3D Subject-Specific Biomechanical Modeling and Simulation of the Oropharynx with Application to Speech Production

by

Negar Mohaghegh Harandi

B.Sc., AmirKabir University of Technology, 2005

M.Sc., Isfahan University of Technology, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Electrical & Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

February 2016

Abstract

The oropharynx is involved in a number of complex neurological functions, such as chewing, swallowing, and speech. Disorders associated with these functions, if not treated properly, can dramatically reduce the quality of life for the sufferer. When tailored to individual patients, biomechanical models can augment the imaging data, to enable computer-assisted diagnosis and treatment planning.

The present dissertation develops a framework for 3D, subject-specific biomechanical modeling and simulation of the oropharynx. Underlying data consists of magnetic resonance (MR) images, as well as audio signals, recorded while healthy speakers repeated specific phonetic utterances in time with a metronome. Based on this data, we perform simulations that demonstrate motor control commonalities and variations of the /s/ sound across speakers, in front and back vowel contexts. Results compare well with theories of speech motor control in predicting the primary muscles responsible for tongue protrusion/retraction, jaw advancement, and hyoid positioning, and in suggesting independent activation units along the genioglossus muscle.

We augment the simulations with real-time acoustic synthesis to generate sound. Spectral analysis of resultant sounds vis-à-vis recorded audio signals reveals discrepancy in formant frequencies of the two. Experiments using 1D and 3D acoustical models demonstrate that such discrepancy arises from low resolution of MR images, generic parameter-tuning in acoustical models, and ambiguity in 1D vocal tract representation. Our models prove beneficial for vowel synthesis based on biomechanics derived from image data.

Our modeling approach is designed for time-efficient creation of subject-specific models. We develop methods that streamline delineation of articulators from MR images and reduce expert interaction time significantly ($\approx$ 5 mins per image volume for the tongue). Our approach also exploits muscular and joint information embedded in state-of-the-art generic models,

while providing consistent mesh quality, and the affordances to adjust mesh resolution and muscle definitions.

Preface

The research presented herein was approved by UBC clinical Research Ethics Board, certificate number: H16-00016. Most of the contributions and ideas described in this dissertation have been presented previously in the publications listed in pages v and vi.

[P1, P3-6]: Negar M. Harandi was the primary author and main contributor to the design of these papers, under the supervision of Dr. Sidney Fels and Dr. Rafeef Abugharbieh. Negar M. Harandi developed the speaker-specific models, performed the simulations, and analyzed the results. Dr. Maureen Stone and Dr. Jonghye Woo provided, and pre-processed the MRI data. Dr. Marek Bucki shared his code for FE registration; Dr. Claudio Lobos shared his implementation of FE meshing ([P1]). Dr. Ian Stavness provided the code for the inverse solver, and the VT skin mesh ([P1, P3, P4]). Dr. Kees van den Doel provided the implementation of the 1D acoustic synthesizer ([P1, P4]). Dr. Maureen Stone assisted with the analysis of the results, and gave editorial feedback to [P3, P6]. The work in [P1] is reflected in Section 3.1, 3.3, and 5.1. The work in [P3] is reflected in Chapter 4.

[P2, P7]: Negar M. Harandi was the primary author and main contributor to the design, implementation, and testing of the methods developed in these papers, under the supervision of Dr. Sidney Fels and Dr. Rafeef Abugharbieh. Dr. Maureen Stone and Dr. Jonghye Woo provided the MRI data. The implementation of the shape-matching algorithm (Gilles and Pai, 2008) was available through the SOFA open-source simulation framework. The work in these papers is reflected in Section 3.2.

[P8]: Negar M. Harandi was the primary author and main contributor to the design of this paper, under the supervision of Dr. Sidney Fels and Dr. Rafeef Abugharbieh. Dr. Daniel Aalto and Dr. Jarmo Malinen provided the VT geometries with resonance frequencies, and assisted with interpretation of the results. The work in this paper is reflected in Section 5.2.

- **Published Journal Manuscripts:**

  [P1] **Harandi NM**, Stavness I, Woo J, Stone M, Abugharbieh R, Fels S. 2015. Subject-specific biomechanical modeling of the oropharynx: towards speech production. Comput Methods Biomech Biomed Eng: Imaging Vis, (ahead-of-print), 1-11.

  [P2] **Harandi NM**, Abugharbieh R, Fels S. 2014. 3D segmentation of the tongue in MRI: a minimally interactive model-based approach. Comput Methods Biomech Biomed Eng: Imaging Vis. 3(4):178–188.

- **Journal Manuscripts in Review:**

  [P3] **Harandi NM**, Woo J, Stone M, Abugharbieh R, Fels S. 2015. Articulation variability in English /s/: a biomechanical modeling approach. Submitted.

- **Peer-Reviewed Conference Papers:**

  [P4] **Harandi NM**, Woo J, Farazi MR, Stavness I, Stone M, Fels S, Abugharbieh R. 2015. Subject-Specific Biomechanical modeling of the Oropharynx with Application to Speech Production. Proceedings of International Symposium on Biomedical Imaging (ISBI); Brooklyn, USA.

  [P5] **Harandi NM**, Woo J, Stone M, Abugharbieh R, Fels S. 2014. Inverse Simulation of Subject-Specific Jaw-Tongue-Hyoid Model from Dynamic MR. Proceedings of International Symposium on Computer Methods in Biomechanics and Biomedical Engineering (CMBBE); Amsterdam, Netherlands.

  [P6] **Harandi NM**, Woo J, Stone M, Abugharbieh R, Fels S. 2014. Subject-specific biomechanical modeling of the tongue: analysis of muscle activations during speech. Proceedings of International Seminar on Speech Production (ISSP); Cologne, Germany.

  [P7] **Harandi NM**, Abugharbieh R, Fels S. 2014. Minimally interactive MRI segmentation for subject-specific modeling of the tongue. MICCAI Workshop on Bio-Imaging and Visualization for Patient-Customized Simulations (BIVPCS); Nagoya, Japan. Best Paper Award.

- **Workshop Articles:**

  [P8] **Harandi NM**, Aalto D, Hannukainen A, Malinen J, Abugharbieh R, Fels S. Spectral analysis of the vocal tract in vowel synthesis: a comparison between 1D and 3D acoustic analysis. In 3^{rd} International Workshop on Biomechanical and Parametric Modeling of Human Anatomy (PMHA 2015), Montreal, Canada.

Table of Contents

Appendices

List of Tables

List of Figures

List of Abbreviations

1D	1-dimensional
2D	2-dimensional
3D	3-dimensional
CBCT	Cone beam Computer Tomography
CSA	Cross-Sectional Area
CT	Computer Tomography
DOF	Degree Of Freedom
E-IDEA	Enhanced Incompressible Deformation Estimation Algorithm
EMA	Electro-Magnetic Articulometry
EMG	Electromyography
FE	Finite Element
FEM	Finite Element Method
GE	Gradient Echo
GUI	Graphical User Interface
HARP	HARmonic Phase
Hz	Hertz
IDEA	Incompressible Deformation Estimation Algorithm
IPA	International Phonetic Alphabet
JR	Jacobian Ratio
MHD	Modified Hausdorff Distance
MMRep	Mesh-Match-and-Repair (algorithm)
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MSCT	Multi-Slice Computer Tomography
OPAL	Oral, Pharyngeal, and Laryngeal (complex)
PCA	Principle Component Analysis
PDE	Partial Differential Equation
POA	Place Of Articulation

RF	Radio Frequency
SE	Spin Echo
SNR	Signal-to-Noise Ratio
STD	Standard Deviation
TE	Echo Time
TF	Time Frame
TLM	Transmission Line-circuit Model
TMJ	Temporomandibular Joint
TR	Repetition Time
VT	Vocal Tract

Oropharyngeal Muscles:

AD	Anterior Digastric
AM	Anterior Mylohyoid
AT	Anterior Temporal
DP	Deep Masseter
GG	Genioglossus
GGA	Genioglossus Anterior
GGM	Genioglossus Medium
GGP	Genioglossus Posterior
GH	Geniohyoid
HG	Hyoglossus
IL	Inferior Longitudinal
IP	Inferior-lateral Pterygoid
MH	Mylohyoid
MP	Medial Pterygoid
MT	Medial Temporal
PD	Posterior Digastric
PM	Posterior Mylohyoid
PT	Posterior Temporal
SH	Stylo-Hyoid
SL	Superior Longitudinal
SM	Superficial Masseter
SP	Superior-lateral Pterygoid
STY	Styloglossus
TRANS	Transverses
VERT	Verticalis

Acknowledgements

My deepest appreciation goes to to my supervisors, Dr. Sidney Fels and Dr. Rafeef Abugharbieh, for making this thesis possible. Your support, advice, and inspiration guided me throughout the past five years. It's been an honour working with you and learning from you.

I am indebted to Dr. Maureen Stone for generous data sharing, many fruitful discussions, and hosting me at her lab and in her own home during my research visits to the University of Maryland Dental School. I admire you as a great woman and academic scholar.

I have had the pleasure of working closely with a number of fine researchers during the course of this PhD. Thanks must go to Dr. Jonghye Woo for a rewarding collaboration, and for his hosting of my visit to Massachusetts General Hospital. Dr. Ian Stavness deserves special thanks for his noble collaboration with me on this project. Many thanks to Dr. Daniel Aalto and Dr. Jarmo Malinen for sharing their knowledge with me during several discussion sessions and lengthy emails. Finally, I would like to thank Dr. John Lloyd for his great support with ArtiSynth.

The positive experience I've had during my graduate studies is due, in part, to my fellow graduate students, with whom I have shared an office, building, or research project. Special thanks go to Antonio Sánchez, and Andrew Ho for many useful conversations, and their assistance with the project.

I want to say a heart-felt thank you to my friends and family for their love and support throughout this journey. I am grateful to my parents for passing on their life-long motivation to me, and my sister for her cheer and laughter. Lastly, I thank my dear Matthew for his daily encouragement, endless love, and his editing eyes which were invaluable to this manuscript.

Dedication

To my dad,
who told me to learn;

and my mom,
who told me to pursue my dreams.

Chapter 1

Introduction

Speech production is a complex neuromuscular human function that involves coordinated interaction of the oropharyngeal structures (shown in Figure 1.1). Speech impairments are widespread and have an adverse effect on the sufferer's quality of life. Currently, articulation, fluency, and voice disorders afflict more than 7.5 million people in the United States. It has proven difficult to characterize the complex, nonlinear relationship between neurological activation units, articulatory motion, and sound generation. Such difficulty hinders efforts in rehabilitation planning greatly. A deeper understanding of speech production would, thus, be invaluable in a clinical setting.

Medical imaging has enabled more detailed observation of the oropharynx; nevertheless, hidden variables, such as muscular forces and activations, remain mostly immeasurable. By predicting such variables, biomechanical models can improve the protocols of diagnosis, and assist in treatment planning. In jaw reconstructive surgery, study of the pre- and post-operative models of the mandible and maxilla (constructed based on computed tomography [CT] scans of the patients) has led to the selection of more suitable operational procedures, and a reduction in the risk of associated complications (Wolánski et al., 2015).

Despite the clinical success of patient-specific bone models in jaw reconstructive surgery, biomechanical modeling of speech remains challenging. Firstly, articulatory motion of oropharyngeal soft tissue is rapid and difficult to capture. Magnetic resonance imaging (MRI) is better suited to depicting soft-tissue, but still falls short due to low contrast and spatial sparsity. Secondly, the structure and physiology of oropharyngeal soft tissue, with its interwoven muscle and tendon fibers, remains a challenge to represent in a biomechanical model. As a result, current models of articulatory movement remain *generic*: that is, they represent an *average* human anatomy and function, thus, failing to provide individualized information. In addition, the making

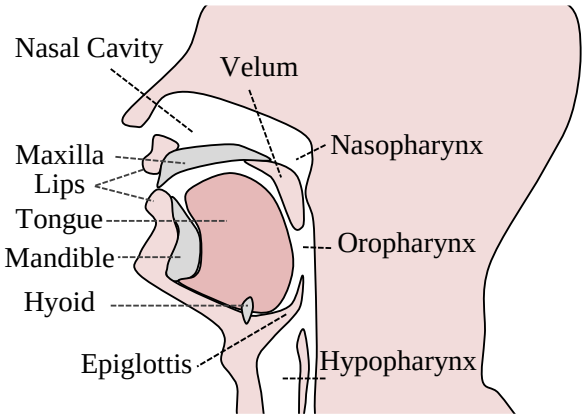


Figure 1.1. A mid-sagittal diagram of a human upper airway, denoting the (oro)pharyngeal structures.

of current generic models relies heavily on expert interaction – a process that is not cost effective when dealing with many individual cases. These factors contribute to a gap in speech research, separating medical imaging from soft-tissue modeling and simulation.

The gap extends further when looking at biomechanical vs. acoustical models of speech. Sound production, the ultimate physical goal of speech, has mostly been addressed independent of biomechanics. Articulatory speech synthesizers generate sound based on the geometry of the vocal tract, which is estimated directly from medical images. The search for an ideal model, one which represents both the acoustical and biomechanical characteristics of the oropharynx, continues to this day.

A unified modeling framework could fill the current gaps in speech analysis by integrating the required constituent units: data processing, subject-specific biomechanical modeling, data-driven simulation and acoustic synthesis. Such a framework would serve as a complementary tool for studying inter- and intra-subject variability in speech production, could potentially lead to the development, modification, and verification of theories of speech strategy across speakers of different gender, age, language, or pathology.

The present dissertation develops subject-specific, oropharyngeal modeling and simulation methods for speech research. The underlying data includes dynamic (cine and tagged) magnetic resonance (MR) images, which capture

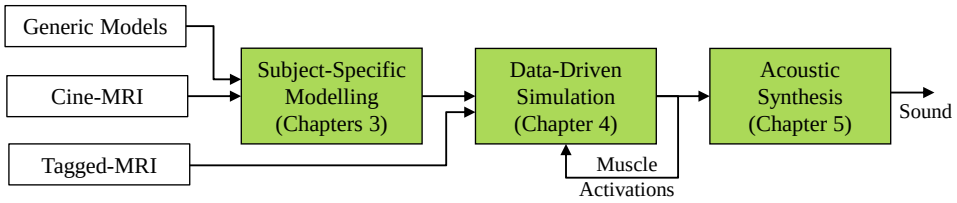


Figure 1.2. Designed work-flow for subject-specific, oropharyngeal modelling and simulation in speech research.

the motion of articulators during the repetition of specific speech utterances. Each chapter of this dissertation details a component of the designed work-flow, as shown in Figure 1.2.

In Chapter 2 we present a review, and identify the challenges facing standard approaches to the characterization of oropharyngeal structures and speech function – with a particular focus on data acquisition and measurement, as well as biomechanical and acoustical modeling.

Chapter 3 details our approach to subject-specific modeling of the oropharynx. We develop methods to address the challenge of MRI segmentation for the articulators, and quantify the efficacy of these methods in regards to time-efficiency, accuracy, and parameter sensitivity. Further, we incorporate standard generic models into our subject-specific modeling approach, to minimize remodeling efforts, before demonstrating the performance of our models in a forward simulation scheme.

In Chapter 4 we enable inverse simulation of our developed models, based on tissue trajectories extracted from the tagged MRI data of each speaker. Through this simulation, we investigate inter-speaker variability in the estimated muscle-activation patterns. A comparison of our results with linguistic theories of speech production validates our findings.

In Chapter 5 our subject-specific models and corresponding data-driven simulations are coupled with a standard 1D acoustical model, in order to synthesize vowel sounds in both running and sustained speech configurations. Spectral analysis of our deformed vocal-tract model reveals discrepancy with that of the recorded audio. We identify the sources of discrepancy by comparing the performance of the 1D and 3D acoustical models, which carry

their computations both in time and frequency domains.

In Chapter 6 we summarize the contributions of this dissertation, describe directions for future work, and provide concluding remarks.

1.1 Contributions

The primary, novel, contributions of this dissertation are highlighted here.

1. Our work fills a gap between medical images of the oropharynx and speech analysis in the biomechanical and acoustical domains, by developing, incorporating, and validating methods that fit our designed framework in Figure 1.2. This is first achieved by efficient processing of dynamic MR images for modeling and simulation purposes (parts of chapters 3 and 4). Following this, we facilitate creation of subject-specific biomechanical models of the tongue, mandible, and hyoid, based on standard generic models (Chapter 3). We then simulate our models by solving the inverse problem for MRI-based tissue displacements (Chapter 4). Finally, we enable sound generation based on the predicted biomechanics, by integrating an articulatory acoustic synthesizer to our biomechanical system (Chapter 5). Components of this contribution have been published in [P1].

2. We significantly reduce the expert interaction time required for 3D segmentation of tongue tissue from MR images, by introducing a real-time, intuitive interaction scheme into a mesh-to-image registration technique (Section 3.2). Each segmentation task requires less than five minutes – compared with two or more hours using standard, semi-automatic tools. The quantitative results show comparable accuracy with human errors. This contribution has been published in [P2].

3. We provide tools for adjusting the spatial resolution and muscle topology of our subject-specific, finite element (FE) tongue models (Chapter 3). By combining the benefits of FE meshing and registration techniques, our modeling approach leverages the biomechanical properties of a standard generic model without being constrained by its mesh configuration or muscle definition. This contribution has been partly published in [P1], and partly submitted for publication in [P3].

4. Using our data-driven simulation, we measure and quantify inter-speaker variability in the muscle-activation patterns responsible for the /s/ sound in front and back vowel contexts (Chapter 4). Results show consistency with theories of speech motor control in predicting the primary muscles involved in tongue protrusion/retraction, jaw advancement, and hyoid positioning. Our findings compare well with published medical measurements in suggesting independent activation units along the genioglossus muscle. This contribution has been submitted for publication in [P3].

5. We identify the challenges of using, and demonstrate the trade-off between, standard 1D and 3D acoustical models for sound generation (Chapter 5). We show that low resolution of dynamic MR images and generic parameter-tuning in acoustical models degrade the accuracy of the computed formant frequencies. We also verify that the stability and accuracy of such formants are adversely affected by ambiguity in 1D representation of the vocal tract, and simplification of acoustic equations required to make 3D analysis possible in frequency-domain. This contribution has partly been published in [P1], and partly presented at the *Parametric Modeling of Human Anatomy (PMHA)* workshop (2015) in Montreal, Canada.

Chapter 2

Background

Current advancements in data acquisition techniques have created a great opportunity to observe articulatory movements. Data analysis and measurement methods enable quantitative assessments of these observations, and are considered inevitable for transition to computational models. This chapter reviews the tools and techniques used for the analysis of speech production. Section 2.1 provides a review of medical imaging methods, such as MRI, for depiction of the oropharyngeal structures. The section follows by addressing the potentials and challenges of physiological data acquisition techniques, such as electromyography. Finally, image segmentation methods are discussed, identifying the challenges in dealing with oropharyngeal medical images.

Despite current advances in medical imaging techniques, measuring the biomechanics of speech (such as chronology and level of muscular force and activation) remains a challenge, mainly due to the structural and physiological complexity of the articulators. Biomechanical models in particular complement the imaging observations, furthering the understanding of speech production. Section 2.2 reviews previously reported generic biomechanical models of the articulators, and identifies the need to move to a subject-specific framework. Finally, inverse simulation techniques are discussed as a solution to the lack of the physiological data.

The ultimate goal of speech mechanics is to generate sound. Established speech synthesis techniques, such as concatenative synthesis, involve the selection of units of speech based on statistical analysis of the sound recorded from a population of speakers. These methods are successful in generating natural sounding speech, but fail to explain the nature of speech production. Articulatory speech synthesis, on the other hand, attempts to simulate speech production by applying the governing rules of physics. Mathematical models of vocal folds and tracts are built, in which sound waves propagate

from the subglottal area to the exterior of the lips. Section 2.3 explains the acoustics characteristics of different speech units, and reviews the current trend in articulatory speech synthesizers.

2.1 Data Acquisition and Measurement

Diversified medical data is acquired to provide insight into the morphology of the oropharynx and physiology of speech. Static imaging techniques capture the tongue in a sustained posture, while dynamic imaging techniques strive to record a chronology of oropharyngeal motion in running speech. The common challenges fall into the following categories: 1) the motion artifact caused by relatively-long scan times; 2) insufficient spatial and temporal resolutions; 3) low signal-to-noise ratio; and 4) poor soft-tissue contrast. In addition, the approved level of the ionization dosage is a matter of active discussion. This section briefly reviews each state-of-the-art imaging modality with respect to the goals of this thesis.

2.1.1 Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) is capable of differentiating between body tissues, by altering magnetic alignment of hydrogen atoms, and measuring the released energy while the protons resume their previous alignment. T1-weighted images are well-suited to depict soft-tissue anatomy and fat, while T2-weighted images are optimal for showing fluid. The major limitation of MRI in depicting the oropharynx is motion artifacts. For years, relatively long acquisition time of conventional MRI made recording of running speech impractical. Nowadays, shorter acquisition times have become possible due to the advances in technical aspects of MR scanners, which allow for parallel reconstruction and/or k-space encoding.

High-resolution MR volumes require a long acquisition time, commonly leading to involuntary movement of the tongue and, hence, introduces severe motion artifacts (Plenge et al., 2012). 2D acquisition can provide a refined depiction of the tongue in the acquired plane, but it results in low through-plane resolution, inadequate for most volumetric analyses. Some previous

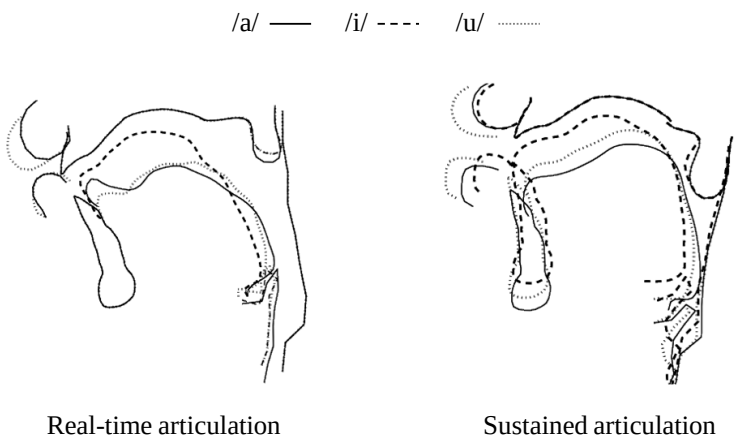


Figure 2.1. Mid-sagittal contours of the articulators derived from sustained and real-time MRI for vowels /a/, /i/, and /u/. ©Engwall (2003b), adapted with permission.

speech studies adjust the orientation of the acquisition plane so it is orthogonal to the axis of vocal tract (VT) (Badin et al., 2002; Engwall, 2003a; Takano and Honda, 2007). This facilitates tongue modeling for a sustained tongue posture; however, sustained articulations cause hyper-articulated configurations, which are not necessarily representative of running speech. A study by Engwall (2003b) suggests that jaw opening is more dominant in sustained vowels; whereas, its contribution in running speech is taken over by larger alterations in tongue shape (see Figure 2.1).

Fast MRI Acquisition

Improvements in magnet quality in recent MR scanners has enabled Gradient Echo sequences to capture higher quality images – comparable to Spin Echo (SE) – resulting in faster acquisitions.

Parallel imaging, as a complementary reconstruction technique, has caused a revolutionary change in the length of MRI scans (Larkman and Nunes, 2007). Multi-array coil configurations augment the obtainable SNR, and provide the spatial information (in the form of receiver coil sensitivity maps) to shorten phase encoding time. In 2011, a 16-channel array receiver coil – custom-designed to provide localized sensitivity in the airway – was used to capture

of the VT at a maximal temporal resolution of 84ms, for a single slice (Kim et al., 2011a) .

Traditional MRI acquires parallel lines of data in the k-space (i.e., Cartesian sampling), mainly to provide simplicity in computations, and robustness to noise in early MRI scanners. Under-sampling of the k-space (in order to reduce scan time) is limited in the Cartesian grid, and introduces a *jumping* effect in consecutive images. Radial encoding enables under-sampling of the k-space, by including both the low and high resolution information in every scan profile. Using radial encoding, Uecker et al. (2010) capture dynamic MR images of the oropharynx, at the rate of 55ms per frame. Through combining under-sampled radial GE MRI, and serial image reconstruction (using parallel imaging), Iltis et al. (2015) quantify the tongue movement in a single slice of real-time MRI acquired at the rate of 10ms per frame.

Synchronization methods

MR images may be acquired following an external stimulus (trigger) with an adjustable delay. This allows for the acquisition and reconstruction of pseudo-moving images, in the case of repetitive periodic motion; thus, permitting under-sampling. Performance can be limited by the reproducibility of the motion; however, it is shown that good reproducibility can be obtained for tongue motions in speech – even in poly-syllabic utterances (Alvey et al., 2008). In a *gated* sequence, speakers are asked to repeat an utterance in time with a metronome (Alvey et al., 2008) or their heartbeat (Ventura et al., 2011), while the audio signal is recorded simultaneous to MR imaging. This allows a higher frame rate and signal-to-noise ratio, compared to the non-gated sequences. The utterance onset from recorded audio signals is used for reconstruction of the MR images (Shimada et al., 2012). The high intensity of audio noise present during MRI acquisition makes use of alternative systems, such as optical microphones and noise cancellation methods, a necessity for simultaneous audio recording (NessAiver et al., 2006; Aalto et al., 2014).

Multi-planar Acquisition

Many MRI studies of speech production are limited to a single acquisition plane. The mid-sagittal plane depicts the dynamics of articulators, but it is not sufficient for identifying some important features of speech, such as grooving/doming, and anatomical asymmetries. Kim et al. (2011b) makes use of slice-interleaving technique to produce multi-slice real-time MR scans of the VT during pronunciation of English fricatives.

Super-Resolution as Post-Processing

Super-resolution reconstruction techniques are introduced to generate isotropic MR volumes from orthogonal slice stacks acquired sequentially (Peled and Yeshurun, 2001; Bai et al., 2004). The imaging process is formulated as an observation model; and the intensity value of each voxel is obtained using an optimization method such as a maximum a posteriori (Bai et al., 2004) or least square (Plenge et al., 2012) estimation. Recently, Woo et al. (2012) apply an edge-preserving data combination technique, based on Markov Random Field, to build super-resolution volumes of the human tongue. Isotropic resolution of 0.94 mm are reported. Figure 2.2 shows the reconstruction result, in comparison to in-plane high resolution data.

Upright MRI

Due to technical limitations of scanners, MR imaging is mostly performed in supine position. However, gravitational force is believed to effect the dynamics and biomechanics of speech. A study by Kitamura et al. (2005) shows that tongue retraction exists in supine position and is more severe in back vowels than front vowels. In case of the former, this may be due to stabilizing the tongue by pressing its sides against hard palate. The study measures noticeable displacement (up to 20 mm) for the tongue tip and identifies the effect of body posture on the larynx, lower jaw, lower lip and posterior pharyngeal wall. The major drawback of upright MRI is that it is currently only available in open-MRI configuration that provides low intensity magnetic fields (typically 0.5 T), due to safety issues. Lower magnetic field translates into longer acquisition time for images of sufficient SNR.

