# A Study on Control Strategy of a Physiological Articulatory Model for Speech Production

by

Xiyu WU

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

*Supervisor:* Jianwu DANG

*School of Information Science*
*Japan Advanced Institute of Science and Technology*

March 2014

# Abstract

In speech production, articulatory apparatus are the final organs that execute efferent motor commands from the central nervous system. Although the articulators play important roles in speech production, the mechanisms of how the motor commands control the articulators to generate speech sounds are not still very clear. Physiological articulatory models together with their control strategy provide a means to investigate the mechanisms of speech production.

In this thesis, a full 3D physiological articulatory model including the tongue, jaw, hyoid bone and vocal tract wall was constructed based on continuum finite element modeling. This model comprises articulatory muscles with realistic properties and geometrical arrangements. In order to control the physiological articulatory model more accurately, not only the extrinsic genioglossus muscle but also some intrinsic muscles are divided into smaller units according to their functions.

A control framework consists of a feedforward mapping, and a feedback learning loop was realized. In speech production, the feedforward mapping is used to find muscle activation pattern directly according to given articulatory targets and feedback learning loop is used to establish and maintain the feedforward mapping. In this control framework, the articulatory targets were defined by the entire posture of the tongue and jaw in the midsagittal plane, which was reduced to a six-dimensional vector with the principal component analysis (PCA).

Different from the musculoskeletal system, in the muscular-hydrostat system angonist-antagonist muscle pairs varied during articulation, which make it difficult to adjust muscle activations to minimize the distances between target positions and realized ones. In this study, the adjustment of muscle activations was guided by a dynamic PCA workspace that was used to predict individual muscle functions in given positions. This dynamic PCA workspace was estimated based on an interpolation of eight reference PCA workspaces.

In order to construct the feedforward mapping, the articulations of five Japanese vowels from magnetic resonance images were used as the targets for the learning process. The articulatory targets of five Japanese vowels were achieved, which proved that the proposed feedback learning loop was effective for the model control. According to the learning process by using the feedback loop, the feedforward mapping was established. This learned mapping function was assessed through an open set test, and reasonable vocal tract shapes were obtained compared with the target as a result. For the minorities that the articulatory targets cannot achieve perfectly, the implementation of the somatosensory feedback loop can further improve the control accuracy. Besides the improvement of control accuracy, the mapping established by a learning process makes the control strategy the ability to adapt to the external forces added as a perturbation. In order to evaluate the adaptation ability, a vertically downward external force was exerted to the jaw when producing Japanese vowels /i/ and /o/, by implementing the feedback loop, the articulatory targets can be re-achieved, which shows the adaptation ability.

The midsagittal contour including the tongue and jaw was used as the articulatory target, instead of using three crucial points [50, 51]. We expect that by using the articulatory posture as a target, the accuracy of model control for speech production will be improved, because the detailed characteristics of speech sounds depend on the whole vocal tract shape rather than the constriction positions alone.

The physiological articulatory model together with the framework of the control strategy can be implemented in the following aspects: 1) Investigating human speech production mechanism including estimating motor commands from observed articulation, exploring the economy of effort, saturation effect, motor equivalence, etc. 2) Medical treatment. 3) Generating speech sounds by simulating the speech production process of human.

Keywords: Speech production, Physiological articulatory model, Muscle activation, Motor control, Motor learning, Somatosensory feedback

# Acknowledgments

I would like appreciate the people who help me generously during my doctoral study. It is impossible for me to finish this doctoral thesis without their help.

First and foremost, I would like to thank my supervisor Prof. Jianwu DANG for his guidance during my doctoral course. Discussions with him inspired me the ideas to conquer the difficulties, and his criticism pushed me to ponder and improve each particular in my research. Biomechanical modeling was a totally new filed for me when I began my doctoral study. Without his considerate teaching, I could not enter into this new field and finish this thesis. Furthermore, I would like to thank Prof. Kiyoshi HONDA, who gave me a lot of valuable comments in my study. He checked the physiological characteristics of the model and pointed out the deficiencies according to his profound knowledge in anatomy of human articulators.

I would also like to appreciate Professor Masato AKAGI, Associate Professor Hirokazu TANATA and Associate Professor Masashi UNOKI, for serving as my defense committees, gave me valuable comments and constructive suggestions for my thesis. Thanks to the former Associate Professor Asao TOKUDA, IIPL Lab at JAIST, for his discussions and suggestions in my research. Thanks to Assistant Professor Atsuo SUEMITSU and Assistant Professor Shin-ichi KAWAMOTO for helping me to find papers and programs necessary for my research and frequent discussions in weekly meeting.

Thanks to Ian Stavness from Saskatchewan University of Canada for teaching me how to use Artisynth tools. I would like to express my thanks to previous and current members at IIPL laboratory for various discussions and suggestions on my research and a lot of helps in daily life.

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 would like to thank my father and mother in law for helping me to take care of our son, which ensured that I can have enough strength to concentrate on my study. Finally, I am highly grateful to my wife, Ms. Shixiang QIN. She is accompanying me through these difficult years with her endless love, gave me enormous support and encouragements. Thanks the love from my son. I could feel his love to me although he was not good at express his feeling. I owe him a lot in these years.

# Contents

**Abstract** **ii**

**Acknowledgments** **iv**

**1 Introduction** **1**

1.1 Speech production process . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

1.2 Investigation of physiological mechanisms in speech production . . . . . . . 4

1.3 Model based exploration of speech production mechanisms . . . . . . . . . . 6

1.4 Other possible applications of physiological articulatory model . . . . . . . 8

1.4.1 Medical application . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

1.4.2 Speech engineering . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9

1.5 Purpose of this study . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

1.6 Organization of the thesis . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

**2 Physiological Articulatory Model** **12**

2.1 Review of our physiological articulatory model . . . . . . . . . . . . . . . . . 12

2.2 Dynamic simulation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 14

2.3 Morphological and anatomical structures of the model . . . . . . . . . . . . 15

2.3.1 Morphological structures . . . . . . . . . . . . . . . . . . . . . . . . . 15

2.3.2 Arrangements of tongue muscles . . . . . . . . . . . . . . . . . . . . . 17

2.3.3 Arrangements of jaw muscles . . . . . . . . . . . . . . . . . . . . . . 17

2.3.4 Control units of the model . . . . . . . . . . . . . . . . . . . . . . . . 23

2.4 Muscle Model . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 26

2.5 Mechanics properties . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28

2.5.1 Individual muscle function . . . . . . . . . . . . . . . . . . . . . . . . 28

2.5.2 Generation of speech sounds . . . . . . . . . . . . . . . . . . . . . . . 31

2.6 Evaluation of present model . . . . . . . . . . . . . . . . . . . . . . . . . . . 31

2.7 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

**3 Model Control for Articulation** **37**

3.1 Speech acquisition process . . . . . . . . . . . . . . . . . . . . . . . . . . . . 37

3.2 Goals of speech production . . . . . . . . . . . . . . . . . . . . . . . . . . . 39
3.3 Framework of model control . . . . . . . . . . . . . . . . . . . . . . . . . . 41
3.4 Control strategies in the literature . . . . . . . . . . . . . . . . . . . . . . . 42
3.4.1 Muscle workspace . . . . . . . . . . . . . . . . . . . . . . . . . . . 43
3.4.2 EP-map . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
3.4.3 Other control strategies . . . . . . . . . . . . . . . . . . . . . . . 46
3.4.4 Summary of previous control strategies . . . . . . . . . . . . . . . 46
3.5 Representation of articulatory target . . . . . . . . . . . . . . . . . . . . . 47
3.6 Object of the control strategy . . . . . . . . . . . . . . . . . . . . . . . . . 48

**4 Algorithm of the control strategy 51**
4.1 Principal component analysis of the articulatory posture . . . . . . . . . . 51
4.1.1 Algorithm of PCA . . . . . . . . . . . . . . . . . . . . . . . . . . . 51
4.1.2 PCA implementation . . . . . . . . . . . . . . . . . . . . . . . . . 52
4.2 Feedback error learning . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55
4.2.1 Flowchart of learning process . . . . . . . . . . . . . . . . . . . . . 56
4.2.2 Dynamic PCA workspace . . . . . . . . . . . . . . . . . . . . . . . 57
4.3 Feedforward mapping . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

**5 Evaluation 61**
5.1 Implementation of feedback error learning . . . . . . . . . . . . . . . . . . . 61
5.1.1 Function of intrinsic muscle . . . . . . . . . . . . . . . . . . . . . . 67
5.2 Evaluation of feedforward mapping . . . . . . . . . . . . . . . . . . . . . . 68
5.3 Behavior of integrated control . . . . . . . . . . . . . . . . . . . . . . . . . 71
5.4 Adaptation experiment . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75
5.5 Summary and discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . 75

**6 Summary and Future Work 79**
6.1 Summary of this thesis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
6.2 Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 80
6.3 Future work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81

**References 83**

**Publications 93**

# List of Figures

1.1 Speech chain in human speech communication (cited from [1]) . . . . . . . 2

2.1 Geometrical structures of the tongue (upper left), vocal tract wall (upper right), mandible (lower left) and hyoid bone (lower right) . . . . . . . . . . 16

2.2 Lateral view (left) and posterior view (right) of extrinsic tongue muscles cited from Elsevier Drake *et al.*, (2010) [106] adapted with permission. . . . 18

2.3 posterior view (left) and sagittal cross-section view (right) of intrinsic tongue muscles cited from Elsevier Drake *et al.*, (2010) [106] adapted with permission. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18

2.4 Arrangement of tongue muscles in the physiological articulatory model . . 20

2.5 Arrangements of muscles in the jaw model . . . . . . . . . . . . . . . . . 21

2.6 Anatomic description of jaw opener muscle, jaw opener muscle include Interior Digastrics, Posterior Digastrics, Lateral Pterygoid, and assistant by strap muscle Sternohyoid, Cited from Elsevier (2010), Drake *et al.* [106] adapted with permission. . . . . . . . . . . . . . . . . . . . . . . . . . . 22

2.7 Anatomic description of jaw closer muscle, jaw closer muscle include Temporalis, Masseter, and Medial Pterygoid, Cited from Elsevier (2010), Drake *et al.* [106] adapted with permission. . . . . . . . . . . . . . . . . . . . . 23

2.8 Control units of the tongue muscles. GGa, GGm and GGp (anterior, middle and posterior portion of genioglossus muscle, respectively); HG (hyoglossus muscle); SG (styloglossus muscle); SLa and SLp (the anterior and posterior portion of superior longitudinal muscle); IL(inferior longitudinal muscle); Va, Vm and Vp (anterior, middle, and posterior portion of vertical muscle, respectively). Ta, Tm and Tp (anterior, middle and posterior portion of transverse muscle, respectively); MH (mylohyoid muscle); GH (geniohyoid muscle). . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

2.9 Profiles of the constructed physiological articulatory model . . . . . . . . . 25

2.10 Muscle modeling: (a) a general model of muscle unit: $k$ and $b$ are stiffness and dashpot, $E$ is the contractile element; (b) generated force varies with stretch ratio $\varepsilon$ (After Dang and Honda [51]). . . . . . . . . . . . . . . . . 27

2.11 Function of individual muscles in the physiological articulatory model. Black solid lines show the equilibrium position after the muscle is activated for 200 ms duration, dotted gray lines correspond to the shape in its rest position. (Unit: cm) . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29

2.12 Muscle activation and equilibrium position. Black thick lines are the rest position of the tongue; gray lines are the equilibrium position driven by different activation levels [0.002, 0.01, 0.03, 0.1, 0.4]. . . . . . . . . . . . . . 30

2.13 Vocal tract configuration with the tongue, the right half of the jaw, palate, pharyngeal wall and larynx for the vowel /a/. . . . . . . . . . . . . . . . 32

2.14 Cross-sections sliced by the planes in Figure 2.13, showing the larynx (left), velum region (middle), and front cavity (right). . . . . . . . . . . . . . . . 32

2.15 An example of area function of vowel /a/ generated by model simulation. . 33

2.16 Equilibrium position by the same muscle activation from different initial positions. Upper left (discrete FEM) and upper right (Continuum FEM) show the equilibrium position of tongue tip when GGp activated with 3.5 Newton. Lower left (discrete FEM) and lower right (Continuum FEM) show the equilibrium positions of tongue dorsum when GGa activated with 3.5 Newton. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

2.17 Comparison of the response time. Blue lines and red lines show the result of continuum model and discrete model, respectively. . . . . . . . . . . . 35

3.1 Schematic diagram of speech production . . . . . . . . . . . . . . . . . . . 38

3.2 Framework of speech motor control implemented in this study. . . . . . . . 42

3.3 Three crucial points used to represent articulatory targets. Three black squares show tongue tip, tongue dorsum and jaw, respectively. . . . . . . . 43

3.4 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 [50]) . . . . . . . . . . . . . . . . . . . . . . . . . . 44

3.5 Coordinates consisting of the equilibrium positions corresponding to the activation forces ranged between 0 and 6 Newton. The net in the right panel consists of the contour lines of the EPs of SG and HG. (After Dang and Honda [51])) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

3.6 Representation of articulatory posture in the midsagittal plane. Black squares on the tongue and jaw are used to represent articulatory posture. . 48

3.7 Agonist-antagonist muscles of elbow [91]. . . . . . . . . . . . . . . . . . . . 49

4.1 Effect of the PCA components. Gray lines show the rest position, dashed lines and dashed lines with squares show the directions of each component with positive and negative coefficient, respectively. . . . . . . . . . . . . . . 54
4.2 Function of individual muscle unit in 3D PCA space. . . . . . . . . . . . 55
4.3 Relationship between muscle activation and PCA vector module (GGa). . . 56
4.4 Flowchart of muscle estimation. . . . . . . . . . . . . . . . . . . . . . . . 57
4.5 Reference PCA workspaces (solid lines) and dynamic PCA workspace (dash lines) in 3D PCA. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 60
4.6 An example of the interpolation surface using four reference points with coordinates (0, 0), (0, 1),(1, 0), (1, 1) and their values 0, 2.5, 7.5, 10. . . . 60

5.1 Processes of muscle activation estimation of vowel /a/ (left) and vowel /o/ (right). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62
5.2 Realized position for five Japanese vowels. Gray dash lines show the rest position. Gray dash lines with square markers show the target postures. Black lines with stars show the realized positions. The average distances (defined in Eq.(4.5)) for /a/, /i/, /u/, /e/, and /o/ are 0.123 cm, 0.145 cm, 0.06 cm, 0.085 cm, and 0.168 cm, respectively. . . . . . . . . . . . . . . . . 63
5.3 Comparison between estimated extrinsic tongue muscle activations and EMG measurement. The vowel positions are referred from [81], where black dots represent the positions of Japanese vowels and gray dots show the positions of partial English vowels that are close to the Japanese vowels in articulatory space. The black bars are the normalized results obtained by the proposed method, and the gray bars are the corresponding normalized EMG measurements. . . . . . . . . . . . . . . . . . . . . . . . . . . 66
5.4 Distance between the realized posture and target posture. Gray bars and black bars show the distances using only the extrinsic muscles and all the muscles, respectively. (Unit: cm) . . . . . . . . . . . . . . . . . . . . . 68
5.5 Comparison of the achievements between using extrinsic muscle and all muscles are available, left and right panels show vowel /a/ and vowel /o/, respectively. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69
5.6 Distribution of the 11 points along the tongue surface in the midsagittal plane of the simulations of target based learning. The area with different colors corresponds to the dispersion of individual tongue nodes. . . . . . 70
5.7 Flowchart of evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . 71
5.8 Distribution of distance between target position and realized ones. . . . . 72
5.9 Distribution of distance between target position and realized ones (integrated control by set the threshold to 0.25cm ). . . . . . . . . . . . . . . 73

5.10 Distribution of distance between target position and realized ones (integrated control by set the threshold to 0.16cm ). . . . . . . . . . . . . . . . . 74
5.11 Setting of perturbation experiment. . . . . . . . . . . . . . . . . . . . . . . . 76
5.12 Adaption of the external force, left and right panels show vowel /i/ and /o/, respectively. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 77

# List of Tables

2.1 Anatomical description of muscle courses and muscle function [56]. . . . . . 19

4.1 Variance of PCA components (%). C1 to C6 are the first six components, VC is the variance of the components, and AVC is the accumulative variance from the first component to the current component. . . . . . . . . . . 53

4.2 Spacial resolution of often used devices (cm). . . . . . . . . . . . . . . . . . . 53

5.1 Average distance and standard deviation of 12 points. . . . . . . . . . . . . 64

5.2 Muscle activation patterns for the five Japanese vowels. The active muscle force of JO and JC is the sum of the active force included in the muscle groups. (Unit: Newton) . . . . . . . . . . . . . . . . . . . . . . . . . . . . 65

5.3 Comparison of formant between observed and generated. (Unit: Hz) . . . . 67

5.4 Average distance to target (calculated by using Eq. 4.5) before perturbation and after adaption. (Unit: cm) . . . . . . . . . . . . . . . . . . . . . . 76

5.5 Comparison of muscle activation patterns before and after jaw perturbation. The active muscle force of JO and JC is the sum of the active force included in the muscle groups. The muscle forces printed in blue color show the muscle forces that have changed. (Unit: Newton) . . . . . . . . . . . . 77

# Chapter 1

# Introduction

Speech is one of the most convenient tools that provide us the ability to communicate with each other, to share experiences and to exchange ideas. Every day, we use speech, a peculiar ability of human beings, without being aware of how speech is generated. Speech production process, involving many levels of human activity, is very complex, and many unknown mechanisms in the process need to be explored. The understanding of the mechanism of speech production will help computer engineers to develop a more natural human-like speech synthesis system by imitating speech production process of human. Moreover, the understanding of speech production mechanism will lead to more effective therapies for the patients with speech disorders.

## 1.1 Speech production process

We first take a look at the speech communication process and then concentrate on aspects of the production process. One of the most famous descriptions of the speech communication process is the speech chain [1]. The schematic diagram of the speech communication process is shown in Figure 1.1. This figure describes the procedures of how a speaker transmits messages to a listener. First, the speaker arranges his/her ideas into linguistic representations by selecting the correct words and phrases that can express his/her ideas. Then the appropriate instructions are issued and sent to the muscles which control the vocal apparatus including the vocal folds, tongue, jaw and lips, etc. The instructions are sent in the form of impulses along the motor nerves to the muscles that control the vocal organs. The movements of the vocal organs generate the proper sound sources and vocal tract configurations. As a result, speech sounds are generated. This process is referred to as speech production.

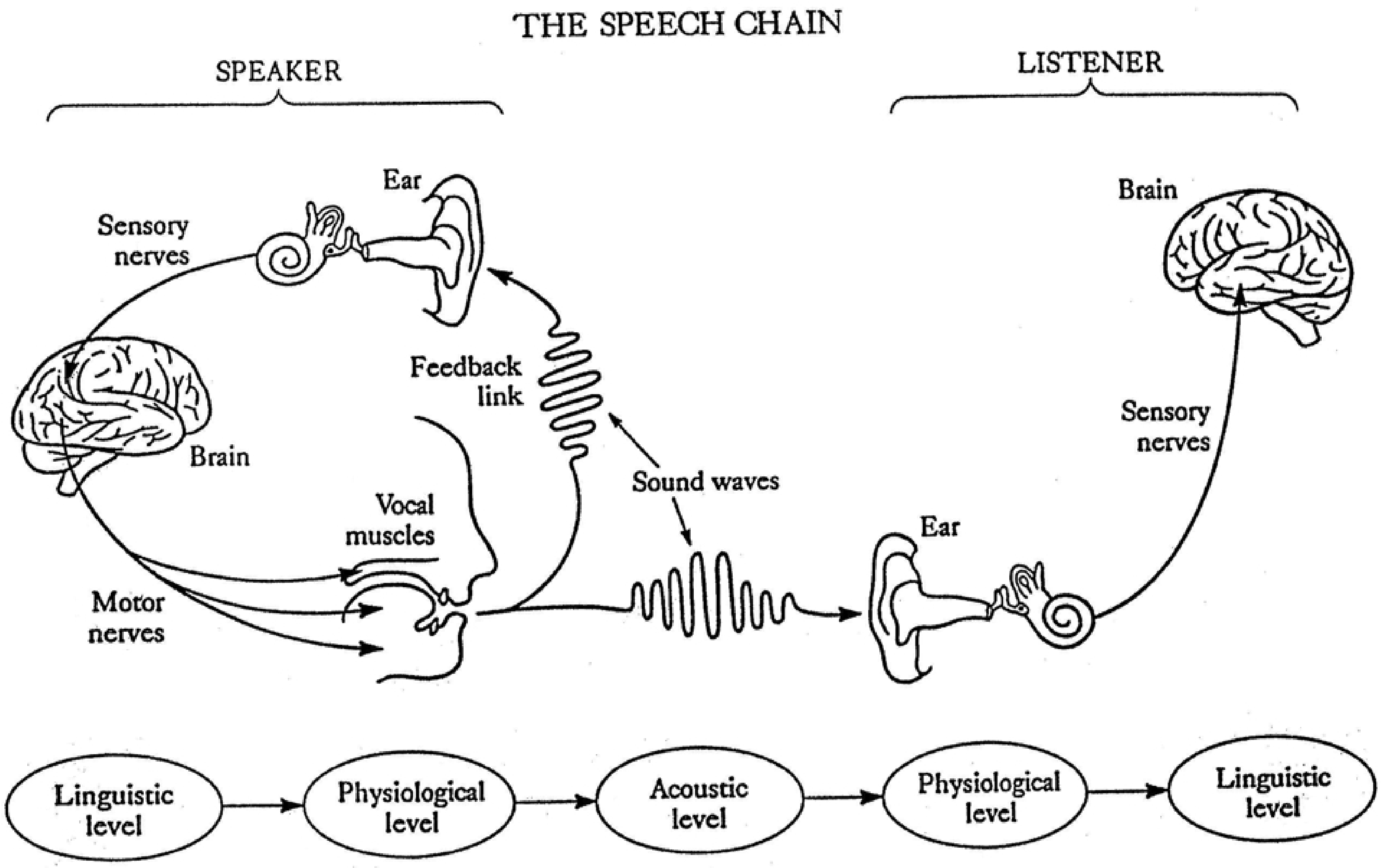

Figure 1.1: Speech chain in human speech communication (cited from [1])

The speech sound wave generated by the speaker travels through the transmission media (usually the air) to the listener. The speech sounds transmitted as the way of air vibrations activate the listeners hearing mechanism and produce nerve impulses. The human brain decodes these impulses received from the sensory nerves, and then recognizes the meaning that the speaker wants to express. This is the perception process.

In speech communication process, speech production and perception are two of the most important aspects. There are mainly two theories indicate that the the exploration of speech production will deepen our understanding of not only speech production but also speech perception: 1) "motor theory of speech perception", and 2) "Mirror neuron system". In the "motor theory of speech perception" [2], the main idea is that "the objects of speech perception are the intended phonetic gestures of the speaker,which are represented in the brain as invariant motor commands that call for movements of the articulators through certain linguistically significant configurations. These gestures commands are the physical reality underlying the traditional phonetic notions, which are the elementary events of speech production and perception". The "mirror neurons" was found in the F5 region of monkeys' frontal premotor cortex[3]. These neurons exhibit the remarkable property of spiking during both the active production and passive observation of certain motor actions. A given mirror neuron may fire spikes, for example, when a monkey grasps a piece of fruit with the hand or when the monkey observes a human grasping fruit in a similar fashion. The mirror neurons related to communicative mouth movements [3] have been found in the region of monkeys' premotor cortex immediately lateral to the region for grasping movements [4]. Some functional MRI studies in humans support the notion that the Brocas area plays a central role in the mirror representation of hand and finger gestures [5], and the Brocas area is classically associated with speech motor control. According to these two theories, we can expect that the investigation of speech production will contribute to the understanding of speech production and speech perception as well.

Because the present thesis focuses on the aspect of speech production process, we would like to describe the aspects of speech production and omit the perception process. As shown in Figure 1.1, the process from the intention of the speaker to the generation of speech sounds can be roughly divided into three different levels. In the brain, the speakers intention is transformed to proper words and sentences. This process can be called as the linguistic level. In the physiological level, the central nervous system generates neural impulses to activate the muscles associated with speech production, and the activation of muscles generates muscle force to control the movements of the vocal organs. The process that speech sounds are generated by the movements of the vocal organs is regarded as the acoustic level. Therefore in speech production, there are three main levels which are the linguistic level, physiological level, and acoustic level. The investigation of the activities in these three levels will further our understanding of the mechanism in speech production.

In order to understand the full mechanism of speech production, all the processes in speech production have to be explored and make these mechanisms clear. Because the acoustic signals can be directly measured, a vast amount of studies have conducted in the acoustic level, and abundant of applicable achievements have been obtained including voice encoder and decoder, speech synthesizer [6, 7], recognizer [9, 10], and so on. In the linguistic level, so far, the researchers mainly focus on how to assemble the linguistic units to express the intention of a speaker according to the rules of the language. These studies focus on the description in the linguistic level including the phonemes, words, grammar, etc., other than the representations of these linguistic elements in the human brain. With the development of the techniques that can measure the activity of the human brain [11], the representation of linguistic elements in the human brain will be investigated in the future. Another possible route to investigate linguistic representations in the human brain may base on an inverse estimation method, where the activities in the human brain can be predicted based on the measurable acoustic signals and the issued motor command. In the present thesis, we focus on exploring the physiological mechanisms of speech production including the movements of articulators and the motor command used to control the articulators, which are issued from human brain.

## 1.2 Investigation of physiological mechanisms in speech production

To fully understand the physiological mechanisms of speech production, four main important aspects need to be investigated: 1) mechanical characteristics of the vocal organs 2) geometrical structures of articulators 3) motor commands that control the movements of articulators 4) the dynamic properties of the articulators.

It is not difficult to measure the mechanical properties of the vocal organs such as the stiffness, Poisson coefficient, and density. For example, the stiffness of the tongue muscles was measured in vivo [19] or from cadaver [20]. The geometrical structures of the vocal organs can be measured by medical imaging techniques, such as MRI (Magnetic Resonance Imaging), Ultrasonography, etc. MRI techniques [13] that use the magnetic properties of hydrogen atoms to create images of soft-tissue structures in the body have been implemented to extract the geometric shape of the tongue and jaw [14, 15]. Ultrasonography is an ultrasound-based imaging device used for visualizing subcutaneous body structures. This device has been implemented to measure the three-dimensional tongue surfaces [16] and the movements of the tongue during speech production [17, 18]. In these four aspects of the physiological mechanisms of speech production, the rather difficult missions are to investigate the motor commands that are issued from the human brain to control the movements of the vocal organs and the dynamic properties of articulators. Because the motor commands activate the articulatory muscles directly, the investigation of motor commands can be regarded as the investigation of muscle activation patterns.

The movements of articulators are driven by the activation of muscles. Therefore, in order to find out the effect of a muscle activation pattern composed by different activation levels of individual muscles, the individual muscle function is very important. The function of individual muscles is determined by its location and orientation. To obtain the location and orientation information of each muscle, Miyawaki examined the tongue muscle fibers by using slices of tongue tissue in the sagittal, coronal, and transverse planes based on cadaver specimens[21]. Takemoto presented detailed descriptions of the muscular structure of the human tongue based on macroscopic and microscopic observations and provided three-dimensional schemata of the tongue musculature [22].

The anatomical description of the orientations and locations of muscles can help to estimate the functions of individual muscles. However, The knowledge of individual muscle functions estimated from anatomical knowledge is not enough to infer the complicated muscle co-activations, because the muscles have a higher degree of freedom than articulatory postures which may result in different muscle activation patterns to generate the same articulatory posture (One-to-Many problem). Different from the inference according to anatomical information, muscle activation can be measured from experimental observations. There are two mainly experimental methods to measure the muscle activation patterns in speech production: 1) measure the muscle activation directly by using EMG (electromyography) 2) estimate the muscle activations according to the observed deformation of muscles. EMG is a technique that detects the electrical potential generated by muscle cells when the muscle is activated. This technique has been successfully implemented to measure the activation of extrinsic tongue muscles simultaneously during producing vowels in the context of /əpvp/ [23]. However, by using EMG, it is difficult to measure the activations for the small intrinsic muscles, such as the verticalis, transversus and longitudinal muscles, which also play important roles in speech production[cite papers if any]. Different from the direct detection method like EMG, muscle activations can be estimate according to their deformation. Deformation of the tongue body has been measured by tagged-MRI [24, 25, 26, 27] and MRI [28]. Nonetheless, sometimes, it is difficult to draw conclusion about muscle activation according to the deformation of muscle fibers, because 1) The shorting of muscle fibers may be caused by the activation of the other associated muscles (passive deformation) rather than caused by its own activation. 2) In specific situation, where some of the muscles can be lengthened even if it is activated [23].

As has been described that the kinematic properties of the articulators including the spacial pathes, velocities and their accelerations, can be measured by using observation devices, however, the movements of the articulators are the result of the actions of muscle activations on the mechanical structures. The characteristics of the movements depend not only on the intrinsic mechanical properties of the articulators but also on their interactions. The observable and measurable mechanical properties including the friction, density, Poisson rate, etc., the geometrical structures of the articulators and the individual muscle functions are only the specific aspects. In order to fully comprehend the dynamic properties of the articulatory system, we have to integrate all the individual properties and their interactions into a system.

To construct a computational physiological articulatory model is necessary, because: 1) As described previously, there are so many inconveniences in exploring the muscle activations in speech production. 2) It is difficult to investigate the dynamic properties of the articulatory systems based on current measurement techniques. If we construct a physiological articulatory model that can simulate the characteristics of human articulators including their anatomical structures and mechanical properties, this model can be used to represent the human articulators, and based on the model speech production mechanisms can be explored.

## 1.3 Model based exploration of speech production mechanisms

Physiological articulatory models can be used to simulate the speech production process, and then explain the cause and effect in speech production. This section introduces the investigation of speech production mechanisms based on the physiological model in the literatures. The implementations of the model based investigation will be introduced in these aspects including muscle activation, economy of effort, saturation effects, etc.

The physiological mechanisms of the articulatory muscles are very poorly understood, although they play critical roles in controlling the movement of articulators and the resonance properties of the vocal tract. As described in the previous section, traditional experimental methods have their own disadvantages in exploring muscle activations in speech production. To suppress these disadvantages, the physiological articulatory model based method can be implemented. The muscle activation patterns of the five Japanese vowels [29] and cardinal French vowels [30] were estimated based on a physiological articulatory model, respectively. The estimated muscle activation patterns have been compared to the muscle activation patterns obtained from EMG experiments, which proved that the physiological articulatory model was feasible to estimate muscle activation in speech production [29]. The results from these two studies include the activations not only for the extrinsic muscles but also for the intrinsic muscles that are difficult to measure by using

EMG. The results of these two studies demonstrate that by model simulation, measurement data can be verified, and potential parameters that cannot be observed by using present techniques can be predicted such as the activation of the intrinsic muscles.

