. in InterSpeech2007. 2007. Antwerp, Belgium.

[86] Badin, P., *et al.*, Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. Journal of Phonetics, 2002. 30(3): p. 533-553.

[87] Webb, A. R., Multidimensional scaling by iterative majorization using radial basis functions. Pattern Recognition, 1995. 28(5): p. 753-759.

[88] Dang, J. and Lu, X. A perspective on the relation between speech production and perception based on a vowel study. in The 8th Phonetic Conference of China (PCC2008) and the International Symposium on Phonetic Frontiers (ISPF2008). 2008. Beijing.

[89] Bishop, C. M., Pattern recognition and machine learning. Information Science and Statistics, ed. M. Jordan, J. Kleinberg, and B. Scholkopf. 2006: Springer.

[90] Kroger, B. J., *et al.*, Towards a neurocomputational model of speech production and perception. Speech Communication, 2008. (In Press).

[91] Gazzola, V., *et al.*, Empathy and somatotopic auditory morror systems in hummans. Current Biology, 2006. 16: p. 1824-1829.

[92] Kohler, E., *et al.*, Hearing sounds, understanding actions: action representation in mirror neurons. Science, 2002. 297.

[93] Lahva, A., *et al.*, Action representation of sound: audiomotor recognition network while listening to newly aquired actions. The Journal of Neuroscience, 2007. 27(2): p. 308-314.

[94] Zienkiewicz, O. and Taylor, R., The Finit Element Moethod: Its Basis & Fundamentals. The 6th Edition ed. 2005: Elsevier Butterworth-Heinemann publicactions.

[95] Fung, Y., Biomechanics - Mechanical properties of living tissue. The 2nd Edition ed. 1993, New York: Spriger-Verlag.

[96] Bathe, K., Finite Element Procedures. 1996, Englewood Cliffs, NJ: Prentice-Hall.

[97] Morecki, A., Modeling, mechanical description, measurements and control of the selected animal and human body manipulation and locomotion movement, in Biomechanics of Engineering - modeling, simulation, control, A. Morecki, Editor. 1987, Spriger-Verlag: New York. p. 1-28.

[98] Tremblay, S., *et al.*, Somatosensory basis of speech production. Nature, 2003. 243(19): p. 866-867.

[99] Zajac, F., Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control. Critical reviews in Biomechanical Engineering, 1989. 17: p. 359-411.

Publications

[1] Fang, Q., Fujita, S., and Dang, J., "Investigation of functional relationship between tongue muscles for model control", Chinese Journal of Phonetics (to appear)

[2] Fang, Q., Fujita, S., Lu, X., and Dang, J., "A model-based investigation on activation of the tongue muscles in vowel production", Journal of Acoustic Science and Technology (accepted)

[3] Fang, Q., Nishikido, A., and Dang, J. "Feedforward control of a 3D physiological articulatory model for vowel production", NCMMCS2009 (submitted)

[4] Fang, Q., Fujita, S., Lu, X., Dang, J. (2008, 9) "A model based investigation of activation patterns of the tongue muscles for vowel production," InterSpeech2008, Brisbane, Australia, pp. 2298-2301

[5] Fang, Q., Fujita, S., Lu, X., and Dang, J. "Investigation of functional relationship of the tongue muscles for model control", The 8th Phonetic Conference of China (PCC2008) and the International Symposium on Phonetic Frontiers (ISPF2008), Beijing, pp.32 (2008/4/19)

[6] Fang, Q., Fujita, S., Nishikido, A., Lu, X., and Dang, J. "Model-based investigation of the activation patterns of the tongue muscles in articulation", The 4th International symposium on biomechanics, healthcare and information science, 2008, Kanazawa, Japan.

[7] Fang, Q., Wei, J., Lu, X., and Dang, J., "A 3D physiological articulatory model for speech synthesis," the Japan-China Joint Conference of Acoustics 2007, P-2-29, June 2007.

[8] Fang, Q., Dang, J., "Speech Synthesis Based on a Physiological Articulatory Model", Lecture Notes in Computer Science, Volume 4274/2006, Springer Berlin / Heidelberg.

[9] Fang, Q., Nishikido, A., Fujita, S., Lu, X., and Dang, J., "Investigation of 3D tongue shapes for model control", Proc. AJS2008 Spring meeting, pp.325-326

[10] Wang, G., Fang, Q., Kitamura, T., Lu, X., and Dang, J., "Studies of Morphological and Acoustical Properties of Mandarin Using MRI," Proc. AJS2008 Spring meeting, pp.327-328

[11] Wei, J., Lu, X., Fang, Q., and Dang, J., "Assessments of a Radial Basis Function based Vocal Tract Normalization," Proc. ASJ2007 Autumn Meeting, 1-P-17, Sep,2007

[12] Fang, Q., Fujita, S., Lu, X., and Dang, J., "Analysis of 3-D tongue shape in speech production," Proc. of ASJ2007 Autumn meeting , P335 336

[13] Wei, J., Lu, X., Fang, Q., and Dang, J., "Normalization of electromagnetic midsagittal articulographic data using Thin-Plate Spline method" Proc. ASJ 2007 Spring meeting, 3-8-14

[14] Fang, Q., Dang, J., and Lu, X., "A language information aided speech inversion," Proc. ASJ 2007 Spring meeting of, 3-8-20

[15] Fang, Q., Dang, J., " synthesize Chinese vowels based on a physiological articulatory model", Proc. ASJ 2006 Autumn meeting, 2-6-4