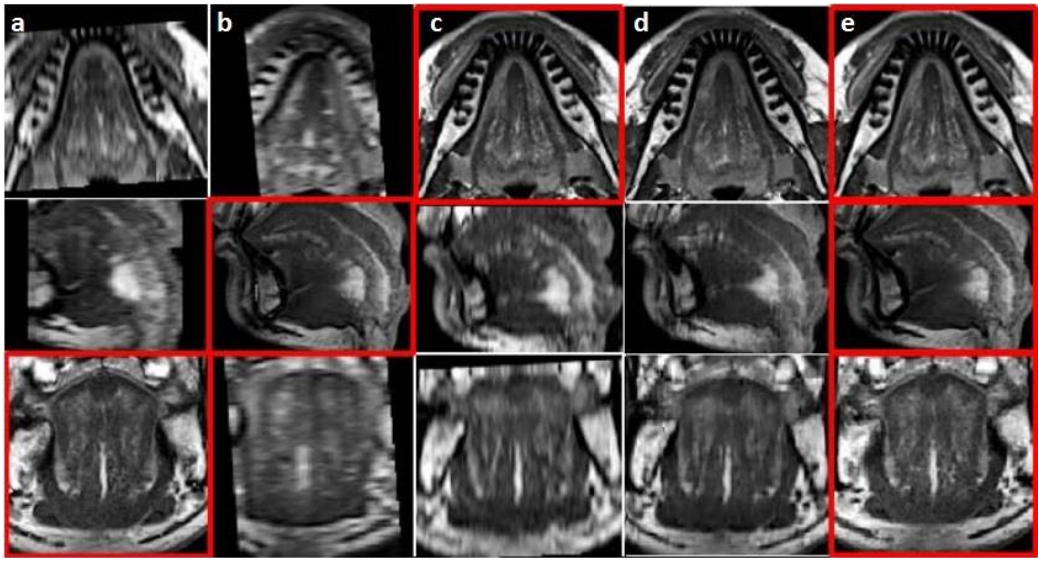


Figure 2.2. Super resolution volume reconstruction for static tongue MRI (Woo et al., 2012); Original axial (a), sagittal (b) and coronal (c) stacks are shown after isotropic re-sampling. Reconstructed volume (d) is compared to the ideal case (e). ©Woo et al. (2012), adapted with permission.

Tagged MRI

Dynamic acquisition fails to provide high image contrast within the soft tissue itself; therefor, internal tissue points are not distinguishable, and their motion is not quantifiable. MR tagging introduces temporary features inside the tissue – by applying a sequence of RF pulses to spatially modulate longitudinal magnetization of hydrogen protons, prior to imaging (Kerwin et al., 2000). In subsequent images, varying magnetization manifests itself as alternating light and dark tag patterns. The induced tags persist through the motion, and are visible in images acquired perpendicular to tagging planes. Tagged MRI is successfully applied to speech imaging – especially using gated cine MRI pulse sequences (NessAiver et al., 2006; Xing et al., 2013).

Harmonic phase (HARP) algorithm is introduced to estimate the 2D motion of the features from tagged MRI slices (Osman et al., 2000). The 2D motion (estimated from slices acquired from different orientations and at different times) is then combined to produce a 3D tracking result. Liu et al. (2012) introduce an incompressible deformation estimation algorithm (IDEA) that

incorporates tongue incompressibility constraint while imposing a smoothing, divergence-free, vector spline in seamlessly interpolating HARP velocity fields across the tongue. The results are shown to be accurate for internal tissue points of the tongue; however, since HARP uses a bandpass filter (that makes the object boundaries blurry) the motion estimated at the tongue surface remains inaccurate. Later on, Xing et al. (2013) propose an enhanced version of the algorithm (E-IDEA) that improves the reliability of the displacement field at the tongue surface, by incorporating 3D deformation of the tongue surface computed from cine MRI.

Diffusion Tensor Imaging

Diffusion Tensor Imaging (DTI) is an MRI method designed to detect the diffusion process of water molecules in biological tissue. Such diffusion reflects the interaction of water molecules with many obstacles (such as muscle fibres), and, hence, reveals microscopic details in tissue architecture. DTI patterns are extracted from the images using tractography methods.

DTI tractography has been used to depict the structure of the human tongue. Gaige et al. (2007) demonstrate the geometric relationships between intrinsic and extrinsic myofiber populations, with a focus on the manner in which key extrinsic fibers merge with the longitudinal, transverse, and vertical intrinsic fibers. Mijailovich et al. (2010) incorporate these fiber structures to derive a finite-element model of lingual deformation during swallowing. Using DTI, Murano et al. (2010) show major changes in the structure of the inferior longitudinal muscle bundle in the tongue anatomy of one control volunteer and one glossectomy patient.

2.1.2 Computed Tomography Imaging

Computed Tomography (CT) imaging combines multiple X-ray projections to reconstruct 3D images of tissue (with spatial resolution as high as 0.3mm per voxel). Bone and airway images have high contrast in CT, since X-ray is absorbed by dense tissue and passes through air. Multi-slice CT (MSCT) scanners have recently been equipped with multiple arrays of X-ray detectors that are able to reconstruct a 3D volume from a single rotation, and thus

reduce exposure time significantly. Cone beam CT (CBCT) is used in studies related to orthodontic treatments and maxillofacial surgeries. Glupker et al. (2015) use CBCT to measure airway volume changes between open and closed jaw positions for patients with temporomandibular joint disorders. Fujii et al. (2011) use a 320-detector-row MSCT scanner to capture a single phase 3D image volume of the oropharynx in less than 0.35 sec. The imaging process is repeated for 29 phases at intervals of 0.1 sec, to generate a fully three-dimensional film of the swallowing on one volunteer. The study is unique in providing a full 3D movie of a fast oropharyngeal function. Very recently, Inamoto et al. (2015) use similar single phase image protocol to investigate the effects of age, gender, and height on the anatomy of the oropharynx in 55 volunteers. High dosage of X-ray exposure is the main drawback of medical CT which hinders its use on healthy volunteers.

2.1.3 Electromyography

Electromyography (EMG) involves transducing electrical signals associated with muscle activations (Miyawaki et al., 1975). EMG assists with understanding speech motor control, by identifying active muscles and their coordination during a speech gesture. However, the relationship between the EMG signal and the mechanical muscle forces is not straightforward, and can be influenced by many factors, such as muscle fiber type, muscle length, and muscle velocity (Sherif et al., 1983). Direct measurement of the muscle force during EMG session has been performed to help investigation of such relationship for musculoskeletal muscles, but is subject to technical challenges (Roberts and Gabaldn, 2008).

Fine-wire needle electrodes are used to measure the activity of the lateral pterygoid muscle (the main jaw opener) mainly in breathing and chewing movements (Murray, 2012). In the early 1980's, Tuller et al. (1981) and Gentil and Gay (1986) compare EMG recordings of jaw muscles for speech and non-speech gestures, identifying the medial pterygoid and superior lateral pterygoid muscles as the jaw elevators, and inferior lateral pterygoid and (anterior belly of) digastric muscles as the jaw depressors during speech. Surface electrodes are used widely for recording facial EMG, in order to facilitate recognition of audible and silent speech. We refer to Wand (2015) for a recent review of the field. Surface EMG of jaw muscles including sub-

mandibular, masseter, anterior temporalis and (anterior belly of) digastric muscles are also reported in conjunction with recordings of jaw movement in control subjects (Kawakami et al., 2012) and patients with related disorders (Ma et al., 2013).

EMG recording of the tongue during speech gestures was performed in the mid 1960's (Mac Neilage and Sholes, 1964); however, the first study to record activity of extrinsic muscles of the tongue (including the anterior and posterior genioglossus, hyoglossus, and genioglossus) did not happen until two decades later (Baer et al., 1988). The process remains to be challenging: the moist surface and highly deformable body of the tongue prohibits excessive use of electrodes on its surface (Yoshida et al., 1982). Measured signals are difficult to interpret reliably, due to their high variability within subjects, across sessions, and across subjects. In addition, deep and small muscles are barely accessible, and their EMG recording is limited by lack of suitable technology in dealing with crosstalk between adjacent channels.

2.1.4 Electromagnetic Articulometry

Electromagnetic articulometry (EMA) systems track the position of articulators by using a set of markers that attach to the surface of the tongue, jaw, teeth and lips. Each marker is a small sensor coil that moves in a known (and varying) electromagnetic field. The electromagnetic field induces a weak current in the sensors, which is further measured and mapped to the coordinates of the markers. A reference marker is often used to compensate for head motion. EMA systems are not invasive, do not require line-of-sight, and are able to capture their data in upright position. Modern systems, such as the Carstens AG500 (`www.articulograph.de`) and NDI Wave (`www.ndigital.com`), can provide high temporal update rates (e.g. around 100Hz) and are self-calibrating; however, their accuracy degrades in the presence of metallic objects (such as mercury tooth fillings) due to field distortion. The tracking is also limited to eight markers in simultaneous acquisition. Most speech studies use the markers on the mid-sagittal line of the articulators, assuming that speech is a bilaterally symmetric phenomenon (Engwall, 2003a; Narayanan et al., 2014).

2.1.5 Image Segmentation

Segmentation is the task of partitioning a medical image into semantically interpretable regions (i.e., organs of interest). The level of difficulty in image segmentation is related to the degree of intensity-variation across tissue boundaries: while segmentation of the mandible from CT can be achieved by applying simple thresholding methods, segmentation of the tongue from MRI remains a challenge and requires intervention from a trained anatomist. Manual segmentation produces accurate results, but is prohibitively time-consuming and tedious – especially where it must be repeated for several datasets. General-purpose interactive tools can ease the task, but still require significant user interaction time.

Prior knowledge of shape and appearance, if incorporated effectively, can assist in dealing with soft-tissue inhomogeneities, noise, and low contrast in medical images (Heimann and Meinzer, 2009). In this regard, shape constraints have been embedded into level-set framework (Leventon et al., 2000; Tsai et al., 2003; Foulonneau et al., 2009), and further equipped with trained distance maps (Bresson et al., 2006). In addition, statistical methods, such as the Active Shape Models (Cootes et al., 1995), have been widely explored to encompass intra- and inter-subject morphological discrepancies. Here, the cardinality of the training set is proportional to the degree of natural variability of the organ shape. For example, Heimann et al. (2007) select 32 of 86 datasets to train their 3D reference model for segmentation of the liver in CT volumes. Acquiring a sufficiently large dataset remains a challenge for MRI volumes of the upper airway.

As an alternative to statistical methods, prior information may be formulated in a single template, and registered to the target image. Saddi et al. (2007) use template matching as a complementary step in their liver segmentation process, in order to compensate for the limitations of their learning set. Somphone et al. (2008) transform their binary template subject to conformity constraints between local patches. In a different approach, Gilles and Pai (2008) use explicit shape representation of the template to segment musculoskeletal structures from MR images. Their mesh deformation is regularized based on an expanded version of a computer animation technique called Shape Matching (Muller et al., 2005). The method is proven to efficiently approximate large, soft-tissue elastic deformations. We refer to Sotiras et al.

(2012) for a detailed review on deformable medical image registration.

Despite successful use of shape prior, automatic segmentation is still challenging in low-contrast medical images of soft-tissue. Clinical applications demand expert supervision and control over the process and final results of the segmentation. Effective and minimal interactivity schemes would provide higher reliability, while lowering the cost of interaction. (Freedman and Zhang, 2005) combine shape prior and interactivity in a graph-cut framework in 2D. Recently, (Mory et al., 2012) incorporate user input as inside/outside labelled points to improve the robustness and accuracy of a non-rigid implicit template deformation.

Lee et al. (2013) propose a semi-automatic seeding method, based on the Random Walker algorithm (Grady, 2006), for 3D segmentation of the tongue in low-resolution dynamic MRI, where the tongue has a uniform intensity. Their dataset consists of stacks of 2D slices, captured and averaged over 26 time-frames for two English speakers. The user provides seeds in some slice images (in space) and frames (in time) which propagate to other frames and slices using a deformable image-to-image registration technique. The seeds are further fed to the Random Walker algorithm (Grady, 2006) to obtain the final segmentations. This method is useful for speech analysis, but still requires excessive amount of user interaction for accurate segmentation of (a high-resolution static) MRI volume.

Other reported works on segmentation of the oropharyngeal structures from MRI data focus on 2D slices. Bresch and Narayanan (2009) propose an unsupervised regional technique (performed in the frequency domain), which captures the shape of the VT. (Peng et al., 2010) use a shape-based variational framework for tracking tongue contour at its surface in mid-sagittal dynamic MRI. Eryildirim and Berger (2011) manage to include physically corresponding surface curves of the tongue in their previously introduced PCA-guided segmentation method. Although these methods provide valuable information for speech studies, they ignore delineation of the tongue at its base, as well as its contact with the epiglottis, hyoid bone, and salivary glands. Segmentation algorithms tend to fail in these areas, due to the fusion of the tongue into neighbouring tissues of similar composition. In addition, 3D reconstruction of the tongue shape from its sparse 2D segmented contours is not straight-forward.

2.2 Biomechanical Modeling

Recent improvements in speech data acquisitions motivate the use of computational approaches to model speech phenomena (Vasconcelos et al., 2012; Ventura et al., 2009, 2013). Biomechanical models aim to simulate the dynamics of speech production following biological and physical assumptions about the articulators and motor control (Fang et al., 2009; Stavness et al., 2012).

Simulation of soft-tissue deformation should be handled based on the physics of continuum mechanics. Finite element (FE) analysis divides the continuous domain into a set of discrete sub-domains, called *elements*, to approximate the solution of the related partial differential equations (PDE). The accuracy and stability of the solution depends on the resolution, as well as on the shape and type of the elements (Nealen et al., 2006). FE models are adopted into the ArtiSynth toolkit (Lloyd et al., 2012) for simulation of the generic oral, pharyngeal, and laryngeal (OPAL) deformable soft-tissues.

2.2.1 Generic Models

The Tongue

Generic models of the tongue, the main articulator in speech production, have been developed previously (Dang and Honda, 2004; Gerard et al., 2006; Buchaillard et al., 2009) and incorporated in the simulation of speech movements (Perrier et al., 2003; Stavness et al., 2012). These models are further enhanced through coupling to the jaw, and hyoid (Stavness et al., 2011), as well as the face and skull (Badin et al., 2002; Stavness et al., 2014a).

The state-of-the-art FE tongue model is developed by Buchaillard et al. (2009), based on the CT images of a single male subject. It consists of 946 nodes, 740 hexahedral elements, and 11 pairs of muscle bundles[1] with bilateral symmetry (see Figure 2.3). The model is imported from the ANSYS environment (`www.ansys.com`) into ArtiSynth (Stavness, 2010). Each muscle

[1]Genioglossus anterior (GGA), medium (GGM), posterior (GGP); hyoglossus (HG); styloglossus (STY); inferior longitudinal(IL); verticalis (VERT); transverses (TRANS); geniohyoid (GH); mylohyoid (MH); superior longitudinal (SL).

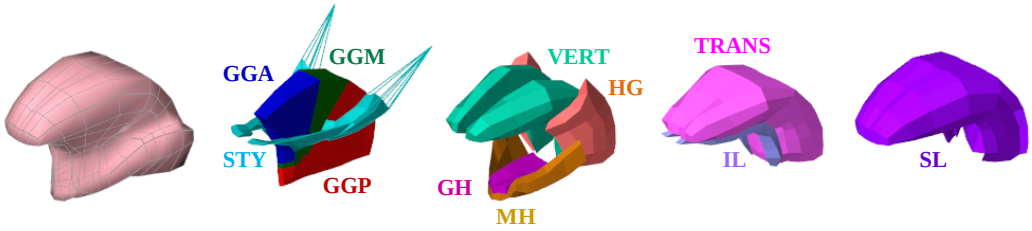


Figure 2.3. The generic FE tongue model, designed by Buchaillard et al. (2009), and its muscle bundles.

bundle in the model is defined as a set of muscle fibers (indicating the direction of the force) and their corresponding elements. The force capacity of each muscle is a function of its cross-sectional area (CSA), and is distributed across its fibers. Table 2.1 shows the maximum force for each muscle bundle as used by Buchaillard et al. (2009).

The model uses a fifth-order Mooney-Rivlin tissue material where strain energy (W) is described as:

$$W = C_{10}(I_1 - 3) + C_{20}(I_1 - 3)^2 + \kappa(\ln J)^2 \tag{2.1}$$

I_1 is the first invariant of the left Cauchy-Green deformation tensor; C_{10} and C_{20} are the Mooney-Rivlin material parameters, and the term $\kappa(lnJ)^2$ reinforces the incompressibility. Values of $C_{10} = 1037$Pa and $C_{20} = 486$Pa are used as measured by Gerard et al. (2006) from a fresh cadaver tongue and scaled by a factor of 5.4 to match the *in-vivo* experiments (Buchaillard et al., 2009). The Bulk modulus is set to $\kappa = 100 \times C_{10}$ to provide a Poisson's ratio close to 0.499. Tongue tissue density is set to 1040 $kg.m^{-3}$, close to

Table 2.1. Max force and CSA for muscle bundles in the generic tongue model (Buchaillard et al., 2009).

Intrinsic Muscles	GGA	GGM	GGP	VERT	TRANS	IL	SL
Max Force (N)	32.8	22	67.2	36.4	90.8	16.4	34.4
CSA(mm^2)	82	55	168	91	227	41	86
Extrinsic Muscles	STY	HG	MH	GH			
Max Force (N)	43.6	118	35.4	32			
CSA (mm^2)	109	295	88	80			

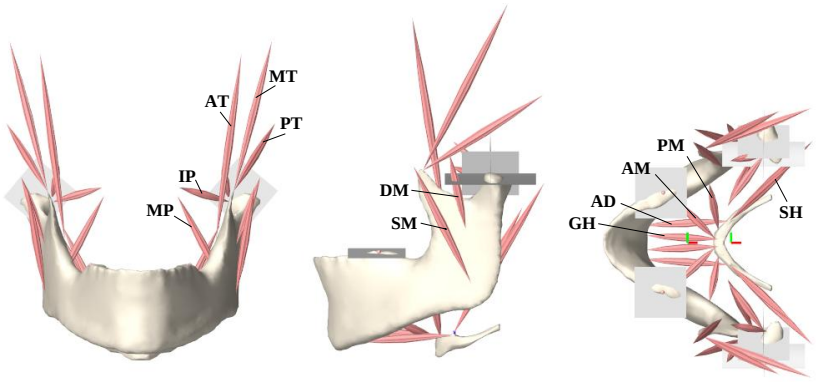


Figure 2.4. The generic jaw-hyoid model, as proposed by Hannam et al. (2008).

water density. The values of C_{10} and C_{20} increase linearly from (1037 Pa, 486 Pa) at no activation, to (10370 Pa, 4860 Pa) at full activation.

The Jaw and Hyoid

The generic jaw-hyoid model in ArtiSynth (Hannam et al., 2008) is coupled to the tongue FE model via multiple attachment points included in the constitutive equations of the system as bilateral constraints. Tongue-jaw attachments include the insertion of the genioglossus and geniohyoid onto the mandibular geniotubercle and the insertion of the mylohyoid along the mandibular mylohyoid ridge. Tongue-hyoid attachments include the entire region around the anterior-superior surface of the hyoid bone, including insertions of the geniohyoid, mylohyoid, and hyoglossus muscles. Eleven pairs of bilateral point-to-point Hill-type actuators are used to represent the associated muscles[2], as shown in Figure 2.4. The temporomandibular joint (TMJ) is modeled by curvilinear constraint surfaces (see Figure 2.4).

The Hill's muscle model provides a 1D mechanical model for the skeletal muscles by defining a relationship between tension and shortening velocity. This relationship accounts for both active and passive forces in a tetanic muscle contraction (Martins et al., 1998). Hill's muscle model can be described by

[2]Mylohyoid: anterior (AM), posterior (PM); temporal: anterior (AT), middle (MT), posterior (PT); masseter: superficial (SM), Deep (DM); pterygoid: medial (MP), superior-lateral (SP), inferior-lateral (IP); digastric: anterior (AD), posterior (PD); stylo-hyoid (SH).

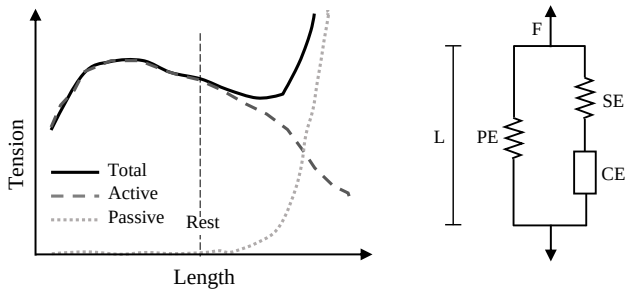


Figure 2.5. Hill's model for skeletal muscles: Force-length relationship (left) and the 3-element representation (right).

a 3-element representation as shown in Figure 2.5. The contractile element (CE) is the active part of the muscle; it shortens when activated, but is freely extensible when unactivated. The series elastic element (SE) allows for rapid transition of the muscle state from inactive to active and works as an energy storing unit. The parallel element (PE) is responsible for the passive behaviour of the muscle when stretched. The overall tension (F) and length (L) of the muscle relates to those of its elements as follows:

$$\begin{aligned} F &= F_{PE} + F_{SE}, \quad F_{SE} = F_{CE} \\ L &= L_{SE} + L_{CE}, \qquad L = L_{PE} \end{aligned} \tag{2.2}$$

where F_{SE} and F_{PE} are non-linear functions of the muscle stretch (change of the length relative to the resting state).

The instantaneous force generating capacity of the muscles in the jaw model vary non-linearly with length, and linearly with shortening velocity. Table 2.2 shows the maximum force and CSA used for the jaw and hyoid muscles in the generic model (Stavness, 2010).

2.2.2 Subject-Specific Modeling

In order to be clinically relevant, the aforementioned generic models should be simulated using neurological or kinematic measurements, such as EMG or EMMA recordings. Unfortunately, available data is often specific to certain subjects that do not share the same geometry as the generic model. A

Table 2.2. Max force and CSA for the muscles in the generic jaw-hyoid model (Stavness, 2010).

Jaw Closers	AT	MT	PT	SM	DM	MP
Max Force (N)	158.0	95.6	75.6	190.4	81.6	174.8
CSA(mm^2)	395	239	189	476	204	437
Jaw Openers	SP	IP	AD			
Max Force (N)	28.7	66.9	40.0			
CSA (mm^2)	72	167	100			

similar issue manifests itself in the validation phase, prohibiting meaningful comparison of numeric simulation results with subject-specific measurements.

To alleviate some of these issues, one option is to perform heuristic registration of subject data to the generic model (Fang et al., 2009; Sánchez et al., 2013), or restrict comparisons to average speech data reported in literature (Stavness et al., 2012). While these approaches are valuable in providing a proof of concept, they are not suitable in a patient-specific medical setting. Subject-specific biomechanical modeling, on the other hand, would address these issues while simultaneously enabling the investigation of inter- and intra-subject variability in speech production. In addition, it facilitates further development of a patient-specific platform for computer-assisted diagnosis and treatment planning of speech disorders.

A volumetric finite element is often represented in a tetrahedral or hexahedral configuration. Tetrahedral meshing is widely explored in the literature (Alliez et al., 2005; George et al., 2002). It is straightforward and applicable to unrestricted topologies; but linear tetrahedra tend to lock and become overly stiff for nearly incompressible materials such as muscles (Hughes , 2000). Hexahedra have better convergence, may vastly reduce the size of the linear system, and are preferable for non-linear analysis of anisotropic materials (Montagnat et al., 2000).

Image-based subject-specific modeling algorithms tend to directly generate a hexahedra-dominant FE mesh from a dataset of stacked images. Keyak et al. (1990) introduce the popular voxel-based method in which each voxel belonging to the organ of interest is identified by thresholding, and then transformed into a cubic element. The method is fully automated, general, and robust; however, it lacks efficient descriptors for identifying soft-tissues.

Other algorithms rely either on contours (Teo et al., 2007) or a surface mesh (Baghdadi et al., 2005; Bucki et al., 2010a), prior-extracted from the image stacks or volume. The modeling process consists of two consecutive phases of segmentation and FE volume mesh generation; and the overall clinical utility depends greatly on the accuracy and the degree of automation of each phase.

Current methods for creating subject-specific biomechanical meshes can be organized into two categories: *meshing* and *registration*. FE meshing techniques tend to generate FE models based solely on a subject's anatomy. The generated mesh often suffers from jagged boundaries, and low quality elements that should be made smooth, and regular. Smoothing, in turn, may cause surface shrinkage and the generation of ill-conditioned elements. Different approaches, such as mesh untangling or interior mesh smoothing, are proposed to cope with the problem (Zhang et al., 2005; Livesu et al., 2015). Still, involved optimization procedures are computationally expensive. A mixed-element FE meshing method, designed by Lobos (2012), has been shown to generate well-behaved meshes that approximate anatomical boundaries effectively. FE meshing techniques provide an adjustable mesh resolution to fit different needs for simulation time and accuracy; but they fail to offer the biomechanical information included in current generic models such as muscle definitions and coupling attachments. This, in turn, introduces prohibitive costs of redesigning these features for each subject model.

The subject-specific FE volume mesh may be acquired by registering a generic FE model to the surface mesh of the anatomy. This will automatically convey the muscle attachments and mechanical properties of the generic model to the volume mesh of a specific subject (Bucki et al., 2010a; Grosland et al., 2009; Sigal et al., 2008). The approach is also well-suited for complex meshes where re-meshing is a burden. Registration is performed using a 3D transformation $Map : x \longmapsto y$, which represents the mapping between the reference, $x = (x_1, x_2, x_3)$, and the actual, $y = (y_1, y_2, y_3)$, coordinate systems of the elements. In the Mesh-Matching approach (Couteau et al., 2000), Map for the internal elements is estimated from the mapping between the source and target surfaces. Due to this estimation, Mesh-Matching methods may generate invalid or poor quality elements, unsuitable for any further FE analysis (Luboz et al., 2005). Figure 2.6 explains the problem in a simple schematic 2D case, where M-M methodology fails to provide high-quality elements for the target configuration.

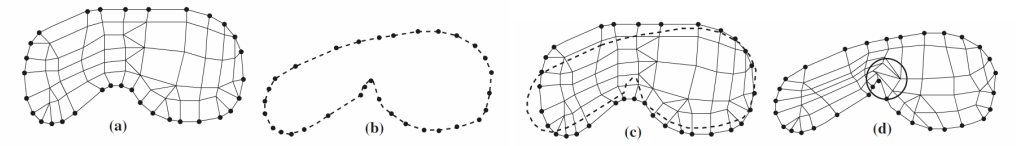


Figure 2.6. Schematic 2D representation of failure in Mesh-Matching methodology, as described by Bucki et al. (2011): the source FE mesh(a), the target surface(b), their aligned configuration(c), and the final mesh(d) after applying the transformation Map to the internal nodes. The circled area represents elements of poor quality. ©Bucki et al. (2011), adapted with permission.

In order to enable FE analysis of hexahedra-dominant meshes, the elements should maintain their convexity and positive volume (Shepherd, 2007). This is assessed through the element regularity measure. The measure is defined as the value of the Jacobian $J(x)$, which is the determinant of the matrix $\partial Map/\partial x$. The element e is considered regular if $J(x) > 0$ for all the element's nodes, x in e.

Hexahedral elements should also keep their shape conformity to prevent uneven discretization of the deformed domain (Knupp, 2000). The popular Jacobian ratio (JR) quality measure is defined as the ratio J_n^e/J_{max}^e, where J_n^e is the value of the Jacobian, at node n, in element e and $J_{max}^e = \mathrm{Max}_{m\in e} J_m^e$. JR ranges from 0 to 1 and gives an indication of the contribution of each node in the element distortion.

To compensate for the aforementioned irregularities, relaxation procedures were introduced, to repair the mesh after deformation. Mesh repair is performed through minimizing some validity and quality energy functions, which are calculated based on the Jacobian value. The state-of-the-art mesh repair method is embedded into the Mesh-Match-and-Repair (MMRep) algorithm (Bucki et al., 2010a). Here, the mesh repair is performed as the follow-up to a multi-scale, iterative, and elastic Mesh-Matching registration process.

2.2.3 Inverse Simulation Methods

Forward simulation consists of tuning muscle-activation signals to produce desired kinematics. There are two main challenges associated with predicting the muscle activations of oropharyngeal structures. Firstly, muscle forces are

difficult to measure, while performing complex motor tasks, such as running speech. Secondly, the associated biomechanics is redundant to the motion space of the system. For example, a tongue model (such as Buchaillard et al. [2009]) contains several pairs of extrinsic and intrinsic muscles, varying activations of which may cause similar motion. This is mostly due to the computational limitations of the models for differentiating all possible kinematic DOFs of the soft-tissue. This phenomenon is called motor redundancy, and has been confirmed for tongue configurations in speech movements (Stavness, 2010). Direct tuning of tongue muscles has been reported previously in the literature (Fang et al., 2009), but the process is mostly manual and relies heavily on trial-and-error. As an alternative, data-driven methods estimate muscle activations based on the measured kinematics, by solving an inverse problem.