In order to explain the causal mechanisms behind the movements of articulators in speech production, researchers have proposed many hypotheses and tried to prove the hypotheses by using measurement techniques. In speech motor control, the economy of effort seems to be a principle that guides speech movement [31, 32]. According to the theory, a specific hypothesis in speech production is that speakers usually produce intelligible sound sequence with an economy of effort by planning the acoustic trajectory to passes through the parts of acoustic goal regions in the sequence that are closest to one another [33]. Nelson proposed five measures of the physical cost associated with accomplishing skilled movement including time, force, impulse energy and jerk, and proved that jaw movement data is likely to be influenced in a very sensible way by physical objectives that relate to the movement's effort or energy [34]. Perkell *et al.* explored the hypothesis that clear speech is produced with greater "articulatory effort" than normal speech by analyzing kinematic and acoustic data gathered in different speaking conditions, including normal, fast, clear, and slow. The analysis of the kinematic data recorded by EMA showed that some speakers increased their effort (reflected in the greater peak speed) in the clear speech condition, some speakers use other combination of parameters to produce speech sounds in clear condition[35, 36]. The deficiency of these studies is that to infer the energy cost from the movement is very difficult because the articulatory muscles are interweaved, and the movement of articulator is the result from a complicated muscle activation process. Simulation based on physiological articulatory model can conquer the deficiency of inferring the energy cost from measured movement by providing diverse parameters, including displacement, relative strain, and relative muscle induced stress. Stavness *et al.* have investigated the cause of two types of articulation patterns of English /r/ based on a physiological articulatory model [37]. About 40% of American English speakers use multiple /r/ variants, where the bunched /r/ was more likely to occur adjacent to the vowel /i/, whereas tip-up postures occurred coupled with /a/ and /o/ [39]. Simulations based on the physiological articulatory model showed reductions in all three measures for the transitions between bunched /r/ and the vowel /i/, and between tip-up /r/ and the vowel /a/, which uncovered the mechanical articulatory factor of the variation. According to this study, we can see that physiological articulatory model based analysis provides the avenues for uncovering the economy of effort speculated in speech production.

Besides the hypothesis of the "economy of effort", the biomechanical mechanism for some other theoretical speculations such as "saturation effect", "motor equivalence", coarticulation, and effect of gravity orientation on speech production can be explored based on a physiological articulatory model. The "saturation effect" is originated from the "quantal effect” [40, 41], which is defined as a nonlinear relation between the input and output parameters in different measurement domains, in which a continuous change of the input parameter results in a rapid change and then a “saturation” of the output. According to this, there is a “saturation” region in which the output parameter is stable even with some inaccuracy of the input parameter. Biomechanical saturation effects usually happen in cases where one articulator contact with another. Because of the contact of articulators, the continuous increase of muscle activation will not change the vocal tract shape. There is a hypothesis that biomechanical saturation effects may help to determine the acoustic goals of the vowel /i/ [42, 43, 35]. Buchaillard *et al.* tried to verify this hypothesis by a simulation based on a physiological articulatory model [30]. In this study, motor commands of the vowel /i/ were modified for 9 muscles independently, and the formants of output speech sounds were obtained following the modification of motor commands. The analysis of the variability patterns and the interpretation in terms of the respective influence of each muscle strongly suggest that there is no saturation effect, which would facilitate the accurate control of the constriction area for vowel /i/.

In human speech, coarticulation is a natural phenomenon, which is affected both at the physiological level (carry over) and motor planning stage (look-ahead). Although it is difficult to investigate the coarticulation in the motor planning stage directly, we can estimate the mechanism in the motor planning stage from the observable articulatory movement. Wei *et al.* [44] proposed a two-layer learning framework to learn the parameters in the motor planning level, and the parameters concerned with the physiological level (carry over) was simulated by a physiological articulatory model.

People may speak in the situation where the gravity orientation or gravity level is changed, and this may alter the vocal tract shape and its control strategy. With the increasing use of MRI systems, numerous speech data are acquired while the subject is lying on his or her back, where the gravity orientation is changed. How the gravity orientation affect speech production was investigated by Buchaillard *et al.* based on a physiological articulatory model [30]. The results of this study are consistent with experimental observations [26, 45], which proves that the model based investigation method is effective.

## 1.4 Other possible applications of physiological articulatory model

In general, physiological articulatory models that possess the morphological structure and mechanical properties of articulators can reflect the properties of human articulators and implemented in different fields as well.

### 1.4.1 Medical application

With the development of computational modeling technology, physiological articulatory modeling would have a wide range of applications in medicine. It can deepen our understanding of motor system dysfunction and consequently aiding to diagnose the possible cause of speech disorder. Computational modeling simulation can help the surgeon to plan surgery and predict the postoperative functions of articulators.

Glossectomy is an effective treatment for the patients with oral cancer, but the prediction of the postoperative effect mainly depend on the empirical knowledge, which is difficult because the tongue has a complex anatomical structure. To address this problem, Fujita, *et al.* developed a computational tongue model, which can produce the movement in the same way as the human tongue, and attempted to use it for predicting post-glossectomy effects under various conditions [46]. In addition, computer simulation allows for iterative refinement of surgical procedures with little cost and risk to patients. Starveness compared the theoretical post-operative deficit for jaw surgery with and without reconstruction by a simulation based on a jaw-tongue-hyoid biomechanical model [47]. This type of analysis could be used on a patient to determine, given the planned extent of tissue resection, whether or not jaw reconstruction would be beneficial.

Simulations based on a physiological articulatory model can benefit by guiding rehabilitation. A model with a new morphological structure and a new definition of related muscles which conform to the status of the patient after surgical operation can be used to simulate the articulatory movement of the patient. Theoretically, muscle activation level has more degrees-of-freedom than the articulatory posture, and articulatory posture has a more degrees-of-freedom than acoustics. A simulation based on the model can tell the patient whether the intended posture can be realized or not by using the muscles after the surgical operation. In order to generate intended speech sound, if the usually used articulatory posture cannot be generated, it is possible to find a new articulatory posture that can generate the desired speech sound by the simulation based on the model.

### 1.4.2 Speech engineering

Although the corpus-based speech synthesis systems are successfully implemented in a wide range of different fields, the quality of the synthesized speech sounds is still far away from human speech. The main deficiency of the corpus-based speech synthesizer is that the mechanism of the synthesis system is different from that used in human speech production. Although most of the synthesis systems have adopted some techniques to smooth the connections between the concatenative units and the naturalness of consequent speech can improve more or less, to solve this problem completely, theoretically, a physiological articulatory model that can imitate the mechanism of human articulators is necessary.

A physiological articulatory model can benefit the pronunciation for patients with speech disorder or normal people in second language learning by providing articulatory visual feedback. Auditory feedback plays a main role in language acquisition process. S-tudies proved that augmented visual feedback can help the patients with speech disorder of apraxia [48] and participants in second language learning [49] to improve their accuracy of speech production. Recently, many instruments that can directly display movement of articulator visually have been developed such as X-ray microbeam, electropalatography (EPG), ultrasound, Magnetic Resonance Imaging (MRI), and electromagnetic articulography (EMA) systems. However, it is difficult to apply these techniques to provide visual feedback because of the following two reasons 1) the high cost of the equipment; and 2) inconvenience to use. The visual feedback information can be acquired by inverse estimation of the articulatory movement from speech sounds generated by the subject [50], which can conquer the disadvantages of using these instruments.

## 1.5 Purpose of this study

As we have described in the previous section, a physiological articulatory model could be implemented in so many fields including investigation of mechanism of speech production, medical treatment, speech synthesis, etc., so far there are two important reasons that obstacle wide implementations of the physiological articulatory model: 1) The physical properties of the physiological articulatory model do not perfectly conform to those of human articulators. 2) The control strategy for the physiological articulatory model is still far away from that used in human.

The object of this study is to construct a physiological articulatory model whose geometrical and anatomical structures, and mechanical properties are as similar as possible to those of the articulators of human, and based on this model realize a high accuracy control strategy.

## 1.6 Organization of the thesis

In Chapter 2, we first describe why the physiological articulatory model need to be upgraded and how it is developed to present status based on a brief introduction of our previous models. The description of the model includes the morphological structures of the articulators, muscle models, control units and equations of dynamic simulation. Then evaluations are conducted to verify the improvement of the present model by comparing it to the previous one. In Chapter 3, a schematic diagram of speech motor learning is presented then introduce where we focus on and the evidences that support our focus. Then introduce the control strategies in the literature and point out the deficiencies of them.

At the end of this chapter, propose the object of our control strategy and meanwhile point out the challenges by comparing the articulator control to arm control. In Chapter 4, the detail algorithm of feedback error learning is described as well as the feedforward inverse model. In Chapter 5, evaluations are conducted to verify whether the objects of control strategy are realized, including the control accuracy and the ability to adapt to perturbation. In this chapter, the functions of intrinsic muscle in vowel articulation are also assessed based on the physiological articulatory model. Finally, in Chapter 6, we summarize the thesis and list the main contributions of this study.

# Chapter 2

# Physiological Articulatory Model

To develop a model which can faithfully represent the characteristics of human articulators, the anatomical structures and the mechanics of articulators have to be constructed precisely. In this chapter, we will introduce why it is necessary to upgrade the previous physiological articulatory models to the current version and which parts are optimized comparing to the previous model. A review on how the physiological articulatory model was developed in the previous version is introduced firstly.

## 2.1 Review of our physiological articulatory model

The purpose of introducing a brief history of our physiological articulatory model is to make it clear where present model is from, and what are the same as the original models and what are improved. For those models unrelated with the model used this study, some of them were reviewed in the first chapter.

Dang and Honda constructed a partial 3D physiological articulatory model by extended-FEM (X-FEM) based on an MRI measurement of a male Japanese subject [51]. 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 the 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 was constructed by connecting the section nodes in the midsagittal plane to the corresponding nodes in the left and right planes, where each mesh was a hexahedron. Thus, the model represents the principal region of the tongue as a 2 cm 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

cm and 1.4 cm from the midsagittal plane on the left side, respectively). The anatomical arrangement of the major tongue muscles genioglossus, hyoglossus, geniohyoid, superior longitudinal, styloglossus and inferior longitudinal were 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, were arranged according to the anatomical literatures [52, 52, 53]. The model of the jaw has four nodes on each side, which were similar to those used by Laboissière *et al.* [54]. And those four nodes were connected by five rigid beams to form two triangles with a shearing beam. The hyoid bone was modeled as three rigid segments corresponding to the body and bilateral greater horns. Each segment had two nodes connected with a rigid beam. The masses were uniformly distributed over the rigid beams. To provide a uniform computational format, rigid beams were also treated as visco-elastic links with a high Youngs 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 muscles cannot be arranged precisely; and (3) it is difficult to account for the interaction between the tongue and surrounding structure accurately.

To overcome the disadvantages of the partial 3D physiological articulatory models, Fujita *et al.* developed the tongue model to a full 3D and implemented it to predict the effect of glossectomy [46]. Furthermore, Fang *et al.* optimized the model by integrating a 3D tongue model [46] with the surrounding structures the jaw and vocal tract wall, and implemented it to estimate muscle activation patterns of the five Japanese vowels [29]. There existed drawbacks in this model and the upgrade of present model is to solve these drawbacks.

According to a large quantity of simulations we found that there are two main deficiencies in the previous model: 1) The muscle activation pattern was not the only determinant factor of the equilibrium position, which makes it difficult to control the model with a high degree of accuracy. 2) The fairly long response time from enrollment of muscle activation to the equilibrium position result in the difficulty when generating running speech based on the physiological articulatory model. 3) The individual intrinsic muscle was controlled as a control unit is not reasonable because different parts of the intrinsic muscles (transversus, verticalis and superior longitudinal) have different functions, division of these muscles into a smaller control unit will improve the degree of accuracy of model control. In order to conquer these deficiencies, the present model is improved in the following aspects: 1) The discrete FEM is substituted by continuum FEM. 2) Some intrinsic muscles are divided into smaller control units to realize accurate control.

## 2.2 Dynamic simulation

According to the analysis, the reason of the unstable of the physiological articulatory model was resulted from the algorithm of the discrete FE model (refer to the appendix of the paper [51]). Therefore, the problem is expected to be solved by implementing the continuum FE model. The present physiological articulatory model was extended to a continuum FE model using the ArtiSynth 3D Biomechanical Modeling Toolkit (www.artisynth.org, University of British Columbia, Vancouver, Canada). ArtiSynth improved a number of aspects of the physiological articulatory model, including volume constraint, computational efficiency, etc.

The ArtiSynth toolkit was used to generate dynamic simulations with the physiological articulatory model. In this section, we describe the pertinent equations and parameters used in this physiological articulatory model. A complete description of the dynamic simulation formulation in ArtiSynth has been published elsewhere (see Section 4 in [76]).

According to the Newton's second law, the equations of motion that govern the dynamic response of the finite element system are given by:

$$\boldsymbol{M}\dot{\boldsymbol{u}} = \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{u}, t), \tag{2.1}$$

where $t$ is time, $\boldsymbol{q}$ and $\boldsymbol{u}$ are the generalized positions and velocities of all dynamical components in the mechanical system, $\boldsymbol{f}(\boldsymbol{q}, \boldsymbol{u}, t)$ is all of the forces acting on the dynamic components, and $\boldsymbol{M}$ is the block-diagonal mass matrix.

The system dynamics are also constrained by bilateral and unilateral constraints. Bilateral constraints are used to attach the tongue to the jaw and hyoid bone, as well as to enforce FEM incompressibility in the FE models (through a mixed u-P formulation [77]). Bilateral constraints form an equality condition on the system velocity $\boldsymbol{u}$:

$$\boldsymbol{G}(\boldsymbol{q})\boldsymbol{u} = 0 \tag{2.2}$$

Unilateral constraints are used to handle contact between the tongue tip, the jaw and palate. Unilateral constraints form an inequality condition on the system velocity $\boldsymbol{u}$:

$$\boldsymbol{N}(\boldsymbol{q})\boldsymbol{u} \geq 0 \tag{2.3}$$

Bilateral and unilateral constraints generate reaction forces $\boldsymbol{G}^T\boldsymbol{\lambda}$ and $\boldsymbol{N}^T\boldsymbol{z}$ respectively, where $\boldsymbol{\lambda}$ and $\boldsymbol{z}$ are the Lagrange multipliers. These reaction forces add to the system forces in Eq. (2.1).

The dynamic equations are solved numerically using a semi-implicit second-order Newmark integrator [?]. The update rules for this integration scheme are:

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + \frac{h}{2}(\dot{\boldsymbol{u}}^k + \dot{\boldsymbol{u}}^{k+1}), \tag{2.4}$$

and

$$\boldsymbol{q}^{k+1} = \boldsymbol{q}^k + \frac{h}{2}(\boldsymbol{u}^k + \boldsymbol{u}^{k+1}), \tag{2.5}$$

where $h$ is the time step.

Solving the equations of motion requires integrating Eq. (2.1) with the update steps given in Eqs. (2.4) and (2.5) subject to the conditions given in Eqs. (2.2) and (2.3). This requires solving the following mixed linear complementarity problem:

$$\begin{pmatrix} \hat{\boldsymbol{M}}^k & -\boldsymbol{G}^{kT} & -\boldsymbol{N}^{kT} \\ \boldsymbol{G}^k & 0 & 0 \\ \boldsymbol{N}^k & 0 & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{u}^{k+1} \\ \boldsymbol{\lambda} \\ \boldsymbol{z} \end{pmatrix} + \begin{pmatrix} -\boldsymbol{b} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \boldsymbol{w} \end{pmatrix},$$

$$0 \leq \boldsymbol{z} \perp \boldsymbol{w} \geq 0, \tag{2.6}$$

where $\boldsymbol{b} \equiv \boldsymbol{M}\boldsymbol{u}^k + h\hat{\boldsymbol{f}}^k$, $\boldsymbol{w}$ is the slack variable in the complementarity condition, and $\hat{\boldsymbol{M}}$ and $\hat{\boldsymbol{f}}$ are the mass matrix and force vector augmented with Jacobian terms due to the implicit integration scheme (see [76] for a full derivation). The complementarity condition, $0 \leq \boldsymbol{z} \perp \boldsymbol{w} \geq 0$, ensures that the unilateral constraint forces are non-zero only when those constraints are active, *i.e.* $\boldsymbol{z}$ is positive if and only if $\boldsymbol{w}$ is zero and *vice versa.*

## 2.3 Morphological and anatomical structures of the model

### 2.3.1 Morphological structures

The initial shape of the tongue was obtained based on the volumetric MR images while producing the Japanese vowel /e/ that is close to the neural position in the vowel space. The jaw and vocal tract wall were superimposed with the images of the lower and upper teeth at the interval of 0.4 cm in the transverse dimension. The mesh structure of the tongue in the lateral 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. In the front view, the tongue was divided into 5 layers with equal intervals. Totally, the tongue tissue consists of 240 hexahedrons. For more details of morphological data and mesh segmentation of the tongue tissue, refer to the previous studies [29, 46].

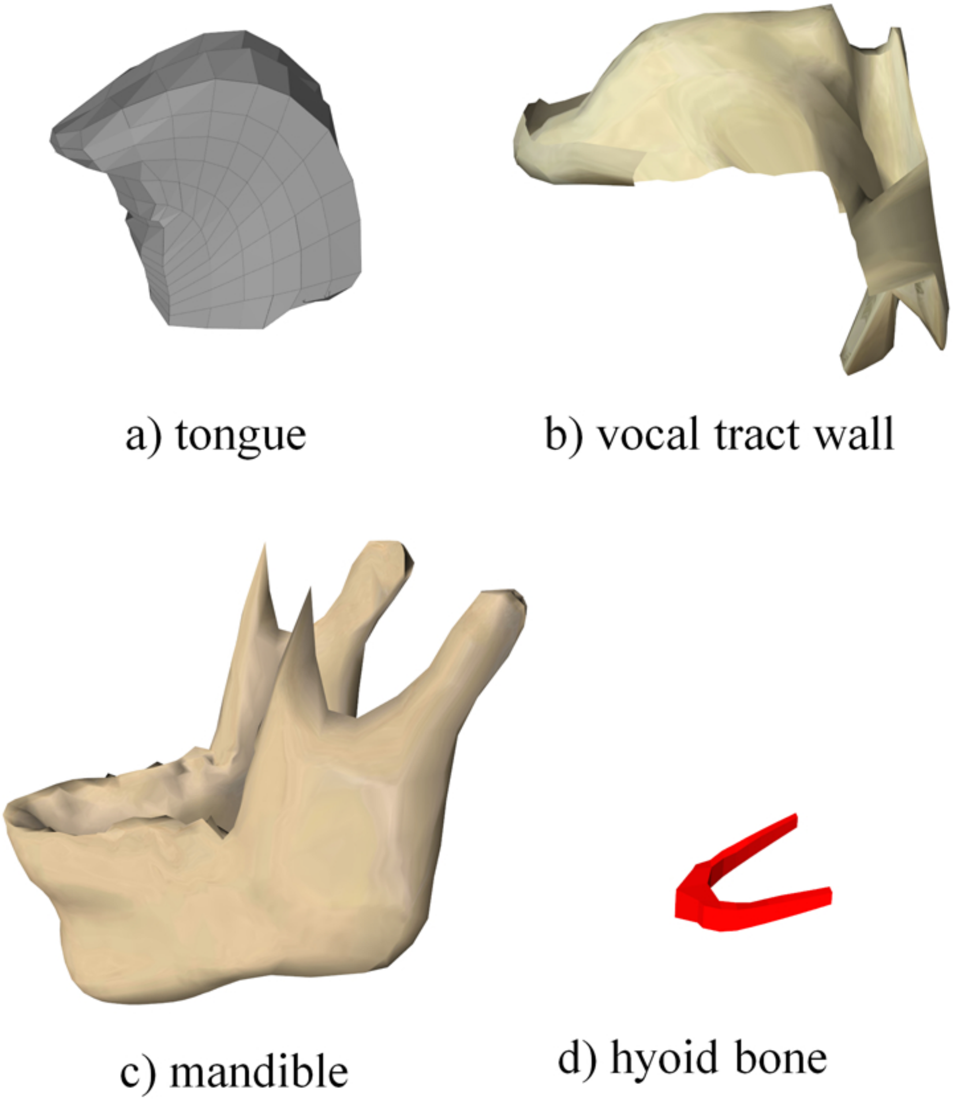

Figure 2.1: Geometrical structures of the tongue (upper left), vocal tract wall (upper right), mandible (lower left) and hyoid bone (lower right)

The geometrical structure of the tongue, vocal tract wall, mandible and hyoid hone are shown in Figure 2.1. Different from Fang *et al.*'s study [29], the geometrical structure of the whole mandible was re-extracted from the MR image to include the upper part of the mandible (condyle, coronoid process and Ramus).

### 2.3.2 Arrangements of tongue muscles

Since the movements of the model is driven by activation of muscles, accuracy of muscle arrangements is very important for model control. Firstly, the original arrangement of muscles based on their anatomical partitions is introduced in this section (tongue) and next section (jaw). Secondly, in order to elaborate model control, some intrinsic tongue muscles are divided into smaller control units, and jaw muscles are combined into groups, according to articulation purpose, which will be introduced in Section 2.3.4.

Three extrinsic muscles, genioglossus, styloglossus, and hyoglossus, are arranged mainly based on the high-resolution MRI analysis [55]. The intrinsic muscles (superior longitudinal, inferior longitudinal, transverse, and vertical) are defined according to the anatomical data [22]. The tongue floor muscles, mylohyoid and geniohyoid, are arranged referring to the anatomical literature [23].

The tongue muscles implemented in the physiological articulatory model include extrinsic tongue muscles (genioglossus, hyoglossus, styloglossus, geniohyoid and mylohyoid) and intrinsic tongue muscles (verticalis, transversus, superior longitudinal and inferior longitudinal). The anatomical structures of the extrinsic tongue muscles and intrinsic tongue muscles are shown in Figure 2.2 and Figure 2.3, respectively. The descriptions of each muscle course are given in Table 2.1, which is mainly cited from an anatomical book [56].

Tongue muscles in the physiological articulatory model are arranged based on the anatomical knowledge of muscle initiation, course and termination. The arrangement of the tongue muscles in this model are shown in Figure 2.4. Because, as has been described previously, the tongue has been segmented into 240 meshes and the muscles are defined by connecting the nodes of the meshes or fixed points of skeletal body, the muscle course is not the exactly the same as humans. Increasing the number of FEM mesh will make the model more realistic, however, the computational cost would dramatically increase with it. The FEM mesh number implemented in this model is a trade-off between computational cost and model verisimilitude.

### 2.3.3 Arrangements of jaw muscles

The muscles that control jaw movements are arranged in the same way as that in our previous partial 3D model [51]. In Figure 2.5, the arrangements of muscles used to control

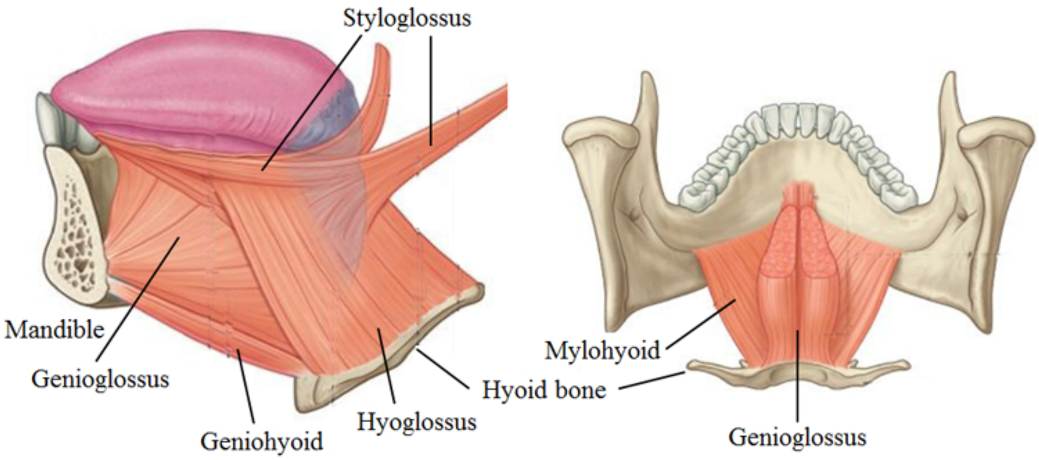

Figure 2.2: Lateral view (left) and posterior view (right) of extrinsic tongue muscles cited from Elsevier Drake *et al.*, (2010) [106] adapted with permission.

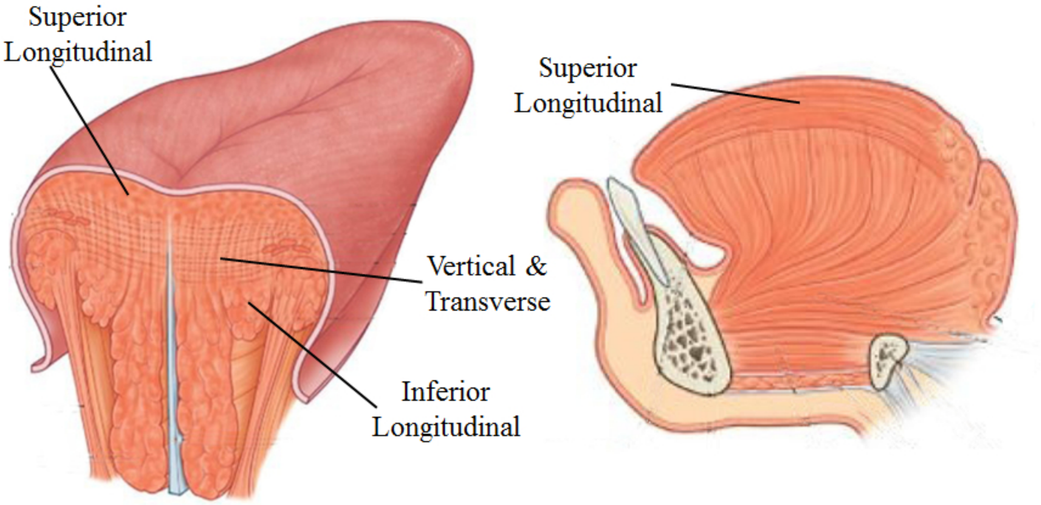

Figure 2.3: posterior view (left) and sagittal cross-section view (right) of intrinsic tongue muscles cited from Elsevier Drake *et al.*, (2010) [106] adapted with permission.