Inverse methods tend to optimize the solution to a set of system equations. These equations incorporate the force and kinematic measurements, subject to specific biomechanical constraints. Static optimization methods calculate the net force of the system for a sustained configuration of the model, such as point-to-point movements of the jaw (Silva and Ambrsio, 2004). In order to break down the net force of the system to those of individual muscles, the process needs to be repeated for each instance of motion. An instantaneous cost function, such as minimum excitation, is used to resolve muscle redundancy. Static optimization may magnify the recording error through differentiation of velocity and acceleration (Erdemir et al., 2007).

Dynamic optimization methods apply more biologically plausible assumptions, such as minimizing metabolic energy expenditure per unit of distance (based on the start and end point of the motion). The approach tends to produce accurate results, but is computationally expensive. Some studies (such as the one by Anderson et al. (2001) on joint movement) show practically equivalent results for static and dynamic inverse optimization, and suggest that, depending on the motor task, the dynamic optimization may not be necessary. However, the dynamic scheme is still useful if, 1) accurate experimental data is not available, 2) activation dynamics is known and plays an important role in the task, or 3) the ability to predict novel movement is desired (Anderson et al., 2001).

The trajectory-tracking simulation is a popular inverse modeling technique widely used for musculoskeletal systems, where the muscles are mainly mod-

eled as mass-less springs (Erdemir et al., 2007). The method is also expanded to quasi-static FE Models for the face (Sifakis et al., 2005). Each input of muscle activations and skeletal configuration is directly mapped to the steady state expression it gives rise to. Stavness (2010) includes the muscular-hydrostatic properties of the tongue (such as incompressibility) to enable dynamic simulation of a FE tongue model. The method shows promise for simulation of soft-tissue (such as those of the tongue) which are activated without the mechanical support of a rigid skeletal structure. It uses per-timestep static optimization (as opposed to optimizing over the full time-varying trajectory), and is computationally efficient; nevertheless, it may lead to suboptimal muscle activations. Estimated activations are fed back to the forward simulation, and the error in the model's trajectories is used to re-adjust the prediction.

2.3 Acoustics of Speech

Speech, the vocalized form of human communication, can be subdivided into small sound units called phonemes. Vowels are voiced phonemes, meaning they are articulated through vibration of the vocal folds, while the VT remains mainly open. Vowels function as the core of speech syllables. An example is the vowel *a* (/æ/) in the English word *cat*, written in the International Phonetic Alphabet (IPA) as /kæt/. In contrast to vowels, consonants mostly serve as the beginning (onset) or end (coda) of a syllable, and are articulated with complete or partial closure of the VT.

2.3.1 Vowel Phonemes

Several articulatory features, referred to as *quality*, are associated with each vowel to distinguish it from others. From the vowel qualities common in phonetic studies, *height* and *backness/frontness* depend on the vertical and horizontal position of the tongue respectively; *roundness* relates to the configuration of the lips; *nasality* is associated with the position of the velum. Figure 2.7 shows the IPA classification of the monophonic English vowels based on their qualities. It also includes a schematic representation of the

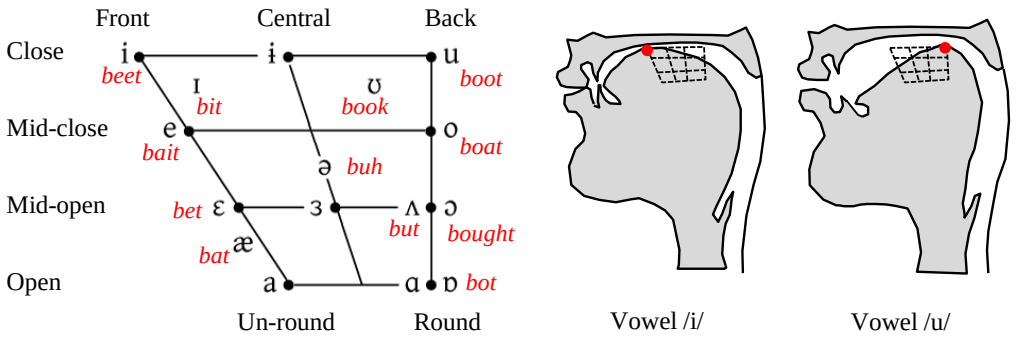


Figure 2.7. IPA classification of monophonic English vowels based on their qualities (left) vs. a mid-sagittal schematic representation of the front vowel /i/ (middle) and back vowel /u/ (left).

articulators in the front vowel /i/ and the back vowel /u/ in the mid-sagittal plane.

Vowels are periodic signals, so they are expected to introduce distinct harmonic peaks in the frequency domain. In particular, the first two peak frequencies, known as F_1 and F_2 formants, are used to define distinct vowels. The value of F_1 is mainly determined by the height of the tongue body; it is higher for an open vowel (such as /a/) and lower for a closed vowel (such as /i/). On the other hand, the value of the F_2 is effected by the backness-frontness of the tongue body; it is higher for a front vowel (such as /i/) and lower for a back vowel (such as /u/) (Ladefoged, 2001). Several studies measure the formant values of the vowels from audio signals across different populations and languages. Figure 2.8 shows the formant diagram for 10 English vowels as proposed in the early 1950's by Peterson and Barney (1952). Later on, Hillenbrand et al. (1995) revisited that study, considering dynamic aspects of vowel production, such as duration and spectral change. Their results suggest numerous differences with those of Peterson and Barney (1952), both in terms of average frequencies of F_1 and F_2, and the degree of overlap among adjacent vowels. Other studies, such as that of Hillenbrand (2003), yield different diagrams for phonetically distinct dialects. We refer to Jacewicz and Fox (2013) for a recent thorough study of the change in formant trajectories in North American English across dialects, gender, and age.

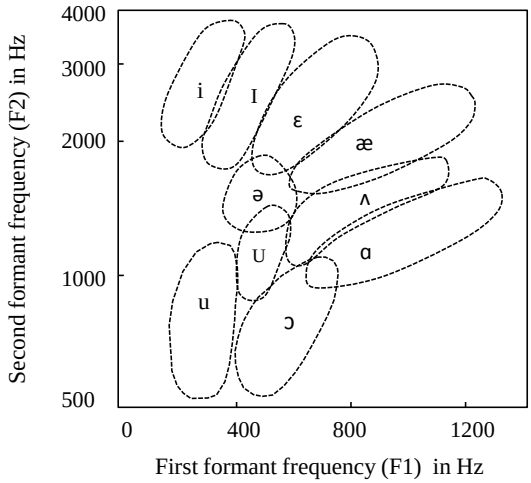


Figure 2.8. Regions of average formant frequencies for 10 English vowels across 76 speakers, as described in Peterson and Barney (1952).

2.3.2 Consonant Phonemes

Consonants are classified by their manner and place of articulation, as well as voiced or unvoiced qualities. By manner, consonants are categorized, as 1) plosives or stops (such as /p/) where air is completely blocked and bursts after release of the constriction, 2) fricatives (such as /s/) where a turbulent air-flow is generated at the point of constriction, 3) nasals (such as /m/ and /n/) generated by lowering the soft-palate, 4) liquids (such as /l/ and /r/) produced by raising the tip of the tongue, and 5) semi-vowels (such as y [/j/] in the word yes) which are phonetically similar to vowels, but function as a syllable boundary rather than its nucleus.

Consonants are also categorized by place of articulation (POA), i.e., the point of contact of the active (e.g., lower lip or tongue) and passive (e.g., upper lip or teeth) articulators. Figure 2.9 illustrates the POA of each category, while including a corresponding example from English sounds. Most of the consonants do not have harmonic frequency spectra, making their acoustic cues more difficult to ascertain than vowels; however, spectral features (periods of silence, voice bars, noise, and effects on adjacent phonemes) can be used to distinguish consonants.

1. Bilabial /p/
2. Labiodental /f/
3. Dental /θ/
4. Alveolar /s/
5. Postalveolar /ʃ/
6. Palatal /j/
7. Velar /k/
8. Uvular —
9. Glottal /h/

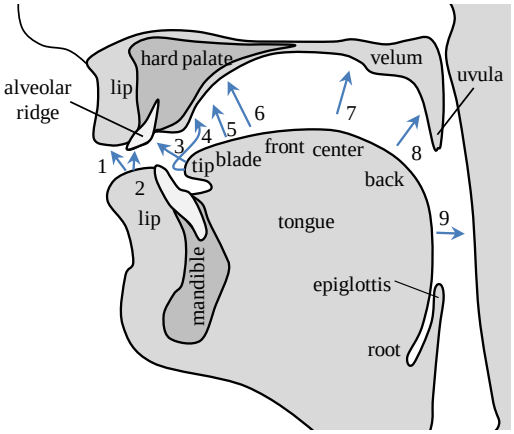


Figure 2.9. Classification of consonants based on place of articulation, each accompanied by an example from English phonetics. Adapted from 1994 Encyclopædia Britannica.

2.3.3 Coarticulation

The term *coarticulation*, also referred to as *assimilation*, has been used to describe the spreading of acoustic and/or articulatory features of a phoneme to its adjacent neighbours in running speech. In general, humans are able to speak as many as five syllables per second: slightly faster than the rate articulators are able to adjust to independent articulatory positions. The brain, however, has an intelligent way of planning an optimal compromised trajectory for adjacent phonemes, and saves time and energy. For example, although the nasal consonant /n/ is normally alveolar, it exhibits a dental POA in the word tenth (tɛnθ), aligning with the following dental sound /θ/. Coarticulation explains the variation in phonetic manifestation of a given sound according to its nearby sounds (Ohala, 1993).

Coarticulation has been hypothesized to occur both in anticipation of an approaching phoneme (*anticipatory*), as well as when the effect of a phoneme is carried over to the following phoneme (*perseverative*). Figure 2.10 shows this effect for three overlapping phonetic gestures. The degree of overlap depends on number of variables, such as the distance of the gestures in articulatory and perception space. Look-ahead models (based on anticipatory hypothesis) predict that some features of a given vowel (such as lip rounding), or a given consonant (such as nasality), affect its preceding phonemes in the

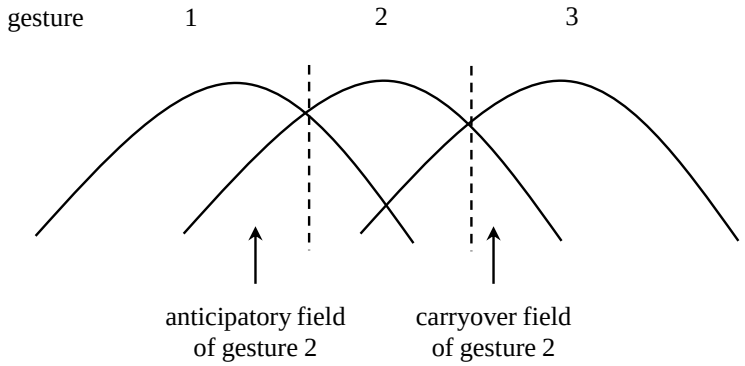


Figure 2.10. Anticipatory and carryover fields for three overlapping phonetic gestures as explained in Fower and Satzman (1993).

speech plan, until another vowel or consonant is reached. Contradicting observations, however, gave rise to constrained-frame models, suggesting that the aforementioned features are time-locked to their associated phoneme, and that the effective field of spreading is smaller than suggested by the look-ahead models (Fower and Satzman, 1993).

2.3.4 Articulatory Speech Synthesis

Speech synthesizers focus on generating the sounds that resemble human speech. Articulatory speech synthesizers, in particular, use a representation of the vocal folds and tract to create the desired acoustics for an observed shape of the oral cavity (Doel et al., 2006; Birkholz et al., 2013). The acoustic theory of speech production suggests that both vowels and fricatives can be generated using a source-filter system (Fant, 1960). For vowels, vibration of the vocal folds, under the expiatory pressure of the lungs, is the only source for the system. The VT, consisting of the larynx, pharynx, oral and nasal cavities, constitutes the filter where sound frequencies are shaped.

Wave propagation in the vocal tract

Traditionally, the acoustic system is approximated by a 1D wave equation, referred to as the *Webster* equation. It associates the slow varying CSA of a rigid tube with the pressure wave for a low-frequency sound. For the sound particle velocity $v(x,t)$, at time t, in a tube with area function $A(x)$, along the axis x, the Webster equation yields the following:

$$\frac{1}{c^2}\frac{\partial^2 v(x,t)}{\partial^2 t} = \frac{\partial^2 v(x,t)}{\partial^2 x} + \frac{1}{A(x)}\frac{\partial A(x)}{\partial x}\frac{\partial v(x,t)}{\partial x} \tag{2.3}$$

where c is the velocity of the sound propagation (Benade and Jansson, 1974). Unfortunately, it is not possible to solve the Webster equation analytically for an arbitrary $A(x)$. For a constant A, however, the equation simplifies to

$$\frac{1}{c^2}\frac{\partial^2 v(x,t)}{\partial^2 t} = \frac{\partial^2 v(x,t)}{\partial^2 x} \tag{2.4}$$

which has a general solution in the form of waves propagating in opposite directions. Using this simplification, Kelly and Lochbaum (1962) approximate the VT with cylindrical segments of constant area. In each segment, the wave gains a propagation delay, and is partially reflected at its junction with the next segment. The Kelly-Lochbaum model can be expressed as a ladder (or a waveguide) filter and, hence, is straightforward to implement. A number of improvements of this basic model have been proposed, to include other features, such viscous and thermal losses along the path of propagation, radiation at the lips, and time-varying nature of the VT (Välimäki and Karjalainen, 1994; Doel and Ascher, 2008).

The complex shape of the VT, with its side branches and asymmetry, has motivated use of higher-dimensional acoustic analysis. The common 3D methods (such as the boundary element method [BEM] [Kagawa et al., 1992], finite element method [FEM] [Vampola et al., 2008a] and the finite-difference time-domain method [FDTD][Takemoto et al., 2010]) produce more accurate results at the price of higher computational cost. 2D methods have been proposed as a compromise, but they must overcome significant errors in the spectra – especially for the first formants – related to the use of circular cross-sections at the mid-sagittal slice of the 3D VT (Arnela and Gausch,

2014). Finally, several studies find that the spectra yielded by 1D acoustic analysis matches closely to those of 3D analysis for frequencies less than 7KHz (Takemoto et al., 2014; Arnela and Gausch, 2014). The discrepancy between the formant frequencies of the filter and the recorded audio is attributed to the insufficient boundary conditions – especially in case of open lips and/or velar port (Aalto et al., 2012).

Coupling the models of vocal folds and tract is necessary for solving the acoustic system. Popular techniques in the literature are based on direct numerical simulation of either a transmission line circuit model (TLM) (Ishizaka and Flanigan, 1972) or a hybrid time-frequency system (Sondhi and Schroeter, 1987). In particular, TLMs provide a descriptive analogy between acoustical and electrical circuits, and are widely adopted in modeling the complex turbulence noise sources for voiceless consonants (Birkholz et al., 2007).

Excitation Source

A steady air-flow passes from the lungs into the trachea, and through the vocal folds (also known as vocal cords) to reach the VT. The primary source of vocal excitation for voiced phonemes is the quasi-periodic vibration of the vocal folds, as explained in the myoelastic-aerodynamic theory. The vocal folds suck together to close the air passage, due to negative supraglottal pressure generating a Bernoulli effect. Subglottal pressure builds up and bursts the vocal folds open, generating a pulse of air pressure at the glottal exit. The cords oscillate before viscosity kills off their vibration. The elasticity of the vocal folds and, hence, the frequency of their vibration is actively controlled by the tension of the thyroarytenoid muscle.

In the late 1960's, Flanagan and Landgraf (1968) propose a simple single mass model to describe the physics of glottal vibration. The preliminary results were satisfactory, but suffered from failure to permit non-uniform, out-of-phase movement of the tissue at the glottal entry and exit. Later on, Ishizaka and Flanigan (1972) introduced a more complex, two-mass model of the vocal-fold vibrations, which was shown to be adequate for synthesis of voiced vowels. The glottis takes a convergent and a divergent shape during each cycle, which creates a rise in the driving pressure and its asymmetry (Scherer et al., 1983), and leads to self-sustained oscillation of the vocal folds (Titze, 1988). Figure 2.11 shows the schematic of the model. The two-

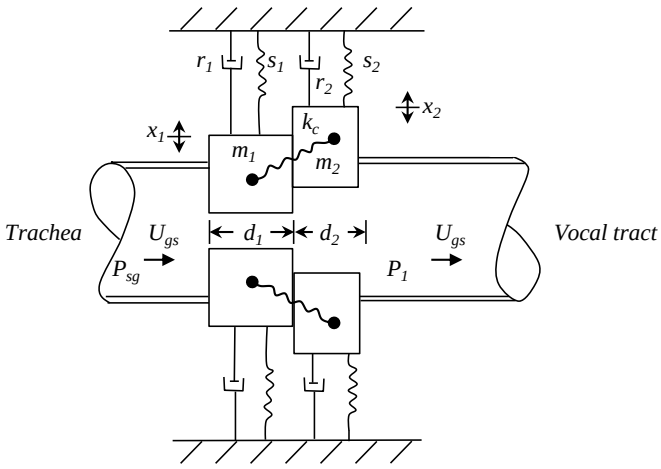


Figure 2.11. Two-mass glottal model, as proposed by Ishizaka and Flanigan (1972).

mass model is proven insufficient for study of glottal pathologies. Several advanced models have been proposed to provide a stronger correlation with biology of the vocal folds (such as their layered structure) at the expense of computational cost (Story and Titze, 1995). We refer to Cveticanin (2012) for a comprehensive review of these models.

Voiceless excitation of the VT is the result of non-acoustic air motions (such as turbulence at articulatory constrictions), and plays an essential role in pronouncing consonants. Accurate physical models to describe this phenomenon currently do not exist. Contemporary methods model the turbulence with a random perturbation in either the pressure or velocity field, usually by inserting lumped noise sources in the TLMs. Good results are achievable, but subject to fine-tuning the parameters of noise sources, such as their numbers, place, level, and spectra (Birkholz et al., 2007).

2.4 Conclusions

Recent advances in acquisition technologies have made it possible to capture abundant data during speech, in terms of audio, medical images and physiological recordings. Fast MRI, in particular, has contributed vastly to understanding the anatomy and motion of the articulators. Computational

models complement such data in describing speech phenomena. Generic biomechanical models of oropharyngeal structures are evolving into complex descriptors of speech behaviour; methods for subject-specific modeling are gaining popularity for describing variability in the physiology of the human body. The acoustics of sound for speech production have been studied for many decades. Mathematical models, as well as their numerical implementations, describe the complex physics of sound propagation in the VT. Articulatory speech synthesizers, in particular, have shown promise in providing valuable insight into understanding the speech process, rather than focusing on speech synthesis.

This chapter has presented an overview of the ongoing research in the fields of speech data acquisition and measurement, biomechanical modeling, and acoustic analysis. We have identified areas of research that require further investigation. The available models of speech articulators are generic and irrelevant for simulation and validation using speaker-specific data. Segmentation of soft-tissue from MRI is challenging and time-consuming, and appears to cause a bottle-neck in subject-specific modeling. There is also a need for a framework that enables acoustic analysis based on the biomechanics of each individual speaker. In the following chapters we describe our contributions to these open research problems.

Chapter 3

Subject-Specific Modeling of the Oropharynx

In this chapter we demonstrate our design for developing subject-specific, biomechanical models of the oropharynx, according to MR images of individual speakers. We use biomechanical information provided by standard generic models, as available in the ArtiSynth simulation framework and described by Stavness et al. (2011, 2012, 2014a,b). Our oropharyngeal model includes a FE model of the tongue, coupled with rigid-body bone structures (such as the mandible, maxilla and hyoid). Later, in Chapter 5, we include a deformable, air-tight model of the vocal tract to enable acoustic synthesis.

The underlying data for construction of the models (for each speaker) is one cine MR image volume of the head-and-neck. This image is the first volume in a sequence of 26 time-frames that capture the utterance /ə-gis/, acquired by our collaborators Woo et al. (2012) at the University of Maryland Dental School, in Baltimore MD, USA. This image volume precedes the phoneme /ə/ and, thus, bears the most resemblance to a neutral tongue posture. We make our subject-specific models for four healthy English speakers, anonymously identified as A, B, C, and D. Figure 3.1 shows the mid-sagittal view of our image volumes. More information on this dataset is provided in Chapter 4,

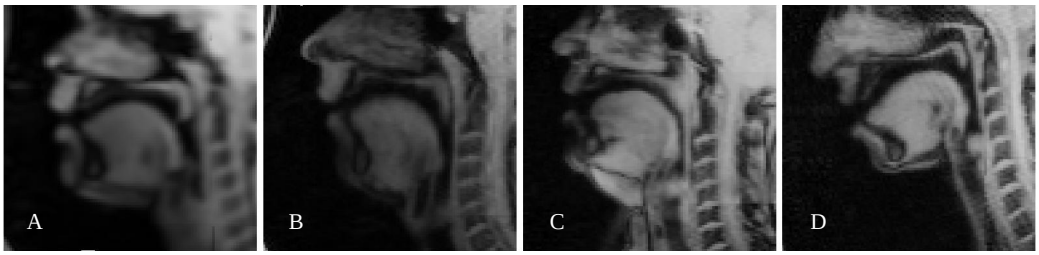


Figure 3.1. Mid-sagittal view of the 1st TF of cine MRI data for speakers A-D.

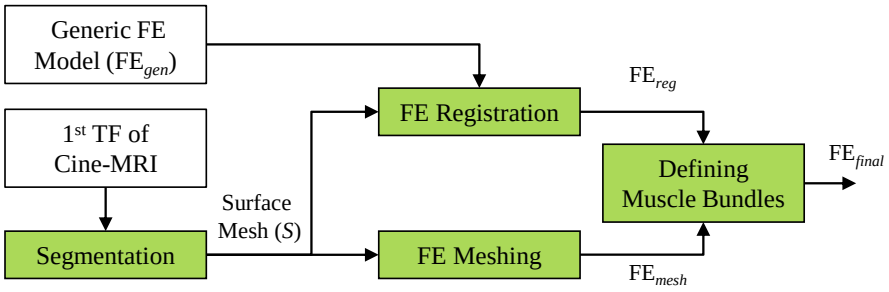


Figure 3.2. Designed pipeline for generation of high-resolution, subject-specific FE modelf of the tongue (FE$_{final}$).

where we simulate our models based on MR images.

3.1 FE Tongue Modeling

To create our subject-specific tongue models, we design and follow the pipeline shown in Figure 3.2. As an input to our pipeline, we use the standard generic FE tongue model developed by Buchaillard et al. (2009), which is described in more details in Chapter 2; this model provides 2493 DOFs (946 nodes and 740 elements), and consists of 11 pairs of muscle bundles with bilateral symmetry. We refer to this generic model as FE$_{gen}$ during the rest of the chapter.

We follow our framework by delineating the surface geometry of the tongue from cine MRI, using the methods we develop and validate later in Section 3.2. We refer to the surface mesh as S. Based on S, we create two versions of FE tongue models, using a registration as well as a meshing technique.

Our **first tongue model** (FE$_{reg}$) is the result of the registration of FE$_{gen}$ to S. We use a multi-scale, iterative, and elastic registration method called Mesh-Match-and-Repair (Bucki et al., 2010a). The registration starts by matching the two surfaces, and then applies the 3D deformation field to the inner nodes of FE$_{gen}$, via interpolation. A follow-up repair step compensates for potential irregularities of the elements. Note that the elements of FE$_{reg}$ – similar to FE$_{gen}$ – are aligned along muscle fibres. Because of this, the size of the elements depends directly on the density of the muscle fibres in each

region of the model. This results in smaller elements at the anterior-inferior region of the tongue (where most fibres originate), and larger elements close to the dorsum of the tongue (where most fibres span into the tongue body). Unfortunately, low resolution elements are located in the region undergoing maximum deformation during speech.

To address our concerns about the resolution of FE_{reg}, we generate our **second tongue model** (FE_{mesh}) following the pipeline shown in Figure 3.2. We use a meshing technique proposed by Lobos (2012) to generate a regular, mixed-element FE mesh (referred to as FE_{mesh}). The meshing algorithm starts with an initial grid (of hexahedral elements) that encloses the surface mesh S. The method then eliminates those elements that present little or no intersection with S, employing a set of mixed-element patterns to fill generated holes at the surface boundary. Further, the quality of the mesh is improved using the Smart Laplacian filter (Freitag and Plassmann, 2000). FE_{mesh} bares our desired resolution and is well-behaved during simulation.

Finally, we augment FE_{mesh} with the definition of muscle bundles available in FE_{reg}; since both FE models are in the same spatial domain, we simply copy the bundle locations from FE_{reg} to FE_{mesh}, replacing the bundle's elements with those of FE_{mesh} which fall into the bundle's spatial domain. Our approach for generating FE_{final} provides multiple fundamental advantages over using FE_{reg}. Firstly, the user has control over the mesh resolution. Secondly, the muscle fibre definitions are no longer tied to the configuration of the elements; therefore, it is possible to modify the muscle fibres based on different linguistic hypotheses and preferences.

The rest of this section describes units of our modeling process in more detail.

3.1.1 FE Registration

To create FE_{reg} from FE_{gen} (from Figure 3.2), we use a FE registration method called *Mesh-Match-and-Repair* (MMRep), which performs iteratively in two consecutive steps of elastic registration and mesh repair (Bucki et al., 2010a). In the mesh registration phase, we first coarsely align our generic source mesh (FE_{gen}) to the target surface (S), by matching the corresponding landmarks between the two. Next, FE_{gen} is embedded in a deformable *virtual* elastic grid. Local elastic registration is then performed by applying

successive elementary deformations in a coarse-to-fine grid scheme. At each grid level, the registration energy (E), is computed and minimized iteratively. Each iteration of the method consists of the following steps (see Figure 3.3 for 2D illustration):

1. Calculate E based on a geometrical similarity measure between corresponding local nodes of FE_{gen} and S.
2. Compute the gradient of E with respect to each grid node position, while other grid nodes are fixed.
3. For each element, find the displacement vector that minimizes E.
4. Convey such displacement vector (from step 3) to the source points located in neighbouring cells, using a distance-based weight function.

After each iteration, the virtual grid returns to its initial (regular) configuration, embedding the newly deformed version of the FE_{gen}. If no substantial energy decrease is observed, the grid is refined by subdividing each cell into eight smaller ones, and the algorithm moves to the next level. The weight functions are designed to ensure C^1 differentiability, bijection, and non-folding property of the total deformation.

To limit the space distortion, the method evenly distributes multiple control points inside the source mesh, and computes a potential elastic energy at them based on the Green Lagrange strain and stress tensors. The preferred elementary deformation at each iteration is the one which can provide the best ratio between registration energy decrease and elastic energy increase.

In the mesh-repair phase, the MMRep algorithm uses a two-fold process to compensate for the regularity and quality of the distorted elements. **Firstly**, all elements are inspected for possible irregularities, and the nodal positions adjusted to regularize the irregular elements. The regularity energy (E_R) in region A is defined as:

$$E_R = \sum_{j \in A,} \phi_k(\mathrm{J}_j) \tag{3.1}$$

where J_j is the Jacobian of the local deformation at node j, and $\phi_k(t) = 1 - exp(-kt)$ defines a penalty function of strength k. E_R is maximized to find a regular configuration. Higher values of k favour the solution in which all Jacobians are positive.

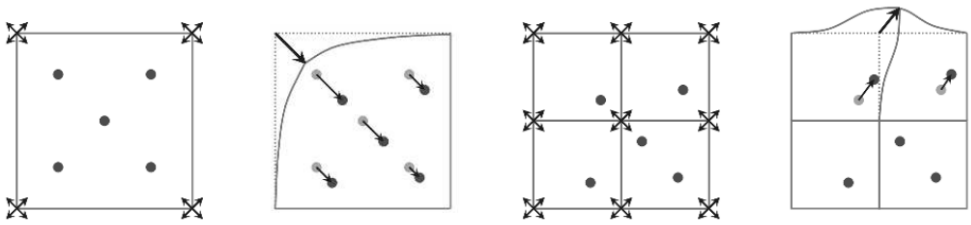


Figure 3.3. Elastic registration in the MMRep algorithm (Bucki et al., 2010a): the source point-set at refinement level 1 (a); after deformation at level 1 (b); at refinement level 2 (c); and after deformation at level 2 (d). ©Bucki et al. (2010a), adapted with permission.