Table 2.1: Anatomical description of muscle courses and muscle function [56].

| Muscle name | Course description | Main function |
|---|---|---|
| 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. | Anterior fibers retract tongue posterior fibers protrude tongue; together, anterior and posterior fibers depress tongue |
| Hyoglossus | Arises from the length of the greater cornu and the lateral body of the hyoid bone, coursing upward to insert into the sides of the tongue between the styloglossus and inferior longitudinal muscles. | Pulls sides of tongue down |
| Styloglossus | Styloglossus originates from the anterolateral margin of the styloid process of the temporal bone, coursing forward and down to insert into inferior sides of the tongue. It divides into two portions: one interdigitates with the inferior longitudinal muscle and the other with the fibers of the hyoglossus. | Draws the tongue back and up |
| Geniohyoid | Originates at the mental spines of the mandible, and projects parallel to the anterior digastricus from the inner mandibular surface to insert into the corpus hyoid. | Depresses the mandible |
| Mylohyoid | originates on the underside of the mandible and courses to the corpus of hyoid. This fanlike muscle courses from the mylohyoid line of the mandible to the median fibrous raphe and inferiorly to the hyoid forming the floor of the mouth. | Depresses the mandible |
| Superior longitudinal | courses along the length of the tongue, comprising the upper layer of the tongue. This muscle originates from the fibrous submucous layer near the epiglottis, the hyoid, and from the fibrous submucous layer near the epiglottis, the hyoid, and from the median fibrous septum. Its fibers fan forward and outward to insert into the lateral margins of the tongue and region of the apex. | Elevates, assists in retraction, or deviates tip of tongue |
| Inferior longitudinal | originates at the root of the tongue and corpus hyoid, with fibers coursing to the apex of the tongue. This muscle occupies the lower sides of the tongue, but is absent in the medial tongue base, which is occupied by the extrinsic genioglossus muscle. | Pulls tip of tongue downward, assists in retraction, deviates tongue |
| Transverse | Originate at the median fibrous septum and course laterally to insert into the side of the tongue in the submucous tissue | Provide a mechanism for narrowing the tongue |
| Vertical | Run at right angles to the transverse muscles and flatten the tongue. Fibers of the vertical muscle course from the base of the tongue and insert into the membranous cover | Pull tongue down into the floor of the mouth |

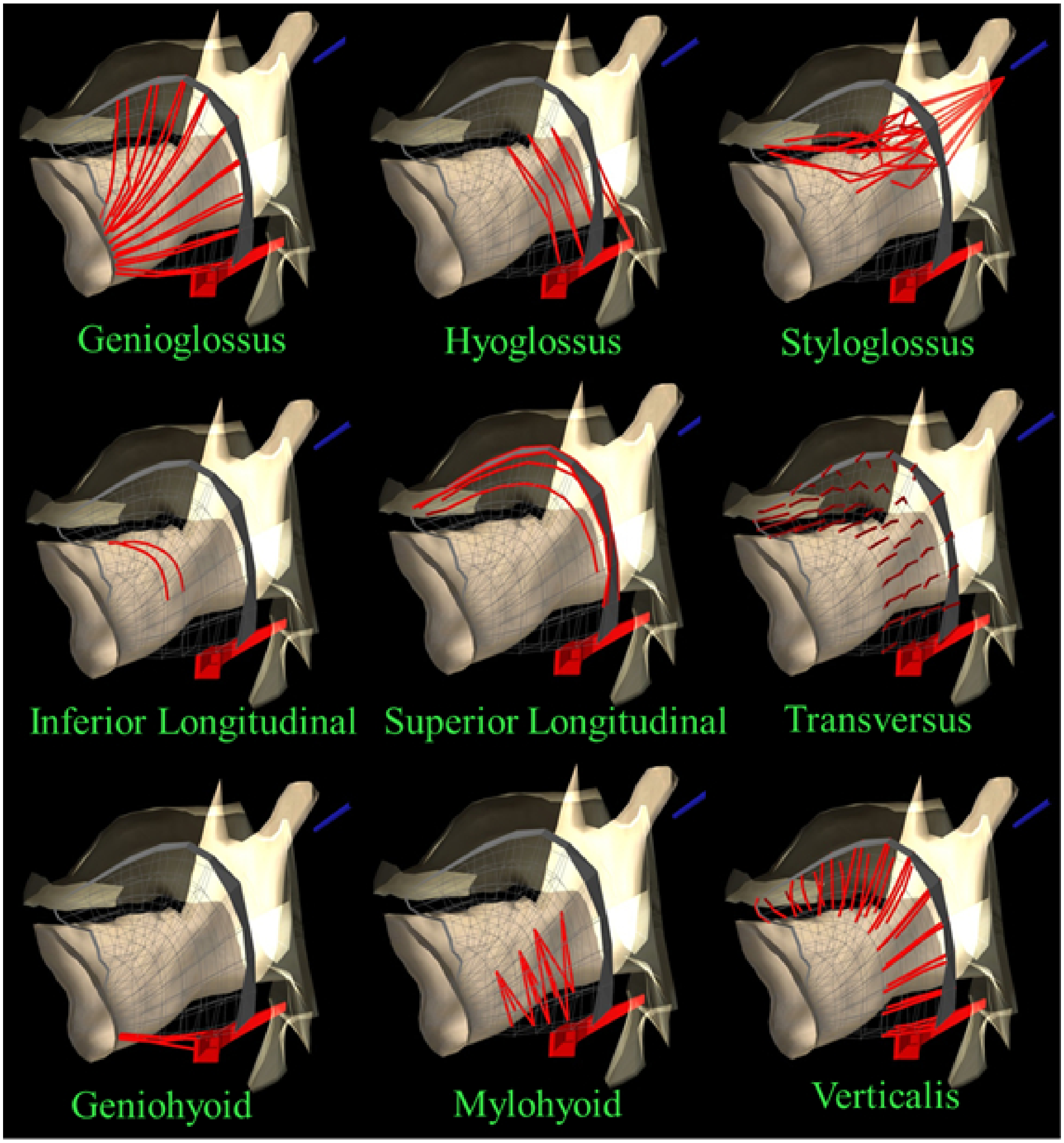

Figure 2.4: Arrangement of tongue muscles in the physiological articulatory model

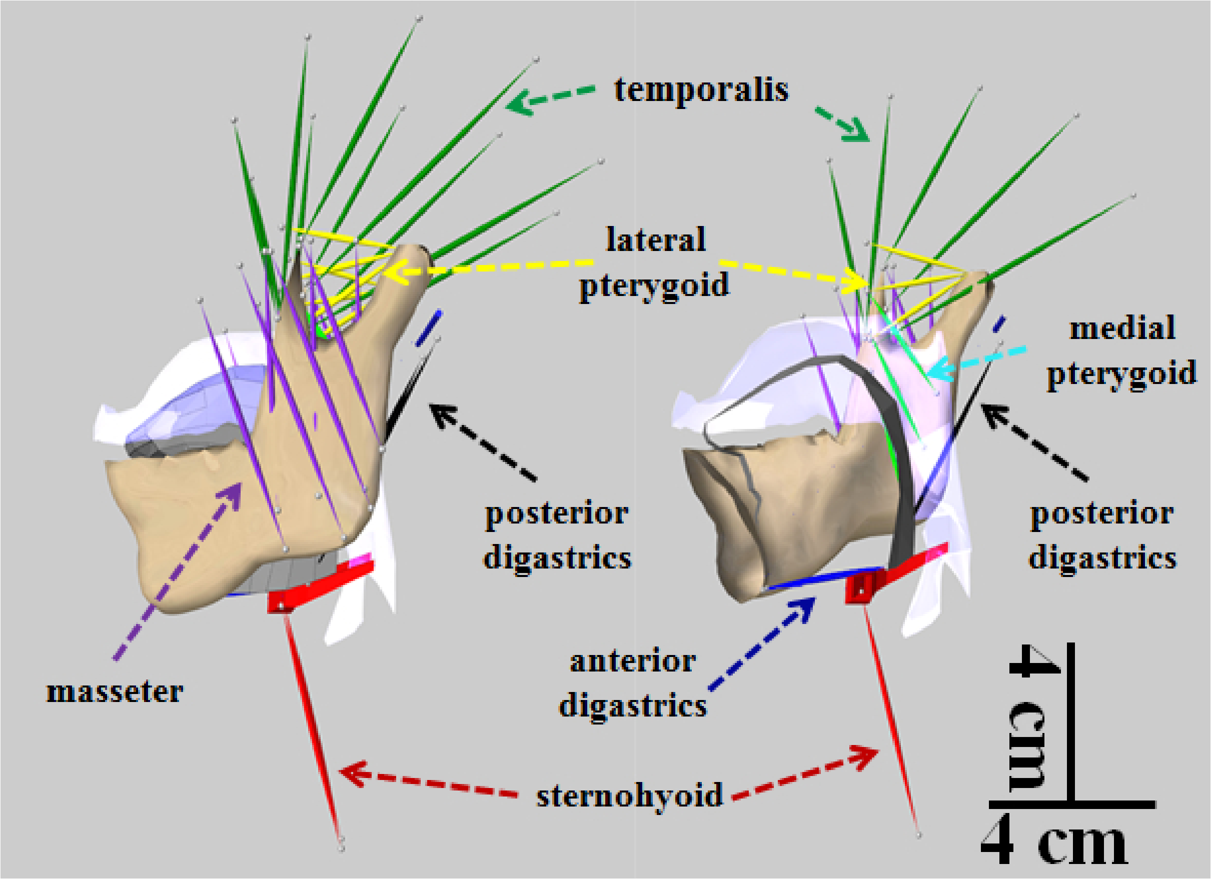

Figure 2.5: Arrangements of muscles in the jaw model

the jaw in the physiological articulatory model are described. The muscles used to control the translation and rotation of the jaw were classified into two muscle groups, the jaw opener (JO) and jaw closer (JC). According to the description of the muscles used to control the jaw [57], the jaw opening muscles include the anterior digastrics, posterior digastrics and lateral pterygoid muscles. The strap muscles such as the sternohyoid also assist jaw opening. The main function of the lateral pterygoid was to move the jaw forward, but the current version of the jaw model only permits hinge-like jaw opening. Therefore, in this study, JO group consists of anterior digastrics, posterior digastrics and sternohyoid. The anatomical description of jaw opener muscle is shown in Figure 2.6. When JO was activated, the muscles in the group active with the same activation level.

The jaw closing muscles include temporalis, masseter and medial pterygoid muscles. Among those muscles, comparatively small muscles are used for speech articulation, while the larger muscles play major roles in biting and chewing [57]. The anatomical description of jaw closer muscles is shown in Figure 2.7. The medial pterygoid plays the main role in speech production, while the temporalis and masseter contribute mainly in the less. According to our simulation, the activation level for the temporalis and masseter were

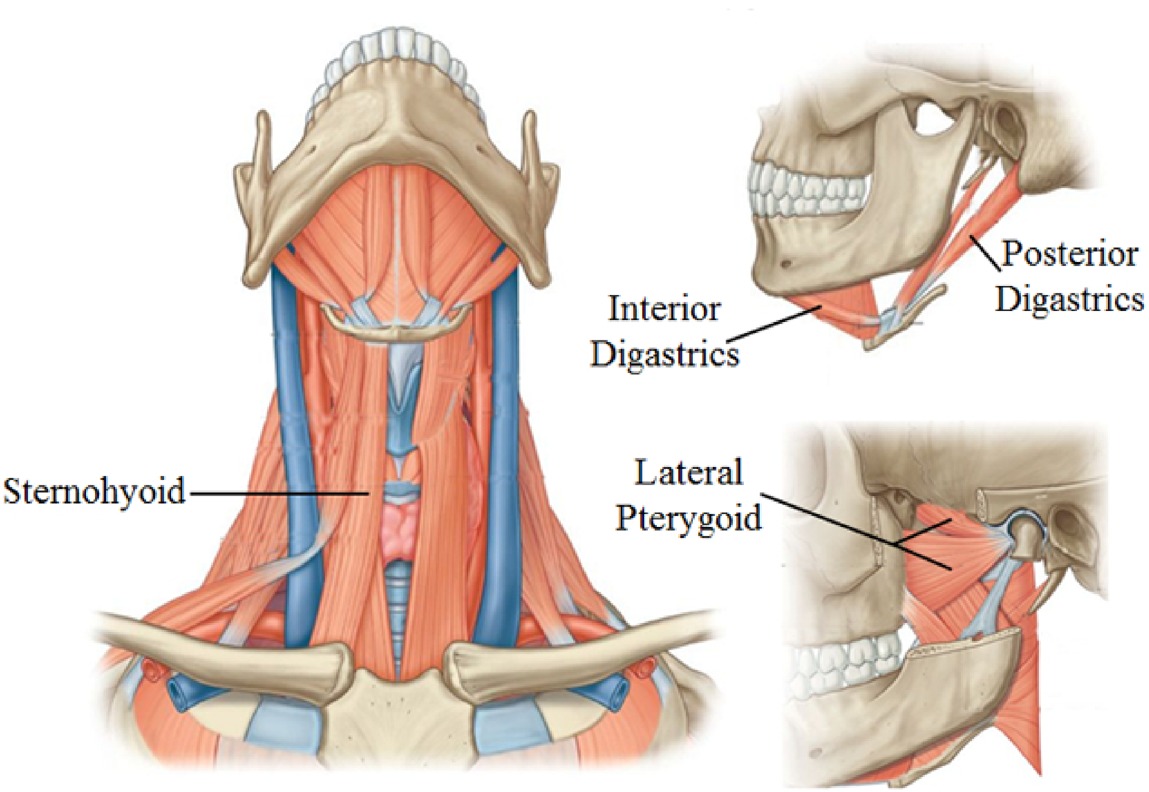

Figure 2.6: Anatomic description of jaw opener muscle, jaw opener muscle include Interior Digastrics, Posterior Digastrics, Lateral Pterygoid, and assistant by strap muscle Sternohyoid, Cited from Elsevier (2010), Drake *et al.* [106] adapted with permission.

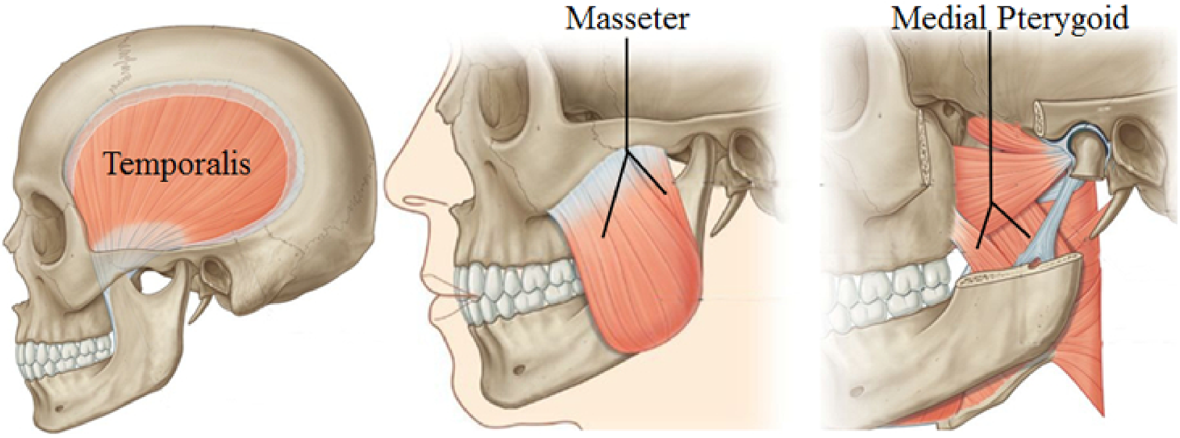

Figure 2.7: Anatomic description of jaw closer muscle, jaw closer muscle include Temporalis, Masseter, and Medial Pterygoid, Cited from Elsevier (2010), Drake *et al.* [106] adapted with permission.

the fourth and fifth of that for the medial pterygoid, respectively.

### 2.3.4 Control units of the model

Muscles in the model were arranged on the basis of their anatomical partitions where different parts of the same muscle may have different functions. In order to simulate fine-grained tongue movements with the model, the muscles were divided into a number of smaller control units according to articulation purposes. Figure 2.4 illustrates the layout of the extrinsic and intrinsic muscles (original or divided) in the 3D physiological articulatory model in a sagittal cut-away view. The genioglossus muscle was divided into three units: anterior (GGa), middle (GGm) and posterior (GGp). This division conforms to previous physiological articulatory models [29, 51]. Different from the previous studies [29, 51], the intrinsic muscles were also divided into several control units according to their functions. The vertical and transverse muscles were functionally divided into three units: anterior (Va, Ta), middle (Vm, Tm) and posterior (Vp, Tp). The superior longitudinal was divided into two units: anterior (SLa) and posterior (SLp). The styloglossus (SG), mylohyoid (MH), geniohyoid (GH) and inferior longitudinal (IL) were controlled as independent units. Altogether, there are 18 muscle control units including the tongue muscles, jaw opener and closer.

The profile of the constructed model is shown in Figure 2.9, where the appearance of the model is shown in the left panel and a sagittal cut-away view is shown in the right panel. The larynx complex is also included in the model, it is not investigated in this study.

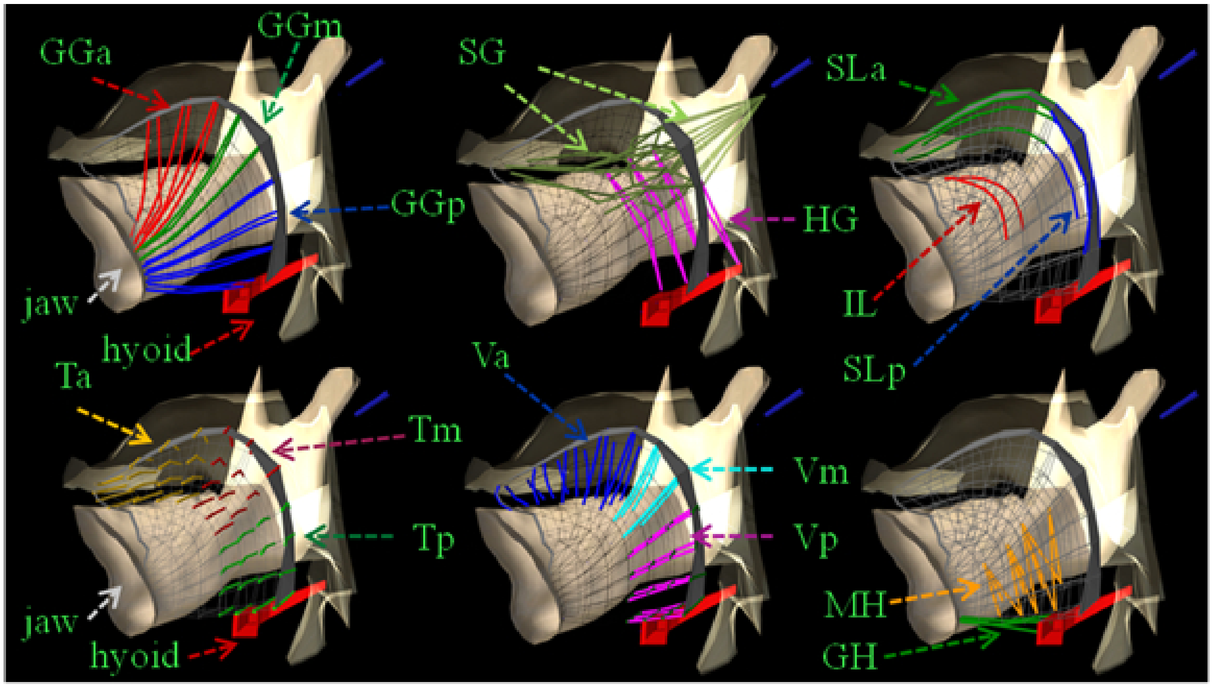

Figure 2.8: Control units of the tongue muscles. GGa, GGm and GGp (anterior, middle and posterior portion of genioglossus muscle, respectively); HG (hyoglossus muscle); SG (styloglossus muscle); SLa and SLp (the anterior and posterior portion of superior longitudinal muscle); IL(inferior longitudinal muscle); Va, Vm and Vp (anterior, middle, and posterior portion of vertical muscle, respectively). Ta, Tm and Tp (anterior, middle and posterior portion of transverse muscle, respectively); MH (mylohyoid muscle); GH (geniohyoid muscle).

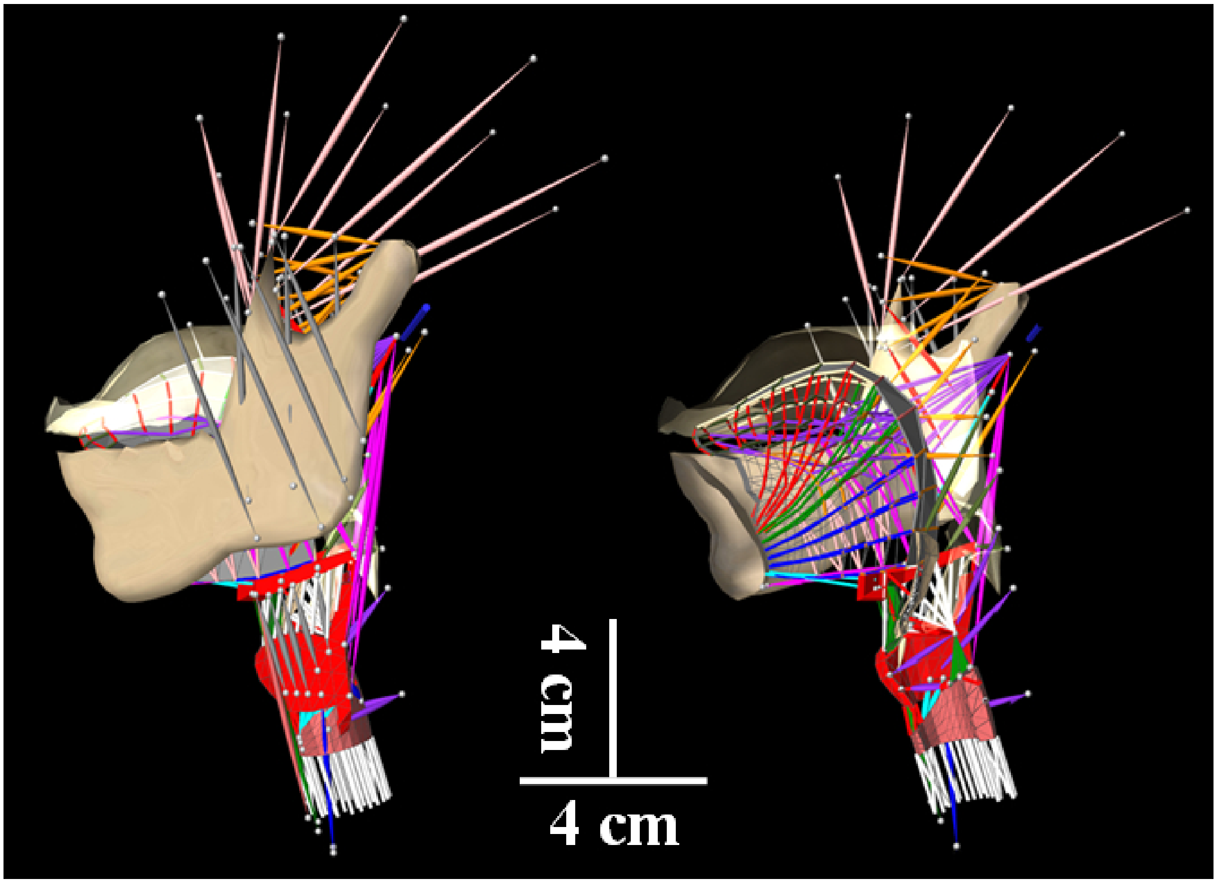

Figure 2.9: Profiles of the constructed physiological articulatory model

## 2.4 Muscle Model

The muscle model implemented in this study was proposed by Morecki [58] which is an extended model of Hill's model [60]. Dang and Honda modified the algorithm of the model to make it more appropriate for computation [51]. Forces generated by muscles include two components: active muscle force which depends on muscle activation and passive muscle force which is independent of muscle activation. Figure 2.10 (a) shows the general description of the muscle model. From this figure, one can see that there are three parts, part 1 is a nonlinear spring $k_1$, which is involved in generating force only when the current length of muscle sarcomere is longer than its original length.

Part 2 consists of a Maxwell body and is always involved in force generation. 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 [54], the force force–velocity characteristic of the muscle is treated as independent of the previous force. 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 generate force as a muscle is activated. The relationship between the stretch ratio of the muscle sarcomere and the generated force including the passive force is shown in Figure 2.10 (b). For details of the equations and parameters used in this model, refer to Dang and Honda [51]. Muscle force was used as a control unit to control their physiological articulatory model. However, the muscle force resulted from muscle length, muscle length changing rate, etc. In present model, the control variable is substituted by muscle activation that ranging from 0 to 1, where 0 means no muscle activation and 1 means that the muscle is fully activated and generates maximum force. To realize it, the parameters and equations is modified more or less.

In model computations, active stress of the muscle sarcomere was generated using force-length function which was derived by matching the simulation and empirical data using the least-square method [58]. In this function, shown in Eq. (2.7), the active stress ($\sigma_{act}$) was calculated using a fourth-order polynomial of the stretch ratio of the muscles, which had a similar shape to that used by Wilhelms-Tricarico [59]. In Eq.(2.7), the muscle length change rate $\varepsilon = (l - l_0)/l_0$ was valid for the range of $-0.185 < \varepsilon < 0.49$, where $l$ and $l_0$ were present muscle length and original muscle length, respectively. Therefore, the active force was set to 0 if $\varepsilon$ was out of the given range.

$$\sigma_{act} = 1.161\varepsilon^4 + 0.243\varepsilon^3 - 1.376\varepsilon^2 + 0.235\varepsilon + 0.164 \tag{2.7}$$

The ability to generate muscle force varies from muscle fiber to muscle fiber depending on their thickness. Therefore, the parameter "thickness" of the muscle fiber was introduced as a coefficient for all the muscles and the thickness decides the capacity of force

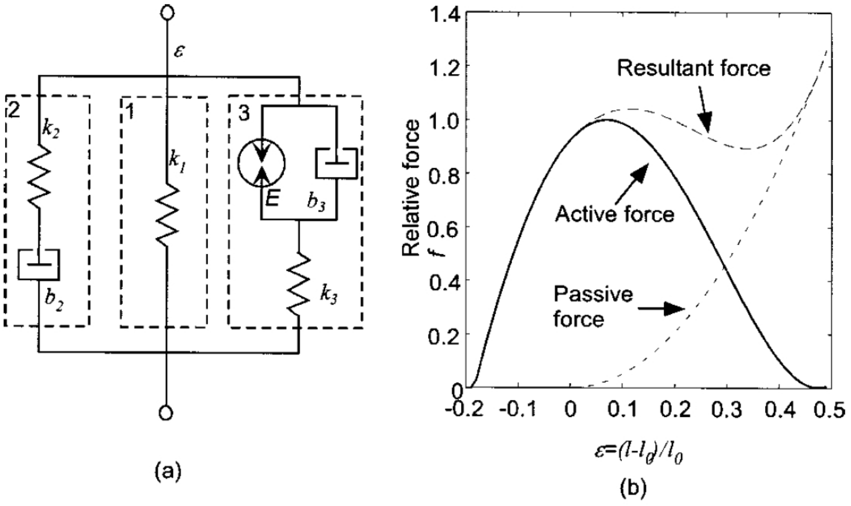

Figure 2.10: Muscle modeling: (a) a general model of muscle unit: $k$ and $b$ are stiffness and dashpot, $E$ is the contractile element; (b) generated force varies with stretch ratio $\varepsilon$ (After Dang and Honda [51]).

generating. The thickness of individual muscles was determined by making the maximum force ($F_{max}$) of the muscles consistent with empirical data [54, 78]. The control variable of individual muscle activation, $a$, was normalized within the interval [0, 1], where 0 means no muscle activation and 1 means that the muscle is fully activated and generates maximum force. The activated muscle force was calculated as:

$$F_{act} = F_{max}\sigma_{act}a, \tag{2.8}$$

where $F_{max}$ is the maximum isometric force capacity of the muscle, $\sigma_{act}$ is the active muscle stress (see Eq. (2.7)), and $a$ is the muscle activation.

Passive muscle force is generated by passively lengthening of muscle. According to common sense, if a muscle is lengthened equal or longer than a threshold $l_{p\_max}$ the passive muscle force will no longer continue increasing with its lengthening and reach the maximum passive force ($F_{p\_max}$) that the muscle can generate. In present physiological articulatory model, $l_{p\_max}$ was set to 1.25 times of original muscle length and $F_{p\_max}$ was related to $F_{max}$ by $F_{p\_max} = 0.015F_{max}$ according to Morecki's muscle model[58]. The passive muscle force was described by:

$$F_{pas} = \begin{cases} 0 & \text{if } l < l_0 \\ F_{p\_max}[(l - l_0)/(l_{p\_max} - l_0)] & \text{if } l_0 < l < l_{p\_max} \\ F_{p\_max} & \text{if } l \geq l_{p\_max} \end{cases} \tag{2.9}$$

The final muscle force was the sum of active muscle force $F_{act}$ and passive muscle force $F_{pas}$.

## 2.5 Mechanics properties

In the present physiological articulatory model, Rayleigh damping was implemented, which took the form:

$$\boldsymbol{D}_F = \alpha \boldsymbol{M}_F + \beta \boldsymbol{K}_F, \tag{2.10}$$

where $\boldsymbol{M}_F$ was the portion of the mass matrix associated with the FEM nodes and $\boldsymbol{K}_F$ was the FEM stiffness matrix. $\boldsymbol{D}_F$ was embedded into the overall system Eq.(2.6) by $\boldsymbol{D_F} = \partial \boldsymbol{f} / \partial \boldsymbol{u}$ . In present model, $\alpha$ and $\beta$ were set to $40\,\mathrm{s}^{-1}$ and $0.03\,\mathrm{s}$, respectively, in order to have a damping close to the critical one in the range of modal frequency from 3 to 10 Hz [30]. For details on how to integrate the Rayleigh damping into Eq. (2.6), refer to paper [79].

A Poisson coefficient of the tongue tissue was set to 0.49, since it was considered to be quasi-incompressible. Density of tongue tissue was set to be $1040\,\mathrm{kg\,m}^{-3}$, and density of mandible and hyoid bone was set as $2000\,\mathrm{kg\,m}^{-3}$. The Young's modulus of the tongue tissue was set to 20 kPa and the bone structures (mandible and hyoid) were approximated as rigid bodies. These parameters were consistent to the previous model [51].

### 2.5.1 Individual muscle function

Activity of each muscle contributes to local deformation or displacement of the articulatory organs. The investigation of speech production mechanism relies on the function of individual muscle unit. For this reason, we investigate the functions of the muscle units, individually. In the simulations, each muscle was activated individually for the duration of 200 ms, which was sufficient for the model to reach its equilibrium position. Figure 2.11 shows the functions of individual muscle units on the midsagittal plane. The functions of the extrinsic and intrinsic muscles were qualitatively assessed based on anatomical description. These assessments show that the role of individual muscles in our model is consistent with anatomical knowledge [21, 28, 24]. By referring to the description of individual muscle function in Table. 2.1, the individual muscle function in the physiological articulatory model is reasonable. The difference of muscle control unit between present model and previous ones [29, 51] is that in the present model some intrinsic individual muscles are divided into smaller control units according to their functions. From Figure 2.11, one can see that different portions of the vertical muscles have different functions, refer to the functions of Va, Vm and Vp. Similarly, the control unit Ta, Tm and Tp have different functions although they belong to the same muscle (transverse muscle). These implied that the divisions of the muscle units were effective.

There is a common feature for FEM-based physiological articulatory models: when a muscle activation pattern is maintained, the model reaches a certain equilibrium position.

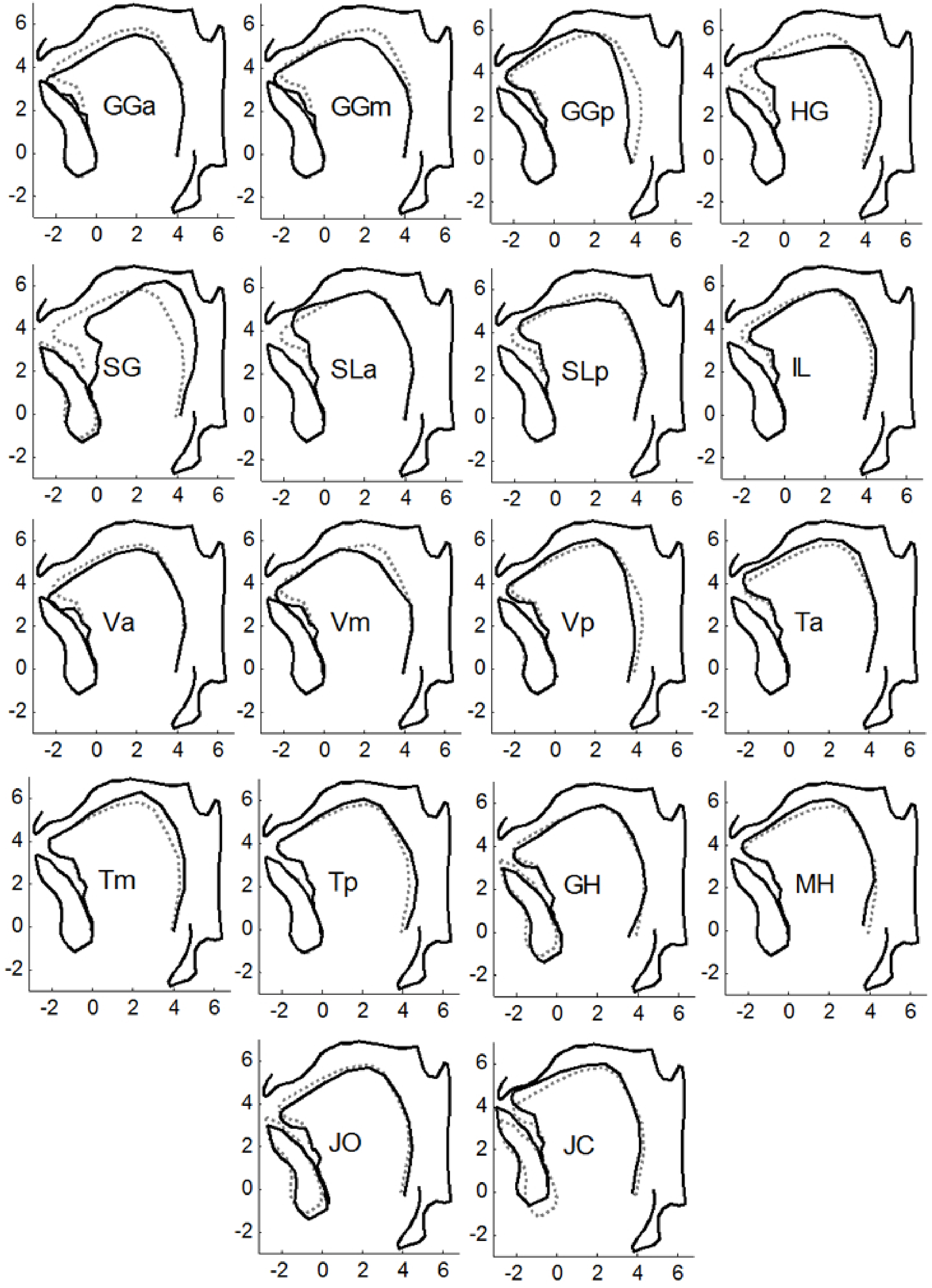

Figure 2.11: Function of individual muscles in the physiological articulatory model. Black solid lines show the equilibrium position after the muscle is activated for 200 ms duration, dotted gray lines correspond to the shape in its rest position. (Unit: cm)

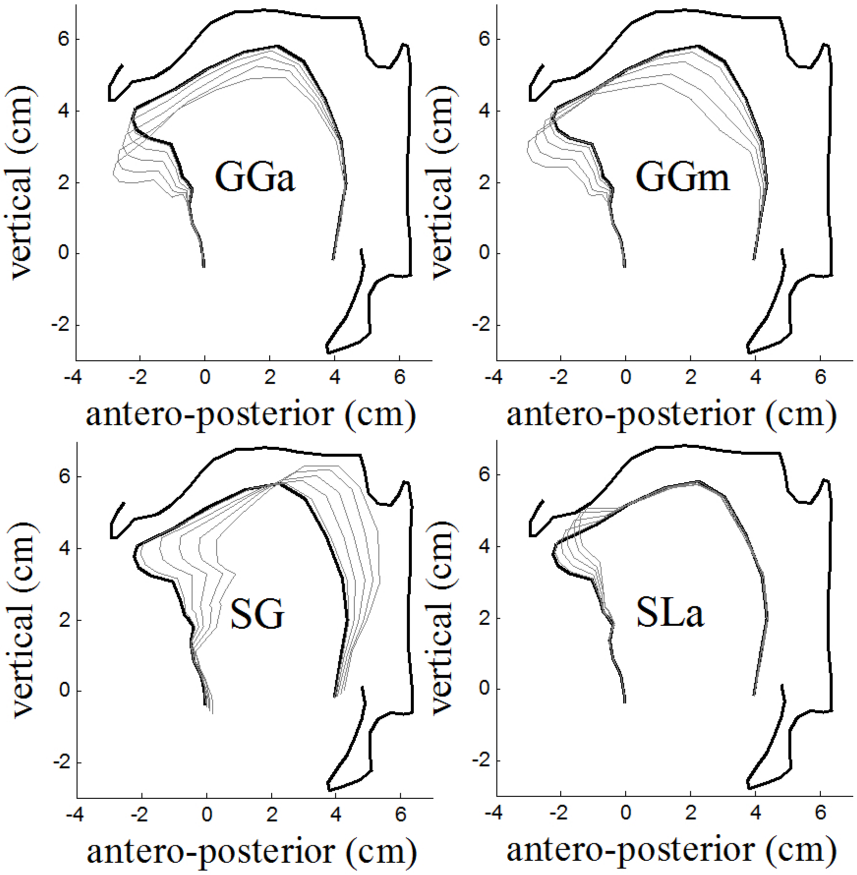

Figure 2.12: Muscle activation and equilibrium position. Black thick lines are the rest position of the tongue; gray lines are the equilibrium position driven by different activation levels [0.002, 0.01, 0.03, 0.1, 0.4].

This equilibrium position is determined only by muscle activation itself, no matter where its initial position is. Figure 2.12 shows the changes of the equilibrium position when activation level is changed. In this figure, GGa, GGm, SG and SLa are active with the activation level [0.002, 0.01, 0.03, 0.1, 0.4], respectively, after 200 ms, the model reaches its equilibrium position. From this figure, one can see that a muscle activation level determines a unique equilibrium position.

Previous model simulations have shown that the relationship between muscle activation level and displacement of the tongue was quasi-logarithmic [51]. To generate displacements with about the same increment, muscle activation was discretized into 11 levels of [0, 0.002, 0.005, 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.4, 0.8]. Section 4.1 provides details on the chosen activation levels.

### 2.5.2 Generation of speech sounds

The acoustic waves in the vocal tract can be regarded as the plane wave in the low frequency region, and the wavefront would be in parallel with the cross-section along the longitudinal midline of the vocal tract. Sound can be generated using a transmission line model with the area function that is made of the cross-section areas of the vocal tract. The area function is time-variant during articulation.

In the previous partial 3D model [51], the area function is approximated from the vocal tract width of the sagittal planes, using an $\alpha$-$\beta$ model [62] [63]. With the 3D physiological articulatory model, it is possible to get the shapes of the cross-sections from the model, and calculate the area function directly and more accurately from the cross-section shapes.

For calculating the area function of the vocal tract, a set of grid planes are prepared based on the configuration of the vocal tract at a given time. The planes are perpendicular to the direction of the sound path in the vocal tract, and have about equal interval from each other along the vocal tract. The cross-sections can be calculated as the intersections of the set of planes and the articulatory organs. The cross-section areas can then be calculated from the cross-sections.

Figure 2.13 shows the articulatory organ configurations when producing the sound /a/ with three grid planes. For calculating the area function in this configuration, a set of 76 planes is used, among which 3 planes are shown in the figure They are located at the larynx, the velum region, and the front cavity. The cross-sections at these three planes are shown in Figure 2.14, in which the blue lines represent the contour of the hard palate, pharyngeal wall and the larynx, the magenta lines represent the contour of the tongue, the red lines represent the contour of the jaw, and the black dashed line represent the boundary of the vocal tract. The corresponding area function is shown in Figure 2.15. The area function is applied to the transmission model to synthesize speech sound.

## 2.6 Evaluation of present model

As we have described previously, that the discrete FEM is substituted by continuum FEM, the performance of the present model was compared to the previous model to verify whether current physiological articulatory model is better than previous model. Two important parameters are implemented in the evaluation, which are convergence and response time.

When the model is control by a muscle activation pattern, the model moves to a specific position and keep equilibrium. The same muscle activation pattern should drive the model to the same equilibrium position no matter where the initial position is. That is to say, the equilibrium position depends on the muscle activation and independent

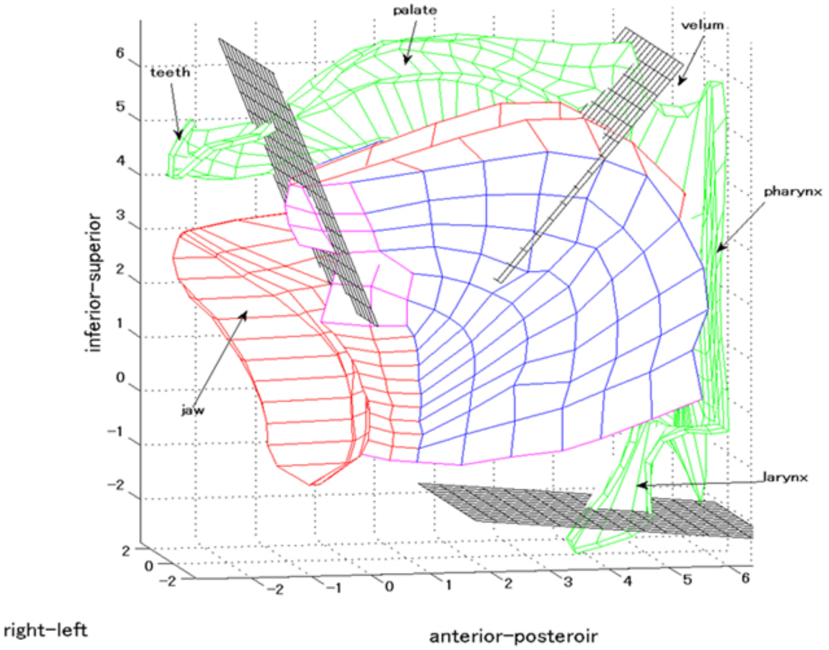

Figure 2.13: Vocal tract configuration with the tongue, the right half of the jaw, palate, pharyngeal wall and larynx for the vowel /a/.

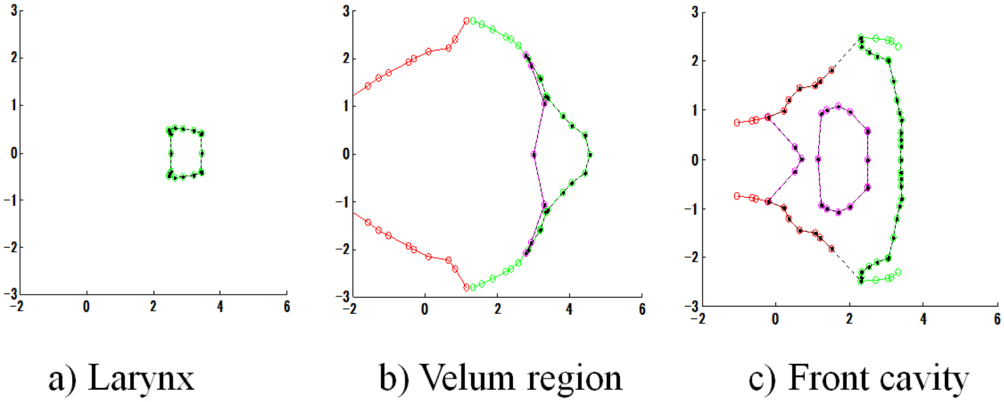

Figure 2.14: Cross-sections sliced by the planes in Figure 2.13, showing the larynx (left), velum region (middle), and front cavity (right).

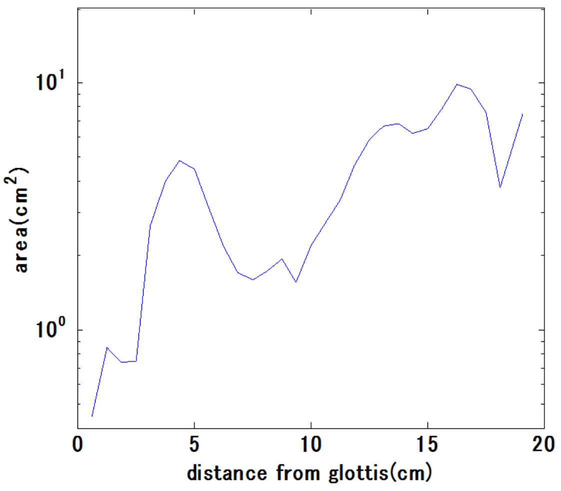

Figure 2.15: An example of area function of vowel /a/ generated by model simulation.

from the initial position. To test this property, first the model was driven to different initial positions by initial muscle activation patterns. Second, the initial muscle activation pattern was released, and meanwhile a new muscle is activated with specific activation level. Finally, we evaluate whether the equilibrium position converge to a small region or not. Figure 2.16 shows the results. From this figure, one can see that in discrete FEM model, the equilibrium positions converge to a region, but in continuum FEM model, the equilibrium positions converge to the smaller region, for both the tongue tip and tongue dorsum. This result indicates that the continuum FEM model is better than discrete FEM in convergence property.

The response time is defined as the duration from muscle activation to the model's reaching its equilibrium position. If the activation of the muscle cannot control the model to its equilibrium position within a designated duration, which will result in emergence of the carry over effect. If the carry over coarticulation effect happens, to control the model to generate continuous speech will be much more complicated. In a fairly long duration the before model really arrives its equilibrium position, the velocity is very small. So how to define the response time turns to be a technique problem. In this evaluation, a parameter is set as the duration from the moment of muscle activation to the model's reaching 80% of its equilibrium position to roughly represent the response time. Figure 2.17 shows the comparison results of the response time between the discrete FEM and continuum FEM. From this figure, one can see that the response time of the continuum FEM model is less than the discrete FEM model. The average response time for the continuum FEM and discrete FEM is 75 ms and 115 ms, respectively. According to the study of vowel duration in running speech [64], in a fast speaking rate, the average vowel duration is less than 100ms. According to this evidence, it is difficult to control the discrete model to generate

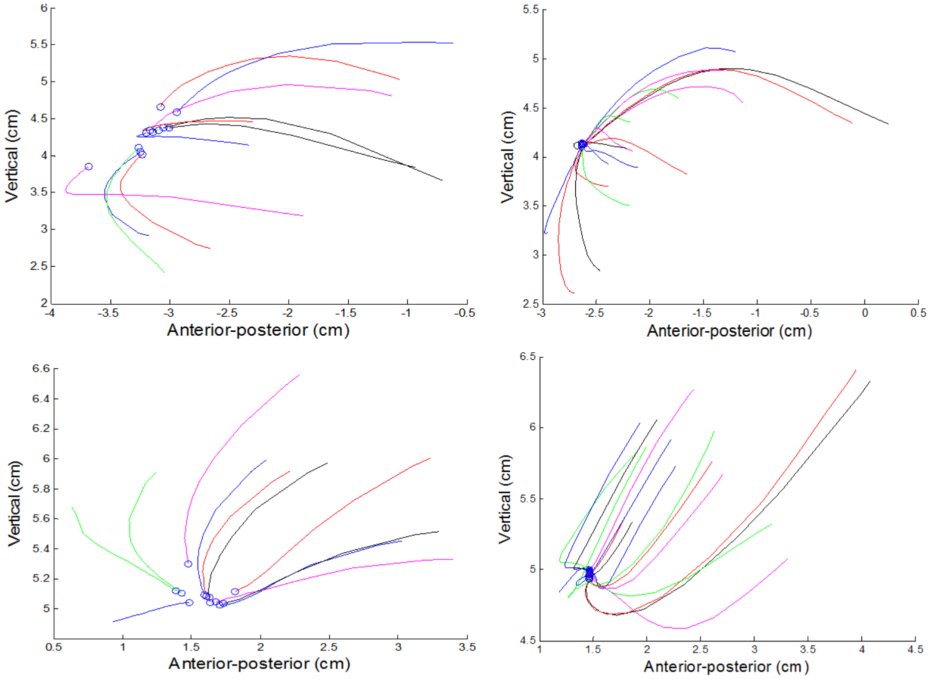

Figure 2.16: Equilibrium position by the same muscle activation from different initial positions. Upper left (discrete FEM) and upper right (Continuum FEM) show the equilibrium position of tongue tip when GGp activated with 3.5 Newton. Lower left (discrete FEM) and lower right (Continuum FEM) show the equilibrium positions of tongue dorsum when GGa activated with 3.5 Newton.

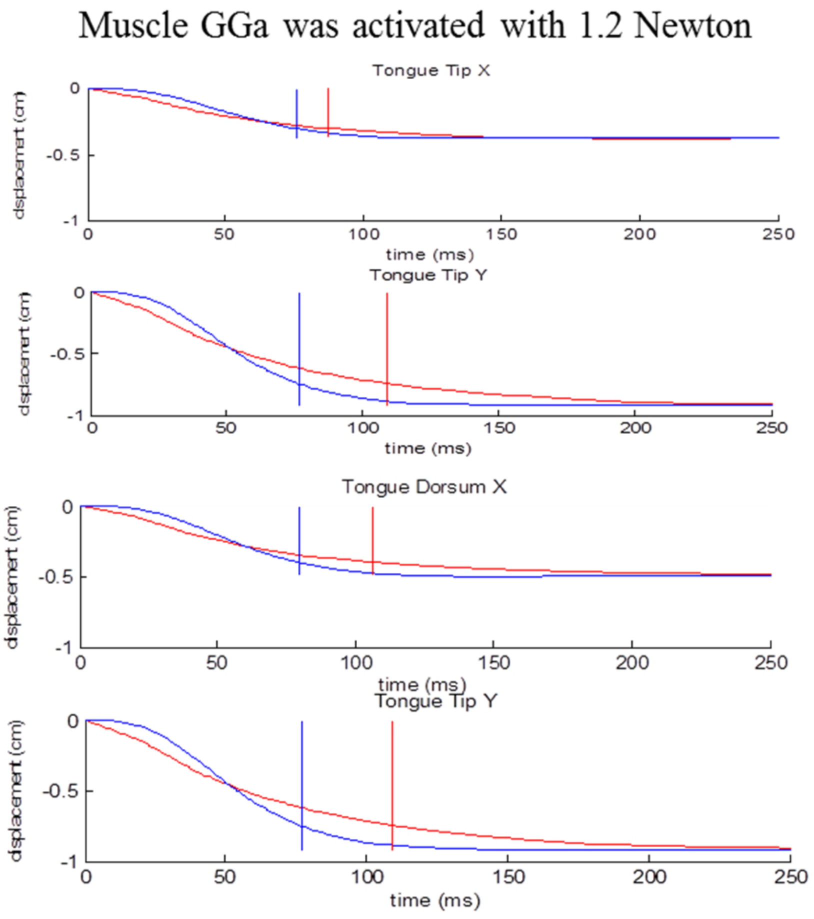

Figure 2.17: Comparison of the response time. Blue lines and red lines show the result of continuum model and discrete model, respectively.

continuous speech with fast speaking rate, however, the continuum FEM can satisfy the requirement. Obviously, it is easier to generate speech in a fast speaking rate using the continuum FEM than using the discrete FEM.

## 2.7 Summary

In this chapter, we first introduce how the present physiological was gradually developed based on the previous studies. Then the two main deficiencies of our previous model were indicated. To overcome these deficiencies existed in the previous model, the discrete FE model was substituted by a continuum FE model. According to the evaluation in the two parameters, response time and convergence, the current model has better performance than the previous model. The improved properties will theoretically improve the control accuracy.

Different from our previous model, the intrinsic muscles are divided into smaller control units to realize accurate control. According to the simulation, different portions of the same muscles (transverse, vertical and superior longitudinal muscle) really have different functions. Theoretically, the smaller the control units the more accurate for model control. If all the muscle fibers are controlled independently, the control accuracy should be the best. In this situation, the degrees of freedom of the control units increases greatly, which results in the increase of the computational cost. The treatment of dividing the intrinsic muscles into two or three control units was a tradeoff between the computational cost and control accuracy.

To explore the mechanism of speech production is one of the most important purposes of the construction of a physiological articulatory model, the process of synthesizing speech sounds based on the model are also described. In order to keep the integrity of the physiological articulatory model in this thesis, the important parts including the muscle models, morphological structures of articulators, and arrangements of the extrinsic muscles were introduced as well.

# Chapter 3

# Model Control for Articulation

In Chapter 2, the physiological articulatory model have been improved and characteristics of the model are more faithful to the human articulator. In order to use the model to investigate the mechanisms of human speech production or to implement it to medical treatment, language learning, etc., a control strategy has to be developed. To do so, in this chapter, we describe how the speech production ability is acquired by a learning process. Then the schematic of speech motor learning implemented in this study is explained. Finally, we will state the main problems need to be solved by reviewing the deficiencies of the previous control strategies.

## 3.1 Speech acquisition process

In order to introduce how the speech ability is acquired, we first describe the speech production process briefly. A schematic diagram of speech production process is shown in Figure 3.1. In this schematic diagram, black thick arrows show the speech production process in normal condition: 1) In order to generate intended speech sounds, the articulatory targets are planned by using the *Motor Plan* module. 2) The *Motor Execution* module is used to issue motor commands to control the articulators to achieve the planned targets. 3) The articulators are controlled by the issued motor commands to achieve the planned articulatory targets. 4) Finally, speech sounds are generated by the movements of articulators.

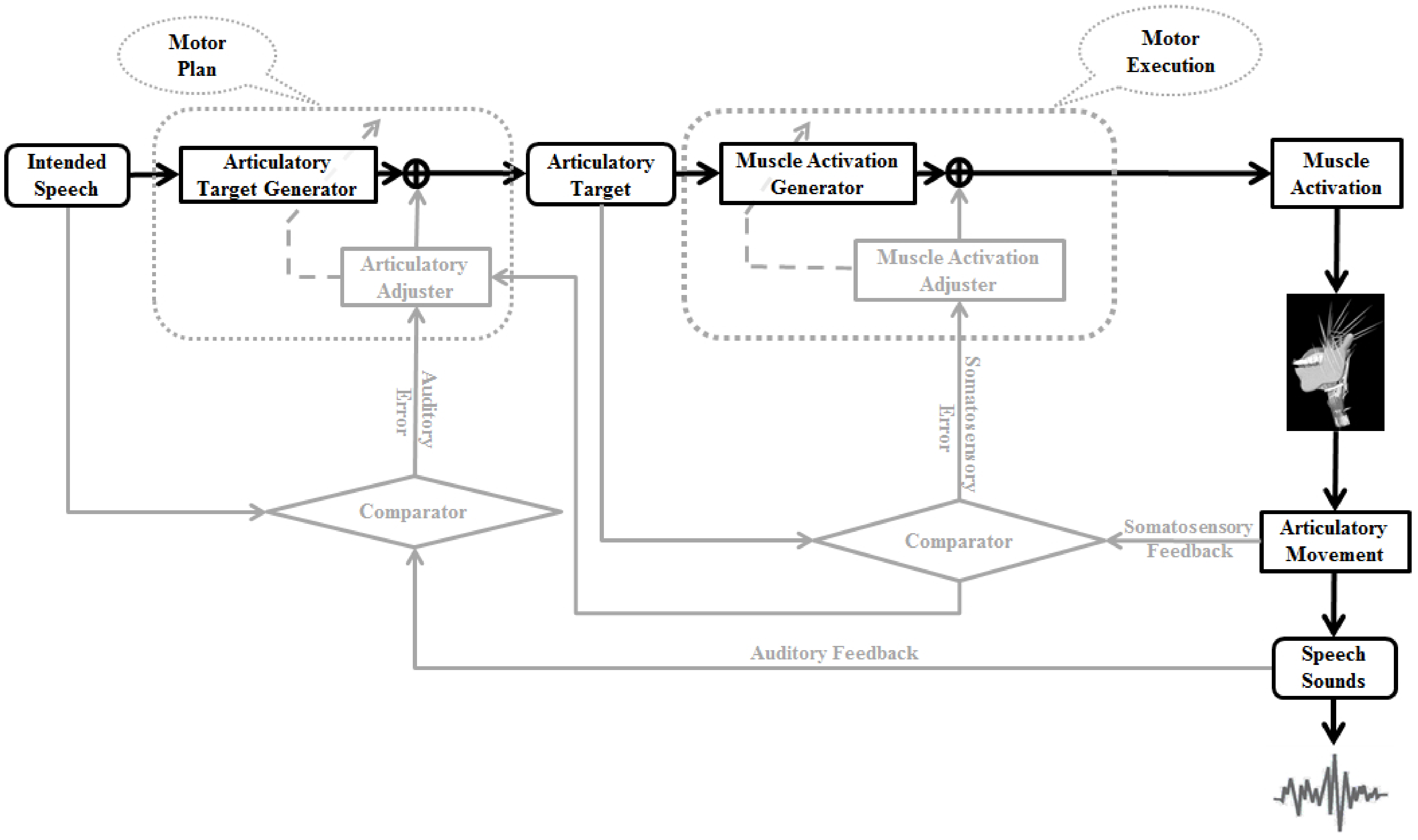

Figure 3.1: Schematic diagram of speech production

In this process, there are two important modules *Motor Plan* and *Motor Execution*. *Motor Plan* is used to generate articulatory targets according to the intended speech sounds, which include a feedforward *articulatory target generator* and an auditory feedback loop. *Motor Execution* is used to issue motor commands to control articulators to achieve planed articulatory targets, which include a feedforward *motor command generator* and a somatosensory feedback loop. In normal condition, these two feedforward mappings, *articulatory target generator* and *motor command generator*, are implemented in generating fluent speech sounds, and feedback loops play a role of verifying the accuracy of intended speech sounds or planed articulatory target. The reason why feedback loops cannot be used to control the articulators to generate speech sounds is because the considerable delay comparing to speech in the feedback loop. According to the somatosensory feedback, if the distance between the planed target and the realized position is considerable, the error information will be sent to the Central Nervous System (CNS) to adjust the motor commands to reduce the error. If the planed articulatory target is well achieved and the difference between the generated speech and intended speech is greater than a threshold according to the auditory feedback, the error information will be sent to the CNS and the planed articulatory target will be adjusted.

How these two feedforward mappings, *articulatory target generator* and *motor command generator*, are constructed? According to Guenther *et al.* [72] and Kröger *et al.* [71], these two feedforward mappings are acquired by a learning process based on auditory and somatosensory feedback in the speech acquisition process. During the learning process, the connections between the neurons used to represent acoustic information and articulatory target information are tuned. Similarly, the connections between the neurons used to represent articulatory information and motor command information are tuned. According to the training process, the knowledge of speech production is stored in these two mappings. The somatosensory and auditory feedback loops play the role of verifying the accuracy of speech production, meanwhile, training and updating these two feedforward mappings.

According to the analysis of speech production process, we can see that there are two targets: acoustic targets and articulatory targets. We would like to analyze the relationship of these two targets and then concentrate on one of them in this dissertation.

## 3.2 Goals of speech production

In speech production, acoustic goal or articulatory goal is still an open debate question. A central hypothesis is that speech goals are defined acoustically and maintained by auditory feedback. Guenther *et al.* believe that the only invariant targets of the speech production process are in auditory perceptual spaces [87]. To prove their hypothesis, they cite some perturbation experiments by imposing a constraint to articulators such as a bite block or lip tube to change the generated acoustic output. In these experiments, subjects try to maintain the auditory perceptual aspects, rather than preserve an invariant articulatory target [82, 83, 86, 88]. Other studies prove this hypothesis by modifying acoustic feedback in real time eventually, the subjects adapted to the modification, and compensate for the change in auditory feedback [84, 85]. These studies prove that the goal of speech production is to produce recognizable phonemes in acoustics and maintained by auditory feedback.

On the contrary, the alternative hypothesis is that speech production is organized in terms of control signals that subserve movements and associated vocal-tract configurations [89, 90]. That is to say, speech goals can be defined in articulatory configuration and maintained by the somatosensory feedback. Indeed, the capacity for intelligible speech by deaf speakers suggests that somatosensory inputs related to movements play a role in speech production. Since, so far, studies that support this hypothesis are not as many as that acoustic goal hypothesis, we would like to elaborate it from the following two aspects: 1) physiological basis of somatosensory feedback 2) Experimental evidences that prove the articulatory goals in speech production.

The somatosensory system is a diverse sensory system comprising the receptors and processing center to produce the sensory modalities such as touch, temperature, proprioception (body position), and nociception (pain). The sensory receptors cover the skin and epithelia, skeletal muscles, bones and joints, internal organs, etc. In speech production, the somatosensory feedback mainly comprises the tactile sensation and proprioceptive sensation. It is easy to prove the existence of the tactile sensation in articulators because when we extend our tongue to contact the palate or lips, we can really feel the contact with pressure. The rest question is whether the proprioceptive sensation exists in articulators because in some of the cases speech articulations do not generate any articulatory contact. As to the proprioception, the muscle spindle and Golgi tendon organ are two of the most important organs [91]. The muscle spindles consist of four to eight specialized intrafusal muscle fibers surrounded by a capsule of connective tissue. The intrafusal fibers are distributed among the ordinary extrafusal fibers of the skeletal muscles in a parallel arrangement. The Golgi tendon organs are encapsulated afferent nerve endings located at the junction of a muscle and tendon. The Golgi tendon organs are in series with the extrafusal muscle fibers. The muscle spindle system is a feedback system that monitors and maintains muscle length, and the Golgi tendon system is a feedback system that monitors and maintains muscle force. The Golgi tendon organs are located at the junction of a muscle and tendon, which indicate that Golgi tendon organs exist in the extrinsic tongue muscles and jaw muscles. Previous studies also proved that the muscle spindles exist in the extrinsic [92] and intrinsic [93] tongue muscles. Although, so far, there is few evidence that the muscle spindles exist in the jaw opener muscles, experimental studies have proved that the stretch reflex is not necessarily mediated exclusively by muscle spindle afference [94]. According to the analysis, we can conclude that the articulators possess the physiological basis of the somatosensory feedback. Another important question is that whether the somatosensory feedback is really implemented in speech production?

In order to prove that the somatosensory information is fundamental to the achievement of speech movements, Tremblay *et al.* designed a perturbation experiment, where the external force applied to the jaw altered jaw movement but has no measurable or perceptible effect on acoustic output [95]. Although there was no change on the acoustic output, subjects still adapted to the external forces to achieve the articulatory targets. The findings indicate that the positions of speech articulators and associated somatosensory inputs constitute a goal of speech movements that is wholly separate from the sounds produced. Furthermore, an experiment was conducted by placing the somatosensory and auditory systems in competition during speech motor leaning [96]. In this experiment, somatosensory and auditory feedback was altered in real time as subjects spoke. As a result, all subjects were observed to correct for at least one of the perturbations, and auditory feedback was not dominant. Indeed, some subjects showed a stable preference for either somatosensory or auditory feedback during speech. These perturbation experiments proved that the articulatory targets can be regarded as the goal in speech production, and the somatosensory feedback is used to learn and maintain this goal.

Based on the evidences described in this section, for simplicity, during the speech motor learning process, only the somatosensory feedback is implemented in this study.

## 3.3 Framework of model control

As described in Section 3.1, two main modules *motor plan* and *motor execution* are constructed during the speech acquisition process. To communicate with speech, humans have to know how to generate the appropriate gestures in their vocal tract, independently of whether these gestures are the ultimate goals of the task or just the obligatory means of achieving the ultimate goals in the acoustics. Therefore, according to the description in Section 3.2, we adopt the hypothesis that articulatory target are regarded as the goal of speech production. The framework of motor control can be simplified, and only the *motor execution* is concerned with, as shown in Figure 3.2. The question comes to "Given an articulatory target, how to find muscle activation that can control the articulators to achieve the target". So far, there is few theories concerning with how humans acquire this ability to control their articulators to achieve the articulatory targets. For the skilled movements of limb, Kawato *et al.* proposed that the feedforward motor command generator is constructed by a learning process by feedback sensory signals [66]. Perrier lead Kawato *et*

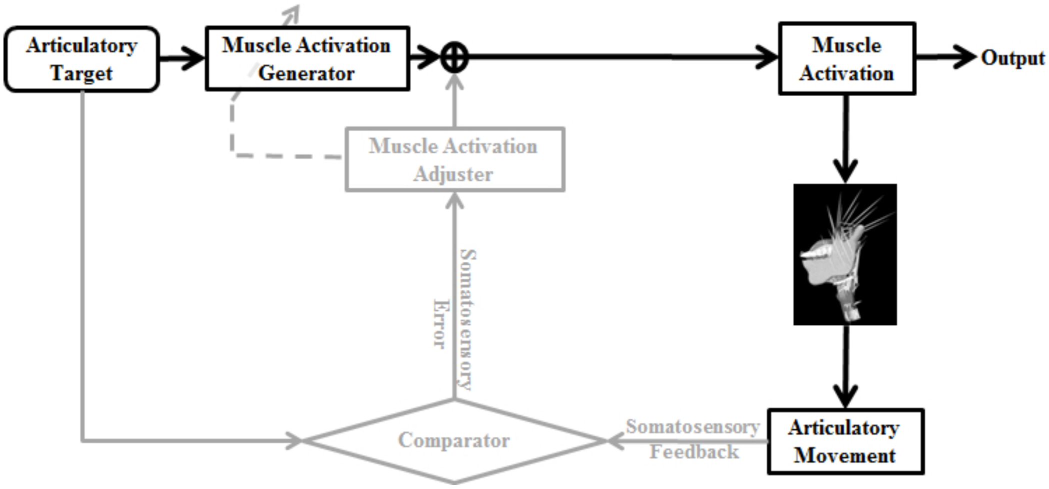

Figure 3.2: Framework of speech motor control implemented in this study.