Secondly, the quality of the mesh is improved by finding a configuration that maximizes the quality energy (E_Q) in region A, defined as:

$$E_Q = \sum_{j \in A, e \in A} \psi_k(\mathrm{JR}_j^e) \tag{3.2}$$

where $\psi_k(t) = 1 - exp(-k(\mathrm{JR}_{min} - t))$, and JR_{min} is a predefined minimum satisfactory level of the Jacobian ratio (JR). Bucki *et al.* use $\mathrm{JR}_{min} = 1/30$ in accordance with standard suggested by ANSYS FE analysis software.

Both of the steps involved in the repair phase can alter and, hence, reduce the geometrical accuracy of the surface. To avoid this issue, the number of repair steps is limited to 50, with a maximum node displacement of 0.1mm in each iteration.

Using the MMRep algorithm, we register FE_{gen} into the tongue surfaces (S) we segmented from the cine MRI data of four healthy speakers. Figure 3.4 shows the registration results (FE_{reg}) in the mid-sagittal plane for speakers A-D. To assess the quality of each hexahedral mesh, we calculate the normalized Jacobian ratio (JR) of its elements as follows: we first compute the Jacobian for each node of the element, and then calculate the quotient between the minimum to maximum Jacobian value found within the element. We average the values of JR^e over all elements in each mesh, to obtain JR values of 0.34 (for speakers A, B and D), and 0.36 (for speaker C). The values of JR^e, however, range anywhere between $1/30 \approx 0.03334$ and 0.9.

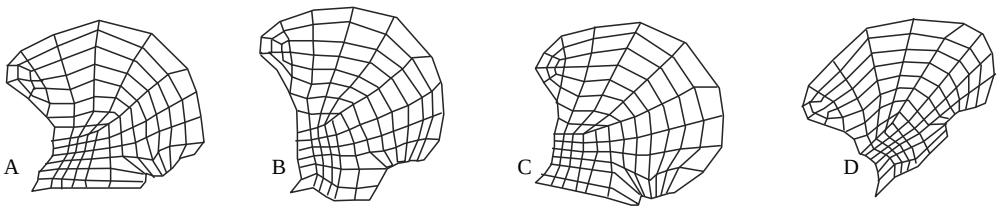


Figure 3.4. Results of FE registration using MMRep (Bucki et al., 2010a): element configuration in the mid-sagittal plane of FE_{reg} for speakers A-D.

3.1.2 FE Meshing

As discussed in Section 2.2.2, hexahedral meshes are preferable for most FE methods in a wide variety of simulation problems; however, because of the cubic shape of the elements, such meshes fail to achieve an adequate approximation of curved domains at the tongue surface. To deal with such issue, Buchaillard et al. (2009) designed an intricate element configuration for their generic model, in which elements align with the direction of the surface. To give a greater degree of freedom to our models (such as adjustable spatial resolution), we adopt a more conventional FE hexaheral meshing developed by Lobos (2012) that deals with the problem of curved domains through limited use of other types of elements (wedges, pyramids and/or tetrahedra). The method introduces a set of mixed-elements patterns – employed at the surface of the target domain – and conserves hexahedra elsewhere. These patterns can combine with any regular or non-regular hexahedral meshing technique, and achieve acceptable representation of the surface, while ensuring proper connectivity among elements.

A hexahedral mesh, built for the surface S, includes elements that are completely outside, completely inside, or at the boundary of S. Elements at the boundary are defined as those that have some nodes inside and others outside S. One way to deal with outside nodes is to project them into S. Such a solution leads to two problems: firstly, a mesh can become tangled because if an element's edges are crossed; secondly, element quality issues may arise due to node proximity. In his method, Lobos (2012) replaces the boundary elements – when necessary – with other type of elements.

Lobos's method uses the face subdivision rules, explained in Figure 3.5 (top), to handle a quadrilateral face that intersects S. The dashed lines are inserted

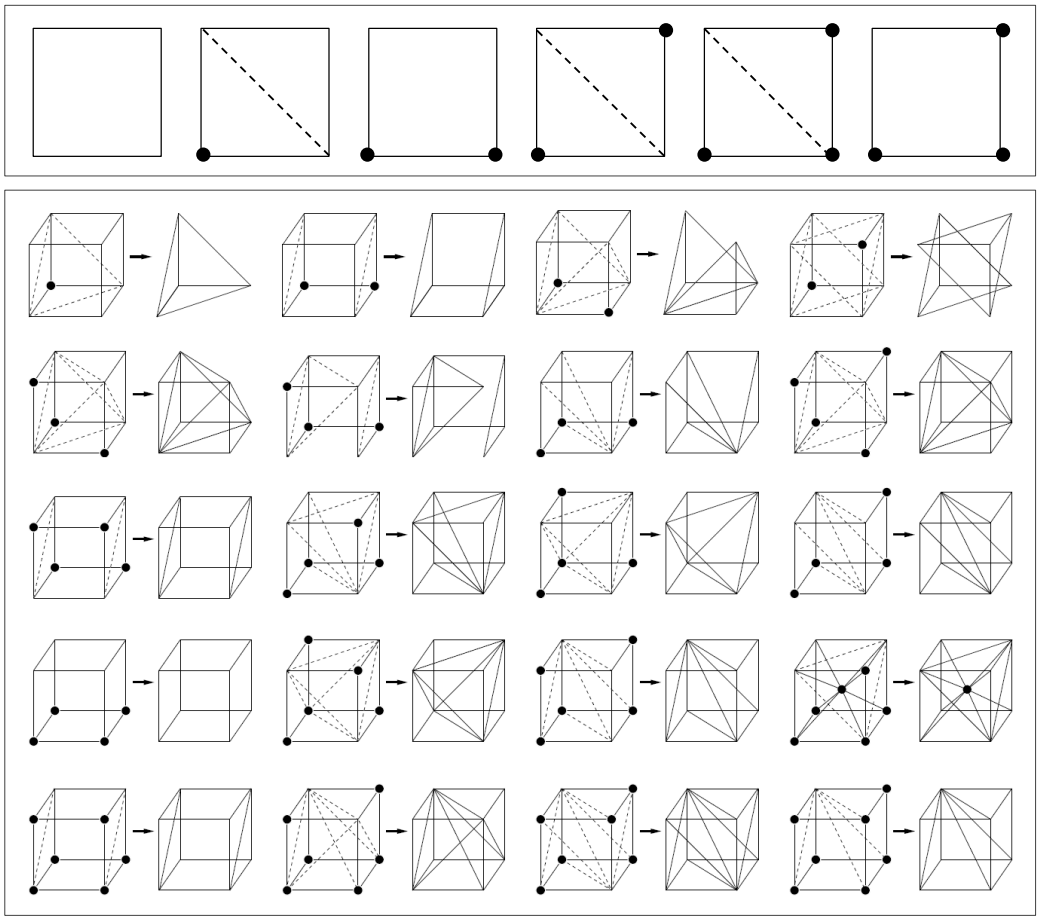


Figure 3.5. Replacing boundary hexahedra with mixed elements: face consistency patterns (top), and surface patterns (bottom). Dashed lines represent diagonals to be inserted, and dots show inside nodes. ©Lobos (2012), adapted with permission.

to split the face into two triangles, ensuring topological consistency between neighbor elements. Use of these basic rules leads to a set of mixed-element patterns shown in Figure 3.5 (bottom).

The steps involved in the algorithm can be summarized as follows:

1. Produce an initial grid with desired resolution that covers S.
2. Eliminate elements that present little intersection with S; For any **inside** node that lies close to the surface, project it to S.

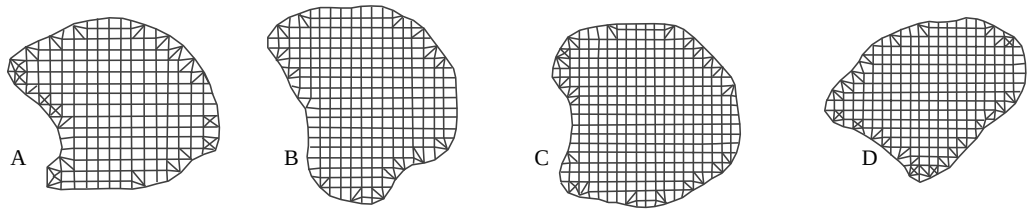


Figure 3.6. Results of FE meshing using mixed-element patterns (Lobos, 2012): element configuration in the mid-sagittal plane of FE_{mesh} for speakers A-D.

3. Replace all remaining boundary elements with mixed elements (Figure 3.5 [bottom]).
4. Project the **outside** nodes of new boundary elements to S.
5. Improve the quality of the boundary elements.

We follow Lobos (2012) in using the Smart Laplacian filter (Freitag and Plassmann, 2000) to improve the quality of the new mixed elements (as instructed in step 5). We apply the 4^{th} level of grid refinement in the algorithm to achieve our desired spatial resolution (with a typical element being approx. 5mm-wide in each dimension). Figure 3.6 shows results of meshing (FE_{mesh}) in the mid-sagittal plane for speakers A-D; Table 3.1 includes the number of tetrahedra, pyramids, wedges, and hexahedra in each model.

At the end, we estimate the mesh quality of each element; for pyramids, wedges and hexahedra, we use the normalized Jacobian Ratio (JR) as formulated by Joe (2008). Since JR=1 for any tetrahedron, we use another quality measure called Aspect Ratio (AR) for the tetrahedral elements (Lobos et al., 2007):

$$\text{AR}^e = \frac{(\frac{1}{6}\sum_{i=1}^{6} l_i^2)^{\frac{2}{3}}}{8.47867 V^e} \tag{3.3}$$

where V^e denotes the volume of the element e, and l_i is the length of the i-th edge in e. AR is 1 for the quadrilateral tetrahedron and $\to \infty$ as e becomes increasingly distorted. Table 3.1 also shows the average quality measure (AR or JR) for each type of element.

Table 3.1. Mesh quality measured for FE_{mesh} in speakers A-D. Number of elements, as well as average values of the quality measure (AR or JR), are presented for each specific element type.

Speaker	Tetrahedra	Pyramids	Wedges	Hexahedra
A	1004	880	529	1414
	AR=2.38	JR=0.86	JR=0.84	JR=0.93
B	917	808	805	2061
	AR=2.37	JR=0.85	JR=0.88	JR=0.94
C	1026	962	757	2741
	AR=2.47	JR=0.86	JR=0.87	JR=0.95
D	1124	1040	626	1763
	AR=2.34	JR=0.84	JR=0.84	JR=0.94

3.1.3 Tongue Muscle Bundles

The top row of Figure 3.7 illustrates the process of defining the muscles in high resolution. The goal is to define a muscle bundle (M_{final}) in FE_{final} that corresponds to a specific muscle bundle (M_{reg}) in FE_{reg}. Since both FE_{reg} and FE_{final} share the same coordinates, the fibers of M_{reg} (indicated in red in Figure 3.7) are simply copied to M_{final}. The elements of M_{final}, however, need to be redefined.

Consider element e in FE_{final}. In a simple and intuitive approach, we can assign e to M_{final} if e falls within a predefined distance (d) to the fibers of M_{final}. However, no single value of d yields satisfactory results. Firstly, in the regions where fibers are very close to each other, their corresponding elements tend to overlap. Overlapping elements may introduce error in the inverse solver, where a wrong muscle may be considered responsible for an unrelated motion. Secondly, in the regions where fibers are relatively far from each other, elements in between fibers tend to fall out of the muscle definition and create *holes* in the muscle. These *holes* may cause inhomogeneity in the force-activation behaviour of the muscle.

We assign e to a certain M_{final} if the elements of the corresponding M_{reg} contain the e. In addition, we incorporate adjacency relationships between the tongue muscles – as in the generic tongue model – to avoid the overlap between non-overlapping bundles.

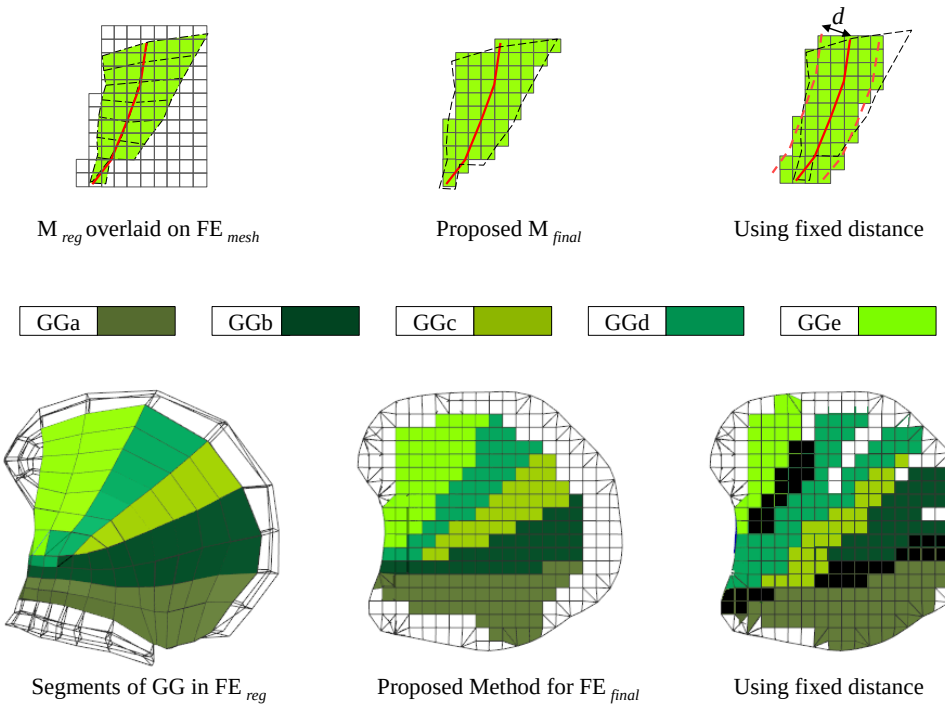


Figure 3.7. Defining muscle elements in the high resolution FE tongue model (top row), as well as functional segments of the genioglossus muscle for speaker C (bottom row). Overlapping elements are shown in black, for the muscle elements defined using a fixed, predefined d distance.

The method is explained in the following steps: For each element e in FE_{mesh}

1. Compute the distance ($Distance_e^i$) from e to M_{reg}^i (the i^{th} muscle bundle in FE_{reg}) for $0 < i < 21$.
2. Initiate a first-in first-out queue (Q); add the indices i to Q in ascending order of $Distance_e^i$; denote the n^{th} value in Q with q_n.
3. Iterate until Q is empty:
   (a) Read q_1 from Q, and add it to the list of bundles for e.
   (b) If q_n $(1 < n < 21)$ does not interdigitate with q_1, remove q_n from Q.
   (c) Remove q_1 from Q.

$Distance_e^i$ (in step 1) is computed as the minimum euclidean distance from the centroid of e to the centroids of all elements in the muscle bundle M_{reg}^i. The number 21 is inclusive of the left and right muscles. Step 3b benefits from a predefined binary matrix that shows the interdigitation between the muscle bundles (e.g., the entry for the TRANS, and VERT is 1, since they share elements in FE_{gen}). We refer to Appendix A for more information on the anatomy of tongue muscles.

The bottom row of Figure 3.7 shows muscle elements of the five functional segments of the genioglossus (GG) muscle for speaker C. The proposed method is compared with the approach using a predefined, fixed distance, while the muscle elements in FE_{reg} serve as the ground truth. Note that the proposed method preserves the boundary of each segment, while preventing overlaps and holes in muscle definition.

The bone attachments in the tongue model (the FE nodes at which the model is biomechanically coupled to the mandible and hyoid rigid bodies) are also transferred from FE_{reg} to FE_{final}. For each attachment node in FE_{reg}, the closest node (according to euclidean distance) in FE_{final} is considered to be the corresponding attachment.

Each muscle bundle in the tongue can be further divided into functionally-distinct fibre groups (referred to as *functional segments*), which are believed to be controlled quasi-independently in a synergistic coordination (Stone et al., 2004). We divide the muscle fibers of the GG, VERT and TRANS into five functional segments (a: posterior to e: interior). This division was initially proposed based on EMG measurements from the GG (Miyawaki et al., 1975), and later reinforced using information from ultrasound imaging and tagged MRI (Stone et al., 2004). We also follow Fang et al. (2009) in dividing the STY into two functional segments: STYa (the anterior part within the tongue), and STYp (originating from the posterior tongue to the styloid process). Note that FE_{gen} includes only three functional segments for the GG and one functional segment for each of TRANS, VERT or STY (Buchaillard et al., 2009).

A more detailed analysis of the tongue's myoarchitecture would require a multiscale conceptualization of tongue muscle mechanics, as in the approach by Mijailovich et al. (2010). Detailed fibre fields could potentially be digitized from cadaver tissue, and registered to subject-specific muscle geometries, as in the method of Sánchez et al. (2014) for muscles in the forearm.

3.1.4 Tongue Muscle Material

The muscles in FE$_{gen}$ (Buchaillard et al., 2009) have been modeled using a hyper-elastic (Mooney-Rivlin) material. Fibers pass through it, indicating lines of action. Applying an external force through the fibers compresses the muscle material according to the Monney-Rivlin constitutive model, as described in Equation 2.1.

For our tongue models, we borrow a term (W_B) from the Blemker muscle model, and add it into Equation 2.1 to modify the strain energy. W_B depends on the fiber stretch (λ) and activation level (a) of the muscle. Using the formulation by Blemker et al. (2005), we compute $W_B(\lambda, a)$ by solving the following:

$$\begin{aligned} \lambda \frac{\partial W_B}{\partial \lambda} &= \sigma_{\max} f^{\text{fiber}}_{\text{total}} \lambda / \lambda_{\text{opt}} \\ f^{\text{fiber}}_{\text{total}} &= a f^{\text{fiber}}_{active}(\lambda) + f^{\text{fiber}}_{\text{passive}}(\lambda) \end{aligned} \tag{3.4}$$

where $\sigma_{\max}$ is the maximum isometric stress and λ_{opt} is the fiber stretch at its optimum length. $f^{\text{fiber}}_{\text{active}}$ and $f^{\text{fiber}}_{\text{passive}}$ correspond to normalized active and passive force-length relationships of a muscle fiber, respectively). The activation level (a) can be any value between 0 (no activation) and 1 (maximal activation). As in Blemker et al. (2005), we assume a piecewise exponential form for the passive force, and a piecewise quadratic form for the active force:

$$\begin{aligned} f^{\text{fiber}}_{\text{passive}} &= \begin{cases} 0 & \lambda \le \lambda_{\text{opt}} \\ P_1(e^{(P_2 \lambda/\lambda_{\text{opt}} - 1)} - 1) & \lambda_{\text{opt}} < \lambda < \lambda^* \\ P_3 e^{\lambda/\lambda_{\text{opt}}} + P_4 & \lambda^* \le \lambda \end{cases} \\ f^{\text{fiber}}_{\text{active}} &= \begin{cases} 9(\lambda/\lambda_{\text{opt}} - 0.4)^2 & \lambda \le 0.6\lambda_{\text{opt}} \\ 1 - 4(1 - \lambda/\lambda_{\text{opt}})^2 & 0.6\lambda_{\text{opt}} < \lambda < 1.4\lambda_{\text{opt}} \\ 9(\lambda/\lambda_{\text{opt}} - 1.6)^2 & 1.4\lambda_{\text{opt}} \le \lambda \end{cases} \end{aligned} \tag{3.5}$$

We follow Blemker et al. (2005) in setting $P_1 = 0.05$, and $P_2 = 6.6$ in accordance with the measurements on muscle tissue (Zajac (1988)). We use $\sigma_{\max} = 1 \times 10^5$Pa, $\lambda_{\text{opt}} = 1.1$ and $\lambda^* = 1.5$ to achieve stability for our models. P_3 and P_4 are set so that $f^{\text{fiber}}_{\text{passive}}$ is C0 and C1 continuous at $\lambda = \lambda^*$. Parameters of the Mooney-Rivlin material are set as in Subsection 2.2.1.

Using the Blemker model with Mooney-Rivlin material, we account for the non-linearity, incompressibility, and hyper-elasticity of the tongue muscle tissue. In addition, we use Rayleigh damping coefficients $\beta = 0.03s$ and $\alpha = 40s^{-1}$ to achieve critically damped response for the model.

3.1.5 Forward Simulation

Using our methods for FE tongue modeling, we first generate a high resolution version (FE_{Final}) of the generic tongue model (FE_{gen}). Both models use the Blemker material, as described in Subsection 3.1.4. We activate each muscle individually and compare the deformation in the low and high resolution models with each other. Figure 3.8 shows the results for 40% activation of the GGP, SL, and IL muscles, after reaching equilibrium. The tip of the tongue is on the left side of the figure; the tongue-jaw and tongue-hyoid attachment points are fixed, and the gravity is set to zero. In the case of the GGA, FE_{Final} shows superior ability to protrude, avoiding unnecessary curling on the top surface of the tongue. In the case of the SL, the high resolution model reaches upper high and back, at the tip, resulting in a small indent at the blade. In the case of the IL, FE_{Final} is able to curl (and depress) the tip and blade to a greater extent. Note that FE_{Final} also enables the user to modify (add, delete and relocate) the muscle fibers, and their element description, to produce desired deformations.

To evaluate the performance of our speaker-specific tongue models, we activate each individual muscle and assess the corresponding deformation. For each speaker, we apply the same level of activation to our model as we do to the (high resolution) generic model, and compare the resulting deformations. Figure 3.9 shows the results for 10% activation in the GGP muscle. Although bearing varying geometry and head posture, all speaker-specific models succeed in protruding.

Figure 3.10 shows the impact of such forward simulation, with 10% muscle activation, on the mid-sagittal contour of the tongue for speaker A. The mid-coronal plane is also included for the TRANS muscle which tends to narrow the tongue from the lateral sides and result in a larger contour in the mid-sagittal plane. After a series of experiments, we conclude that the speaker-specific models exhibit similar muscular behaviour to the generic model.

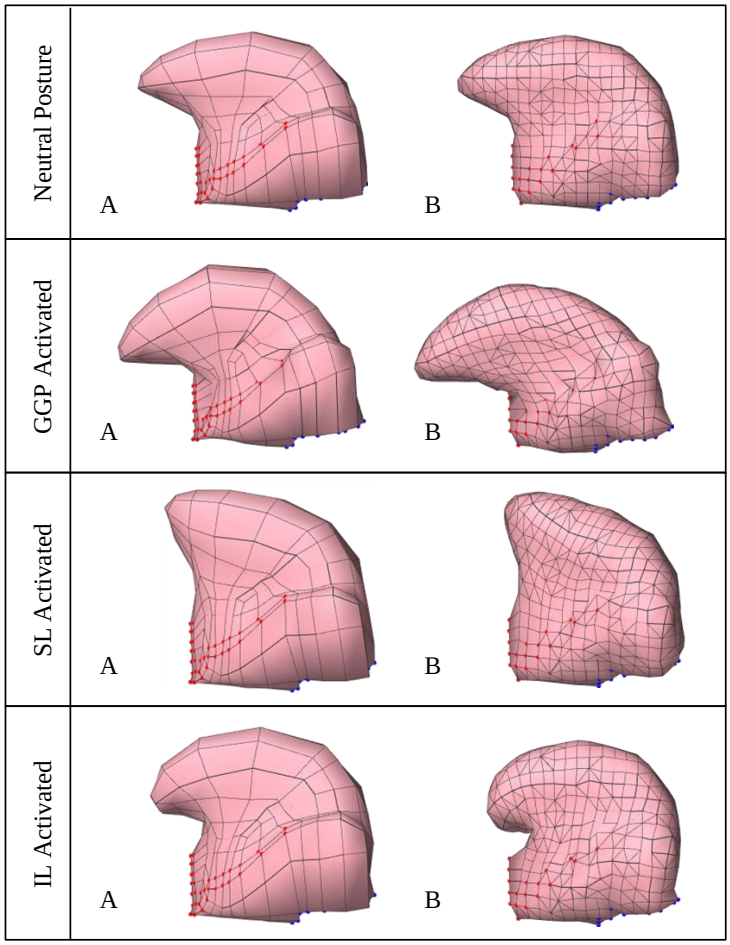


Figure 3.8. Impact of muscle activation on deformation of the generic tongue model, in original resolution (A) vs. the high resolution (B), depicted in lateral view. The GGP, SL, and IL muscles are activated up to (40%) in each case. Red/blue dots show the tongue-jaw/hyoid attachment points.

We finally activate each functionally-distinct segment of the posterior GG, VERT, TRANS, and STY muscles, to observe their individual impact on the shape of our tongue models. Figure 3.11 shows the deformation in the mid-sagittal plane, with *a*: most posterior to *e*: most anterior portion of each muscle. The STYa and STYb denote the intrinsic and extrinsic muscle fibers, respectively.

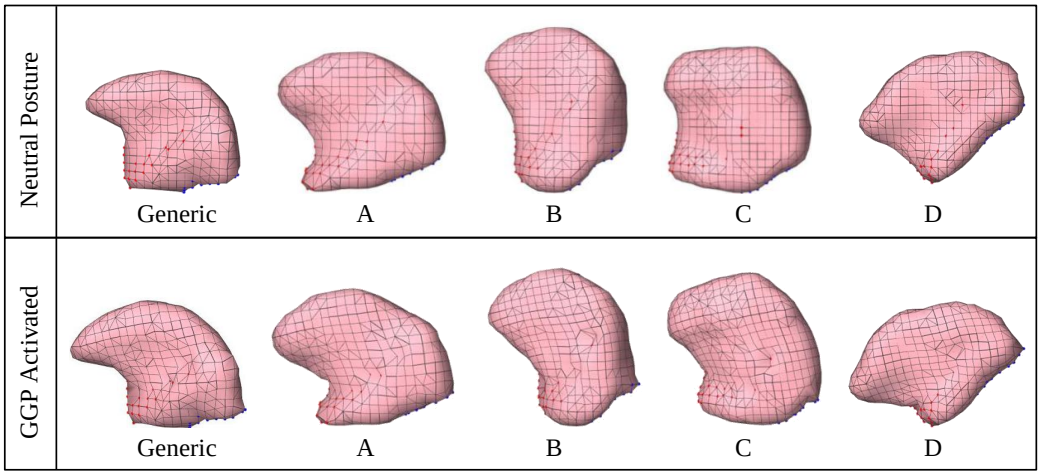


Figure 3.9. Impact of the GGP activation (10%) in the generic, and speaker-specific tongue models. Red/blue dots show the tongue-jaw/hyoid attachment points.

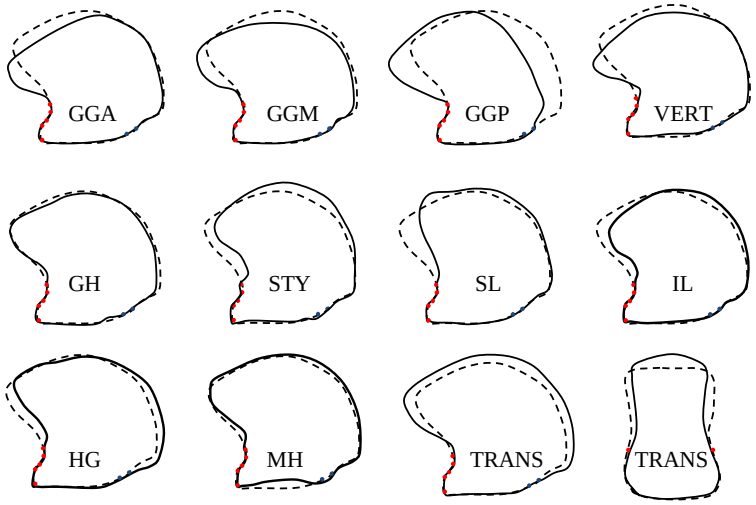


Figure 3.10. Impact of muscle activations (solid) vs. neutral posture (dashed) on the tongue contour in the mid-sagittal plane, for Speaker A. The mid-coronal plane is included for the TRANS muscle. Muscle are activated up to (10%) in each case. Red/blue dots show the tongue-jaw/hyoid attachment points.