*al.*'s model into speech production, similarly the feedforward mapping is acquired by a learning process through the somatosensory feedback loop [97]. Perrier divided the learning process into four stages: 1) The CNS learns a forward model to represent the relation between motor commands and sensory signals by a biological feedback. 2) To control fast movements, the forward model and its internal feedback is used as feedback control. 3)Inverse model is learned by using the forward model. 4) The inverse model has been learned and it is used to infer directly the sequence of motor commands from the desired trajectories. In the first three stages, there is a forward model whose function is to predict the articulatory positions according to the motor commands, the essence of the forward model is to accelerate the learning process because the biological feedback has a noticeable delay. If we do not consider this delay, the four stages can be simplified into two. The inverse model (*muscle activation generator*) is learning by biological somatosensory feedback, and finally, the inverse model is used to generate motor command according the desired articulatory target. In Figure 3.2, black thick arrows show the feedforward control route and gray arrows show the feedback back loop.

It seems that the framework of the speech motor learning process is not very complicated. Why the previous control strategies for the physiological articulatory model could not realize this framework? What is the difficulty in realizing the framework? These questions will be answered by reviewing the previous control strategies.

## 3.4 Control strategies in the literature

In this section, we will review the systematic control strategies for the physiological articulatory model, and those control the model using EMG signals will be neglected.

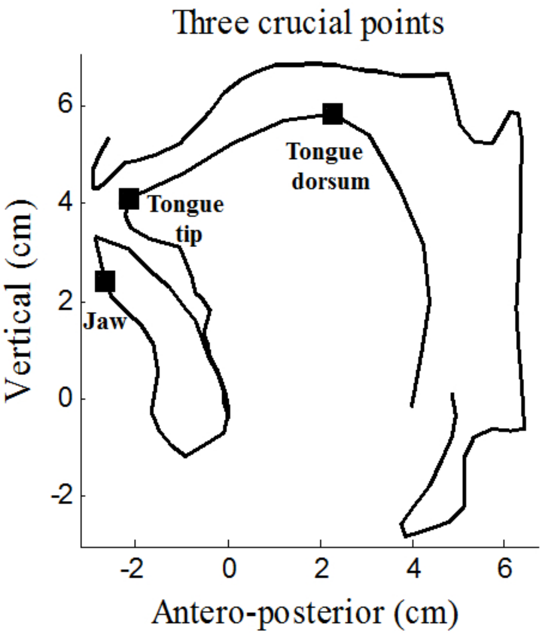

Figure 3.3: Three crucial points used to represent articulatory targets. Three black squares show tongue tip, tongue dorsum and jaw, respectively.

### 3.4.1 Muscle workspace

Dang and Honda 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 [50]. In their method, three control points, shown in Figure 3.3, (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 multi-points 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 the 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 3.4). And the muscle workspace at an 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

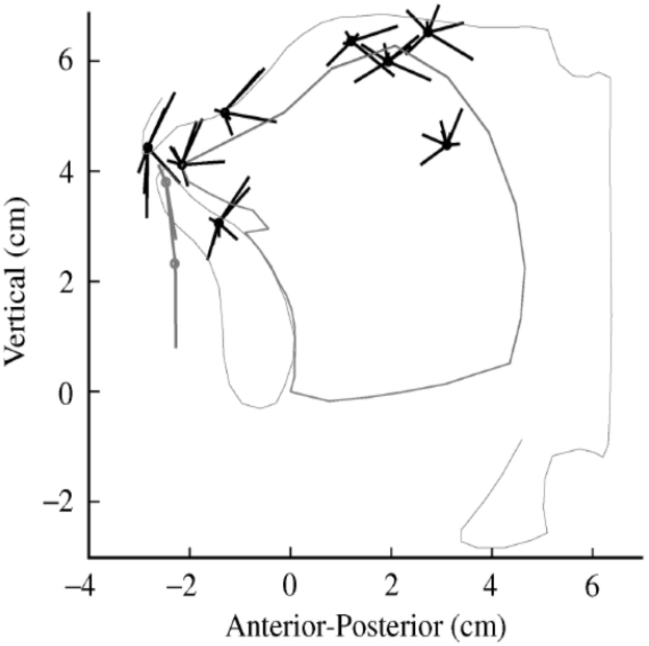

Figure 3.4: Typical muscle workspaces for three control points. Four muscle workspaces were built for tongue tip (dark lines surround the tongue tip) and tongue dorsum (light lines surround the dorsum), and two for the jaw (light lines). (After Dang and Honda [50])

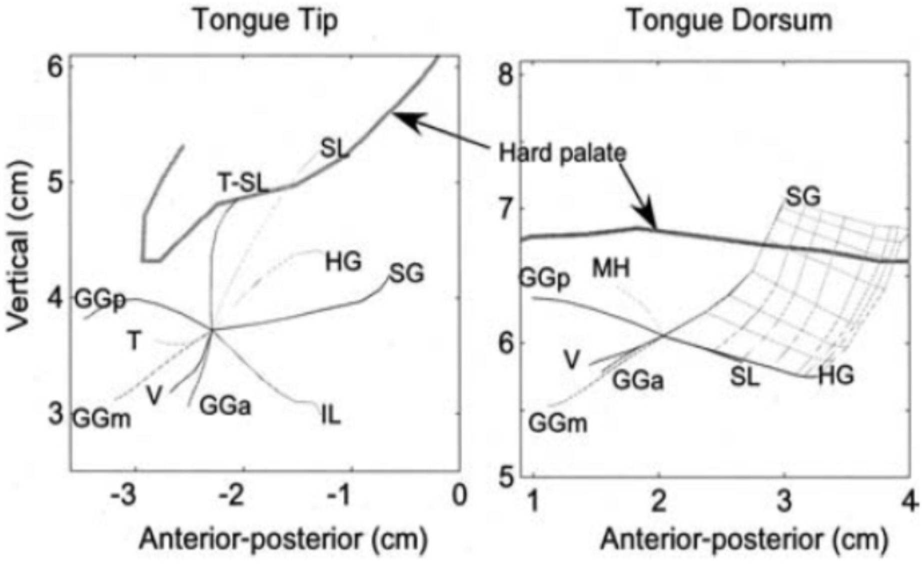

Figure 3.5: Coordinates consisting of the equilibrium positions corresponding to the activation forces ranged between 0 and 6 Newton. The net in the right panel consists of the contour lines of the EPs of SG and HG. (After Dang and Honda [51]))

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

### 3.4.2 EP-map

Dang and Honda found that the control points, converge to sufficient small regions no matter where the initial position is, if the activation duration is sufficiently prolonged [51]. 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.

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 Newton), and elaborated a coordinate based on the EPs for each control point. Figure 3.5 (left) and Figure 3.5 (right) show 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 3.5 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.

### 3.4.3 Other control strategies

The Muscle workspace and EP-map introduced previously are multi-points control strategies, where the control points of the model are controlled independently. This independent control may result in confliction that when the muscle is enrolled to control the tongue tip, it may more or less affect the movement of tongue dorsum, and vice versa. Moreover, in speech production, the phonetic qualities of speech sounds depend on the whole vocal tract shape rather than the size and location of vocal tract constriction only at the tongue tip or tongue dorsum. In order to avoid this disadvantage, Fang *et al.* proposed posture control by manually adjust muscle activation patterns for articulatory target posture [29]. However, because the trial-and-error method depends on the experiential knowledge of the researcher, it is difficult to estimate muscle activation patterns for specific postures. To conquer this deficiency, Fang *et al.* use a regression neural network to predict muscle activation pattern for a given articulatory target [65].

### 3.4.4 Summary of previous control strategies

From the control point of view, the control strategies in the literatures can be divided into two categories, manually control and automatic control. Because the manually control (“trial-and-error”) method [29, 30] depends on the experiential knowledge of the researcher, it is difficult to estimate muscle activation patterns for specific postures. So far, the automatic control strategies control the physiological articulatory model by using crucial control points [51, 50, 38]. However, in speech production, the phonetic qualities of speech sounds depend on the whole vocal tract shape rather than only the size and location of the vocal tract constriction at the tongue tip or tongue dorsum. Therefore, the automatic estimation of muscle activations for a given articulatory posture is necessary for exploring speech motor control.

Let’s classify the previous control strategies into feedforward control or feedback control as shown in Figure 3.2. One can see that the EP-map and Fang *et al.*’s method belongs to feedforward control. The muscle workspace and Stavness *et al*’s method [38] belongs to feedback control. The independent use of feedforward or feedback control has the following deficiencies: 1) Producing fluent speech by using the feedback control is difficult because of the long delay in the feedback loop. 2) It cannot adapt to the perturbed external force added to the model by using feedforward control. 3) It is difficult to correct muscle activation when the degree of accuracy do not meet the requirement for some cases by feedforward control. So far, there is no integrate control strategy that include both feedforward and feedback control. The unsolved problem is that the feedback control is independent crucial points control rather than posture control. That is to say, given the difference between the target posture and realized one, how to adjust muscle activations to reduce the difference is an unsolved problem.

## 3.5 Representation of articulatory target

Two different schemata are available regarding motor templates for executing speech production: target oriented and trajectory oriented. Lindblom proposed that the speech sounds were generated by achieving successive targets of phonemes [67]. However, Van Bergem proposed that speech production would base on generating dynamic trajectories [68]. These two theories result in two different control objects: the targets or the trajectories. In previous model control studies, Payan *et al.* [69] adopted articulatory trajectories to generate vowel-to-vowel sequences. Other studies (Pascal *et al.* [70], Dang and Honda [51], Wei et al.[44] . Fang et al. [29] ) controlled their physiological articulatory models based on the target theory. Because these two theories are still in open debate, the choice of theories relies on the object of studies and the convenience for researchers. In current stage, we adopt target theory and develop the target based control strategy.

In the previous studies [50, 51], articulatory targets were defined by isolated control points (tongue tip, tongue dorsum and jaw), and these points are used to control the constriction position of the vocal tract. Since the acoustic characteristics of speech sounds depend on the whole vocal tract configuration, the contour of the tongue and jaw is a proper target. In this study, we use midsagittal contour to describe the articulatory posture, which can represent most of phonemes except for some lateral ones. In addition, it is convenient to measure the movement on the midsagittal plane based on observation techniques, such as Electromagnetic Articulography (EMA), X-ray microbeam, and MR imaging. The articulatory posture defined in the midsagittal plane will facilitate the comparison of model shapes to the measurement data. Consequently, eleven points on the midsagittal tongue surface and one point on the lower incisor are used to represent the articulatory posture of our model, as seen in Figure 3.6 .

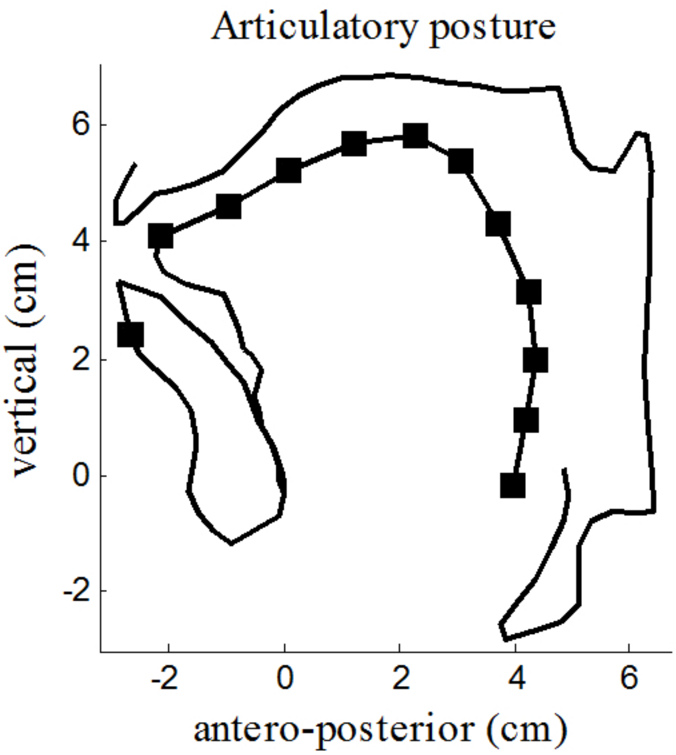

Figure 3.6: Representation of articulatory posture in the midsagittal plane. Black squares on the tongue and jaw are used to represent articulatory posture.

## 3.6 Object of the control strategy

In order to overcome the disadvantages of independent use of feedforward control and feedback control, this study attempts to realize a control strategy that integrates the feedforward and feedback control, where the feedback control served in the learning process for establishing and updating feedforward mapping, see Figure 3.2. The feedforward mapping constructed and maintained by feedback error learning will render the control strategy with the capacity of adaption. When the environment varies, a correctional feedforward model can be reconstructed through a new learning process.

Kawato *et al.*'s work was for controlling musculoskeletal movements such as the arm movement, while in this study we deal with more complex movements including skeletal and muscular-hydrostat movements of the speech organs. In order to elucidate the difference between the musculoskeletal system and muscular-hydrostat system, we use elbow as an example. As shown in Figure 3.7, biceps is the agonist muscle and the contract of biceps will flex the forearm; triceps is the antagonist muscle and its contraction will extent the forearm. The agonist-antagonist pair (biceps and triceps) does not vary during the movement of the arm. In human body, the majority of the six hundred musculoskeletal systems exist agonist-antagonist muscles.

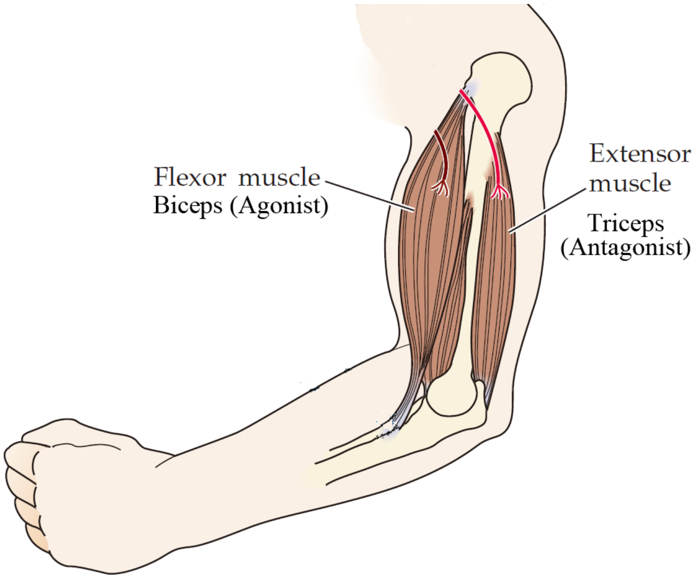

Figure 3.7: Agonist-antagonist muscles of elbow [91].

As to the tongue, agonist–antagonist pair is not as clear as that musculoskeletal system. Dang *et al.* have investigate the agonist–antagonist property of tongue muscle pairs based on their partial 3D physiological articulatory model [51]. According to their study, agonist-antagonist muscle pairs are found for crucial control points, e.g. GGm and HG are agonist-antagonist muscles for tongue tip. Furthermore, Fang *et al.* explore the agonist-antagonist tongue muscle more systematically based on a full 3D model [80]. They found that the muscle pairs GGm–SL, GGm–HG and GGa–HG act as agonist–antagonist for tongue tip; GGp-HG, GGp-SL and GGm-SG act as the antagonist for tongue dorsum. According to their studies, one can see that the agonist–antagonist pairs is regarded to independent points, tongue tip or tongue dorsum. With regard to the whole posture, of the tongue, this agonist–antagonist pairs would not exist. From their studies, one can se that even if for the crucial points the agonist–antagonist is not very clear when the tongue move far away from its rest position.

In short, for the musculoskeletal system the agonist–antagonist muscle pair does not change during the movement, while for the muscular-hydrostat system (tongue) agonist–antagonist muscle pairs is not clear and vary during articulation. This make it very difficult to control the articulators during articulation. The solution of this problem will be elucidated in Section 4.2.2.

The expectation of the integrated control strategy is that 1) Feedforward control and feedback control can be used to generate muscle activation patterns independently. 2) Feedback control can be implemented to refine the muscle activation patterns which are generated from feedforward mapping and control the model to the targets with higher accuracy. 3) When external force is added to the model, the control strategy can adapt to the perturbation.

# Chapter 4

# Algorithm of the control strategy

In Chapter 3, the main control strategies in the literatures were introduced and based on the analysis of the disadvantages of the previous control strategies, we implement a framework of "feedforward mapping constructed by feedback error learning". In this chapter, the detail algorithms of the control strategy will be introduced.

## 4.1 Principal component analysis of the articulatory posture

As described previously, each articulatory posture is depicted by 12 points with the horizontal and vertical coordinates. Therefore, each articulatory posture is represented by a 24 dimensional vector. However, these points have significant redundancy and correlativity. To reduce the redundancy, there are a lot of methods, such as, principal component analysis (PCA), linear component analysis (LCA), etc. In order to represent the posture uniquely, the relationship between the new components must be perpendicular. Furthermore, the new components have to represent as much as the variance of the existing data. Therefore, PCA is appropriate for this implementation. We suppose that the functions of individual muscles can be decomposed into these orthogonal components, and their effect can be superimposed.

### 4.1.1 Algorithm of PCA

In this section, the main algorithm implemented in this study is introduced [104]. Let $Pos = [TX_1, TX_2 ... TX_{11}, JX, TY_1, TY_2 ... TY_{11}, JY]$ denote a posture vector, where $TX$ and $TY$ are the X and Y coordinate of tongue node, respectively, subscript number represent the index of the node; $JX$ and $JY$ are the X and Y coordinates of the Jaw point. In order to construct a PCA space, first the covariance matrix $C$ is constructed according to a set of simulations, Equation 4.1, where $cov(x, y)$ is the covariance of variable x and y.

$$C = \begin{bmatrix} cov(Pos_1, Pos_1) & cov(Pos_1, Pos_2) & ... & cov(Pos_1, Pos_{24}) \\ cov(Pos_2, Pos_1) & cov(Pos_2, Pos_2) & ... & cov(Pos_2, Pos_{24}) \\ \vdots & \vdots & \vdots & \vdots \\ cov(Pos_{24}, Pos_1) & cov(Pos_{24}, Pos_2) & ... & cov(Pos_{24}, Pos_{24}) \end{bmatrix} \tag{4.1}$$

Let $C\xi = \lambda\xi$, $\lambda$ and $\xi$ are the eigenvalue and eigenvector, respectively. In order to obtain the eigenvector and eigenvalue, solve the equation, Eq. 4.2, where $I$ is a identity matrix. If and only if Eq. 4.3 is satisfied, Eq. 4.2 have non-zero solution.

$$(C - \lambda I)\xi = 0 \tag{4.2}$$

$$|C - \lambda I| = 0 \tag{4.3}$$

Solve Eq. 4.3, 24 eigenvalues will be obtained, and substitute the $\lambda$ into Eq. 4.2 the eigenvector $\xi$ will be calculated. Note that the eigenvectors are normalized to *unit* eigenvectors, ie. their lengths are 1.

### 4.1.2 PCA implementation

To generate a data set for PCA, our objective was to create simulations that cover most of the possible postures by using reasonable muscle combinations considering the agonist-antagonist properties of muscles. With reference to the previous study [80] about the agonist-antagonist muscles and muscle combinations, 9703 articulatory postures that cover most of the possible postures were obtained for PCA.

According to the PCA, 24 components were obtained. In order to acquire the important components and abandon less important components, we have to decide a threshold. The ideal situation is that the threshold is decided by the spacial resolution of somatosensory feedback of articulatory organs. However, so far, we do not know the exact somatosensory resolution of articulators. The spacial resolutions of devices that are commonly used to measure the movement of articulators are used as a reference Table. 4.2, because the construction and evaluation of the model are from the data set extracted from these devices. The variance of each component and the accumulated explanations of variance are shown in Table. 4.1. From this table, one can see that the first six components can explain 99.33% of the variance, which indicates that the articulatory posture can be determined by the first six components within 0.7% error. In order to evaluate the error result from using only the first six components to represent the posture, all the

Table 4.1: Variance of PCA components (%). C1 to C6 are the first six components, VC is the variance of the components, and AVC is the accumulative variance from the first component to the current component.

| Component | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| VC | 79.58 | 12.54 | 3.01 | 2.32 | 1.42 | 0.46 |
| AVC | 79.58 | 92.12 | 95.13 | 97.45 | 98.87 | 99.33 |

Table 4.2: Spacial resolution of often used devices (cm).

| Tagged cine-MRI [24] | EMA [100] | Ultrasound [101] |
|---|---|---|
| 0.1 | 0.05 | Axial: 0.1 Lateral: 0.19 |

9703 articulatory postures were transformed to six dimensional PCA space, and then six dimensional PCA were reversely transformed to twenty four dimensional postures. The distance between the transformed posture and original posture were calculated by Eq. 4.5, the average distance is 0.039 cm and standard deviation is 0.024 cm. By comparing to Table 4.2, one can see that the deviation resulted from using only the first six components is acceptable.

The contribution of the first four components is shown in Figure 4.1. Figure 4.2 shows the functions of individual muscles by transforming the articulatory postures, which were resulted from the individual muscle activation, into the PCA space consisting of the first three components, where some muscles with small impact are not shown. In this figure, each curve shows the function of a single muscle unit in PCA space, where the dots on the curves indicate the results using different muscle activation levels. From this figure, one can see that the equilibrium positions travel from the rest position in PCA space as the activation level increases. There is a monotonic relationship between the muscle activation level and displacement increment: increasing the activation level will drive the articulator to move away from rest position, whereas decreasing the activation level will make the model return towards the rest position. This monotonicity is crucial for the estimation of muscle activations because we can increase activation if the realized position does not reach the target and decrease the activation if the realized position exceeds the target.

In the PCA space, the relationship between the module of the six dimensional PCA vector and the activation of individual muscle is quasi-logarithmic according to simulations. This relationship can be represented by a fitting curve shown in Eq. (4.4), where $x$ and $y$ are two undetermined coefficients, $a$ is the activation and $ML$ is the module of the PCA vector. In order to calculate the undetermined coefficients of Eq. (4.4) for each muscle, the individual muscles were activated with 20 equal scales between 0 and 1 to

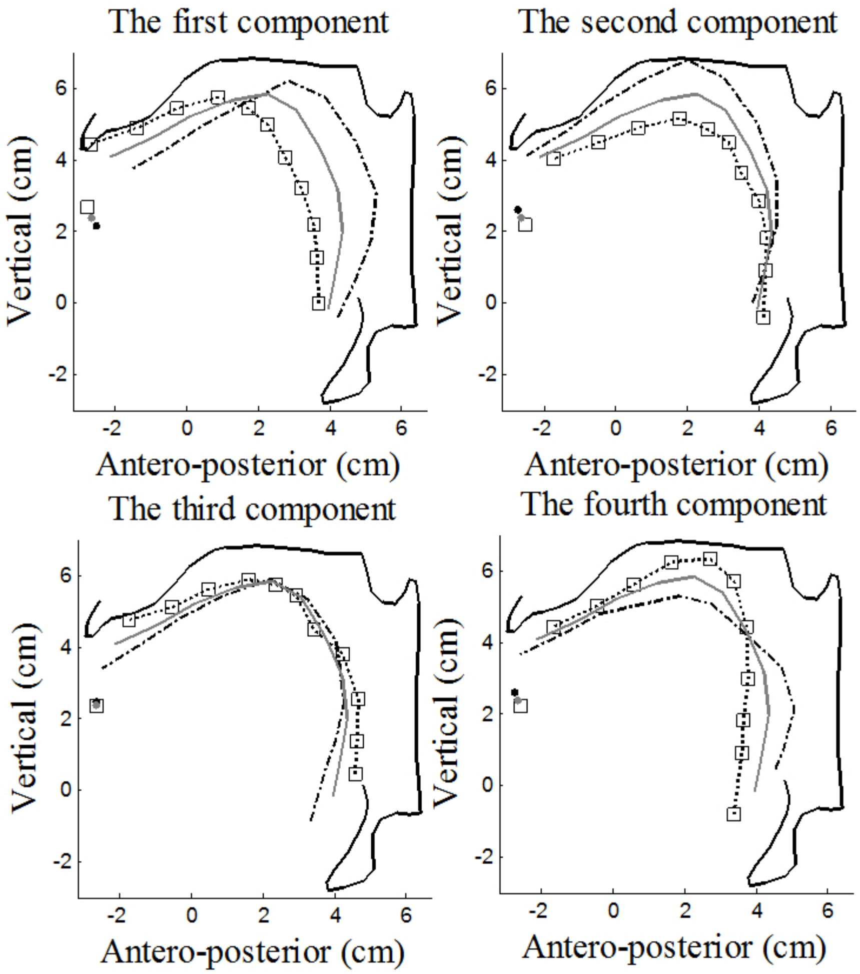

Figure 4.1: Effect of the PCA components. Gray lines show the rest position, dashed lines and dashed lines with squares show the directions of each component with positive and negative coefficient, respectively.

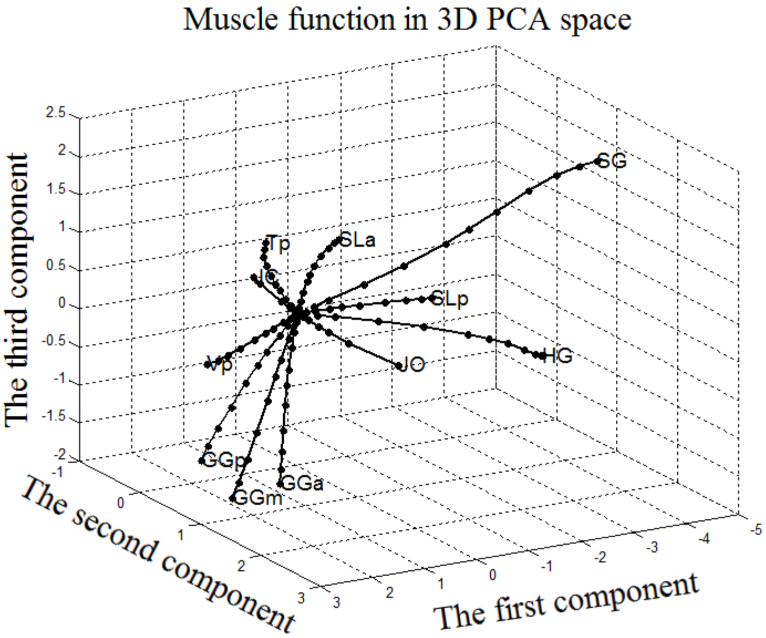

Figure 4.2: Function of individual muscle unit in 3D PCA space.

obtain the corresponding data pair of muscle activation and PCA vector module. According to this data, the undetermined coefficients were obtained for individual muscle. Figure 4.3, shows a fitting curve for the GGa muscle, where $x$=0.53501 and $y$=230.8211. For each muscle, 1/10 of the PCA vector module generated by 0.8 activation was defined as a scale. The increase in muscle activation that make the vector module increase by a scale is defined as a *unit increment* of muscle activation. The *unit increment* of muscle activation for each muscle can be calculated by Eq. (4.4). In Figure 4.3 asterisk show the ten *unit increments* of muscle activation and their PCA modules of GGa muscle.

$$ML = x\log_e(ya + 1) \tag{4.4}$$

## 4.2 Feedback error learning

As we have introduced in Chapter 3, the feedforward mapping is constructed by a learning process using feedback learning loop. In this section, we will describe in detail on how to correct adjust activation patterns to control the model approaching to the target and meanwhile how to train the feedforward mapping.

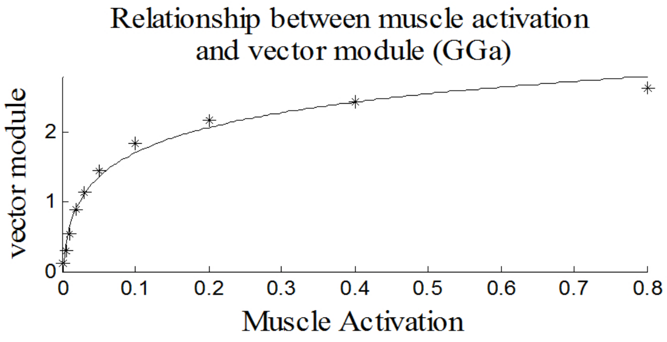

Figure 4.3: Relationship between muscle activation and PCA vector module (GGa).

### 4.2.1 Flowchart of learning process

An iteration method was used to find muscle activation patterns by gradually minimizing the difference between the target posture and realized position. The distance ($D$) between the target posture and realized posture was calculated by Eq. (4.5), where $Rx_p$ and $Ry_p$ were the horizontal and vertical coordinate values of the $p^{th}$ point used to represent the realized posture, and $Tx_p$ and $Ty_p$ were the coordinate values of the corresponding target points.

$$D = \frac{1}{12}\sum_{p=1}^{12}\sqrt{(Rx_p - Tx_p)^2 + (Ry_p - Ty_p)^2} \tag{4.5}$$

The flowchart of the estimation procedure is shown in Figure 4.4. A target posture is projected into a 6-dimensional PCA space described above in order to obtain a PCA vector. The muscle activation pattern is initiated based on the difference between the rest posture and target posture, and the *Counter* used to count the failed iteration times is initiated to 0. The muscle activation initiation method will be introduced in Section 4.2.2 The muscle activation is input to the physiological articulatory model, and then the model moves to a certain position and reaches equilibrium. If the distance in the $k^{th}$ iteration $D^k$ is smaller than the distance generated from the previous iteration $D^{k-1}$ , the muscle activation is accepted and we move on to decision (2), otherwise we increase the failed counter and go to *Muscle Activation Adjuster* module. In decision (2), we output the muscle activation, if the distance is smaller than the threshold ( $D^k < TH$ ) or the *Counter* is greater than the number of muscle units $Counter \geq n = 18$; if the output conditions are not satisfied, set the *Counter* to 0 and go to *Muscle Activation Adjuster* module. The most important module in this procedure is the *Muscle Activation Adjuster*, which will be introduced in the next section.

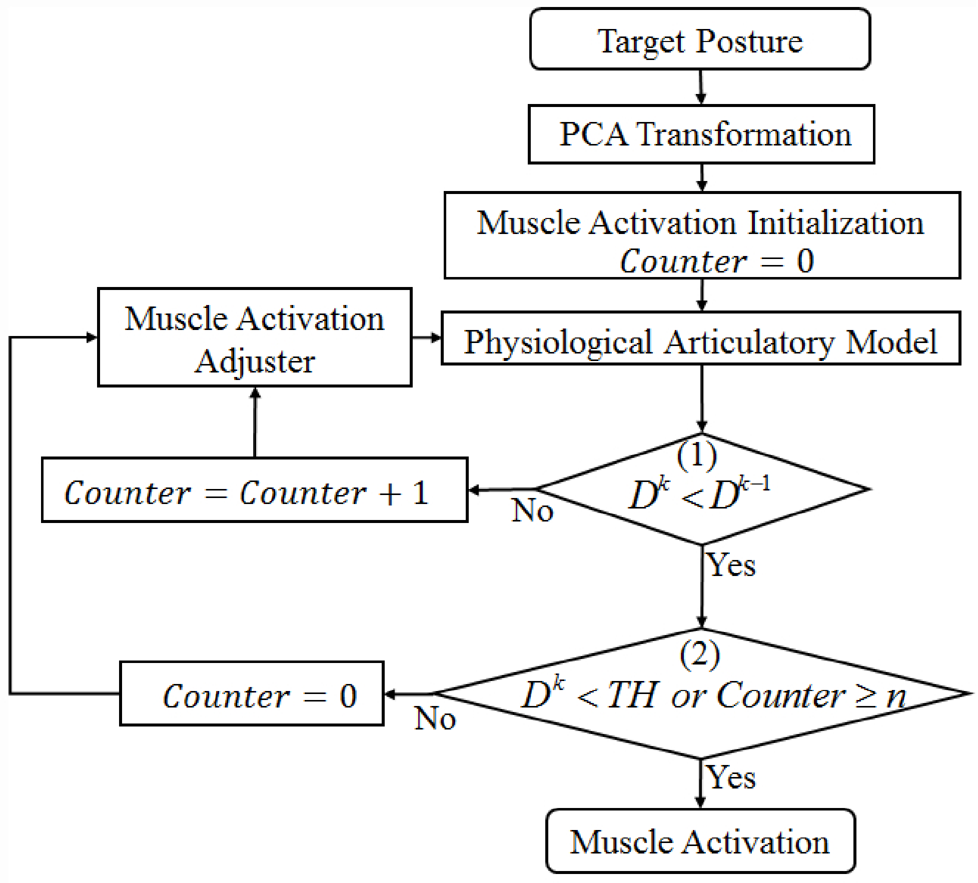

Figure 4.4: Flowchart of muscle estimation.

### 4.2.2 Dynamic PCA workspace

In each iteration step, the muscle activation is adjusted by the following equation:

$$\mathbf{a}^k = \mathbf{a}^{k-1} + \Delta\mathbf{a}^k, \tag{4.6}$$

where $\mathbf{a}^k$ and $\mathbf{a}^{k-1}$ are the muscle activation used in current and previous iteration step,respectively. The muscle activation vector $\mathbf{a} \equiv [a_1 a_2 ... a_n]^T$ is constituted by the activation of individual muscle $a_i$, and the number of muscle units $n = 18$.

The main work in each iteration step is to find the adjustive muscle vector $\Delta\mathbf{a}^k \equiv [\Delta a_1^k \ \Delta a_2^k ... \Delta a_n^k]$. $\Delta a_i^k$ is related to the contribution of individual muscle function vector to the target vector $C_i^k$ as follows:

$$\Delta a_i^k = |\boldsymbol{V}_{mi}^k| \cos\theta = |\boldsymbol{V}_{mi}^k| \frac{\boldsymbol{V}_{mi}^k \cdot \ \boldsymbol{V}_t^k}{|\ \boldsymbol{V}_{mi}^k||\boldsymbol{V}_t^k|} = \frac{\boldsymbol{V}_{mi}^k \cdot \boldsymbol{V}_t^k}{|\boldsymbol{V}_t^k|}. \tag{4.7}$$

$C_i^k$ is calculated by projecting individual muscle function vector to target vector, where $\boldsymbol{V}_{mi}^k$ is the individual muscle function vector of the $i^{th}$ muscle, $\boldsymbol{V}_t^k$ is target vector and $\theta$

is the angle between $\boldsymbol{V}_{mi}^{k}$ and $\boldsymbol{V}_{t}^{k}$. Target vector $\boldsymbol{V}_{t}^{k}$ is defined as

$$\boldsymbol{V}_{t}^{k} = \boldsymbol{P}_{t} - \boldsymbol{P}_{r}^{k-1}, \tag{4.8}$$

where $\boldsymbol{P}_t$ is target posture and $\boldsymbol{P}_{r}^{k-1}$ is the realized posture after the previous iteration.

The muscle function vector $\boldsymbol{V}_{mi}^{k}$ is defined by the effect when the activation of the $i^{th}$ muscle has a *unit increment*:

$$\boldsymbol{V}_{mi}^{k} = \boldsymbol{P}_{i+1} - \boldsymbol{P}_{r}^{k-1}, \tag{4.9}$$

where $\boldsymbol{P}_{r}^{k-1}$ is the posture realized by activation $\mathbf{a}^{k-1}$, and $\boldsymbol{P}_{i+1}$ is the posture realized by *unit increment* of muscle activation for the $i^{th}$ muscle in $\mathbf{a}^{k-1}$. *Unit increment* was explained in Section 4.1

It should be noted that the posture used here is defined by a six-dimensional PCA vector. The $i^{th}$ muscle ($i = \arg\max_i(C_i^k)$) with the maximum contribution $C_{max}^{k} = \max(C_i^k)$, will have a *unit increment.* For the other muscles, the increased activations are less than *unit increment* and their proportion to *unit increment* is calculated by $(C_i^k/C_{max}^k)$. The increment of muscle activation of *unit increment* is calculated by a constructed fitting curve for individual muscle in Eq. (4.4). Note that, after adding the $\Delta\mathbf{a}^k$ to $\mathbf{a}^{k-1}$ in Eq. (4.6), if the activation of a muscle $a_i^k$ is smaller than 0, it will be set to 0, because there is no negative muscle activation.

In the rest position, the individual muscle function vector $\boldsymbol{V}_{mi}^{k}$ was built by activating individual muscles, as shown in Figure 4.2. However, during articulation, muscle orientations vary along with the movement of the jaw and tongue, which will result in the variation of muscle function vector. To solve this problem, Dang and Honda [50] proposed a method to estimate the muscle function orientation dynamically. Following their idea, we constructed a set of reference PCA workspaces in some extreme locations by moving the origin to given extreme locations. When speech organs move to an arbitrary position, a dynamic PCA workspace can be interpolated based on the reference PCA workspace.

We first construct seven reference PCA (r-PCA) workspaces by the following procedures: 1) move the PCA center to seven extreme locations in PCA space by a set of selected muscle activation patterns; 2) in the given PCA center construct a r-PCA workspace by increasing a *unit increment* of individual muscle, refer to Section 4.1 Together with the r-PCA workspace in the rest position, we have eight r-PCA workspaces. The r-PCA workspaces in 3D PCA is shown in Figure 4.5, where No. 1 is the original PCA workspace in the rest position, and No. 2-8 show the other r-PCA workspaces in different reference positions. In order to show the r-PCA workspace clearly, only four r-PCA workspaces with extrinsic muscles are shown in this figure.

The dynamic PCA workspace (d-PCA) for a given position is interpolated based on their distance to the eight reference PCA workspace by using the following equation:

$$\boldsymbol{V}_{mi}^{k} = \frac{\sum_{s=1}^{w} L_s \boldsymbol{V}_{si}}{\sum_{s=1}^{w} L_s}; \quad L_s = \prod_{\substack{j=1 \\ j \neq s}}^{w} l_j^2 \tag{4.10}$$

where $\boldsymbol{V}_{mi}^{k}$ denotes a muscle function vector in d-PCA, $\boldsymbol{V}_{si}$ is the muscle function vector in the $s^{th}$ r-PCA workspace, $l_j$ is the Euclidean distance from current position to the origin of $j^{th}$ r-PCA workspace, $w = 8$ is the number of reference PCA workspaces. The coefficient $L_s$ of the $s^{th}$ r-PCA is the product of the distance from current position to the origin of the other $(w-1)$ r-PCA workspace. The characteristic of the interpolation method is shown in Figure 4.6, which demonstrates that the interpolation has a quadratic surface with a relatively flat characteristic surrounding the reference points. Figure 4.5 shows an example of the d-PCA workspace in the dash lines, which was generated by using the interpolation method. The d-PCA workspace reflects individual muscle function vector in current position. Occasionally, by adding the adjustive muscle vector $\Delta\mathbf{a}^k$ cannot control the model closer to the target, the adjusted vector $\Delta\mathbf{a}^k = [\Delta a_1^k \ \Delta a_2^k \ ... \ \Delta a_n^k]$ will be adjusted by setting $\Delta a_i^k$ to 0, where $|\Delta a_i^k|$ is the smallest nonzero value in the vector, and the *Counter* in Figure 4.4 will be increased by 1.

## 4.3 Feedforward mapping

As shown in Figure 4.4, during the learning process using feedback loop, each simulation can generate a muscle activation pattern and its corresponding articulatory posture. This corresponding simulation results can be used to train the feedforward mapping. The input is the articulatory target in 6-dimensional PCA space and the output is the muscle activation patterns. Artificial neural networks have been successfully implemented in neurocomputational models of speech production to simulate the neuronal connections of synapses in human brain cortex [71, 72]. In this study, a two-layer artificial neural network was trained to build up the feedforward mapping.

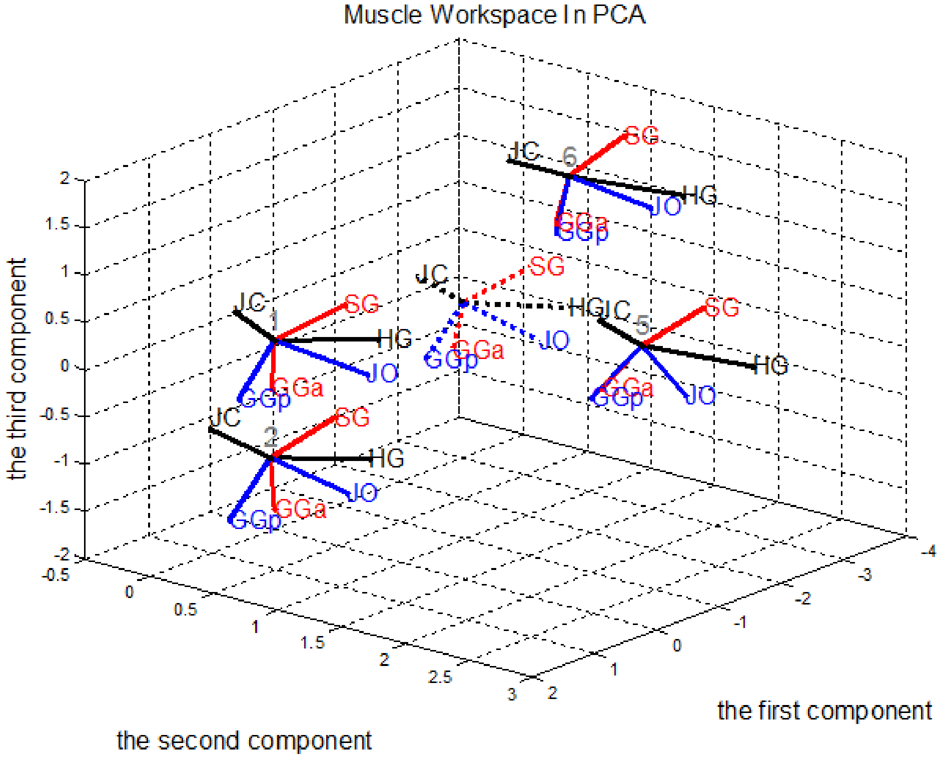

Figure 4.5: Reference PCA workspaces (solid lines) and dynamic PCA workspace (dash lines) in 3D PCA.

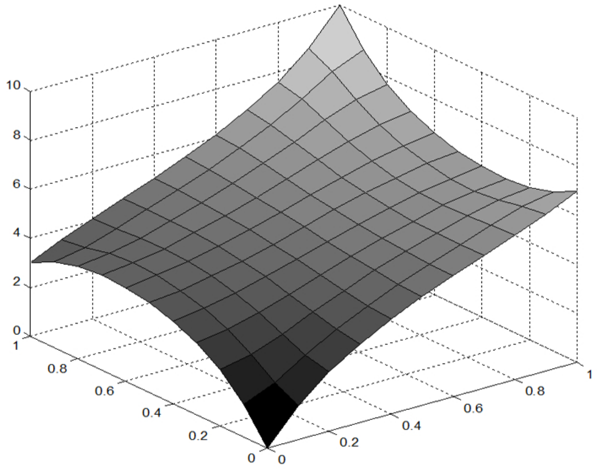

Figure 4.6: An example of the interpolation surface using four reference points with coordinates (0, 0), (0, 1),(1, 0), (1, 1) and their values 0, 2.5, 7.5, 10.

# Chapter 5

# Evaluation

So far, we have finished introducing the flowchart of using a feedback error learning loop to construct feedforward mapping. In this chapter, we will evaluate the control strategy from the following aspects: 1) Evaluate the feedback loop. 2) Evaluate the feedforward mapping. 3) Evaluate the integrated control strategy. 4) Evaluate the adaption ability.

## 5.1 Implementation of feedback error learning

The proposed feedback learning loop was evaluated by using the five Japanese vowels obtained from magnetic resonance images as the targets to estimate muscle activation patterns. Since the prototype subject of the physiological model was the same as the subject serving in the MRI data, we can directly compare them without any normalization processing. To find muscle activation pattern for the target posture an iteration method was used by gradually minimizing the difference between the target posture, and realized position. In the iteration process the muscle activation was guided by the dynamic PCA workspace introduced previously. Two examples of the iteration processes approaching to the targets are shown in Figure 5.1, where the upper left and upper right panel shows the example of the vowel /a/ and vowel /o/, respectively. The lower left (vowel /a/) and lower right (vowel /o/) panel show the distance between the target posture and currently realized posture with the adjustment of muscle activation. The difference between the target posture and the realized posture is measured by Equation 4.5.

From Figure 5.1, one can see that there are some knee points, which indicate the adjustment of muscle activation in the current step cannot drive the model closer to the target. Although, it is seen that some ripples appeared along with the distance curve, the muscle adjusting method can adjust the muscle activation patterns automatically, and eventually control the model to achieve the target.

The target postures of the five Japanese vowels were well achieved, as shown in Figure 5.2. The difference, calculated by Equation 4.5, was ranged from 0.06 cm to 0.17 cm.

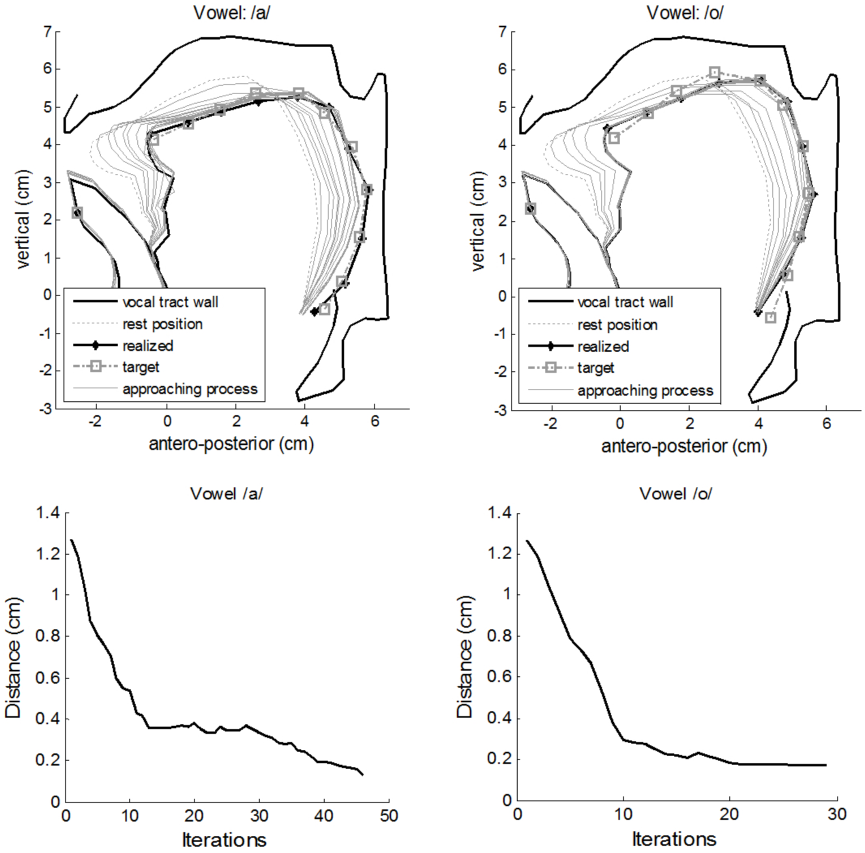

Figure 5.1: Processes of muscle activation estimation of vowel /a/ (left) and vowel /o/ (right).

Because the average distance and standard deviation from the average distance are shown in Table. 5.1. From this figure one can see that the posture targets of the five Japanese vowels are achieved. The obtained muscle activation patterns (active muscle forces) are shown in Table 5.2.

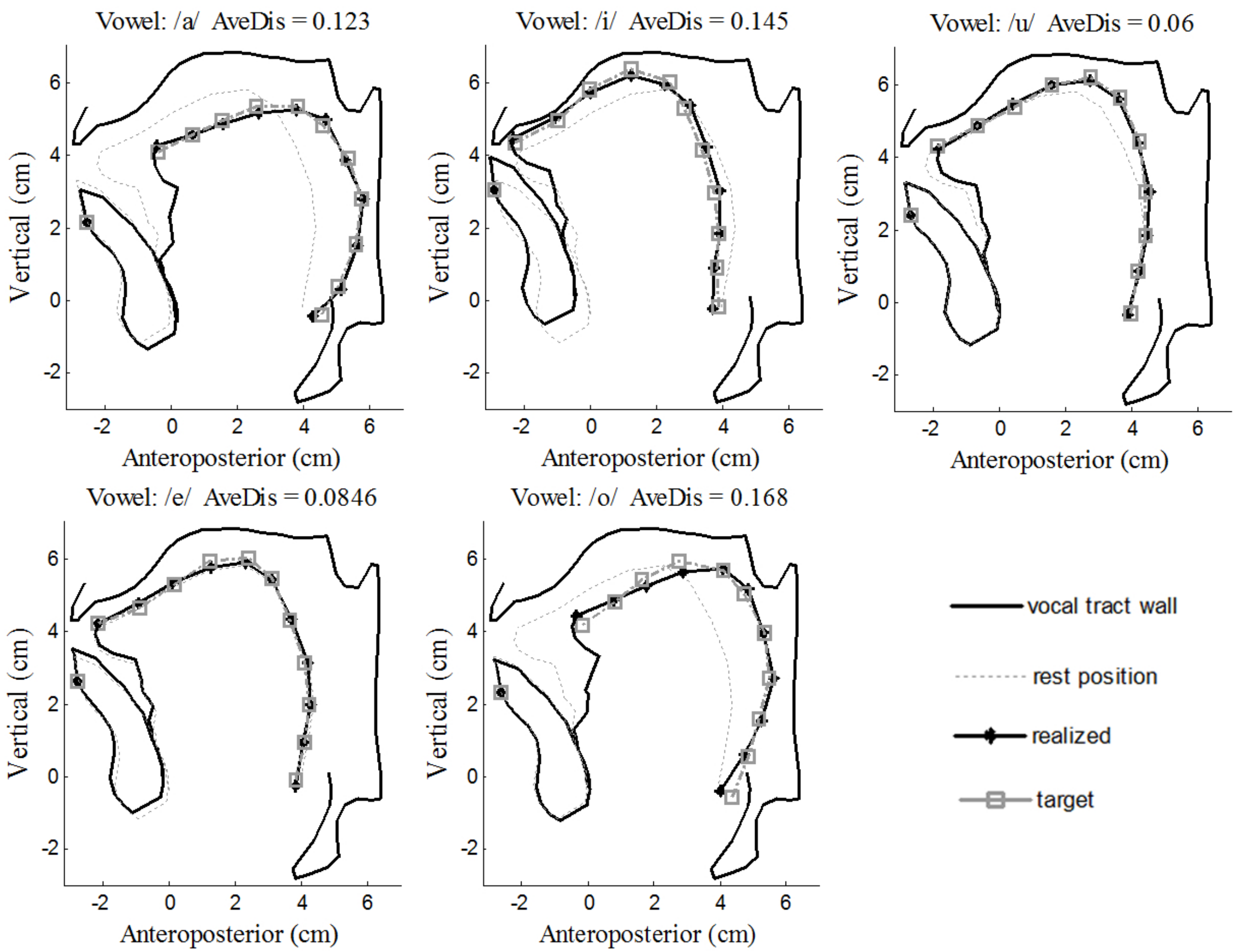

Figure 5.2: Realized position for five Japanese vowels. Gray dash lines show the rest position. Gray dash lines with square markers show the target postures. Black lines with stars show the realized positions. The average distances (defined in Eq.(4.5)) for /a/, /i/, /u/, /e/, and /o/ are 0.123 cm, 0.145 cm, 0.06 cm, 0.085 cm, and 0.168 cm, respectively.

Table 5.1: Average distance and standard deviation of 12 points.

| Vowel | /a/ | /i/ | /u/ | /e/ | /o/ |
|---|---|---|---|---|---|
| Mean Distance (cm) | 0.123 | 0.145 | 0.06 | 0.085 | 0.168 |
| Standard Deviation (cm) | 0.08 | 0.051 | 0.036 | 0.053 | 0.091 |

To evaluate the obtained muscle activation patterns, we first compare activations of the extrinsic tongue muscles to the normalized EMG (Electromyography) measurements [23]. Note that the EMG signals used here were extracted from English vowel articulations because there are no EMG signals for Japanese vowels so far. The EMG signals and muscle activation were normalized to the value between 0 and 1, according to their maximum value in the activation pattern, the maximum values were normalized to 1. Figure 5.3 shows the comparison between estimated extrinsic tongue muscle activations and EMG observations. One can see that, the estimated muscle activation patterns are consistent with the EMG patterns for vowels /a/, /o/, and /i/. In Figure 5.3, there are outstanding differences for vowel /e/ and /u/. Japanese /e/ was the prototype of the model. Accordingly, there should be no muscle activation in the estimation for vowel /e/. A slight activation of GGp in the estimation was probably caused by the difference of the prototype /e/ and the reference /e/ used in this study. The difference shown in vowel /e/ does not show any significant effects. For vowel /u/, as well known, unlike English /u/, Japanese /u/ does not have lip protrusion. The articulatory positions are different between Japanese vowels and English vowels. From this figure, one can see that the vowels used to compare have similar but not exactly the same articulatory positions. This articulation difference may cause some compensation on the tongue shape but not only on the tongue dorsum. This may be a reason of the difference for /u/.

As we have described in Chapter 3, somatosensory feedback as well as auditory feedback are used in the speech motor learning process. In this study we focus on the somatosensory feedback loop, we prefer to evaluate the learning process by synthesizing acoustic signal to inspect whether the learning process by using somatosensory feedback can result in correct speech sounds. At the current moment, the lips are not physiologically modeled. 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 [105]. Table. 5.3 shows the first three formants of the five vowels, where the Observed formants are formants of the speech sounds obtained from MRI experiments and the Synthesized formants are the formants of the synthesized speech. From this table one can see that the formant frequencies between Observed and Synthesized are a little bit different, which may due to two reasons: 1) The midsagittal plane cannot represent the whole vocal tract shape. 2)

Table 5.2: Muscle activation patterns for the five Japanese vowels. The active muscle force of JO and JC is the sum of the active force included in the muscle groups. (Unit: Newton)

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| /a/ | Muscle | GGa | GGm | GGp | HG | SG | SLa | SLp | IL | Va |
| | Force | 0 | 1.78 | 1.71 | 6.51 | 6.04 | 0 | 0.11 | 4.16 | 1.30 |
| | Muscle | Vm | Vp | Ta | Tm | Tp | GH | MH | JO | JC |
| | Force | 2.67 | 0 | 1.76 | 1.72 | 1.14 | 0 | 0 | 9.00 | 0 |
| /i/ | Muscle | GGa | GGm | GGp | HG | SG | SLa | SLp | IL | Va |
| | Force | 0.62 | 0.51 | 3.36 | 0 | 0 | 0.02 | 0 | 0.78 | 0 |
| | Muscle | Vm | Vp | Ta | Tm | Tp | GH | MH | JO | JC |
| | Force | 0.41 | 0 | 0 | 0.52 | 0 | 0 | 3.01 | 0 | 14.0 |
| /u/ | Muscle | GGa | GGm | GGp | HG | SG | SLa | SLp | IL | Va |
| | Force | 0 | 0 | 0 | 0 | 3.12 | 0 | 0.09 | 0 | 0 |
| | Muscle | Vm | Vp | Ta | Tm | Tp | GH | MH | JO | JC |
| | Force | 0 | 0.67 | 0 | 0.87 | 0 | 0 | 0 | 0 | 3 |
| /e/ | Muscle | GGa | GGm | GGp | HG | SG | SLa | SLp | IL | Va |
| | Force | 0 | 0 | 1.79 | 0 | 0 | 0 | 0 | 0.76 | 0 |
| | Muscle | Vm | Vp | Ta | Tm | Tp | GH | MH | JO | JC |
| | Force | 0 | 0 | 0 | 0.82 | 0 | 0 | 0 | 0 | 0 |
| /o/ | Muscle | GGa | GGm | GGp | HG | SG | SLa | SLp | IL | Va |
| | Force | 0 | 0.48 | 0 | 5.13 | 10.0 | 0 | 0.07 | 0 | 0 |
| | Muscle | Vm | Vp | Ta | Tm | Tp | GH | MH | JO | JC |
| | Force | 0 | 0 | 0 | 1.68 | 1.50 | 0 | 0 | 0 | 0 |

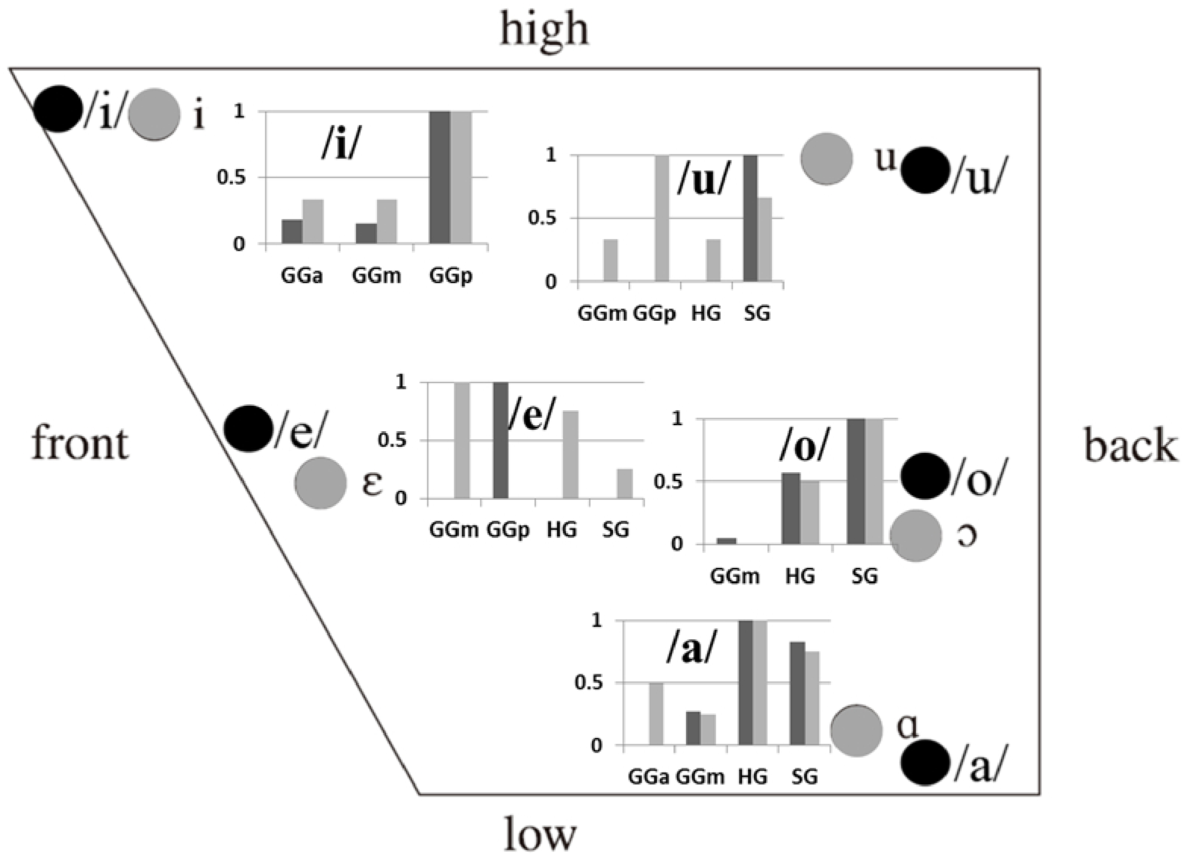

Figure 5.3: Comparison between estimated extrinsic tongue muscle activations and EMG measurement. The vowel positions are referred from [81], where black dots represent the positions of Japanese vowels and gray dots show the positions of partial English vowels that are close to the Japanese vowels in articulatory space. The black bars are the normalized results obtained by the proposed method, and the gray bars are the corresponding normalized EMG measurements.