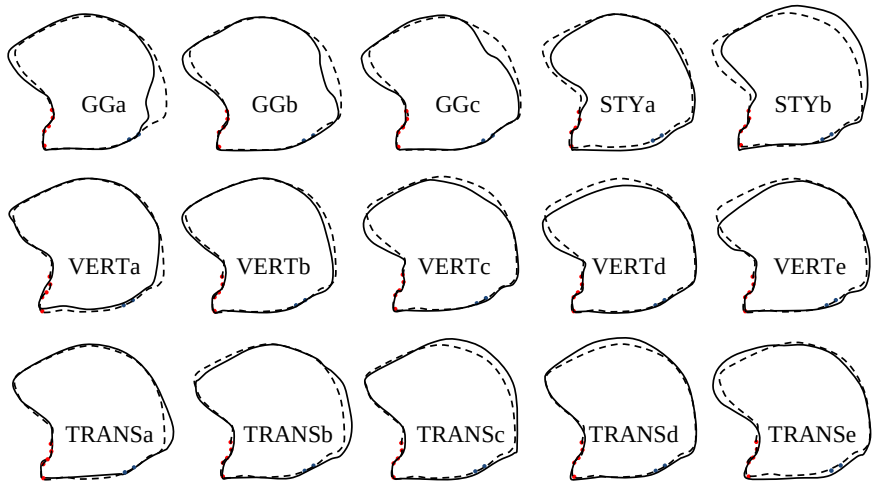


Figure 3.11. Impact of activation of functionally-distinct muscle segments on the tongue contour (solid) vs. neutral posture (dashed) for speaker A in the mid-sagittal plane. Red/blue dots show the tongue-jaw/hyoid attachment points. The subscripts a to e range from the posterior to the anterior tongue. For the STY, a and b include the intrinsic and extrinsic fibers, respectively.

3.2 Tongue Segmentation from MRI

In this section, we develop a real-time, force-based, user interaction platform, coupled with a mesh-to-image registration technique (Gilles and Pai, 2008), to delineate tongue tissue from MR image volumes. The developed method expands the application of such methodology from musculoskeletal structures to highly deformable soft-tissue. We also extend the state-of-the-art by considering delineation of the tongue from the epiglottis, hyoid bone, and salivary glands (depicted in high resolution static MRI).

Both shape and intensity priors are incorporated in the form of a source image volume and its corresponding surface mesh, which is delineated by a dental expert. The choice of the source dataset is arbitrary. We use a discrete surface mesh representation to deal with regularity and shape constraints. The overall pipeline of the developed method is shown in Figure 3.12. The current position of the surface nodes is stored in the *Mechanical State* module. Loop 1 includes the modules that handle mesh-to-image registration. The mesh is deformed according to local intensity similarity between the source

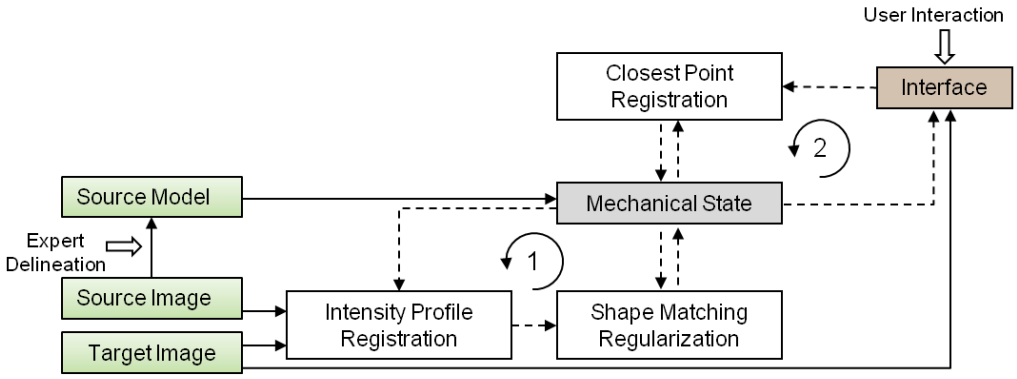


Figure 3.12. Designed segmentation pipeline. Iterative loop (1) includes the mesh-to-image registration modules. Iterative loop (2) incorporates potential user-interactive boundary labelling. Both loops update the mechanical state of the deforming mesh, simultaneously.

and target volumes. The deformation is regularised using an extended version of a shape matching algorithm (Gilles and Pai, 2008). In Loop 2, we deploy an effective minimal user interaction mechanism to help attain higher clinical acceptance. Both loops shown in Figure 3 have access to and are able to update the mechanical state of the mesh, simultaneously. This provides real-time visualisation of the surface evolution.

Our developed method has been fully implemented under the Simulation Open Framework Architecture (SOFA @`www.sofa-framework.org`), an open-source modular framework based on C++. This allows for the registration algorithm to be interpreted as a real-time simulation process, during which the source model iteratively deforms to match the target configuration, starting from its initial position.

3.2.1 Methods

Mesh-to-image registration (Loop 1 in Figure 3.12) is handled by applying external and internal forces to the vertices of the mesh. The external force $\boldsymbol{f}^e(t)$ steers the mesh toward the target boundaries. The internal force $\boldsymbol{f}^i(t)$ keeps the mesh regular and close to the prior shape. Instead of summing up the internal and external forces, we use a Pair-and-Smooth approach,

originally proposed by Cachier and Ayache (2001), in which the external and internal forces are applied sequentially. This approach minimizes propagation of noise from image features to the final result. Let $\boldsymbol{x}(t)$ denote the vector of positions $\boldsymbol{x}$ for all the vertices on the surface, at time t. We also define $\boldsymbol{x}^r = \boldsymbol{x}(0)$ as the reference vertex positions. The mesh is deformed in two steps.

1. Intensity profile registration: the image-based external force, $\boldsymbol{f}^e(t)$, is calculated for each node, as described later in this subsection; then, the position of each vertex is augmented with the vertex's external force,in a unitary time step $(\boldsymbol{x}(t) + \boldsymbol{f}^e(t))$.
2. Shape Matching Regularization: a smoothing internal force is applied on the augmented position of the vertices, and results in the vector of the regularized goal positions, defined as $\tilde{\boldsymbol{x}}$. The details of the smoothing process follow later in this subsection.

The deformation is done by moving the vector of reference vertex positions $\boldsymbol{x}^r$ to the vector of current positions $\boldsymbol{x}(t)$. Afterwards, the deformation is smoothened by applying the relevant internal forces:

$$\boldsymbol{f}^i(t) = \alpha_i(\tilde{\boldsymbol{x}} - \boldsymbol{x}(t))$$

where $\tilde{\boldsymbol{x}}$ is the vector of new regularized goal positions. The two steps involve an iterative search for $\boldsymbol{x}(t)$ and $\tilde{\boldsymbol{x}}$ respectively. The details of each module are described in the following section.

Intensity Profile Registration

For each node at position $\boldsymbol{x}$ on the surface, the external force at time t is calculated by

$$\boldsymbol{f}^e(t) = \alpha_e(\boldsymbol{x}^{'} - \boldsymbol{x}(t)) \tag{3.6}$$

where α_e is the stiffness and $\boldsymbol{x}^{'}$ denotes the new location of the node. The search for $\boldsymbol{x}^{'}$ is performed within a pre-defined range of inward and outward steps at the direction normal to the surface. At each iteration, $\boldsymbol{x}^{'}$ is selected to be the point which maximizes a local similarity measure between the source and target image volumes. Our algorithm matches the 1D gradient intensity profiles of pre-defined length L in the direction normal to the surface. Let $\boldsymbol{G}_{tar}(x)$ be the gradient profile of the target image, centred at point x, and

let $\boldsymbol{G}_{src}$ denote the corresponding gradient profile in the source image. The optimum value of x in each time step, denoted by $x^{'}$, is calculated using the normalized cross correlation as the similarity metric:

$$x^{'} = \underset{x}{\operatorname{argmax}} \left\langle \frac{\boldsymbol{G}_{tar}(x) - \overline{\boldsymbol{G}}_{tar}}{\parallel \boldsymbol{G}_{tar} - \overline{\boldsymbol{G}}_{tar} \parallel}, \frac{\boldsymbol{G}_{src} - \overline{\boldsymbol{G}}_{src}}{\parallel \boldsymbol{G}_{src} - \overline{\boldsymbol{G}}_{src} \parallel} \right\rangle \tag{3.7}$$

where $\overline{\boldsymbol{G}}$ is the average value of $\boldsymbol{G}$ and $<,>$ and $\parallel \, . \, \parallel$ denote the inner product and L^2 norm respectively.

Shape Matching Regularization

To regularize the mesh deformation, we apply the extended version of the shape matching algorithm, previously introduced in the context of musculoskeletal structures (Gilles and Pai, 2008). The underlying mesh is subdivided into overlapping clusters of nodes, defined around each vertex (i) on the surface. The cluster for vertex i is defined as

$$\zeta_i = \{j : d(x_i, x_j) < s\} \tag{3.8}$$

where d is the Euclidean distance and s is the predefined cluster size (or radius). Then, for each cluster ζ_i, the algorithm approximates the local deformation of the nodes with a rigid transform $(\mathbf{T}_i)$, applied on the reference position. The least square estimation of $\mathbf{T}_i$ is obtained by

$$T_i = \underset{T}{\operatorname{argmin}} \sum_{j \in \zeta_i} m_j \parallel \mathbf{T}\boldsymbol{x}_j^r - \boldsymbol{x}_j - \boldsymbol{f}_j^e \parallel^2 \tag{3.9}$$

where m_j represents the mass weight of particle j in the cluster; and $\boldsymbol{f}_j^e$ is computed from Equation 3.6. This, in turn, will update the goal position of each node in the cluster from $\tilde{\boldsymbol{x}} = \mathbf{T}\boldsymbol{x}^r$.

Due to the overlapping nature of the clusters, each vertex may obtain different goal positions from the different clusters it belongs to. These goal positions are subsequently combined into an average position for each vertex. The final (goal) position is used to calculate the corresponding internal forces, which are then averaged and applied to all the vertices of each cluster. Here, shape matching acts as an elastic force that is proportional to the strain;

whereas, updating the reference positions at each time step would simulate plastic deformations. We follow Gilles and Pai (2008) in using the simple gradient descent scheme with unitary time step

$$\boldsymbol{x}(t + dt) = \boldsymbol{f}^i + \boldsymbol{x}(t) \tag{3.10}$$

To summarize, one iteration of the mesh-to-image registration (in Loop 1) involves the following steps:

1. Calculate external forces $\boldsymbol{f}^e$ using Equation 3.6 and 3.7.
2. Calculate shape-matching forces $\boldsymbol{f}^i$.
   (a) For each cluster ζ_i, compute $\mathbf{T}_i$ from Equation 3.9.
   (b) For each vertex i, average goal positions as in $\tilde{\boldsymbol{x}}_i = \sum_i(\mathbf{T})\boldsymbol{x}_i^r/|\zeta_i|$.
3. Evolve node positions from $\boldsymbol{x} = \tilde{\boldsymbol{x}}$.
4. Update reference positions to simulate plasticity from $\boldsymbol{x}_r = \boldsymbol{x}$.

For the initialization mode, we model the underlying mesh with one cluster. Hence, the bodily movement of the mesh would be purely rigid, containing three translational and three rotational DOFs. If desired, we enable the user to guide initialization towards what he may deem as a better position (in a simple *mouse-click and drag*). This step inserts a spring force from the mesh toward the cursor. The initialization scheme compensates for large displacements between the initial and final tongue positions (see Figure 3.13). At any time, the user can make the transition to the deformation mode by increasing the number of clusters (through entering the desired number in a dialogue box and clicking a button).

User Interaction

We incorporate an effective minimal user interaction mechanism to guarantee a satisfactory result for the end user. The procedure is shown in Loop 2 in Figure 3.12. At any time during the registration process, the user is free to inspect the orthogonal cut-planes of the deforming mesh, overlaid on the corresponding 2D sections of the target image. The user may provide additional boundary labels by simply clicking in any area where automatic segmentation is deemed inadequate.

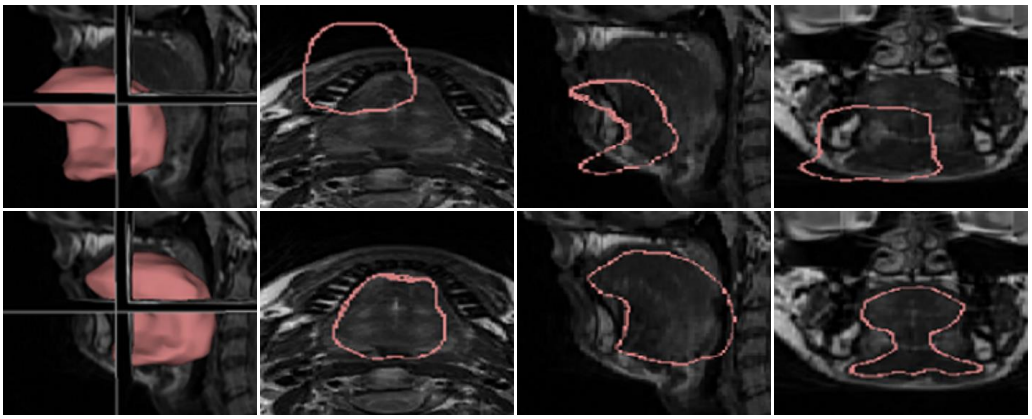


Figure 3.13. Initialization mode. From left to right: the position of the tongue in the volume, mid-axial, mid-sagittal, and mid-coronal plane, before (top) and after (bottom) initialization. No user guidance force was necessary here.

Since boundary constraints are handled through forces in SOFA, our interaction is also force-based. As soon as a new boundary voxel is clicked, the algorithm searches for the closest surface node on the mesh and inserts a spring force between these two points. The closest points on the surface will update in each iteration of the mesh deformation. We empirically use a predefined stiffness of about 10^4 in all implementations. Stiffness values of a higher order of magnitude may cause instability and, hence, are avoided.

All parameters are fixed in all the experiments. We unify the number of surface nodes to number 2502, in order to capture the sub-millimetre details of the tongue's shape. For the intensity profile registration module, the length of the profiles is set to 50 pixels, centred on the investigated voxel. The search range is five voxels, inward and outward, in the normal direction to the mesh surface. The stiffness coefficient, α_e, is set to 1. For shape matching regularization, we set the number of clusters to 300. To attain high flexibility, the radii of all clusters are set to 1.

3.2.2 Results

We apply our segmentation method on MRI scans of 18 normal subjects. For each dataset, three sagittal, coronal, and axial stacks of MRI slices are

acquired, with the tongue at the rest position, using a T2-weighted Turbo Spin Echo pulse sequence. A Siemens 3.0 T Tim Treo MRI scanner is used with an 8-channel head and neck coil. The size of each dataset is $256 \times 256 \times z$ (z ranges from 10 to 24) with 0.94mm $\times$0.94mm in-plane resolution and 3mm slice thickness. A MRF-based edge-preserving data combination technique is applied to build super-resolution volumes of the tongue with isotropic resolution of 0.94mm. Details of data acquisition and reconstruction techniques is described by Woo et al. (2012).

The developed method is evaluated for 18 normal subjects (eight females, 10 males). All 18 datasets are manually segmented under the supervision of our dental expert collaborator, using the TurtleSeg interactive tool (Top et al., 2011). The results are used, both as the ground truth and as the source model in inter-subject registration, as described later in this subsection. The segmented surface for each volume includes all of the internal muscles of the tongue, as well as the digastric, geniohyiod, and hyoglossus muscles (see Figure 3.14); it excludes the hyoid bone, mandible bone, epiglottis, and salivary glands. We cut the mylohyoid, palatoglossus and styloglossus muscles, following the contour of the tongue. In addition, tongue tissue above the line between the epiglottis and hyoid bone is included. The process takes about five to seven hours for each dataset.

Figure 3.15 shows 3D representation of the result for subject 5 as the source and subject 2 as the target.

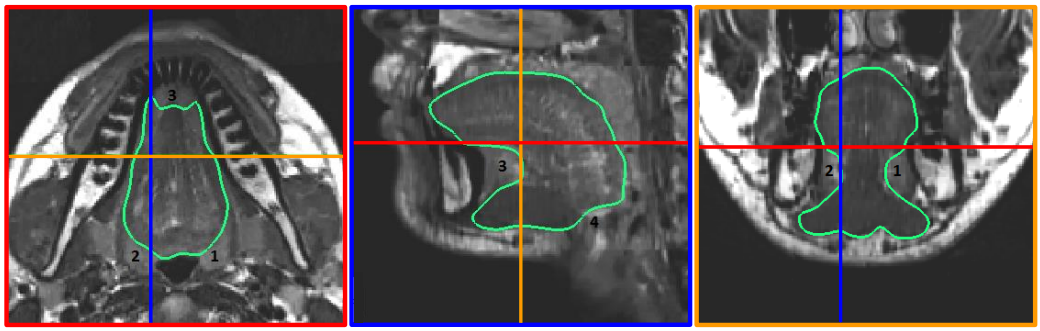


Figure 3.14. Ground truth segmented by a dental expert in TurtleSeg (Top et al., 2011). Axial (right), sagittal (middle), and coronal slices are shown in red, blue, and orange respectively. Salivary glands (labels 1-3), hyoid bone (label 4) and epiglottis (label 5) have been excluded.

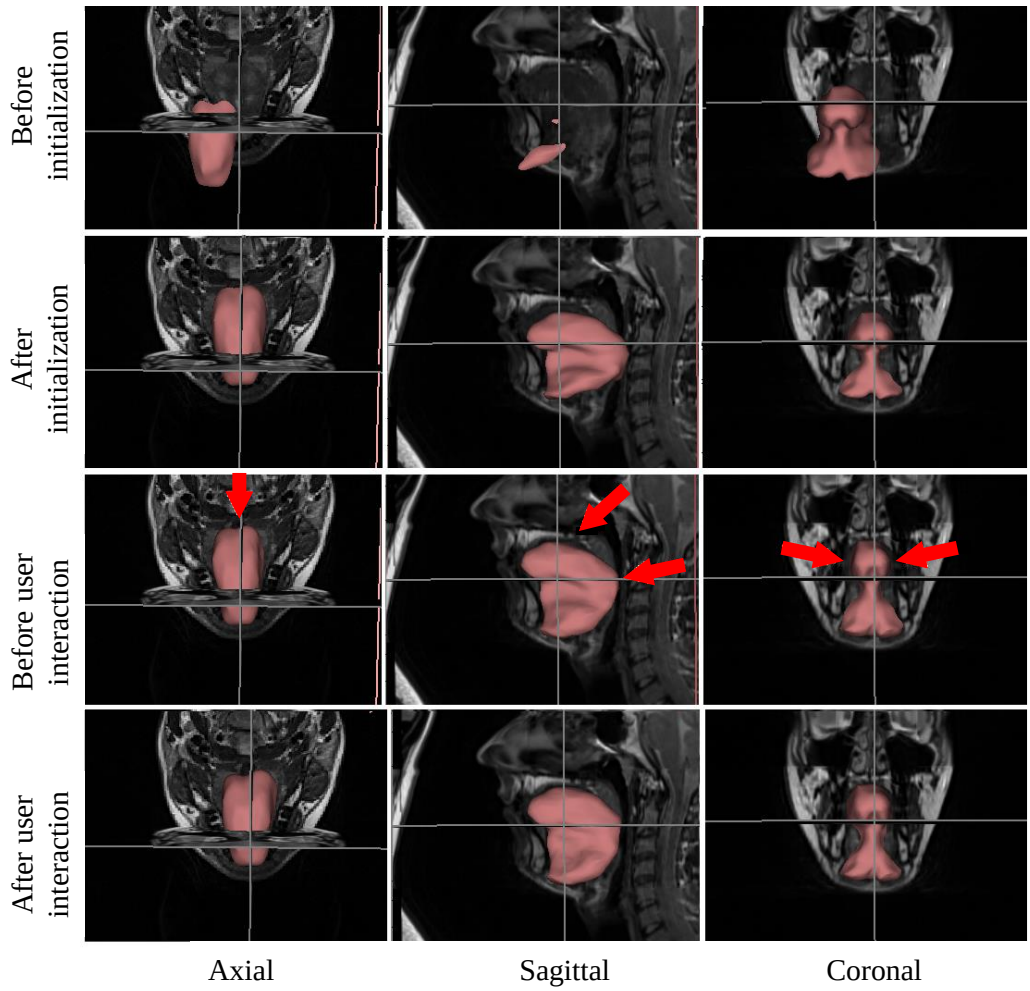


Figure 3.15. 3D representation of the mesh during the segmentation process. Red arrows show the areas that require extra, user-provided boundary labels.

Measures of Volume Overlap. After categorizing the subjects into two groups of (*female* and *male*) anatomy, we noticed that anatomy of one male subject was a closer match with the female group; therefore, he was excluded from the male group and added to the female group (F9 in figures 3.16 and 3.17). For each dataset in each group, the segmentation was repeated by iterating the source on other members of the corresponding group, resulting in $(9 \times 8) \times 2$ experiments in total. In each case, the dental expert was asked to interact with the segmentation for 1-3 minutes. The distance and

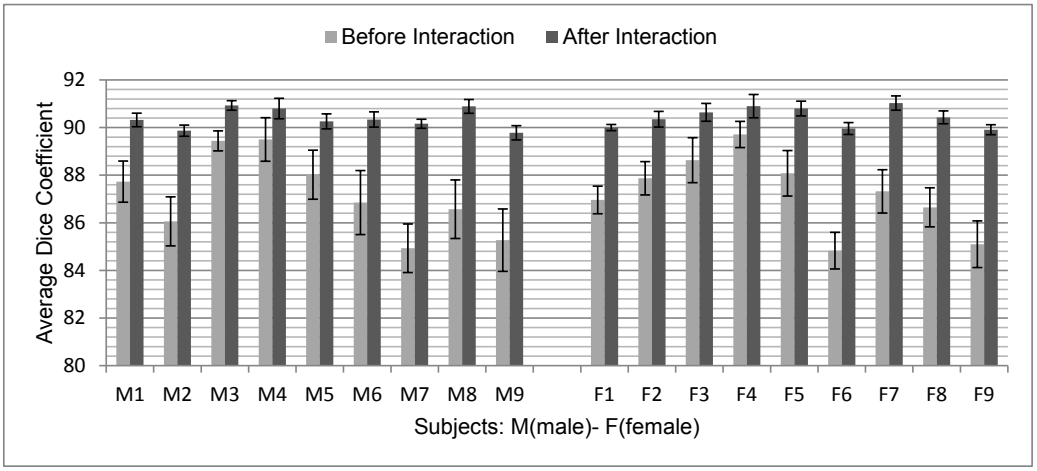


Figure 3.16. Average Dice coefficient for subjects in the male (M) and female (F) group, before (light gray) and after (dark gray) expert interaction time of 2 ± 1 minutes.

the volume overlap between result (A) and ground truth (B) were calculated before and after the interaction. We used the Dice coefficient as a measure of the volume overlap, reported as a percentage:

$$Dice(A, B) = 2\frac{|A \cap B|}{|A| + |B|} \times 100 \tag{3.11}$$

Figure 3.16 shows the Dice measure calculated before and after expert interaction, for both the male and female groups. The average Dice, measured on all the datasets in the male group, improved from 87.15 ± 1.65 to 90.37 ± 0.42 after expert interaction. In addition, the mean of the inter-subject standard deviation (STD) dropped from 1.02 ± 0.28 to 0.29 ± 0.07. The average overlap in the female group is 87.23 ± 1.58, before interaction, and 90.44 ± 0.42, after interaction. The mean of the measured STD also changes from 0.80 ± 0.16 to 0.29 ± 0.10.

Measures of Surface Distance. For calculating the distance between the two surfaces, we used the Modified Hausdorff Distance (MHD) as the measure

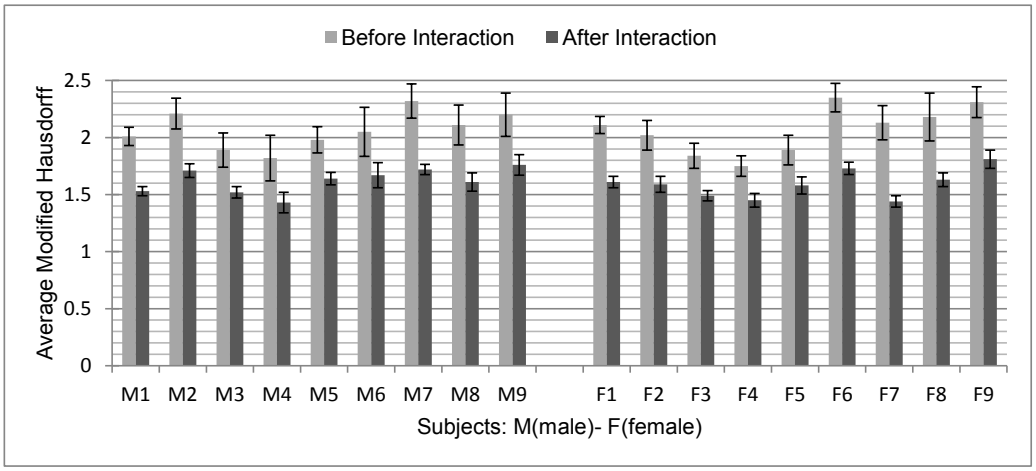


Figure 3.17. Average Modified Hausdorff distance for subjects in the male (M) and female (F) groups, before (light gray) and after (dark gray) expert interaction time of 2 ± 1 minutes.

of object-matching (Dubuisson and Jain, 1994):

$$\begin{aligned} \mathrm{MHD}(A,B) &= \max\big(d(A,B), d(B,A)\big) \\ (A,B) &= \frac{1}{N_a}\sum_{a\in A} d(a,B) \\ d(a,B) &= \min_{b\in B} d(a,b) \end{aligned} \tag{3.12}$$

where N_a is the set of all the nodes (a) on the surface A, and $d(a,b) = ||a-b||$ is the simple Euclidean distance between vertices a and b. Figure 3.17 shows the average MHD. The mean of the measure in the male group changes from 2.06 ± 0.16 mm to 1.62 ± 0.10 mm after interaction. The mean of the measured STD also changes from 0.15 ± 0.04 mm to 0.07 ± 0.02 mm. For the female group, the mean of MHD is 2.06 ± 0.21 mm, before, and 1.59 ± 0.12, after interaction. The mean of the STD also decreases from 0.13 ± 0.04 mm to 0.06 ± 0.01 mm.

Sensitivity Analysis. In order to justify our choice of parameters, we performed a sensitivity analysis on four main parameters of the algorithm: cluster number, cluster radius, search range, and stiffness coefficient. For each experiment, we changed the value of one of the parameters, while the others were fixed to their default values. To minimize the control variables,

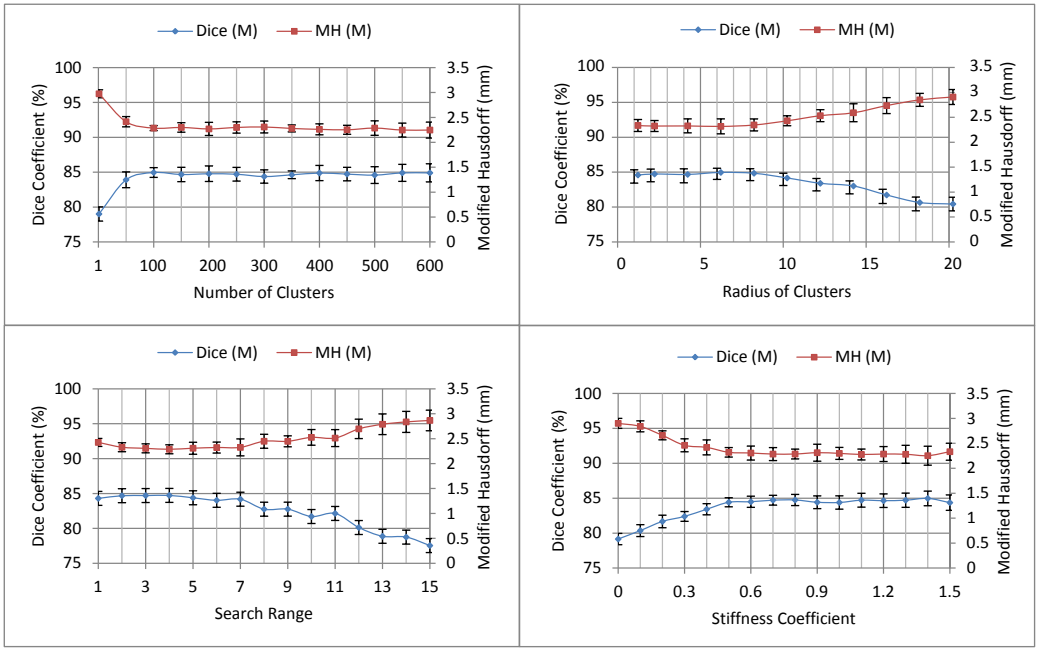


Figure 3.18. Sensitivity analysis on four main parameters of the developed algorithm: Dice and modified Hausdorff measures for subject #7 in the male group.

we excluded the initialization and user interaction stages, running the analysis only for the mesh-to-image registration. In each group (e.g. male/female), we chose the target image with the worst segmentation accuracy measured in the previous experiments (e.g. M7 and F6). Other members of the group were iteratively used as the source of registration. Hence, we conducted $(1 \times 8) \times 2$ experiments for each value of each parameter. The average and variance of the Dice and MH measures, for the male and female subjects, are shown in figures 3.18 and 3.19, respectively.

- **Number of clusters.** Proposed default value: 300. Small numbers of clusters result in a more rigid model and decrease segmentation accuracy. More clusters results in a higher degree of freedom for the deformation, but values greater than 600 cause convergence failure.
- **Radius of clusters.** Proposed default value: 1. Radii of less than 10 produce similar results. The deformable model becomes more rigid as the radius (e.g. overlap) of the clusters increases, decreasing segmentation accuracy.

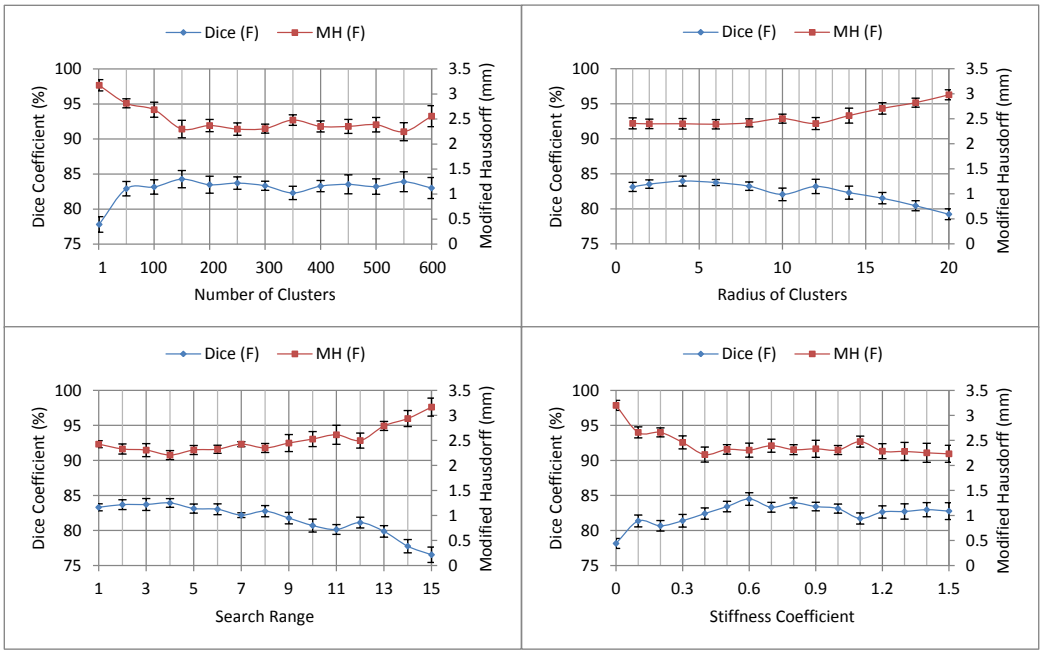


Figure 3.19. Sensitivity analysis on four main parameters of the developed algorithm: Dice and modified Hausdorff measures for subject #6 in the female group.

- **Search range.** Proposed default value: 5. Greater search range will enable the model to deform more during each time step; however, for values greater than 10 pixels inward/outward, the model will have difficulties in convergence, and will fluctuate back and forth close to the optimum solution.
- **Stiffness Coefficient.** Proposed default value: 1. Low values of stiffness decrease the effect of image-based forces, and are therefore inadequate for matching the model to the target image. Values higher than 1.5 will introduce instabilities in the model.

3.2.3 Discussion

The manual segmentation used as the ground truth is likely to be fuzzy and uncertain in problematic regions, such as at the boundary of the hyoid bone and salivary glands (see Figure 3.14). However, modeling requisites

prohibit the segmentation to include bones and non-muscle soft tissues. To assess such uncertainty, we designed an experiment in which the same dental expert was asked to repeat the same manual segmentation after 10 days, without referring to his first attempt. The result showed a volume overlap of Dice= 91% and a surface-to-surface MHD= 1.52 mm between the two manually segmented volumes. We argue that this uncertainty imposes similar limits (91% for the volume overlap, and 1.5 mm for the surface-to-surface distance) on the measures achievable by automated segmentation. In fact, while expert interaction resulted in about 7% improvement for Dice values as low as 83% (0.6 mm decrease in distance value), the improvement was less than 1% for the Dice values as high as 90% (0.3 mm decrease in distance value). The same ambiguities cause the user interaction to be inefficient after a certain time limit, justifying our choice of restricted interaction time for reporting the results.

3.2.4 Summary

In this section, we have tackled the challenging problem of semi-automated 3D segmentation of the tongue from isotropic MRI volumes. Previous works have included delineation of tongue contours at its surface in 2D MRI slices. We adapted an inter-subject registration framework using a shape matching-based regularization technique. This method was combined with an instant force-based user interaction mechanism, which attracts the model towards user-provided boundary labels. We were able to achieve segmentation accuracy with an average Dice coefficient of $90.4 \pm 0.4\%$, and an average distance = 2 ± 0.2 mm, within an expert interaction time of 2 ± 1 minutes. Thus, we conclude that our human-in-the-loop approach, using a variation of the shape matching technique (Gilles and Pai, 2008), provides an effective method to segment complicated soft tissue areas, like the tongue. Our future work will focus on integrating the developed segmentation scheme within a more comprehensive biomechanical model, suitable for modeling of speech and swallowing.

3.3 Jaw and Hyoid Modeling

Our speaker-specific model of the mandible and hyoid is similar to the ArtiSynth generic model (Stavness et al., 2011) in its biomechanics: it is coupled to the tongue FE model via multiple attachment points, included in the constitutive equations of the system as bilateral constraints. Eleven pairs of bilateral point-to-point Hill-type actuators, as listed in Subsection 2.2.1, are used to represent the associated muscles, and the TMJ is modeled by curvilinear constraint surfaces. The bone density is set to 2000 kg/m^3, as suggested by Dang and Honda (2004). For each speaker, the geometries of mandible and hyoid bone rigid bodies are replaced with the corresponding surfaces, segmented from the first TF of cine MRI data (Subsection 3.3). The bone-tongue attachment points are computed based on the generic tongue model, as described in Section 3.1.3.

3.3.1 Bone Segmentation from MRI

To build our speaker-specific models, we need to delineate the surface geometry of the articulators from cine MRI data. Unfortunately, cine MRI only provides partial bone visibility, which makes the results of manual segmentation inadequate for detection of the location of muscle insertion sites and the temporomandibular joint (TMJ).

Static MRI, however, provides higher resolution and a better representation of bone surfaces. Woo et al. (2015) create a high resolution static MRI atlas that includes speaker data used in the present study, as shown in Figure 3.20. D_i, in the figure, denotes the deformation from static MRI of speaker i onto the atlas space. We first build a segmentation mask for the mandible in the atlas space, and then morph the mask onto the static MRI of the speaker (using the inverse of D_i [D_i^{-1}]). Finally, we perform an image-based elastic registration (Vercauteren et al., 2009) between the static and cine MRI images of each speaker, to generate the mask in the cine MRI (at first TF). In the figure, this final registration is denoted by R_i.

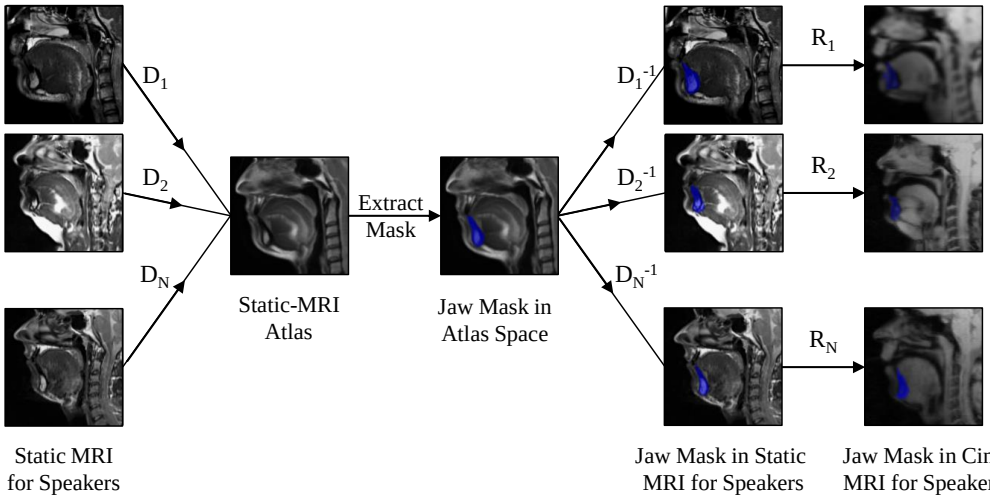


Figure 3.20. Atlas deformation for jaw segmentation. D_i denotes the deformation from static MRI of speaker i onto the atlas space; R_i stands for the elastic registration from the static to cine MRI space. Jaw masks are shown in blue.

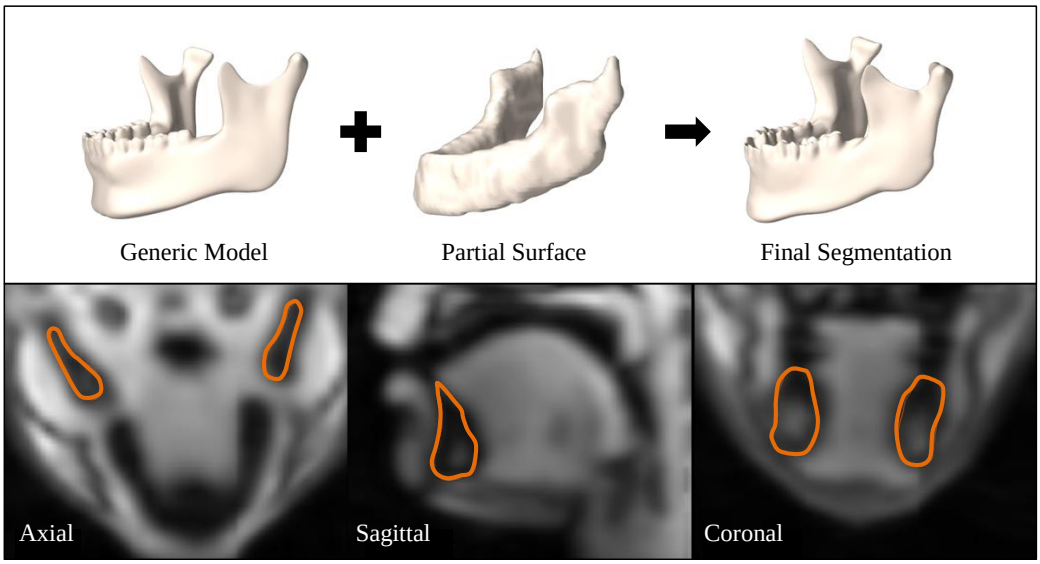


Figure 3.21. Mandible segmentation (speaker A). The generic model is sculpted to match the partial surface, while its intersection with the image is inspected. Orange contours (bottom row) show the final result at mid-views of cine MRI data.

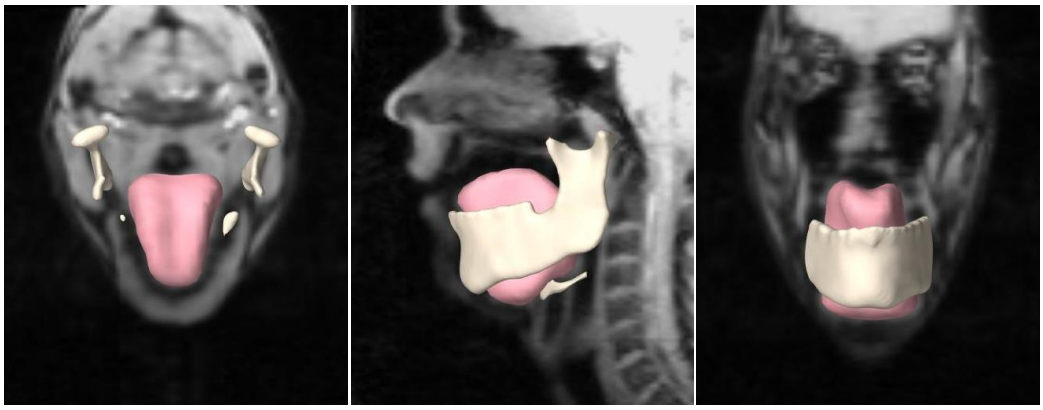


Figure 3.22. Geometries of the tongue, jaw and hyoid overlaid on the mid-axial, mid-sagittal, and mid-coronal slices of cine MRI data, for speaker B.

The final mask (in the cine MRI space) yields a partial mandible surface, as shown in Figure 3.21 (for speaker A). We deploy this partial surface as the guide for (manual) sculpting of a generic mandible mesh (available in ArtiSynth). For sculpting, we use BlendSeg, a customized plug-in for the Blender mesh editing software (`www.blender.org`) that allows inspection of the mesh intersection with the image data, throughout the sculpting process (Ho et al., 2014). Figure 3.22 shows the final geometries of the jaw, hyoid, and tongue, overlaid on cine MRI data for speaker B.

As mentioned earlier, cine MRI data does not provide sufficient resolution for depicting the sites of muscle origin and insertion. The insertion sites on the mandible are transformed from the generic jaw model to speaker geometry, through the sculpting process. Using the relative distance between the origin and insertion sites in the generic model, we calculate rough coordinates of the origin sites in the speaker space. These coordinates are further fine-tuned according to the generic model/speaker MR data, in the ArtiSynth's graphical user interface (see Figure 3.23). Making the image data visible, during the process, helps estimate the length of the muscles, ensuring the origin sites do not lie outside of the speaker's skull.

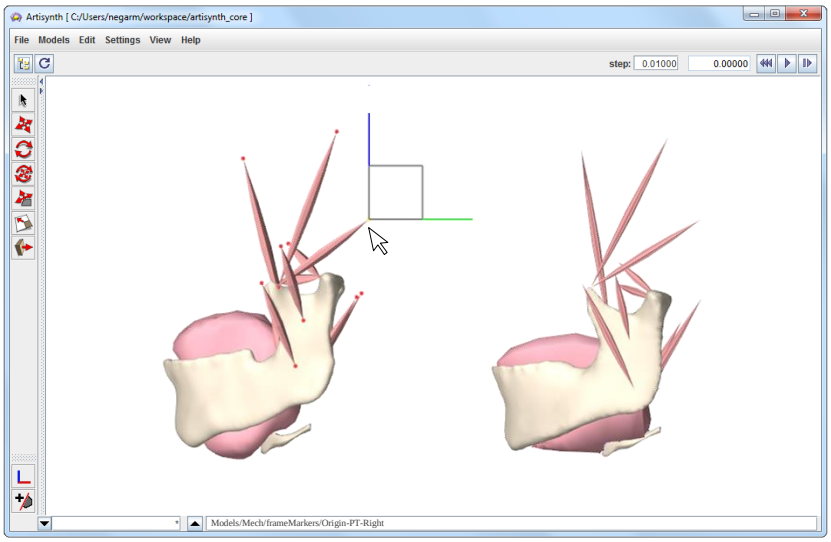


Figure 3.23. Using ArtiSynth GUI to adjust muscles of the jaw and hyoid. Configuration of the muscles for speaker B (left) is adjusted based on the generic model (right). Origin and insertion sites of the muscles are defined as frame markers (red dots) that can be moved freely in space (using the transition handle). A matching cine MRI image data can be made visible, if necessary.

3.3.2 Forward Simulation

Finally, we attach our tongue models to the mandible and hyoid through the attachment points calculated in Subsection 3.1.3. We further assess the impact of jaw muscles in each speaker-specific jaw-tongue-hyoid model in a forward simulation scheme. Figure 3.24 shows the models after activation of the jaw-opener (SP, IP, AM, PM, AD), and jaw-closer (SM, DM, MT, PT, AT, MP) muscles, for speakers A-D. The active muscles that are visible, are depicted in darker shades of red. Less excitation is used for the jaw-closer muscles (compared to jaw openers), since the jaw-closers consist of more, larger muscles, and generate a higher magnitude of force per excitation unit. Figure 3.25 shows mid-sagittal contours of the tongue, mandible, hard palate, and hyoid after activation of the jaw-closer and jaw-opener muscles, for speaker B. The maxilla is considered to be fixed, and we avoid collision between the tongue and mandible, or maxilla, in our simulations.

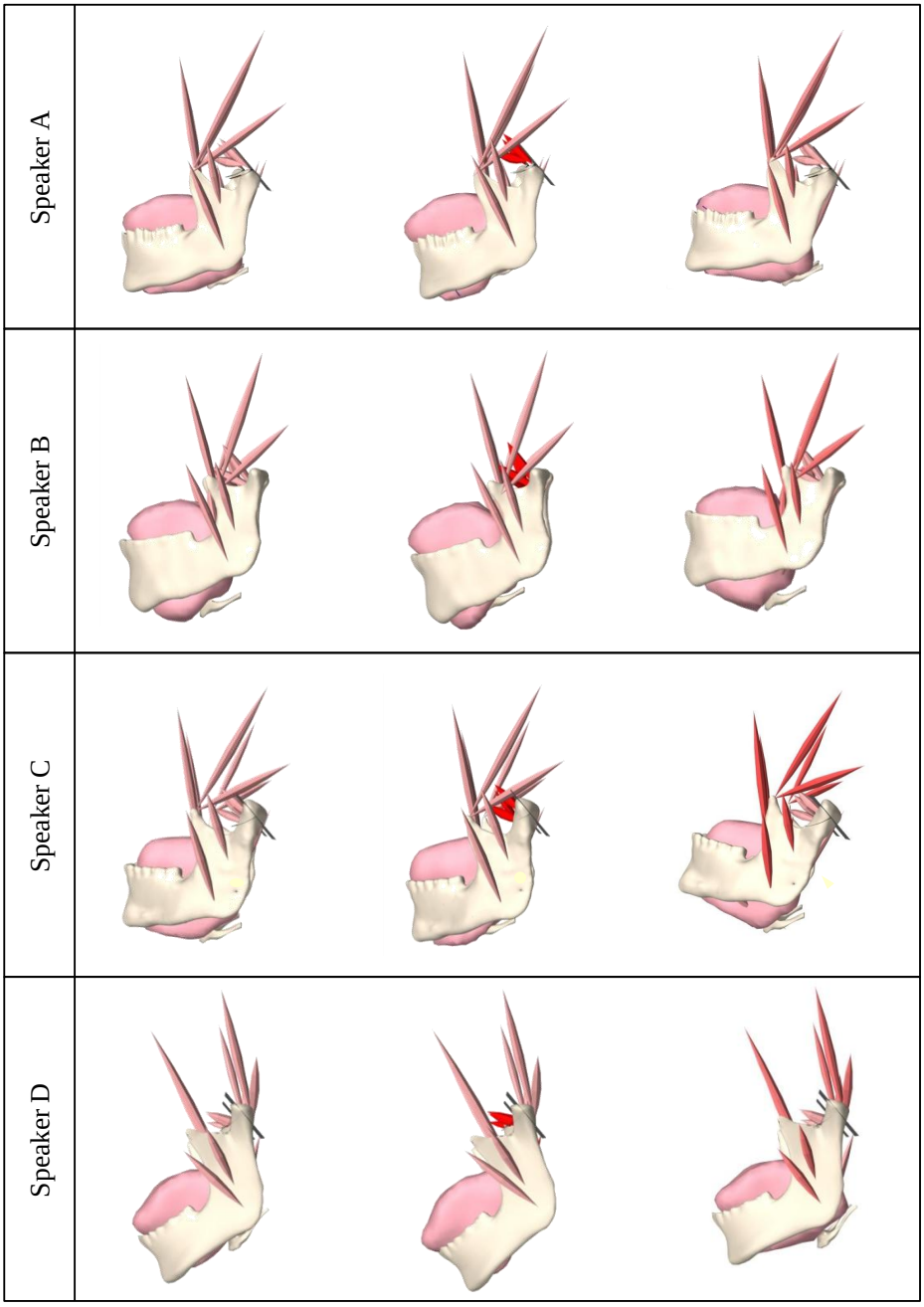


Figure 3.24. Impact of activation of jaw-opener (middle) and jaw-closer (right) muscles vs. the neutral posture (left) for speakers A-D. The jaw-opener/closer muscles are activated linearly up to ten/five percent.

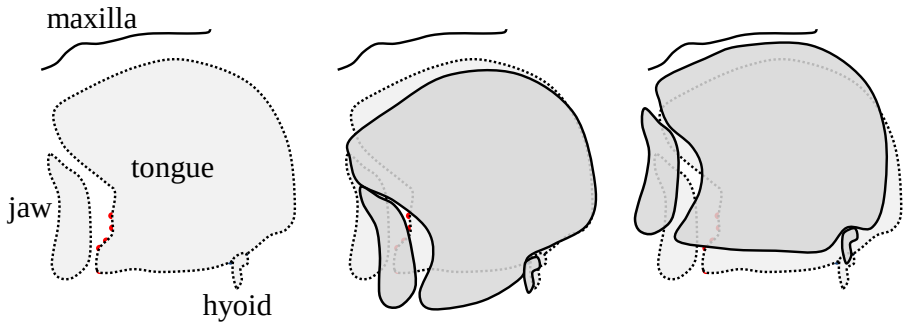


Figure 3.25. Impact of activation of jaw-opener (middle) and jaw-closer (right) muscles vs. the neutral posture (left) for speaker A in the mid-sagittal plane. The jaw-opener/closer muscles are activated linearly up to ten/five percent.

3.4 Conclusions

In this chapter we have detailed the development of our subject-specific, biomechanical models of the tongue-jaw-hyoid based on MR data. We start by delineating the articulators from the MR volume. Our method for tongue segmentation benefits from a real-time user interaction that significantly improves time-efficiency. We also suggest methods to enable segmentation of the mandible and hyoid, notwithstanding the poor contrast of bone in MRI. Based on the delineated surfaces, we then create our models. Our approach for creating tongue models combines meshing and registration techniques, to benefit from a state-of-the-art generic model (Buchaillard et al., 2009), while providing the opportunity to adjust the resolution and modify muscle definitions. To conclude, we test the performance of our models in a forward simulation scheme by activating individual muscles and observing the corresponding tissue motion/deformation.

We believe that our approach for generating $\mathrm{FE}_{\mathrm{final}}$ offers benefits that can be investigated further in future. Firstly, we suggest that a higher resolution tongue model provides the opportunity to simulate more complex and longer speech utterances that exhibit additional variability in tongue shape. Swallowing is another example where more local tongue motions are observed. Second, our proposed approach offers structural independence between the configuration of muscle fibres and elements. This enables the user to modify, add or delete individual muscle fibres, in order to accommodate more sub-

tlety in neural innervation (as suggested by Slaughter et al. (2005) for the IL and SL muscles and by Mu and Sanders (2000) for the GG). A finer fibre structure is also useful in studying different languages where sounds are similar but not identical. In addition, ability to edit the fibres is beneficial for simulation of speech, in disorders such as glossectomy, where the innervation pattern varies based on the missing tissue (Chuanjun et al., 2002). Finally, as resolution of dynamic MRI data improves, we will be able to capture finer shapes of the tongue. Hence, our model should be positioned to present more details.

Chapter 4

Data-Driven Simulation for Speech Research

In this chapter, we enable data-driven simulation of our subject-specific models, and demonstrate its application in investigating motor control to the tongue. We especially look at the fricative sound /s/ because it is made in a region of the vocal tract where minor changes in tongue position and shape are audible and can easily compromise the pronunciation of /s/. For this reason, /s/ is used as a test of articulatory accuracy by oral surgeons, prosthodontists, and speech pathologists. At least two prevalent /s/ gestures are identified in the literature: the *apical* /s/, which uses the tongue tip to contact the alveolar ridge; and the *laminal* /s/, which uses the tongue blade (Dart, 1991). A 2D tagged MRI study by Reichard et al. (2012) identifies a trend in the tongue tip moving faster during apical /s/, with no significant effect observed in the tongue body. A similar study confirms a difference in tongue shape for the two gestures, reporting the occurrence of a shallower groove at the velopalatal juncture in apical /s/. A possible correlation between palate height and /s/-type is also noted (Stone et al., 2012).

This study examines the motor control and motion patterns of /s/ in two back and front vowel contexts, in the utterances /ə-gis/ and /ə-suk/, to exploit the differential effects of neighboring sounds on /s/ realization. By simulating speaker-specific models – based on the MRI data of multiple speakers – this chapter explores possible answers to two questions: namely, *What are the key muscles responsible for the motion into the two laminal and apical gestures of /s/?* and *How do vowel context and /s/-type affect activation pattern among different speakers?*

We base our simulations on 3D tagged and cine MRI data, which captures the motion of the tongue's tissue-points during the production of /s/. Using this quantified, tissue-point motion, Xing et al. (2015) calculate internal

motion patterns, as well as degree of shortening and lengthening of individual muscles. However, the data alone provides an incomplete picture of the motor control to the tongue. For example, active and passive shortening of a muscle can cause similar motion, and co-contraction of antagonist muscles may result in no shortening. It is, therefore, difficult to disambiguate the causes of muscle shortening from MRI alone. In this study, however, we simulate our speaker-specific biomechanical models using tissue-point motion, extracted from tagged MRI, to first infer which muscles are actively shortening (using an inverse model), and then to actively shorten those muscles to predict tissue-point motion (forward model). Following this, we compare the results with the tagged MRI trajectories, in order to fine-tune the predicted muscle activations. Our results (as presented in subsection 4.4 and discussed in subsection 4.5) supplement and enhance current knowledge of how muscle activation is related to tongue motion patterns.

Figure 4.1 shows a schematic representation of the MRI data used in this study. The cine and tagged MRI slices are recorded during synchronized repetition of the desired speech utterance, and averaged over time to produce a volumetric representation of the oropharynx for each time frame (TF), as described in Section 4.1. The MR images are then fed into the work-flow presented in Chapter 3. The internal tissue displacements are calculated from tagged MRI and further enhanced with the tongue surface information of cine MRI data (subsection 4.2). The biomechanical models of the tongue, mandible, hyoid, and maxilla are then constructed for each speaker based on the surface geometries segmented from the first TF of cine MRI, as described in Chapter 3. The speaker-specific models are then simulated based on the tissue displacements (subsection 4.3). We use the Artisynth platform, which supports both forward and inverse simulations. Forward simulation yields kinematic trajectories of the model based on muscle activations, and the inverse simulation provides estimates of muscle activation patterns, based on the tissue trajectories measured at specific control points from the data.