Table 5.3: Comparison of formant between observed and generated. (Unit: Hz)

| Vowel | | /a/ | /i/ | /u/ | /e/ | /o/ |
|---|---|---|---|---|---|---|
| Observed | The first formant | 620 | 359 | 344 | 414 | 415 |
| | The second formant | 1243 | 1981 | 1431 | 1657 | 967 |
| | The third formant | 2258 | 2748 | 2150 | 2312 | 2200 |
| Synthesized | The first formant | 687 | 327 | 400 | 564 | 572 |
| | The second formant | 1037 | 2109 | 1032 | 1586 | 823 |
| | The third formant | 2252 | 2506 | 2173 | 2412 | 2443 |

The degree of accuracy of transmission line model used to calculate the resonance is not enough. However, the object of this evaluation in the acoustic space is to inspect whether the learning process using somatosensory feedback (articulation target) can generate similar acoustics to some extent. A small scale perceptual experiment was conducted by three native Japanese speakers. As a result, all the subjects can recognize the vowels correctly. This experiment proved that the learning process using the somatosensory feedback is effective.

### 5.1.1 Function of intrinsic muscle

The tongue muscles can be divided into two categories: extrinsic and intrinsic muscles. Extrinsic muscles originate from outside of the tongue (mandible, hyoid bone, styloid process, and the soft palate) and insert into the tongue. Intrinsic muscles are contained entirely within the tongue without external attachments. It has been suggested that the movements of the tongue body are accomplished by the contraction of the extrinsic lingual muscles, while the shape of the tongue surface is determined by the contraction of the intrinsic lingual muscles [98]. So far, most of the experimental studies concentrated on the functions of extrinsic muscles [23, 28, 14]. Only few of them shed light on the functions of intrinsic muscles [26], because it is difficult to measure the activations by using measurement devise such as EMG. Based on a physiological articulatory model, Fang *et al.* investigate the functions of intrinsic muscles by manually adjust muscle activations for Japanese vowels, and found that the co-contraction of transverse and vertical muscles plays the role of elongating the tongue in longitudinal direction [29]. However, as we have mentioned in Section 3.4.4, it is difficult to explore muscle functions systematically by using the "trail-and-error" method. The automatic estimation method proposed in this study provides us a convenient approach to investigate the functions of intrinsic muscles. As one can see from Table 5.2 that intrinsic muscles were activated for all the five Japanese vowels, which showed their importance for vowel production. However it is still difficult to imaging the degree of importance of intrinsic tongue muscles for vowel articulation.

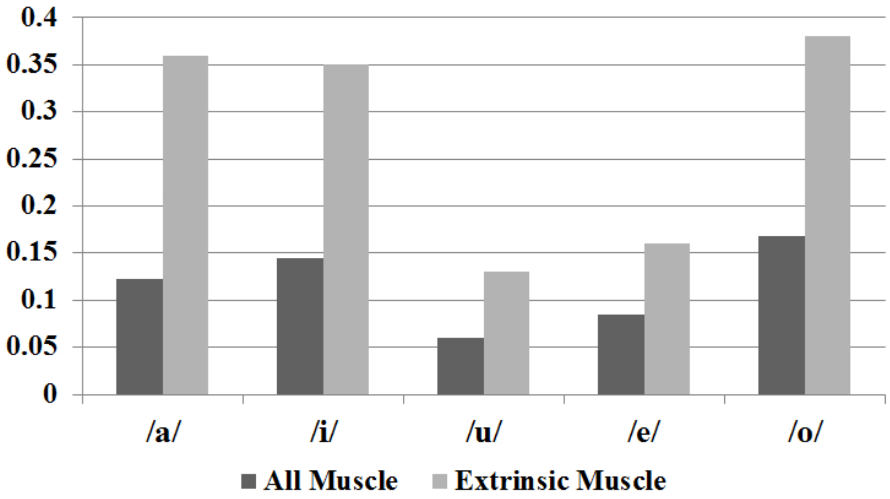

Figure 5.4: Distance between the realized posture and target posture. Gray bars and black bars show the distances using only the extrinsic muscles and all the muscles, respectively. (Unit: cm)

Therefore the function of intrinsic muscles in vowel production is explored quantitatively.

A numerical experiment was conducted to test the accuracy of realizing the postures by using only the extrinsic muscles, which include GGa, GGm, GGp, HG, SG, GH, MH, JO and JC. The distances between target posture and realized posture by using the extrinsic muscles and all muscles are shown in Figure 5.4. From this figure, one can see that the extrinsic muscles can control the articulators closer to the targets and by adding the intrinsic muscles the targets can be achieved with high accuracy. Figure 5.5 shows the realized positions by using only extrinsic muscles and all muscles are available. From this figure, one can see that by the extrinsic muscles can drive the model closer to the targets, when the intrinsic muscles are used, the targets can be realized with high accuracy. To realize all the five Japanese vowels, the intrinsic muscles: the superior longitudinal, inferior longitudinal, verticalis and transversus muscles are activated with different activation levels. The results shown in Figure 5.4 proved that the intrinsic muscles are very important in fine control of the vowel postures.

## 5.2 Evaluation of feedforward mapping

The previous section showed that articulatory targets have been achieved well for five Japanese vowels using the target based learning process. During the learning process, each iteration step generates one set of muscle activation pattern and its corresponding articulatory posture. Combing the learning data set together with the data set that was used for PCA, we obtained 8630 simulations, and used these data to train the feedforward

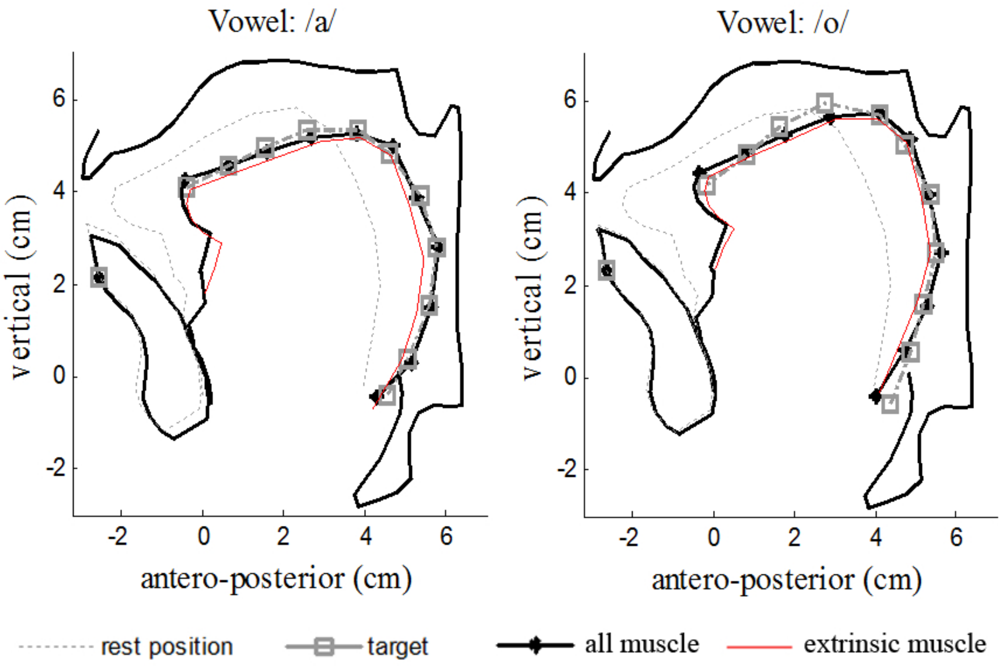

Figure 5.5: Comparison of the achievements between using extrinsic muscle and all muscles are available, left and right panels show vowel /a/ and vowel /o/, respectively.

mapping from the articulatory posture to muscle activation pattern. The distributions of 11 points on the midsagittal tongue surface are shown in Figure 5.6. From this figure one can see that the data set covers almost all the possible articulatory postures.

In this study, a two-layer artificial neural network was used to obtain the mapping from articulatory target to muscle activation. For training the neural network, the input vector was the 6-dimensional PCA components, and the output vector was the 18-dimensional muscle activations. About 70% of the simulations were randomly selected as training data set, and the remained 30% were used as test data set. A set of experiments were conducted using a number of neural networks with different configurations. The numbers of neurons in the hidden layer were set as 5, 10, 15, 20 and 25, respectively. The transfer functions for the hidden layer and output layer were set as a different combination of tansig (hyperbolic tangent sigmoid transfer function) and purelin (Linear transfer function). The best configurations were obtained by choosing the smallest prediction error in the opening test. As a result, the following configuration was used for the neural network to achieve the best performance. The transfer functions of tansig and purelin were used for the hidden layer and output layer, respectively, and the number of neurons was set to 20 in the hidden layer. In the opening test, predicting muscle activation pattern from articulatory target using the trained neural network, the average error was 0.003 in the muscle activation level. The average error is small, but it is difficult to assess the extent of the discrepancy

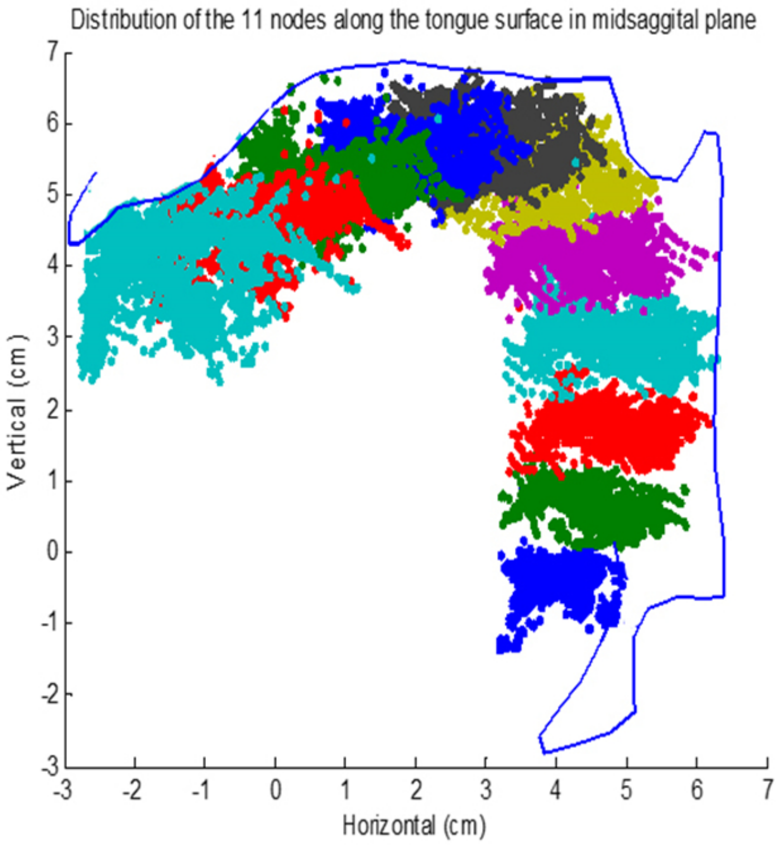

Figure 5.6: Distribution of the 11 points along the tongue surface in the midsagittal plane of the simulations of target based learning. The area with different colors corresponds to the dispersion of individual tongue nodes.

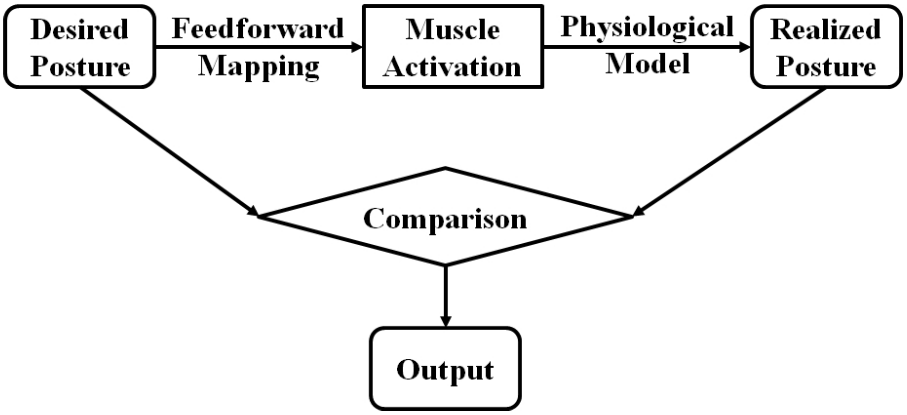

Figure 5.7: Flowchart of evaluation

from this error. We proposed the flowchart to evaluate the validity of the feedforward control (see Figure 5.7). Thirty percent of the data set (2580 simulations) not included in the training are used to evaluate the constructed feedforward mapping.

The evaluation is conducted as follows. Firstly, the feedforward control is used to generate muscle activation pattern from desired articulatory target. Secondly, the generated muscle activation pattern is input to the physiological articulatory model so that the articulatory model generates an articulatory posture. Finally, the distance between the articulatory target and generated posture is calculated. The distance is calculated using Equation 4.5. Figure 5.8 shows the distribution of the difference between the desired target and realized posture, where the horizontal axis shows the difference, the vertical axis on the left shows the occurrence of the distance, and the vertical axis on the right is the integration of distribution. The blue columns correspond to the left vertical coordinate, and the lines correspond to the right coordinate. Gray columns and dash line show the results using feedforward mapping. If assuming that the difference less than 0.25 cm is regarded as achieving the articulatory target, 88.9% of the test simulations satisfy the requirement. For those simulations in which the targets were not achieved, the feedback control can be implemented to control the model to the desired target and the evaluation is done in the next sub section.

## 5.3 Behavior of integrated control

The advantage of the integrated control strategy is that in the case the articulatory target is not achieved, the feedback control can assist the model to achieve the target by adjusting muscle activations. In this evaluation, the average distance larger than 0.25 cm was regarded as errors, muscle activations of 11.1% of the simulations need to be adjusted by feedback control. The muscle activation patterns were initiated by the output of

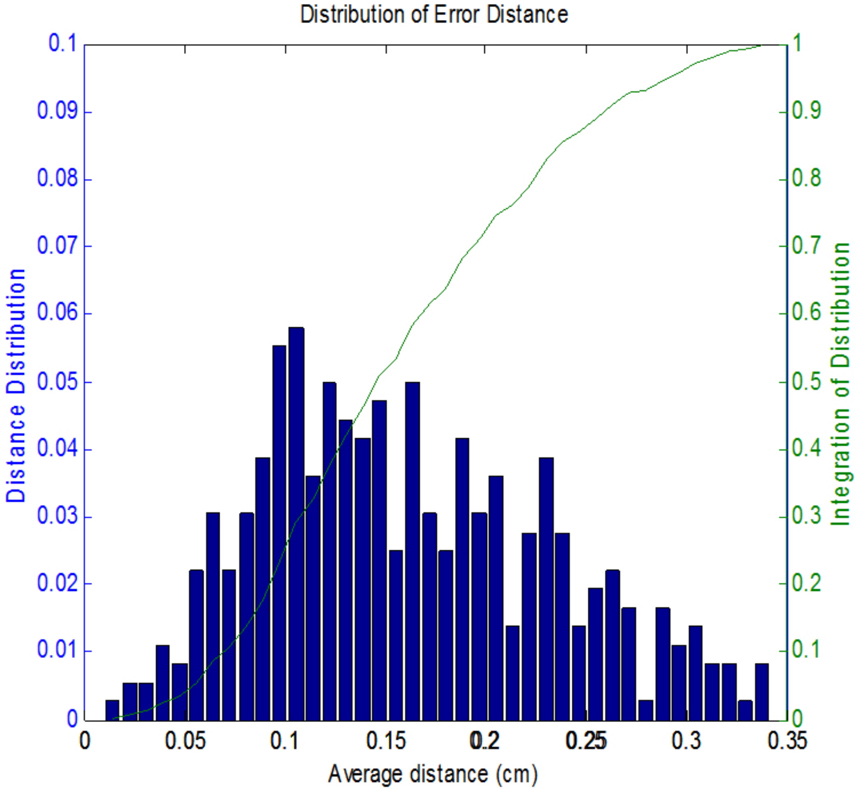

Figure 5.8: Distribution of distance between target position and realized ones.

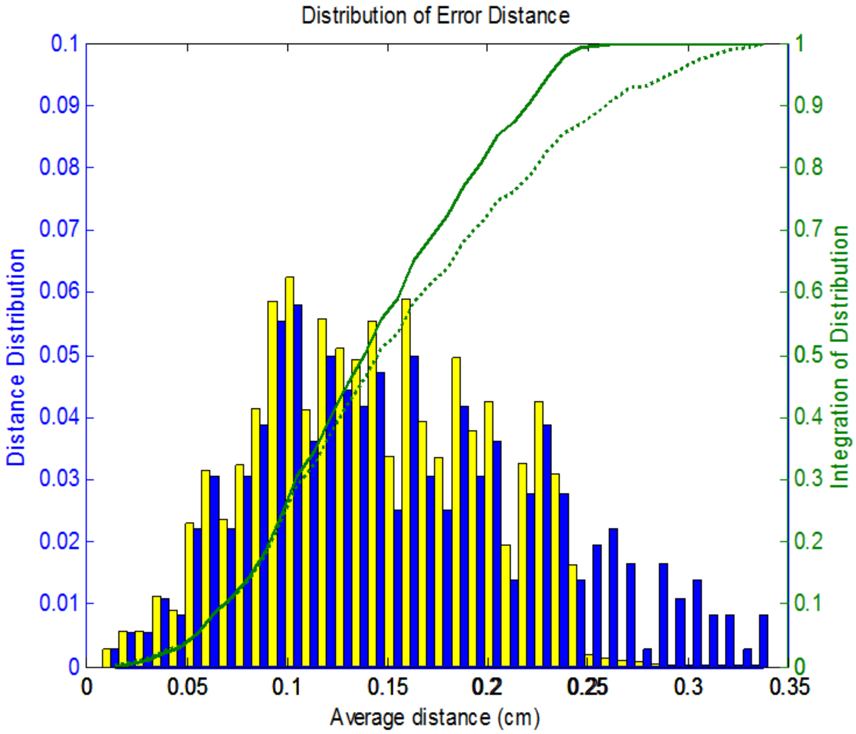

Figure 5.9: Distribution of distance between target position and realized ones (integrated control by set the threshold to 0.25cm ).

feedforward mapping. Figure 5.9 shows the comparison of distance distribution between feedforward control and integrated control. In these two figures, blue bars show the occurrence of the distance between desired posture and realized posture using feedforward control; yellow bars show the corresponding results of using integrated control. One can see that the occurrence of the distance less than 0.25cm or 0.16 cm is higher for the case with than without the feedback loop. Dash line and solid line show the integration of distribution of using the feedforward control and integrated control, respectively, and the latter one reached about 99.8% at the distance equal to 0.25 but the former reached about 88.9%.

If the threshold was set to 0.16cm, 36.4% of the simulations need to be adjusted by feedback control. A similar experiment was conducted as the threshold was set to 0.25. Figure 5.10 shows the result. From these results one can see that when the threshold decreased the accuracy by using integrated control can continue increasing. Accordingly, it is obvious that the integrated control greatly improves the control accuracy.

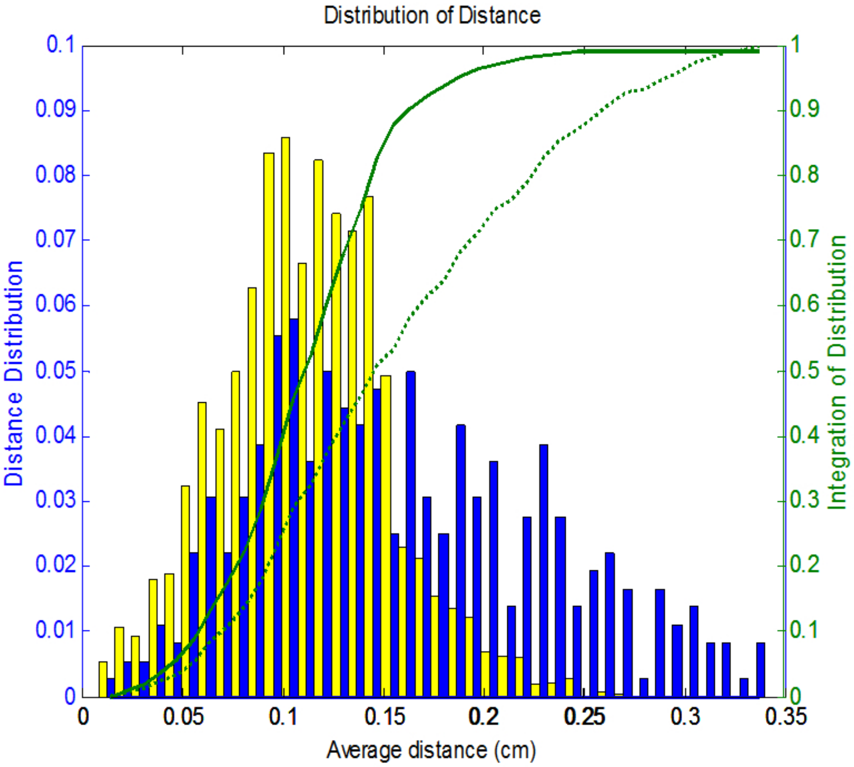

Figure 5.10: Distribution of distance between target position and realized ones (integrated control by set the threshold to 0.16cm ).

## 5.4 Adaptation experiment

In speech production, one of the most important ability of the human is that when the environment is changed human can adapt to the new environment. Perturbation experiments have proved that human can adapt to the external forces and generate the same articulatory target after learning [95, 109]. In this study, the feedforward control is acquired from feedback learning. As we have declared previously that the dynamic learning strategy enable the model the adaption capability when some perturbation is added. In order to evaluate this ability, a perturbation experiment was conducted by simulating the previous experiment [109]. Figure 5.11 shows the configuration of the experiments. When the articulatory targets are achieved, a sudden vertically downward external force of 5 Newton are exerted to the jaw. Because of the perturbation, the articulators would deviate from its target. In this situation, the somatosensory feedback loop will launch to refine the muscle activation to adapt to the external force.

An experiment was conducted for vowel /i/ and vowel /o/. When the model reaches its target for vowel /i/ and vowel /o/ by the muscle activation patterns shown in Table 5.2, an external vertically download force of 5 Newton were added to the model. The model deviates from its target due to the external force, then the feedback loop works to refine the muscle activation pattern and adapt to the external force. The adaption results are shown in Figure 5.12, in this figure, green lines with squares show the target; black lines with stars show that the target postures were achieved before adding external force; after the external force was added the model moved to the position shown as dash lines; by using feedback learning loop, the target could be achieved once more by a newly learned muscle activation pattern. Table 5.4 shows the comparison of the average distance between the situations of “Before perturbation” and “After adaption”.

The newly obtained muscle activation patterns are shown in Table 5.5, where the numbers printed with blue color show muscle forces that have changed. The jaw closer muscle group increase the muscle force to compensate for the external downward force. For vowel /i/ the activation of Va and MH are increased to elevate the tongue tip and tongue body. As to vowel /o/, muscle GGp, SG, Tm and Tp increase their activation to elevate the tongue body. From this experiment, we can conclude that the control strategy can adapt to external force.

## 5.5 Summary and discussion

In this chapter, we evaluate the proposed control strategy by assessing the objects proposed in Chapter 3. We have implemented the idea proposed by Kawato *et al.*[66] that feedforward mapping function is constructed by a learning process using a somatosensory

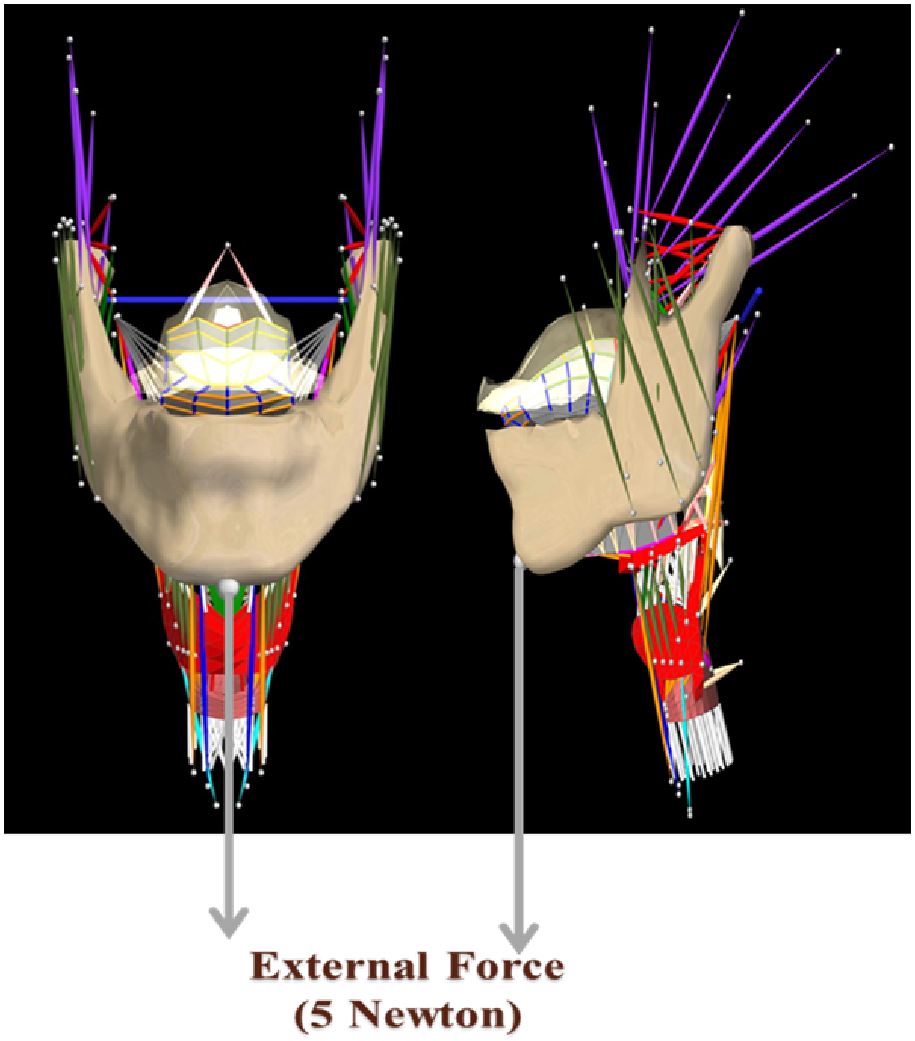

Figure 5.11: Setting of perturbation experiment.

Table 5.4: Average distance to target (calculated by using Eq. 4.5) before perturbation and after adaption. (Unit: cm)

| Vowel | Before perturbation | After adaption |
|---|---|---|
| /i/ | 0.145 | 0.151 |
| /o/ | 0.168 | 0.162 |

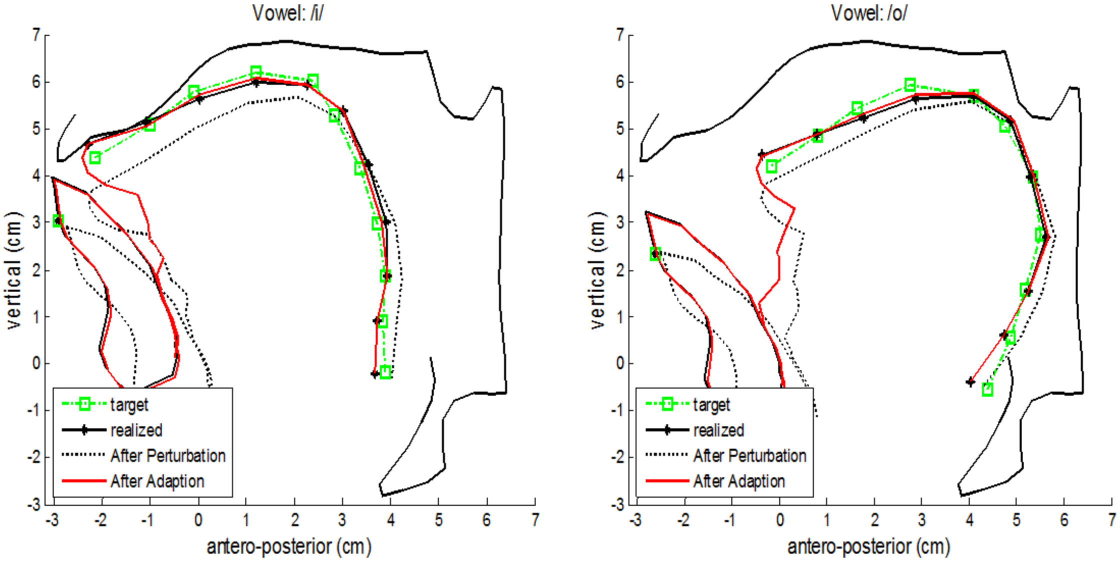

Figure 5.12: Adaption of the external force, left and right panels show vowel /i/ and /o/, respectively.

Table 5.5: Comparison of muscle activation patterns before and after jaw perturbation. The active muscle force of JO and JC is the sum of the active force included in the muscle groups. The muscle forces printed in blue color show the muscle forces that have changed. (Unit: Newton)

| /i/ | Muscle | GGa | GGm | GGp | HG | SG | SLa | SLp | IL | Va |
|---|---|---|---|---|---|---|---|---|---|---|
| | Normal | 0.62 | 0.51 | 3.36 | 0 | 0 | 0.02 | 0 | 0.78 | 0 |
| | Adaption | 0.62 | 0.51 | 3.36 | 0 | 0 | 0.02 | 0 | 0.78 | 0.35 |
| | Muscle | Vm | Vp | Ta | Tm | Tp | GH | MH | JO | JC |
| | Normal | 0.41 | 0 | 0 | 0.52 | 0 | 0 | 3.01 | 0 | 14.0 |
| | Adaption | 0.41 | 0 | 0 | 0.52 | 0 | 0 | 4.23 | 0 | 21.3 |
| /o/ | Muscle | GGa | GGm | GGp | HG | SG | SLa | SLp | IL | Va |
| | Normal | 0 | 0.48 | 0 | 5.13 | 10.0 | 0 | 0.07 | 0 | 0 |
| | Adaption | 0 | 0.48 | 0.3 | 5.13 | 11.5 | 0 | 0.07 | 0 | 0 |
| | Muscle | Vm | Vp | Ta | Tm | Tp | GH | MH | JO | JC |
| | Normal | 0 | 0 | 0 | 1.68 | 1.50 | 0 | 0 | 0 | 0 |
| | Adaption | 0 | 0 | 0 | 1.73 | 1.66 | 0 | 0 | 0 | 8.4 |

feedback loop.

Feedback learning loop was evaluated using articulatory targets of five Japanese vowels, as a result, the obtained muscle activation patterns are consistent with the anatomical knowledge and measurement EMG signals. Feedforward mapping function was constructed by the feedback learning loop and was evaluated to be feasible according to an open set test. As declared in Chapter 3, the integrated control strategy should have a higher accuracy than feedforward mapping. This hypothesis was proved to be true according to the comparison of accuracy by the feedforward mapping and integrated control. According to the perturbation experiment, one can see that the control strategy has the ability to adapt to the external force, which is similar to the human mechanism.