4.1 MRI Data

Our MRI data captures four normal, Caucasian American English speakers with mid-Atlantic dialect – two with apical /s/, and two with laminal

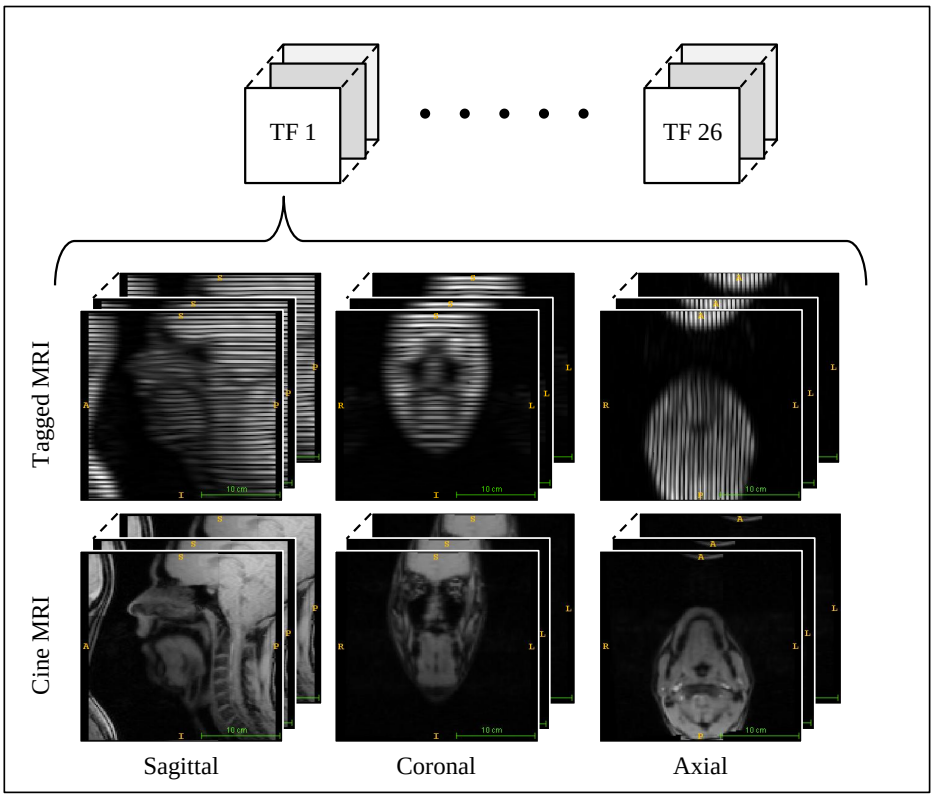


Figure 4.1. Schematic representation of the MRI data used in this study. Stacks of tagged and cine MRI are acquired at each time-frame (TF) in sagittal, coronal, and axial views.

/s/. Each speaker repeats the utterances /ə-gis/ and /ə-suk/ in time with a metronome. Both cine and tagged MRI data is acquired using a Siemens 3.0T Tim-Treo MRI scanner with a 16-channel head and neck coil. The in-plane image resolution is $1.875mm \times 1.875mm$ with a slice thickness of 6 mm. Other sequence parameters include the following: repetition time (TR) 36 ms, echo time (TE) 1.47 ms, flip angle 6, and turbo factor 11. The axial, sagittal, and coral stacks of cine MRI slices are combined to form isotropic super-resolution volumes for 26 TFs, using a Maximum A Posteriori-Markov Random Field method with an edge-preserving regularization scheme (Woo et al., 2012). Table 4.1 summarizes the information of each individual speaker. The phonemes of interest (/ə/, /g/, /i/, /s/ and /ə/, /s/, /u/, /k/) are identified in specific TFs in each utterance. Each

Table 4.1. Speaker information for this study: sex, age, and time frames associated to individual sounds in /ə-gis/ and /ə-suk/ utterances.

Speaker	Sex	Age	/s/-type	/ə-gis/ TFs				/ə-suk/ TFs			
				ə	g	i	s	ə	s	ʊ	k
A	M	23	apical	8	12	16	21	8	13	19	21
B	M	22	apical	6	10	18	20	7	10	16	19
C	F	43	laminal	8	10	14	23	4	9	15	18
D	F	21	laminal	5	9	13	19	7	10	17	19

vowel is identified at the TF before the tongue begins to move toward the next consonant, i.e., that is, the maximum vowel position. Each consonant is identified at the TF when the tongue first contacts the palate, i.e., the initial frame, rather that the maximum consonant position.

These frames were chosen because they were easily identifiable from the MRI movies. Figure 4.2 shows the mid-sagittal slice of cine MRI at the TF associated with /s/ for each speaker in both utterances.

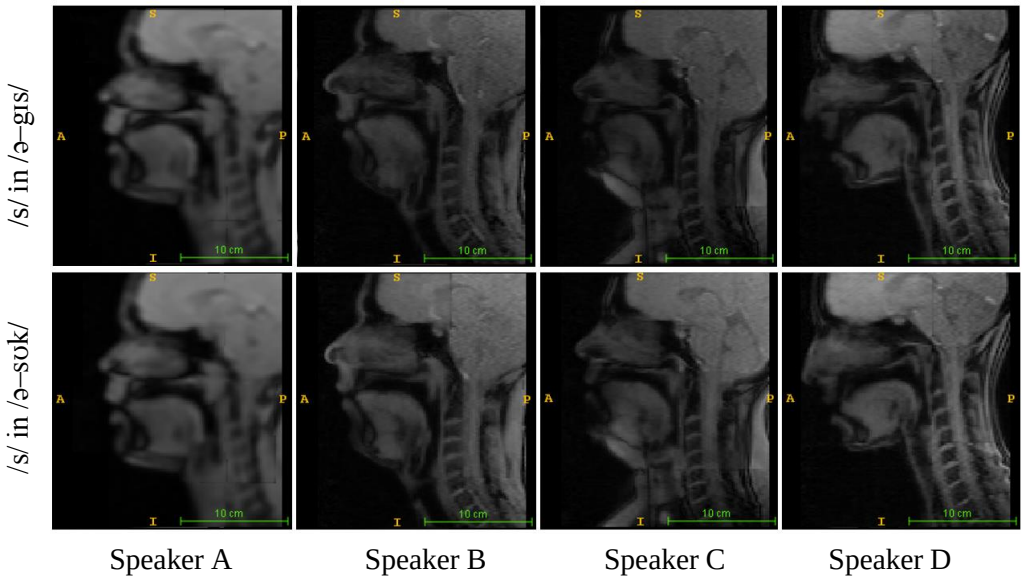


Figure 4.2. Mid-sagittal slice of cine MRI at /s/ in /ə-gis/and /ə-suk/. Speakers A and B show the apical /s/, and speakers C and D show the laminal /s/ gesture.

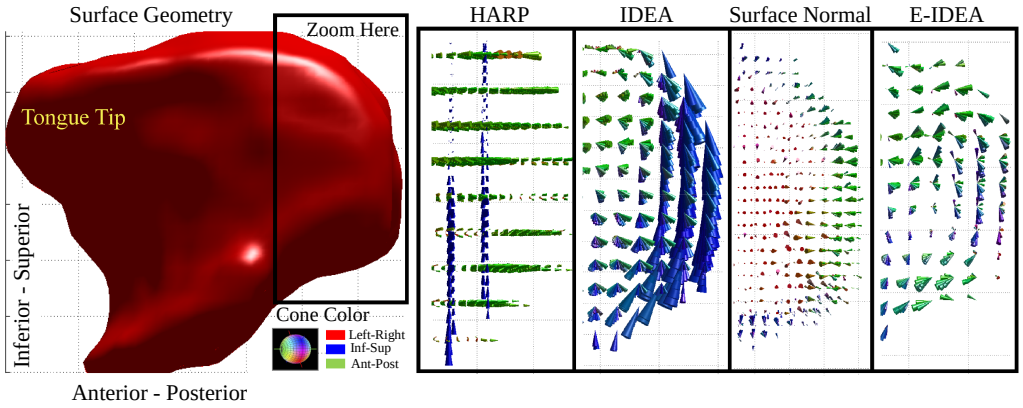


Figure 4.3. Tissue displacements, calculated from tagged MRI, using HARP (Osman et al., 2000), IDEA (Liu et al., 2012), and enhanced by surface normals from cine MRI in E-IDEA (Xing et al., 2013). ©Xing et al. (2013). Adapted with permission.

4.2 Tissue Displacement

The 2D motion of the tongue tissue points is estimated from tagged MR image slices using the Harmonic Phase (HARP) algorithm (Osman et al., 2000). We further utilize the Enhanced Incompressible Deformation Estimation Algorithm (E-IDEA), to combine 2D motion data and produce the 3D tracking result with an incompressibility constraint (Liu et al., 2012; Xing et al., 2013). E-IDEA imposes a smoothing, divergence-free, vector spline to seamlessly interpolate velocity fields across the tongue. In addition, it improves the reliability of the displacement field at the tongue surface by incorporating 3D deformation of the tongue surface computed from cine MRI. Figure 4.3 demonstrates the effectiveness of E-IDEA in improving the accuracy of the tissue displacements at the tongue surface.

Both HARP and E-IDEA calculate the displacement field (at each TF) with reference to the first TF (when tags were initially applied). However, in order to simulate our models, we have to calculate displacements between successive TFs. To get from the n^{th} to the $(n+1)^{\text{th}}$ TF, we first move from the n^{th} to the first TF – via the inverse of the nth velocity field – and then move from the first to the $(n+1)^{\text{th}}$ TF by adding the $(n+1)^{\text{th}}$ velocity field. We adopt a simple fixed-point algorithm (Chen et al., 2008) to invert the

E-IDEA velocity fields.

In this study, we perform spatial and temporal regularization to reduce possible noise in the estimated motion caused by registration errors or surface ambiguities; in the spatial domain, the displacement vectors are averaged in a spherical region of predefined radius around each point of interest; in the time domain, a cubic interpolation is performed between successive TFs to smooth the trajectories and find the intermediate displacements.

4.3 Inverse Simulation

Forward dynamic simulation requires fine-tuning of the muscle activations over time. EMG recordings of the tongue have been used previously (Fang et al., 2009), but they suffer from a lack of suitable technology to deal with the moist surface and the highly deformable body of the tongue (Yoshida et al., 1982). Also, the relationship between EMG signals and muscle forces is not straightforward. As an alternative, muscle activations can be predicted from available kinematics by solving an inverse problem. The results may be further fed to a forward simulation system to provide the necessary feedback to the inverse optimization process. The forward-dynamics tracking method was initially introduced for musculoskeletal systems (Erdemir et al., 2007); later on, Stavness et al. (2012) expanded the method to the FE models (such as the tongue) with muscular hydrostatic properties that are activated without the mechanical support of a rigid skeletal structure.

In ArtiSynth, the system velocities ($\mathbf{u}$) are computed in response to the active and passive forces:

$$\mathbf{M}\dot{\mathbf{u}} = \mathbf{f}_{active}(\mathbf{q}, \mathbf{u}, \mathbf{a}) + \mathbf{f}_{passive}(\mathbf{q}, \mathbf{u}) \tag{4.1}$$

$$\mathbf{f}_{active}(\mathbf{q}, \mathbf{u}, \mathbf{a}) = \Lambda(\mathbf{q}, \mathbf{u})\mathbf{a} \tag{4.2}$$

where $\mathbf{M}$ is the mass matrix of the system, and Λ denotes a nonlinear function that relates the system positions ($\mathbf{q}$) and the system velocities ($\mathbf{u}$) to the active forces. In the case of Blemker muscles, the value of $\mathbf{f}_{active}(\mathbf{q}, \mathbf{u}, \mathbf{a})$ is calculated from Equation 3.5. The inverse solver uses a sub-space ($\mathbf{v}$) of the total system velocities as its target:

$$\mathbf{v} = \mathbf{J}_m\mathbf{u} \tag{4.3}$$

where the target velocity sub-space ($\mathbf{v}$) is related to the system velocities ($\mathbf{u}$) via a Jacobian matrix $\mathbf{J}_m$. The inverse solver computes the normalized activations ($\mathbf{a}$) by solving a quadratic program subject to the condition $0 \leq \mathbf{a} \leq 1$:

$$\mathbf{a} = argmin(\|(\mathbf{v} - \mathbf{H}\mathbf{a})\|^2 + \alpha\|\mathbf{a}\|^2 + \beta\|\dot{\mathbf{a}}\|^2) \tag{4.4}$$

Here $||\mathbf{a}||$ and $\dot{\mathbf{a}}$ denote the norm and time-derivative of the vector $\mathbf{a}$; the matrix $\mathbf{H}$ summarizes the biomechanical characteristics of the system, such as mass, joint constraints, and force-activation properties of the muscles; α and β are coefficients of the ℓ^2-norm regularization and damping terms, respectively. The regularization term deals with muscle redundancy in the system and opts for the solution that minimizes the sum of all activations. The damping term secures system stability by prohibiting sudden jumps in the value of activations. We follow Erdemir et al. (2007) in opting for ℓ^2-norm (rather than ℓ^1-norm) as ℓ^2-norm is easier to implement. As a result, the regularization favors similar activation values over sparsity, across different muscles. The solution converges after iterating between inverse and forward dynamics in a static per time-step process, where the system is made linear in each integration step. This method is computationally efficient compared to the static methods; however, it may lead to sub-optimal muscle activations (Stavness et al., 2012).

4.3.1 Definition of the Control Points

As mentioned above, the inverse solver in ArtiSynth uses a sub-space of the total system kinematics as its target. This means that the solver tracks the velocities of certain points in the model (referred to as *control* points). In this study, we define a control point as a marker that attaches to an element of the tongue model, at a desired location. The initial location of control points is defined according to a subset of FE nodes in the generic tongue model (and hence, in FE_{reg} from Section 3.1). As a result, for all four speakers, control points have similar location relative to their tongue geometries in the first TF of cine MRI.

The biomechanical models in this study are generated based on /ə-gis/ data. Although both the /ə-gis/ and /ə-suk/ utterances were recorded consecutively in the same MRI session, their image data does not necessarily match

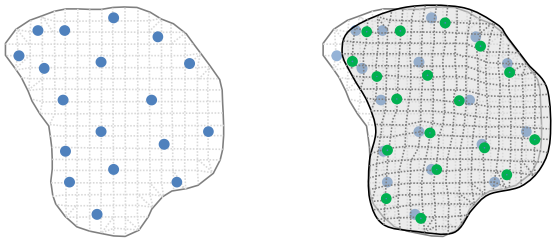


Figure 4.4. Initializing simulation of the /ə-suk/ sequence for speaker B. Mid-sagittal view of the FE tongue model after rigid registration from the 1st TF of /ə-gis/ (left) vs. result of inverse simulation to match the 1st TF of /ə-suk/. Blue/green circles show target tracking points before/after inverse simulation. The mandible, hyoid, and maxilla are included in the model, but are not shown in the figure, for the sake of simplicity.

at the first TF. This can either be due to, (1) repositioning of the head between the two acquisitions (rigid transform), or (2) having a slightly different tongue posture at /ə/ for the two utterances (non-rigid deformation). Therefore, the tagged MRI trajectories of /ə-suk/ do not hold a direct association with the control points of the FE tongue model. Building a new model for /ə-suk/ is not optimal, since it doubles the labour cost of modeling, and makes comparison of the muscle activations between the two utterances less meaningful. To deal with this issue, we compensate for the head motion by applying a rigid transformation on our model (Sharp et al., 2002), and then use the inverse solver to estimate the muscle activations that put the tongue in the correct position at the first TF of /ə-suk/. Figure 4.4 shows this initialization process for the /ə-suk/ sequence in speaker B. The mandible, hyoid, and maxilla are included in the model, but are not shown in the figure, for the sake of simplicity.

In this study, the tongue and bone models of the speakers fit the exact surface geometry extracted from the first TF of cine MRI (which is not perfectly symmetrical). However, to reduce the computational cost of the inverse problem, we assume bilateral symmetry in motion; the left and right muscles are activated together and with the same intensity. The control points (32 FE markers) are distributed in left half of the tongue.

Table 4.2. Absolute tracking error (mm) for speakers A-D, averaged over all control points and all time frames in /ə-gis/ and /ə-suk/.

		A	B	C	D
/ə-gis/	Mean	1.80 ± 0.68	1.95 ± 0.75	1.85 ± 0.64	1.70 ± 0.66
	Std	0.83 ± 0.35	0.71 ± 0.26	0.67 ± 0.22	0.73 ± 0.26
/ə-suk/	Mean	1.90 ± 0.55	1.88 ± 0.81	1.94 ± 0.60	1.90 ± 0.69
	Std	0.49 ± 0.19	0.79 ± 0.27	0.97 ± 0.28	0.62 ± 0.21

4.4 Results

We estimate the muscle activation patterns using the inverse simulation, with kinematic trajectories from the MRI data. Table 4.2 shows the tracking error average over all the control points, at 26 TFs, in the two utterances, /ə-gis/ and /ə-suk/. Figures 4.5 and 4.6 show the muscle activation patterns. The four speakers are shown in columns A-D with TFs (1 to 26) along the x axis.

Speakers A and B use an apical /s/, while speakers C and D use a laminal /s/. The muscles of the tongue include the following: genioglossus (GG), hyoglossus (HG), styloglossus (STY), verticalis (VERT), transverses (TRANS), geniohyoid (GH), mylohyoid (MH), and longitudinal (inferior [IL], superior [SL]). The GG, VERT and TRANS muscle bundles are further divided into five smaller functionally-distinct segments (a: posterior to e: interior), as suggested by Miyawaki et al. (1975) and Stone et al. (2004). We also follow Fang et al. (2009) in dividing the STY muscle into two functional segments (p:posterior and a:anterior). The muscles of the jaw and hyoid include the following: temporal (anterior [AT], middle [MT], posterior [PT]), masseter (superficial [SM], deep [DM]), pterygoid (medial [MP], superior-lateral [SP], inferior-lateral [IP]), digastric (anterior [AD], posterior [PD]) and stylo-hyoid (SH).

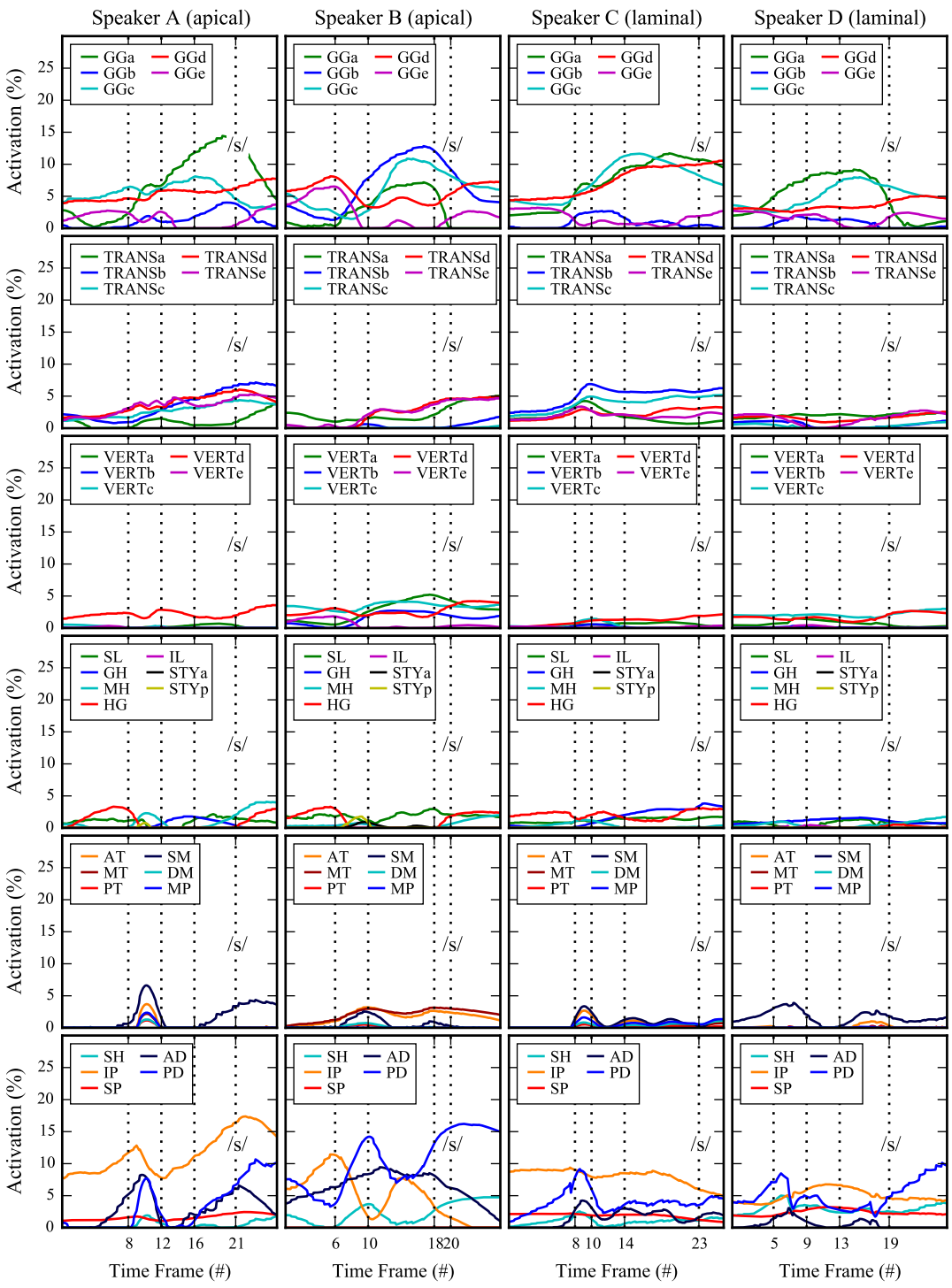


Figure 4.5. Muscle activations estimated by the inverse solver during the utterance /ə-gis/ for speakers A-D.

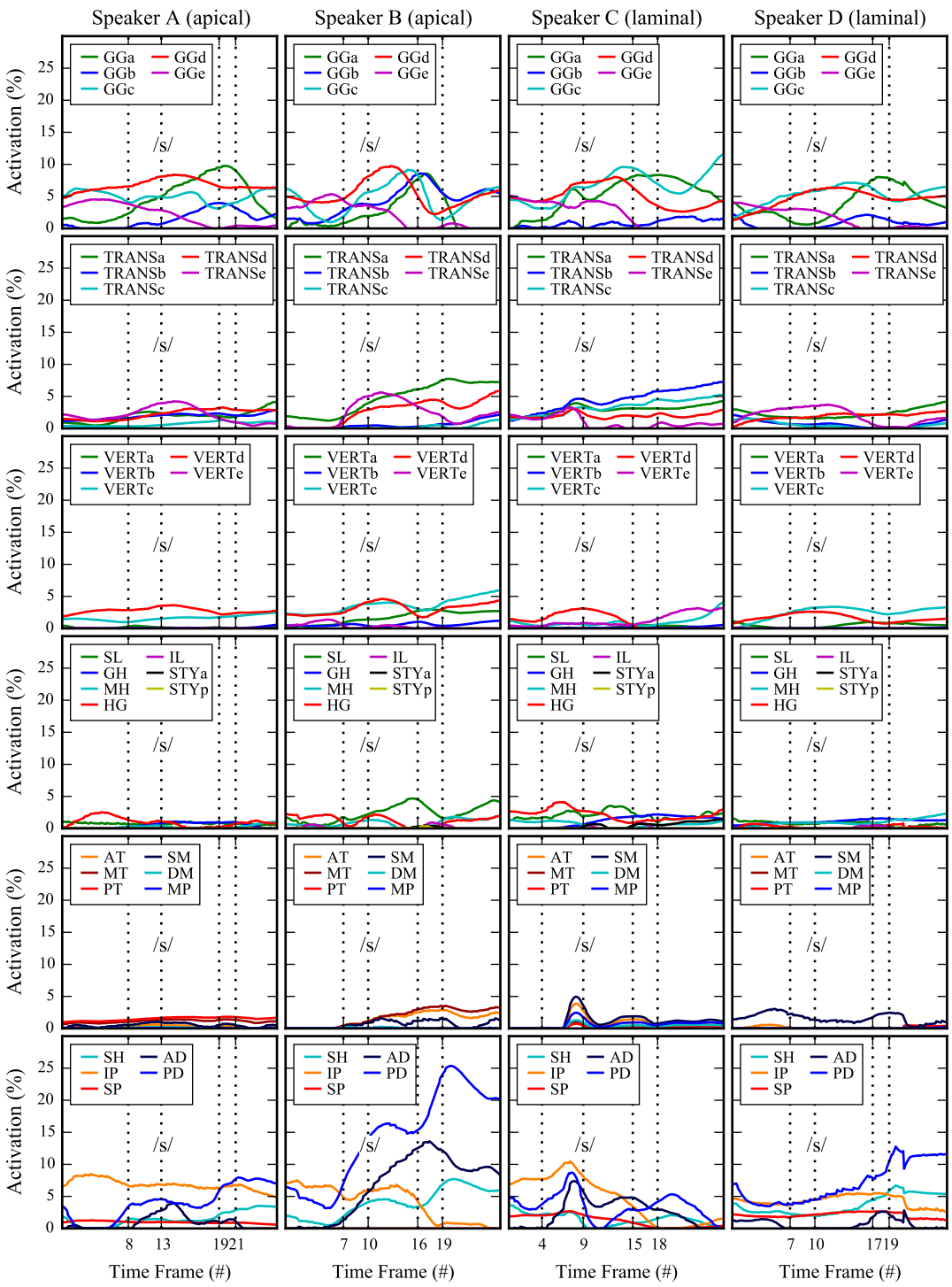


Figure 4.6. Muscle activations estimated by the inverse solver during the utterance /ə-suk/ for speakers A-D.

4.4.1 Tongue-Protruder Muscles

Tongue protruder muscles include the posterior region of the genioglossus muscle (GGa/b/c), which pulls the tongue forward, as well as the TRANS and VERT muscles. The GGa/b/c and TRANS muscles also elevate the tongue body. The floor muscles GH and MH assist in tongue elevation and protrusion.

Our results, as demonstrated in Figure 4.5 and 4.6, show that for both utterances, the GGa/b became more active over time and was maximally active prior to the final consonant. The exception to this pattern was the upper pharyngeal region of the GG (GGb) for speakers C and D, which had little activity and no pulse toward the end of the utterance. The GGc pulse occurred during both vowels. Speaker B used the GGb more than the other speakers, to position the vowel. The TRANS tended to increase in activation throughout the course of the utterance, with slightly more activation for the velar stops (/g/ or /k/) than the alveolar /s/. The TRANSd/e increased activity just before /s/, consistent with local tongue tip protrusion, more so for the apical speakers (A and B). Speakers varied in terms of which region of the TRANS was more active, but the net effect was protrusion. The activation patterns across the five segments of the TRANS, and VERT, were more alike (though not identical) than the activation patterns across the five segments of the GG. The GH pulls the hyoid forward, and shows small activations local to the vowel for speaker A, and throughout the utterance for speakers C and D. Speaker B does not use it at all. The MH activates more during the consonants, mainly to elevate the tongue.

4.4.2 Tongue-Retractor Muscles

The tongue is retracted by the extrinsic muscles, STY and HG, which pull the tongue backward/upward and backward/downward, respectively. Two intrinsic muscles, the SL and IL, also retract the tongue; they additionally elevate (SL) and lower (IL) the tip. Finally, the anterior fibers of the genioglossus (GGd/e) lower the upper body and blade of the tongue, causing backward motion of the tongue body.

More activation of the retractor muscles was expected for /suk/ than /gis/, since the tongue moves backwards during /suk/. For /suk/, the SL was

active for all subjects, with speakers B and C increasing the activation until the /u/ was reached. A continuous low-level activation of the SL was used by speakers A and D during /suk/, and by all four subjects during /gis/. The IL was not used at all during /gis/; but it was used for the /uk/ by three of the subjects, consistent with retracting the tongue tip. The largest activations in both utterances were seen in the GGd/e for all four speakers (5-10% activation). The GGd muscle – the most active – lowers or stabilizes the tongue dorsum, and the GGe further lowers the tongue blade. For /ə-gis/, the speakers used the GGd throughout the utterance, with smaller activations at /g/ than /i/ and /s/. During /ə-suk/, the GGd was most active during /u/. The GGe was active for /ə/ and the breath, with occasional activation for the first consonant, irrespective of what it was.