Theoretically, the same articulatory posture may be generated by different muscle activation patterns because muscle activations have more degrees-of-freedom than articulatory posture. In order to obtain the optimal activation, economy of energy is typically used as the optimality criterion. Stavness *et al.* proposed a method to find muscle activations to control the tongue tip to move along given target trajectories by considering minimum muscle activation as a constraint [38]. In this study, although it is difficult to guarantee that the obtained muscle activation has a minimum activation cost, the result can be regarded as a good approximation of the minimum because in each iteration step the added muscle has the greatest contribution to the target vector.

# Chapter 6

# Summary and Future Work

## 6.1 Summary of this thesis

As we have introduced in the Chapter 1, it is difficult to investigate the mechanism of speech production by current techniques. If we can model the biomechanical characteristics of articulatory system, it is possible to investigate the speech production mechanism by model simulation. The unknown mechanism can be uncovered by control the model to achieve the observable articulations.

Therefore, we have to construct a physiological articulatory model which can model the biomechanical characteristics of articulators and musculatures. The construction of the physiological articulatory model was not from zero but constructed based on the previously studies. The improvements of the model are as follows: 1) The intrinsic tongue muscles are subdivided into smaller control units according to their functions, which would essentially improve the control accuracy. 2) The discrete FEM model was substituted by continuum FEM model, which make the mechanical characteristics inside the meshes closer to realities. The evaluations on response time and convergence indicated that the present physiological articulatory model has a better performance than the previous model.

To investigate the mechanism of speech production based on a physiological articulatory model, we need not only a model that can faithfully model the characteristics of human articulators but also a control strategy that closer to that implemented in humans.

To overcome the disadvantages of model control strategies proposed in the literatures, the integrated control strategy including both feedforward mapping and feedback learning loop is realized, where feedforward mapping was constructed by using feedback learning loop. According to model simulation, feedforward mapping can be constructed by the feedback learning loop. Feedback learning loop was evaluated by using articulatory targets of five Japanese vowels, as a result, the obtained muscle activation patterns are consistent with the anatomical knowledge and measurement EMG signals. Feedforward mapping function was evaluated to be feasible according to an open set test. As declared in Chapter 3, the integrated control strategy should have a higher degree of accuracy than feedforward mapping. This hypothesis was proved to be true according to the comparison of accuracy by using feedforward mapping and integrated control.

Perturbation experiment showed that control strategy has the ability to adapt to external force, which make the control more robust to different environment.

## 6.2 Contributions

Due to the work addressed in the previous chapters, there are a number of contributions made for revealing the mechanisms of speech productions based on a physiological articulatory model. Firstly, articulation is easier to be observed than corresponding motor command. Investigation of muscle activation patterns based on a physiological articulatory model was first proposed by Fang *et al.*[29]. In their method, the muscle activation patterns were obtained by a trial-and-error method, which strongly depends on the experience of the researcher, and it is difficult to explore all the articulations if not impossible. In this thesis, the proposed feedback learning loop is an automatic method, which can be used to find muscle activation pattern for any given articulation. Using this method it is more convenient to investigate motor command for arbitrary articulation. Furthermore, because there is one-to-many problem that different muscle activation patterns can generate the same articulations, malfunction of a specific muscle may be compensated for by the combination of the other muscles. The proposed automatic method can be easily implemented to investigate the compensation mechanism. In biomedical application, if specific muscle is malfunctioned for some reasons, the proposed method can be implemented and explore whether and how the coordination of the other muscles can compensate for it.

Secondly, in a muscular-hydrostat system, such as the tongue, muscle orientations change during articulation, which results in a variation of individual muscles function. In this study, a dynamic PCA workspace was constructed to estimate individual muscle functions during articulation. This dynamic PCA workspace was proved effective by using it to estimating muscle activations for the five Japanese vowels.

Thirdly, the midsagittal contour including the tongue and jaw was used as the articulatory target, instead of using three crucial points. We expect this to improve the accuracy of model control for speech production, because the detailed characteristics of a speech sound depend on whole vocal tract shape rather than constriction alone. Contour control can avoid the possible conflictions that may happen in independent crucial points control [50, 51], because the components of articulators are coupled with each other, and the muscle that control the tongue tip may affect the tongue dorsum and vice versa.

Fourthly, the integrated control strategy can greatly improve the control accuracy than the feedforward mapping. As we have mentioned previously, if the degree of accuracy cannot satisfy the requirement, the feedforward mapping can do nothing on it. By using the integrated control strategy, the feedback learning loop can be implemented to find muscle activation patterns to realize precise control.

Fifthly, the integrated control strategy can compensate to the perturbed external force, which endows the model the adaptation ability to the change of environment.

Finally, in the neural computational models constructed for simulating speech acquisition process, such as the models by built by Guenther *et al.* [72] and Kröger *et al.* [71], geometric articulatory models [74, 75], employed to imitate the auditory feedback and somatosensory feedback can be substituted by the physiological articulatory model, because physiological articulatory model possesses the physical properties of articulators and muscles can provide not only the auditory (acoustic) feedback but also somatosensory feedback.

## 6.3 Future work

In the present thesis, a 3D physiological articulatory model has been elaborated, and an integrated control strategy including feedforward mapping and feedback learning loop was proposed and implemented in model control. To further investigate the mechanism of speech production, the following work should be done.

Firstly, in the speech production framework described in Chapter 3, auditory feedback is very important as well as somatosensory feedback. In the current study, only the somatosensory feedback is implemented. In the future, the auditory feedback will be implemented to correct the articulatory target according to the difference between the desired speech sounds and generated speech sounds. Similar to the integrated control strategy implemented in this study, the motor plan mapping can be constructed by the auditory feedback learning process. When both of the auditory feedback and somatosensory feedback are implemented in the control strategy, the control strategy will be even closer to what human implemented.

Secondly, in this study, a two layer feedforward neural network is implemented. In order to promote the prediction accuracy, more appropriate neural network will be implanted to represent the feedforward mapping function.

Thirdly, we need to extend the current control strategy for the control of consonants. In the current study, 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.

Fourthly, in order to elaborate the physiological articulatory model multiple observable articulatory parameters like electropalatographic data need to be compared to the model. According to the discrepancy of model simulation and observable parameters, the model parameters can be optimized.

Fifthly, as we have introduced in Section 2.3.1, the tongue tissue is segmented into 240 meshes in the model. The course of the muscles defined in the physiological articulatory model have to pass through the nodes of the meshes, which make them approximate not exactly the same as humans. There are two main reasons which restrict the accuracy of the model: 1) The increasing of the FE mesh number will increase the computational cost of model simulation. 2) The anatomical knowledge cannot provide detail information of muscle courses. With the development of computer science, the computational ability of PC will improve greatly, and the division of more meshes will result in the course of muscles more faithful to human. The detail locations and orientations of tongue muscle fibres can be obtained by dissection of tongue from cadavers [99], this kind of studies will further our understanding of anatomical structures of tongue muscles and lead to a more accurate model.

It had been believed that the neuromotor control of all extrinsic lingual muscles and intrinsic lingual muscles (except the palatoglossus muscle) were innervated by the hypoglossus nerve alone [102]. However recently, Saigusa *et al.* reported that the superior longitudinal and the inferior longitudinal lingual muscles were innervated by motor fibers of the trigeminal nerve in the human adult subjects [103]. If so, impairment of trigeminal nerve will result in the disfunction of superior longitudinal and interior longitudinal muscle. The effect of trigeminal nerve impairment in articulation can be investigated based on the physiological articulatory model.

# Bibliography

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

[2] A. M. Liberman and I. G. Mattingly: "The motor theory of speech perception revised," Cognition, Vol. 21, pp. 1–36 (1985).

[3] P. F. Ferrari, V. Gallese, G. Rizzolatti and L. Fogassi: "Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex," European Journal of Neuroscience, Vol. 17, pp. 1703–1714 (2003).

[4] G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese and G. Rizzolatti: "Understanding motor events: a neurophysiological study," Experimental Brain Research, Vol. 91, pp. 176–180 (1992).

[5] M. Iacoboni and M. Dapretto: "The mirror neuron system and the consequences of its dysfunction," Nature Reviews Neuroscience, Vol. 7, pp. 942–951 (2006).

[6] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura: "A hidden semi-markov model-based spech synthesis system," IEICE Trans. Inf. & Syst., Vol. E90–D, No. 5, pp. 825–834 (May 2007).

[7] H. Zen, K. Tokuda and A. W. Black: "Statistical parametric speech synthesis," Speech Communication, Vol. 51 Issue 11, pp. 1039–1064, (Nov. 2009).

[8] T. Schultz and M. Wand: "Modeling coarticulation in EMG–based continuous speech recognition," Speech Communication Vol. 52, pp. 341–353 (2010).

[9] M.J.F Gales: "Maximum likelihood linear transformations for HMM–based speech recognition," Computer Speech & Language, Vol. 12, Issue 2, pp. 75–98 (April 1998).

[10] B. H. Juang and L. R. Rabiner: "Hidden Markov models for speech recognition," Technometrics Vol. 33, Issue 3, pp. 251–272 (1991).

[11] S. A. Huettel, A. W. Song and G. McCarthy: "Functional Magnetic Resonance Imaging (2 ed.)," Massachusetts: Sinauer, ISBN 978–0-87893–286–3, (2009).

[12] B. Denby, T. Schultz, K. Honda, T. Hueber, J.M. Gilbert and J.S. Brumberg: "Silent speech interfaces," Speech Communication, Vol. 52, pp. 270–287 (2010).

[13] W. R. Hendee and C. J. Morgan: "Magnetic Resonance Imaging Part I–Physical Principles," West J Med. 141 (4), pp. 491–500 (1984).

[14] K. Honda, H. Takemoto, T. Kitamura, S. Fujita and S. Takano: "Exploring human speech production mechanisms by MRI," IEICE Info. Syst. E87–D, pp. 1050–1058 (2004).

[15] O. Engwall: "Combining MRI, EMA and EPG measurements in a three-dimensional tongue model," Speech Communication, 41(2-3), pp. 303–329 (2003).

[16] M. Stone and A. Lundberg: "Three-dimensional tongue surfaces from ultrasound images," MedicalImaging: Physiology and Function from Multidimensional Images, 2709, pp. 168–179 (1996).

[17] M. Stone: "A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data," Journal of the Acoustical Society of America, Vol. 87(5), pp. 2207–2217 (1990).

[18] B. C. Sonies, T. H. Shawker, T. E. Hall, L. H. Gerber and S. B. Leighton: "Ultrasonic visualization of tongue motion during speech," Journal of the Acoustical Society of America, Vol. 70, pp. 683 (1981).

[19] F. A. Duck: "Physical Properties of Tissues: A Comprehensive Reference Book," (Academic, London) (1990).

[20] J. M. Gerard, J. Ohayon, V. Luboz, P. Perrier and Y. Payan: "Non-linear elastic properties of the lingual and facial tissues assessed by indentation technique. Application to the biomechanics of speech production," Med. Eng. Phys., Vol. 27, pp. 884–892 (2005).

[21] K. Miyawaki: "A study on the musculature of the human tongue–observations on the transparent preparations of serial sections," Ann. Bull. RILP 8, pp. 23–49 (1973).

[22] H. Takemoto: "Morphological analyses of the human tongue musculature for three-dimensional modeling," J. Speech Lang. Hear. Res., Vol 44, pp. 95–107 (2001).

[23] T. Baer, J. Alfonso and K. Honda: "Electromyography of the tongue muscle during vowels in /əpvp/ environment," Ann. Bull. RILP, 7, pp. 7–18 (1988).

[24] M. Stone, E. P. Davis, A. S. Douglas, M. NessAiver, R. Gullapalli, W. S. Levine and A. Lunberg: "Modeling the motion of the internal tongue from tagged cine–MRI images," J. Acoust. Soc. Am., Vol. 109, pp. 2974–2982 (2001).

[25] M. Kumada, M. Niitsu, S. Niimi and H. Hirose: "A study on the inner structure of the tongue in the production of the 5 Japanese vowels by tagging snapshort MRI," Ann. Bull. RILP, Vol. 26, pp. 1–13 (1992).

[26] S. Niimi, M. Kumada and M. Niitsu: "Functions of tongue-related muscles during production of the five Japanese vowels," Ann. Bull. RILP, Vol. 28, pp. 33–40 (1994).

[27] M. Niitsu, M. Kumada, S. Niimi and Y. Itai: "Tongue movement during phonation: A rapid quantitative visualization using tagging snapshot MRI imaging," Ann. Bull. RILP, Vol. 26, pp. 149–156 (1992).

[28] S. Takano and K. Honda: "An MRI analysis of the extrinsic tongue muscles during vowel production," Speech Commun., Vol. 49, pp. 49–58 (2007).

[29] Q. Fang S. Fujita, X. Lu and J. Dang: "A model-based investigation of activations of the tongue muscles in vowel production," Acoust. Sci.& Tech., Vol. 30, pp. 277–287 (2009).

[30] S. Buchaillard, P. Perrier and Y. Payan: "A biomechanical model of cardinal vowel production: muscle activations and the impact of gravity on tongue positioning" J. Acoust. Soc. Am. Vol. 126, pp. 2033–2051 (2009).

[31] F. H. Guenther: "A modeling framework for speech motor development and kinematic articulator control," Proceedings of the XIIIth International Congress of Phonetic Sciences, Vol. 2, pp. 92–99 (1995).

[32] F. H. Guenther, M. Hampson and D. A. Johnson: "Theoretical investigation of reference frames for the planning of speech movements," Psychological Review, Vol. 105, pp. 611–633 (1998).

[33] S. J. Perkell, F. H. Guenther, H. Lane, M. L. Matthies, P. Perrier, J. Vick, R. Wilhelms-Tricarico and M. Zandipour: "A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss," Journal of Phonetics, Vol. 28, pp. 233–272 (2000).

[34] W. L. Nelson: "Physical Principles for Economies of Skilled Movements," Biol. Cybern., Vol. 46, pp. 135–147 (1983).

[35] J. S. Perkell, M. Zandipour, M. L. Matthies and H. Lane: “Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues,” J. Acoust. Soc. Am., Vol. 112 (4), pp. 1627–1641 (2002).

[36] J. S. Perkell and M. Zandipour: “Economy of effort in different speaking conditions. II. Kinematic performance spaces for cyclical and speech movements,” J. Acoust. Soc. Am., Vol. 112 (4), pp. 1642–51 (2002).

[37] I. Stavness, B. Gick, D. Derrick and S. Fels, “Biomechanical modeling of English /r/ variants,” J. Acoust. Soc. Am., Vol. 131 (5), pp. 355–360 (May 2012).

[38] I. Stavness, J. E. Lloyd and S. Fels: “Automatic prediction of tongue muscle activations using a finite element model,” Journal of Biomechanics, Vol. 45, pp. 2841–2848 (2012).

[39] J. Mielke, A. Baker and D. Archangeli: “Variability and homogeneity in American English /*r*/ allophony and /*s*/ retraction,” in Laboratory Phonology 10 (Phonology and Phonetics 4–4), edited by C. Fougeron, B. K*ü*uhnert, M. d’ Imperio and N. Vall*é*e (de Gruyter Mouton, Berlin), pp. 699–730 (2010).

[40] K. N. Stevens: “The quantal nature of speech: evidence from articulatory-acoustic data,” In Human communication, a Unified View (P. B. Denes & E. E. David, editors), New York: McGraw-Hill, pp. 51–56 (1972).

[41] K. N. Stevens: “On the quantal nature of speech,” Journal of Phonetics, Vol. 17, pp. 3–45 (1989).

[42] O. Fujimura and Y. Kakita: “Remarks on quantitative description of lingual articulation,” in Frontiers of Speech Communication Research, edited by B. Lindblom and S. *Ö*hman (Academic, San Diego), pp. 17–24 (1979).

[43] P. Badin, P. Perrier, L. J. Boë and C. Abry: “Vocalic nomograms: Acoustic and articulatory considerations upon formant convergence,” J. Acoust. Soc. Am., Vol. 87, pp. 1290–1300 (1990).

[44] J. Wei, X. Lu and J. Dang: ”A model-based learning process for modeling coarticulation of human speech,” IEICE, Vol. E90-D Issue 10, pp. 1582–1591, Oct., (2007).

[45] R. Otsuka, T. Ono, Y. Ishiwata and T. Kuroda: “Respiratoryrelated genioglossus electromyographic activity in response to head rotation and changes in body position,” Angle Orthod., Vol. 70, pp. 63–69 (2000).

[46] S. Fujita, J. Dang, N. Suzuki and K. Honda: “A computational tongue model and its clinical application,” Oral Sci. Int., Vol. 4, pp. 97–109 (2007).

[47] I. K. Stavness: "Byte Your Tongue: A Computational Model of Human Mandibular-Lingual Biomechanics for Biomedical Applications," Doctor thesis, THE UNIVERSITY OF BRITISH COLUMBIA (2010).

[48] M. R. McNeil, W. F. Katz, T. R. D. Fossett, D. M. Garst, N. J. Szuminsky, G. Carter and K. Y. Lim: "Effects of Online Augmented Kinematic and Perceptual Feedback on Treatment of Speech Movements in Apraxia of Speech," Folia Phoniatr Logop, Vol 62, pp. 127–133 (2010).

[49] J. S. Levitt and W. F. Katz: "The Effects of EMA-based Augmented Visual Feedback on the English Speakers' Acquisition of the Japanese Flap: a Perceptual Study," INTERSPEECH, Makuhari, Chiba, Japan, pp. 26–30 (Sep. 2010).

[50] J. Dang and K. Honda: "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," J. Phonetics, Vol. 30, pp. 511–532 (2002).

[51] J. Dang and K. Honda: "Construction and control of a physiological articulatory model," J. Acoust. Soc. Am., Vol. 115(2), pp. 853–870 (2004).

[52] K. Miyawaki: "A study on the muscularture of the human tongue," Ann. Bull. RILP, Vol. 8, pp. 23–49 (1974).

[53] J. Warfel: "The Head, Neck, and Trunk," Philadelphia and London: Led & Febiger (1993).

[54] R. Laboissière,D. J. Ostry and A. G. Feldman: "The control of multimuscle system: Human jaw and hyoid movement," Biol. Cybern., Vol. 74, pp. 373–384 (1996).

[55] J. Dang and K. Honda: "A physiological model of a dynamic vocal tract for speech production," Acoust. Sci. & Tech., Vol. 22, pp. 415–425 (2001).

[56] J. A. SeiKel, D. W. King and D. G. Drumright: "Anatomy & Physiology for speech, language, and hearing," Thomson Delmar Learning, (2005).

[57] K. Honda: "Physiological Processes of Speech Production," Handbook of speech processing, Springer, pp. 7–26 (2007).

[58] A. Morecki: "Modeling, mechanical description, measurements and control of the selected animal and human body manipulation and locomo-tion movement," in Biomechanics of EngineeringModeling, Simulation, Control, edited by A. Morecki (Springer, New York), pp. 1–28 (1987).

[59] R. Wilhelms-Tricarico: "Physiological modeling of speech production: Methods for modeling soft-tissue articulators," J. Acoust. Soc. Am., Vol. 97, pp. 3085–3098 (1995).

[60] V. Hill: “The heat of shortening and the dynamic constants of muscle,” *Proc. R. Soc. London, Ser. B* **126**, pp. 136–195 (1938).

[61] S. Fels, F. Vogt, K. van den Doel, J. Lloyd, I. Stavness and E. Vatikiotis-Bateson: “Artisynth: A biomechanical simulation platform for the vocal tract and upper airway”, Computer Science Dept., University of British Columbia, Tech. Rep. TR-2006-10 (2006).

[62] H. Hirai, J. Dang and K. Honda: “ A physiological model of speech organs incorporating tongue-larynx interaction,” J. Acoust. Soc. Jpn, Vol. 52 (12), pp. 918–928, (1995) (in Japanese)

[63] J. Dang and K. Honda: “Speech production of vowel sequences using a physiological articulatory model,” the $5^{th}$ International Conference on Spoken Language Processing (1998).

[64] J. E. Flege: “Effects of speaking rate on tongue position and velocity of movement in vowel production,” J. Acoust. Soc. Am. Vol. 84 (3), pp. 901–916 (Sep. 1998).

[65] Q. Fang, A. Nishikido and J. Dang: “Feedforward control of a 3D physiological articulatory model for vowel production,” Tsinghua Science And Technology, ISSN 1007-0214 11/18, Vol. 14(5), pp 617–622 (Oct. 2009).

[66] M. Kawato, K. Furukawa and R. Susuki: “A hierarchical neural-network model for control and learning of voluntary movement,” Biological Cybernetics, Vol. 57, pp. 169–185 (1987).

[67] B. Lindblom: “Spectrographic study of vowel reduction,” Journal of the Acoustical Society of America, Vol. 35, pp. 1773–1781 (1963).

[68] DR. V. Bergem: “Acoustic vowel reduction as a function of sentence accent , word stress and word class,” Speech Communication, Vol. 12, pp. 1–23 (1993).

[69] Y. Payan and P. Perrier: “Synthesis of V–V sequences with a 2D biomechanical tongue shape in vowel production,” Speech Commun. Vol. 22, pp. 185–206 (1997).

[70] P. Perrier, L. Lœvenbruck and Y. Payan: “Control of tongue movements in speech: the Equilibrium Point Hypothesis perspective,” Journal of Phonetics. Vol. 24, Issue 1, pp. 53–75, January (1996).

[71] B. J. Kröger, J. Kannampuzha and C. Neuschaefer-Rube: “Towards a neurocomputational model of speech production and perception,” Speech Communication, Vol. 51, pp. 793–809, (2008).

[72] F. H. Guenther, S. S. Ghosh and J. A. Tourville: "Neural modeling and imaging of the cortical interactions underlying syllable production," Brain and Language, Vol. 96, pp. 280–301 (2006).

[73] F. H. Guenther, T. Vladusich: "A neural theory of speech acquisition and production," Journal of Neurolinguistics, Vol. 25, Issue 5, pp. 408–422 (Sep. 2012).

[74] P. Birkholz, D. Jackèl: "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," In: Proc. Internat. Conf. on Speech and Language Processing, Interspeech, Jeju, Korea, pp. 1125–1128 (2004).

[75] S. Maeda: "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," In W. J. Hardcastle & A. Marchal (Eds.), Speech production and speech modeling, Boston: Kluwer Academic Publishers, pp. 131–149 (1990).

[76] J. Lloyd, I. Stavness, and S. Fels: "ArtiSynth: a fast interactive biomechanical modeling toolkit combining multibody and finite element simulation," In Soft Tissue Biomechanical Modeling for Computer Assisted Surgery, Springer-Verlag, **11**, pp. 355–394 (2012).

[77] TJR. Hughes: "The Finite Element Method: Linear Static and Dynamic Finite Element Analysis," Dover Publications: New York (2000).

[78] V. Sanguineti, R. Laboissière and Y. Payan: "A control model of human tongue movements in speech," *Biol. Cybern.* **77**, pp. 11–22 (1997).

[79] I. Stavness, J. E. Lloyd, Y. Payan and S. Fels: "Coupled hard-soft tissue simulation with contact and constraints applied to jaw-tongue-hyoid dynamics," *Int. J. Numer. Meth. Biomed. Engng.*, **27**, pp. 367–390 (2011).

[80] Q. Fang and J. Dang: "Physiological articulatory model for investigating speech production modeling and control," *VDM Verlag Dr. M*üller, pp. 117–118 (2009).

[81] P. Ladefoged and K. Johnson: "A course in phonetics," *Cengage Learning* (2011).

[82] J. H. Abbs: "Invariance and variability in speech production: A distinction between linguistic intent and its neuromotor implementation," In J. S. Perkell and D. H. Klatt (Eds.), Invariance and variability in speech processes. Hillsdale NJ: Erlbaum. pp. 202–219 (1986).

[83] J. H. Abbs and V. L. Gracco: "Control of complex motor gestures: Orofacial muscle responses to load perturbations of lip during speech," Journal of Neurophysiology, 51, pp. 705–723 (1984).

[84] J. F. Houde and M. I. Jordan: "Sensorimotor adaptation in speech production," Science 279, pp. 1213–1216 (1998).

[85] J. A. Jones and K. G. Munhall: "Perceptual calibration of F0 production: Evidence from feedback perturbation," J. Acoust. Soc. Am. 108, pp. 1246–1251 (2000).

[86] J. S. Perkell, M. L. Matthies, M. A. Svirsky and M. I. Jordan: "Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot 'motor equivalence' study, " J. Acoust. Soc. Am. 93, pp. 2948–2961 (1993).

[87] F. H. Guenther, M. Hampson and D. A. Johnson: "Theoretical investigation of reference frames for the planning of speech movements," Psychol. Rev. 105, pp. 611–633 (1998).

[88] B. Lindblom, J. Lubker and T. Gay: "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation," Journal of Phonetics, 7, pp. 147–161 (1979).

[89] E. L. Saltzman and K. G. Munhall: "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. Vol. 1, pp. 333–382 (1989).

[90] C. P. Browman and L. Goldstein: "Articulatory phonology: An overview," Phonetica 49, pp. 155–180 (1992).

[91] D. Purves, G. J. Augustine, D. Fitzpatrick and W. C. Hall: "Neuroscience," Fifth Edition, Sinauer Associates, Inc. (2012).

[92] LB. Walker and MD. Rajagopal: "Neuromuscular spindles in the human tongue," Anat Rec 133: 438 (1959).

[93] S. Cooper: "Muscle spindles in the intrinsic muscles of the human tongue," J. Physiol., 122: pp. 193–202 (1953).

[94] D. J. Ostry, P. L. Gribble, M. F. Levin and A. G. Feldman: "Phasic and tonic stretch reflexes in muscles with few muscle spindles: human jaw-opener muscles," Experimental Brain Research, Volume 116, Issue 2, pp 299–308 (Sep. 1997).

[95] S. Tremblay, D. M. Shiller and D. J. Ostry: "Somatosensory basis of speech production," Nature, Vol. 423 (19), pp.866–869 (June 2003).

[96] D. R. Lametti, S. M. Nasir and D. J. Ostry: "Sensory Preference in Speech Production Revealed by Simultaneous Alteration of Auditory and Somatosensory Feedback," The Journal of neuroscience, 32(27), pp. 9351–9358 (July 2012).

[97] P. Perrier: "Gesture planning integrating knowledge of the motor plant's dynamics: a literature review from motor control and speech motor control," Speech Plann. Dyn. pp. 191–238 (2012).

[98] WJ. Hardcastle: "Physiology of speech production," Academic, Press, London (1976).

[99] H. Saigusa, K. Tanuma, K. Yamashita, I. Aino, M. Saigusa and Seiji Niimi: "Fiber arrangements of the vertical lingual muscle in human adult subjects," Eur J Anat, 16 (3), pp. 177–183 (2012).

[100] Peter Branderud, Robert McAllister and Bo Kassling: "Methodological Studies of Movetrack: Coil placement procedures and their consequences for accuracy," Working Papers 43, Dept of Linguistics and Phonetics, Lund, Sweden, pp. 28–31 (1994).

[101] Maureen Stone and Thomas H. Shawker: "An ultrasound xxamination of tongue movement during swallowing," Dysphagia, Vol. 1 pp. 78–83 (1986)

[102] TW. Sadler: "Langmans medical Embryology," ($8^{th}$ Ed.), Lippincott Williams & Wilkins, Baltimore (2000).

[103] H. Saigusa, Tanuma, K. Yamashita, M. Saigusa and S. Niimi: "Nerve fiber analysis of the lingual nerve of the human adult tongue," Surg Radiol Anat, Vol. 28, pp. 59–65 (2006).

[104] R. O. Duda, P. E. Hart and D. G. Stork: "Pattern Classification," $2^{nd}$ A Wiley–Interscience Publication Hohn Wiley & Sons, Inc, pp. 114–117 (2001).

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

[106] R. L. Drake, W. Vogl, A. W. M. Mitchell, R. Tibbitts and P. Richardson: "Gray's anatomy for students," Churchill Livingstone, Elsevier Inc., Philadelphia, PA, 2nd edition, (2010).

[107] R. A. Schmidt and C. A. Wrisberg: "Motor learning and performance," Champaign, IL: Human Kinetics. ISBN 978-0-7360-4566-7. OCLC (2004).

[108] K. E. Bouchard, N. Mesgarani, K. Johnson and E. F. Chang: “Functional organization of human sensorimotor cortexfor speech articulation,” Nature, Vol. 495, pp. 327–332 (March 2013).

[109] JA. Kelso, B. Tuller, E. Vatikiotis-Bateson and CA. Fowler: “Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures,” J. Exp. Psychol. Hum. Percept. Perform. Vol. 10(6) pp. 812-32(Dec. 1984).

# Publications

## Journal

[1] **X. Wu** and J. Dang: “A Control Strategy of a Physiological Articulatory Model for Speech Production,” Journal of Chinese Linguistics, 2014 (Accepted in March 2014).

[2] **X. Wu**, J. Dang and I. Stavness: “An iterative method to estimate muscle activations with a physiological articulatory model,” Acoust. Sci. & Tech. (Accepted in January 2014).

[3] **X. Wu**, J. Wei and J. Dang: “Study of Control Strategy Mimicking Speech Motor Learning for a Physiological Articulatory Model,” Journal of Signal Processing (July 2011).

## International Conferences

[4] D. Huang, **X. Wu**, J. Wei, H. Wang, C. Song, Q. Hou and J. Dang: “Visualization of Mandarin Articulation by using a Physiological Articulatory Model,” Proc. APSIPA, Kaohsiung (October 2013).

[5] **X. Wu**, Q. Li, J. Wei and J. Dang: “Simulation of Speech Babbling Based on a Physiological Articulatory Model,” Proc. NCSP, pp. 183-186, Tianjin (2011)

[6] **X. Wu**, Q. Fang and J. Dang: “Investigation of Muscle Activation in Speech Production Based on an Articulatory Model,” Proc. ISCSLP, 330-334, Tainan (2010).

## Domestic Conferences

[7] **X. Wu**, Q. Fang and J. Dang: “Simulation of Speech Motor Learning Based on a 3D Physiological,” Proc. ASJ, pp. 353-354, Tokyo (Spring 2011).

[8] **X. Wu** and J. Dang: ”Investigation of speech motor control using a 3D physiological articulatory model,” Proc. ASJ, pp. 469-470, Tokyo (Spring 2010).

[9] **X. Wu**, Y. Wang and J. Dang: ”Investigation of speech production using a 3D physiological articulatory model,” Proc. ASJ, pp. 335-338, Koriyama (Autumn 2009).