Of the two extrinsic retractors, the STY was fairly quiescent for both utterances, with speaker B using it during /gis/, and speaker C during /suk/. The HG, on the other hand, was active for 3 speakers (A, B and C), mostly during /ə/ and /s/ in both utterances. Of the two intrinsic retractors, the SL was active for all subjects, mostly during /ə/, /is/ (gis) and /uk/ (suk). During /suk/, speaker A and D had minimal SL activity. The IL was mostly quiescent, with very slight, occasional activity during the velar stops /g/ and /k/.

4.4.3 Other Muscles

Row 5 in Figure 4.5 and 4.6 contains the jaw closing muscles, which globally elevate the tongue. For /ə-gis/, these muscles have large peaks of activity during closure into /g/, and smaller ones during the motion into /s/ (consistent with tongue elevation for those sounds). For /ə-suk/, the activations tend to be smaller than for /ə-gis/, with only speakers C and D, having a closure activation into /s/. The IP and SP in row 6 are jaw protruding and closing muscles. In /ə-gis/, they behave like the jaw closing muscles in row 5, with two peaks of activity, one preparatory for /g/ and a second during the /i/ or /s/. The SH and PD pull the hyoid back and up; the AD pulls the hyoid forward and tongue up. These muscles, like the jaw closers, are most active during the /g/ and /s/, although speaker B had a fairly active AD throughout. These activations could allow fine tuning of jaw position; They could also be related to pitch changes, as the position of the larynx

varies with pitch. During /ə-suk/, the IP is active throughout the utterance (subjects A, D) or during the first half (subjects B, C). The SP has little activity in either utterance. The hyoid-positioner muscles again show significant activity in the PD associated with /k/. Speaker C appears to use both the jaw closers, openers, and hyoid-positioners in the transition from /ə/ to /s/.

4.5 Discussion

This study investigates differences in the key muscle activations and their overall activation pattern, during apical vs. laminal /s/ production, and as a function of differences in vowel context, using speaker-specific biomechanical models. We discuss the results below.

4.5.1 Apical vs. Laminal Speakers

Speakers A, B used an apical /s/, and speakers C, D used a laminal /s/. Both the VERTd and TRANSd/e were more active for the apical /s/. This difference is not seen in the GG data; however, it should be noted that for the TRANS and VERT, region *e* extends into the tongue tip, whereas the GGe stops at the tongue blade. It is possible that these additional activations create a very subtle difference in tongue positioning. The differences involved in creating an apical vs. laminal /s/ may require less active effort than one would expect. For example, Stone et al. (2012) find that palate shape has a strong effect on choice of /s/-type, and some of the difference in tongue tip shape may reflect palate shape. Moreover, thus far, only a slightly faster tip motion in apical /s/ has been found to distinguish the two motions (Reichard et al., 2012). Perhaps the simultaneous activation of the VERTd and TRANSd/e protrudes the tip slightly more in apical /s/, and the palate constraint reduces the overall activation needed.

4.5.2 Mechanisms of Tongue Elevation

One of the original questions asked in the study was whether the /s/ sound in /gis/ and /suk/ would differ because of different neighboring sounds and different locations in the utterance. Interestingly, a large difference was observed in muscle activation patterns related to the location of the /s/ in the utterance. This was due, however, to the vowel preceding the /s/ more than the /s/ position.

Figures 4.5 and 4.6 show that the pharyngeal portions of the GG (a/b/c) are active into the last consonant of each utterance, regardless of whether it is /s/ or /k/, while the jaw muscles appear more active at the beginning of the utterance. This can be explained by the context. The /ə/ at the start of the utterance requires an open jaw. The jaw closure into the following consonant is large, as seen in the activation of the jaw closure muscles at or after the /ə/, and may do the lion's share of the tongue elevation/fronting needed for both the /s/ and the /k/. When these same consonants appear at the end of the utterance, however, the jaw is already quite closed for the preceding vowel (/i/ or /u/), and so the tongue must internally elevate and front itself, increasing activation in the GGa/b/c.

4.5.3 Commonalities Across Speakers

Since tongue muscle activity measured from EMG usually shows variability among subjects, it is not surprising to see individual differences among speakers in our simulation results. However, there are some similarities that can be observed among all speakers. The first commonality across speakers is the large amount of muscle activation in the largest tongue muscle, genioglossus (GGa/b/c/d/e), followed by the jaw advancement muscle (the internal pterygoid [IP]), and the hyoid positioner muscles (the digastric [AD, PD] and the stylo-hyoid [SH]). The GGa/b/c was the most active muscle of protrusion/elevation for all subjects, with as much as 15% activation. The GGd/e was the most active muscle of retraction/lowering, with up to 10% activation. The GGa was always activated during articulation of the consonants, to elevate the tongue to the palate absent jaw assistance. The GGd was continually active in both utterances – possibly to stabilize the upper tongue surface so it did not hit the palate inadvertently. Equally active were

the IP, AD, PD and SH. IP was more active during the forward moving /gis/, but was still quite active in /suk/, where it was most active for /s/ and tapered off for /k/.

Jaw position is critical and inflexible for /s/. In both utterances, the IP was quite active at or before the /s/. The hyoid positioning muscles, AD, PD and SH, were active in both utterances, often with pulses for the consonants. The hyoid is a particularly unstable bone, as it is the only bone in the human body that does not articulate with another bone. It is stabilized entirely by muscles. The AD pulls it forward, The PD pulls it back and up, The SH pulls it down. The PD and SH were often active synchronously, sometimes with AD and sometimes without. These muscles may be so active because they have three roles that occur simultaneously in speech. Firstly, they position the hyoid to allow anterior-posterior tongue body motion during vowels. Secondly, they resist the anterior pull on the hyoid of the GGa during /s/ and /k/ or /g/. Thirdly, they assist in changing pitch, as hyoid/thyroid position varies with pitch in speaking.

The second commonality among speakers was a similar variety of activation patterns across the GG regions (a/b/c/d/e), consistent with independent activation of fibers throughout the GG. Stone et al. (2004) and Miyawaki et al. (1975) found independent regions of compression and activation in this muscle. Anatomical studies have shown very high innervation of these and all muscles of the tongue (Sokoloff et al., 1992). The other muscles that make up a structural unit with the GG, namely the TRANS and VERT (see Takemoto et al. (2001)), show considerably less activation ($< 5\%$) and may be used to fine-tune the position and surface shape of the tongue. Some behavioral differences in these muscles were consistent with differences in apical vs. laminal /s/. The floor muscles, GH and MH, have little activation during these utterances and may be more important for swallowing.

4.6 Conclusions

In this chapter we have enabled data-driven simulation of our speaker-specific biomechanical models to investigate inter- and intra-subject variability in speech motor control. We use MRI data of four normal subjects speaking the /s/ sound in two vowel context. Two of the speakers use apical and

the other two use laminal /s/. The results indicate the dominant use of the genioglossus muscle over the other muscles of the tongue. As expected, the posterior portion of the genioglossus is found to be more active than its anterior when /s/ follows /i/ in /ə-gis/. The reverse is true when /s/ precedes /u/ in /ə-suk/. The transverse muscle is also found to be moderately active more at the anterior, but also at posterior for two of the speakers. The activations of other muscles of the tongue seem to be subtle, perhaps used to fine-tune tongue posture. Our results also identifies the anterior-digastric muscle, as well as the inferior-lateral portion of the pterygoid muscle, to be the key active muscles of the jaw and hyoid during /s/. The results do not indicate any substantial difference between apical and laminal /s/ types.

Chapter 5

Acoustic Analysis for Speech Synthesis

Due to recent advances in acquisition technologies, speech data, including audio signals and medical images is abundant today. Such data motivates the use of computational approaches for modeling speech phenomena (Vasconcelos et al., 2012; Ventura et al., 2009, 2013). On one hand, biomechanical models of the oropharynx aim to simulate the dynamics of speech production under simplified – but biologically and physically valid – assumptions; on the other hand, articulatory speech synthesizers focus on generated sound as an end product, by designing a representation of the vocal tract and folds that is capable of generating the desired acoustics of an observed shape of the oral cavity (Doel et al., 2006; Birkholz et al., 2013). The search for an ideal model – that represents both the acoustical and biomechanical characteristics of the oropharynx – continues to this date.

5.1 Synthesis of Vowels in Running Speech

In this section, we expand our subject-specific modelling and simulation framework from chapters 3 and 4 to include a real-time acoustic synthesizer. The biomechanical models are enhanced with an air-tight VT mesh that deforms along with the movement of the articulators. The VT geometry is then processed in real-time to extract the 1D representation required for solving Navier-Stokes equations for vowel synthesis (Doel et al., 2006).

The vocal folds oscillate during the articulation of vowels and voiced-consonants (e.g., /b/), but are wide open and have little effect on articulation of fricatives (e.g. /s/) and stops (e.g. /k)/. Constrictions or obstructions at certain points in the tract create turbulence that generates the high frequency noise

responsible for making the fricatives and stops. The synthesis of fricatives depends to a large extent on lung pressure and the noise characteristics of the system. Due to the lack of voicing information, we focus our acoustic analysis solely on the synthesis of the vowels, specifically /i/ in /ə-gis/, and /u/ in /ə-suk/. The reduced vowel /ə/ is only used to help the subject maintain a neutral tongue posture at the start of the speech utterance.

5.1.1 Biomechanical Model of Vocal Tract

We model the VT as a deformable, air-tight mesh – referred to as geometric *skin* – which is coupled to the articulators, as proposed by Stavness et al. (2014b). Each point on the skin is attached to one or more master components, which can either be 3-DOF points, such as finite-element nodes, or 6-DOF frames, such as rigid body coordinates. The position of each skin vertex ($\mathbf{q_v}$) is calculated as a weighted sum of contributions from each master component:

$$\mathbf{q}_v = \mathbf{q}_{v_0} + \sum_{i=1}^{M} w_i f_i(\mathbf{q}_m, \mathbf{q}_{m_0}, \mathbf{q}_{v_0}) \tag{5.1}$$

where $\mathbf{q}_{v_0}$ is the initial position of the skinned point, $\mathbf{q}_{m_0}$ is the collective rest state of the masters, w_i is the skinning weight associated with the i^{th} master component, and f_i is the corresponding blending function. For a point master (such as a FE node) the blending function f_i is the displacement of the point. For frames (such as rigid bodies) f_i is calculated by linear, or dual-quaternion linear, blending. To provide two-way coupling between the skinned mesh and articulators, the forces acting on the skin points are also propagated back to their dynamic masters.

To create the skin, we initially segment the shape of the VT from the first time-frame of cine MRI data (described in Section 4.1). The skin is attached to and deforms along with the motion of the mandible rigid-body and tongue FE model. We also restrict the motion of the VT to the fixed boundaries of the maxilla and pharyngeal wall.

5.1.2 Time-Domain Acoustical Model

For our 1D acoustic analysis, we describe the VT with an area function $A(x,t)$, where $0 \leq x \leq L$ is the distance from the glottis on the tube axis and t denotes time. We take a similar notion of Doel and Ascher (2008) in defining the variables $u(x,t)$ and $p(x,t)$ as the scaled versions of volume-velocity $\hat{u}$ and air density $\hat{\rho}$, respectively:

$$u(x,t) = A(x,t)\hat{u}/c \tag{5.2a}$$
$$p(x,t) = \hat{\rho}/\rho_0 - 1 \tag{5.2b}$$

where ρ_0 is the mass density of the air and c is the speed of sound. We solve for $u(x,t)$ and $p(x,t)$ in the tube using derivations of the linearized Navier-Stokes equation (5.3a), and the equation of continuity (5.3b), subject to the boundary conditions described in Equation 5.3c:

$$\frac{\partial(u/A)}{\partial t} + c\frac{\partial p}{\partial x} = -d(A)u + D(A)\frac{\partial^2 u}{\partial x^2} \tag{5.3a}$$
$$\frac{\partial(Ap)}{\partial t} + c\frac{\partial u}{\partial x} = -\frac{\partial A}{\partial t} \tag{5.3b}$$
$$u(0,t) = u_g(t), \quad p(L,t) = 0 \tag{5.3c}$$

The right hand side of Equation 5.3a is the damping term, added to model the frequency dependant wall-loss. It is beneficial to note that setting the damping term to zero, and combining equations 5.3a and 5.3b together, yields the classic simple wave equation (if $A = 1$) as in the following:

$$\frac{\partial^2 u}{\partial t^2} = -c^2\frac{\partial^2 u}{\partial x^2} \tag{5.4}$$

For monochromatic waves of the form $e^{iw(t-x)/c}$, the damping term in Equation 5.3a results in the following:

$$-\left[d(A) + D(A)w^2/c^2\right]u \tag{5.5}$$

We follow Doel and Ascher (2008) in using $d(A) = d_0 A^{-3/2}$ and $D(A) = D_0 A^{-3/2}$ with the wall loss coefficients $d_0 = 1.6\ m/s$ and $D_0 = 0.002\ m^3/s$, to match previously reported hard-wall loss at frequencies 500 Hz and 2000 Hz; some more scaling was applied to these damping coefficients ($\times 4$ and

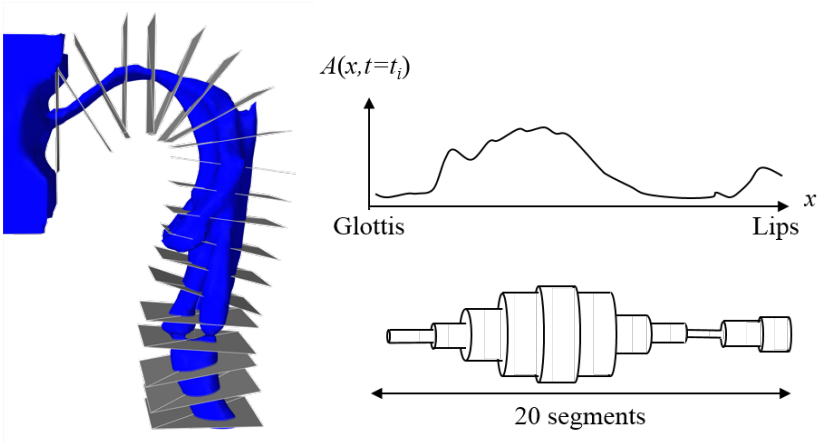


Figure 5.1. Intersecting planes superimposed on VT geometry at time $t = t_i$ (left), vs. analog area function $A(x, t_i)$ and its schematic discretized representation, using 20 segments of equal length (right).

$\times 8$ depending on discretization factor) during implementation, in order to yield the desired bandwidth. Figure 5.1 shows the intersecting planes used to calculate a discretized area function representation of the VT geometry at time $t = t_i$.

In Equation 5.3c, $u_g(t)$ is the source volume velocity at the glottis. We couple the VT to a two-mass glottal model (Ishizaka and Flanigan, 1972). We refer to Doel and Ascher (2008) for full details of the implementation.

5.1.3 Results and Discussion

In this section, we first use the muscle activations calculated in the previous chapter (Section 4.4) to run our models in the forward simulation scheme. This time we add the skinned VT mesh to our models, to find the deformed VT geometries in each time-frame. More information on the MRI data is presented in Section 4.1. Figure 5.2 shows the mid-sagittal cut of the models, superimposed on the image data in time-frames that correspond to /i/ and /u/. The red arrows in the figure indicate the areas of mismatch that often occur around the lips, velum, and epiglottis.

The mismatch occurs primarily because our VT model is based on segmen-

tations around the initial position of these structures, which, at times, is different from what we see in the /i/ or /u/ time-frames.

Next, we fit a center-line to our VT geometries, and set up 20 planes, angled normal to the center-line, that start from the lips and end at the epiglottis. The intersection of these planes with the VT gives a discrete representation of $A(x,t)$ at any time (t) during the simulation. These area functions are fed into the acoustic synthesizer in real-time, to generate sounds from which we then calculate the formant frequencies (as the peaks of the spectrum).

As for the ground truth, we first look at the recorded audio signals, which are available for three of our speakers (A, B, and D), while they repeat /ə-gis/ and /ə-suk/, synced to the metronome in a lab environment. Figure 5.3 shows the acoustic profile and spectrum of a single repetition of /ə-gis/, as spoken by speaker B. The formant frequency is calculated at the mid-point of the time interval associated with the vowel (here /i/). We average the results over all repetitions of the utterance.

In addition, we extract the airway from the associated TF of the cine MRI data (using the semi-manual segmentation tool ITK-SNAP by Yushkevich et al. [2006]) to obtain the corresponding area functions that we later feed

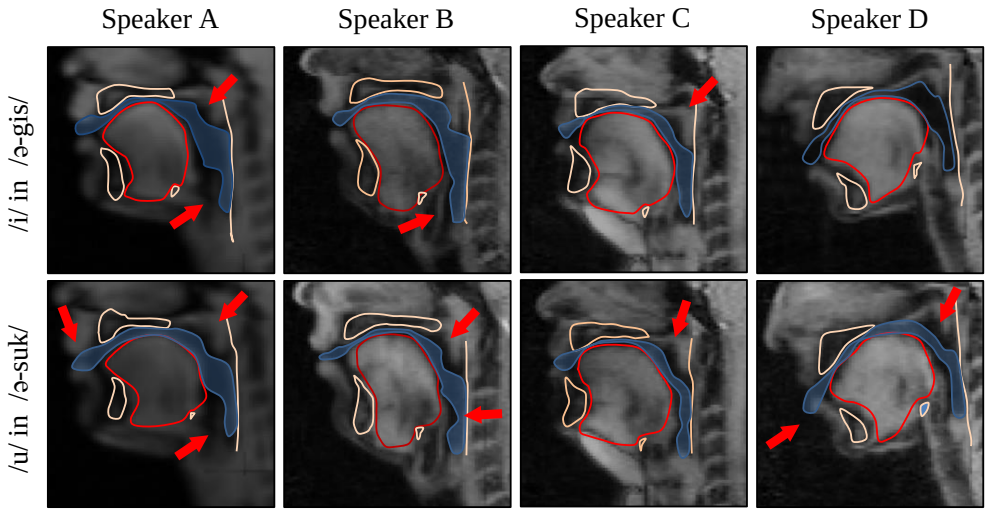


Figure 5.2. Deformed VT mesh (blue) superimposed on mid-sagittal cine MR images at /i/ and /u/. The red arrows indicate the areas of mismatch.

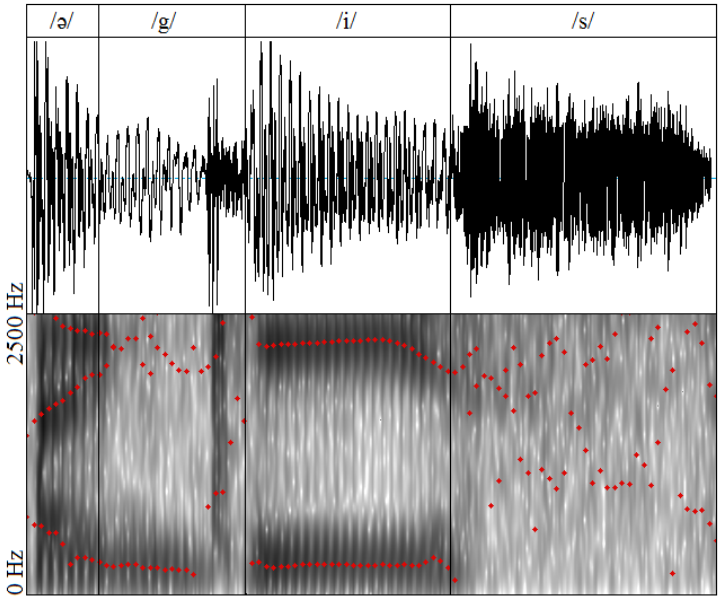


Figure 5.3. The audio signal and spectra for one repetition of the speech utterance /ə-gis/, as spoken by speaker B. The formants are shown as red dots associated with each time instant of audio, using Praat phoneme analysis software (Boersma and Weenink, 2005).

into the acoustic synthesizer. The resultant formants provide a maximum limit to the accuracy obtainable in our simulations – given the contrast and resolution of the cine MRI data, and the fidelity of the acoustic synthesizer.

Table 5.1 shows the first three formants for both /i/ and /u/ in speakers A-D, as calculated from our simulations, the audio signals, and cine MRI images. Figure 5.4 represents the average values, and standard deviation of these formants, to ease the overall comparison. Note that for both /i/ and /u/, F_1 is often higher in our simulations, compared to the audio signals. The opposite happens for F_2, where the simulations fall short. F_3 shows a mixed behaviour where it is noticeably lower than the audio in /i/ of speaker A, and higher than the audio in /u/ of speakers B and D. Looking at cine MRI measurements, we realize that, for /i/, the values of F_2 are also lower than the audio, with a large difference for /i/ of speaker D. For the same case, F_1 is unusually low in cine MRI.

We speculate that one reason for such discrepancies lies in the fact that our

Table 5.1. Formant frequencies (F_1, F_2, F_3) of /i/ and /u/ in speakers A-D, as computed from our simulations, audio signals, and cine MRI data (values are in Hz).

	Speakers	Simulation	Cine MRI	Audio
/i/	A	(317, 1652, 2613)	(240, 1826, 3062)	(243, 2264, 3077)
	B	(256, 1905, 3086)	(267, 2055, 3012)	(268, 2272, 3032)
	C	(327, 1600, 3091)	(272, 1943, 3033)	Not Available
	D	(424, 1871, 2886)	(196, 2086, 2720)	(347, 2675, 3032)
/u/	A	(379, 1516, 2321)	(373, 1699, 2771)	(300, 1636, 2541)
	B	(311, 1698, 2696)	(317, 1995, 2704)	(333, 1814, 2317)
	C	(405, 1633, 2899)	(371, 1859, 2842)	Not Available
	D	(430, 1990, 2921)	(321, 2440, 2836)	(380, 2022, 2674)

dynamic MRI images provide incomplete VT visibility. In particular, due to the fuzzy boundaries between the teeth and airway, we tend to exclude some of the VT volume in the mouth opening (mostly between the teeth). The low spatial resolution of the images also negates our efforts to extract the VT posterior pharyngeal side branches. In addition, we notice that the average length of the VT, as reported in the literature for adult speakers, is around 17cm. However, the longest airway extracted from our MRI images does not exceed 14.5cm, leaving out a few centimeters around/below the epiglottis. If these areas are not visible in the first TF of cine MRI sequence, our models can not recover later. It is important to note that synthetic increase of the VT length (i.e., as an independent parameter to the acoustic synthesizer) lowers the value of the three formants, with greater effect on F_2 and F_3. Such scaling of the center-line does not provide a solution though, since it remaps the position of the intersecting planes to an arbitrary position that no longer matches the image data.

The second hypothesis to explain discrepancies in Table 5.1 is that the 1D implementation of Navier-Stokes equations, chosen here for the sake of real-time simulations, does not capture the complexities of 3D VT geometry, and is not enough for accurate calculation of the formant frequencies. We will address this hypothesis in the next section, by comparing the methods to the results of some 3D analysis.

In a separate experiment, we perform the acoustic analysis of our simulations using two different resolution for our tongue models, as described in Chapter 3). FE_{reg} is of lower resolution and matches the element configuration of an standard generic model (Buchaillard et al., 2009), while FE_{final} provides a mixed-element alternative with higher resolution. We carry our analysis on /i/ of speaker B, which, as shown in Table 5.1, demonstrates a better match in formants with the cine MRI and audio data. The cine MRI value

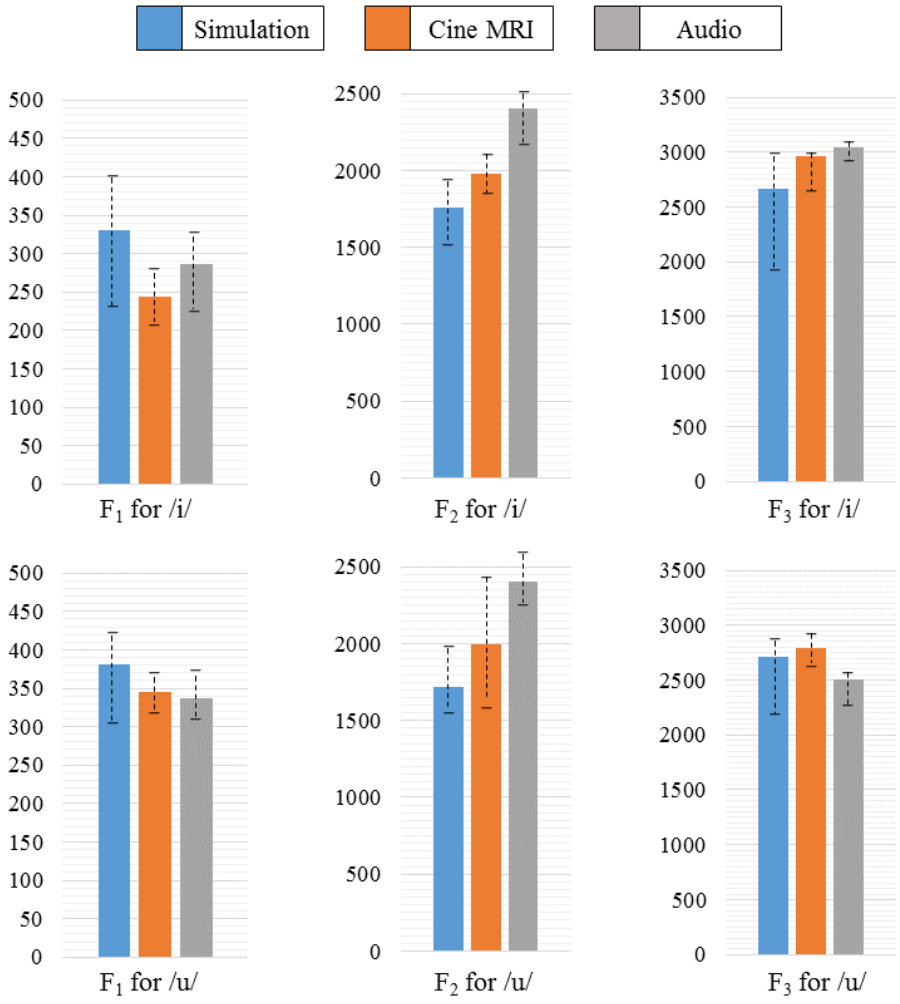


Figure 5.4. Average formant frequencies for /i/ (top row) vs. /u/ (bottom row) as computed from our simulations, audio signals, and cine MRI data.

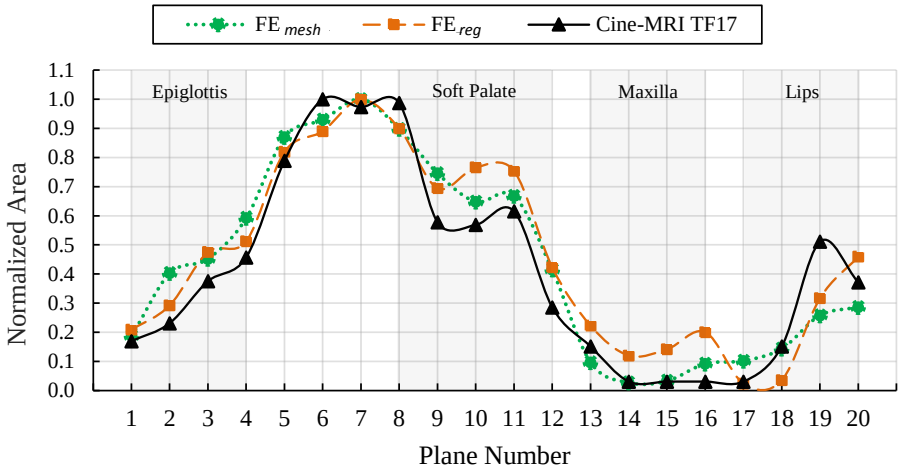


Figure 5.5. Simulation results for speaker B. A normalized area profile along the VT for /i/ compared to the cine MRI at time-frame 17.

Table 5.2. Formant frequencies for the simulations using FE_{reg} and FE_{mesh}, compared to the audio and cine MRI data for /i/ of speaker B.

	Audio	Cine MRI	FE_{reg}	FE_{mesh}
F_1(Hz)	268	267	262	256
F_2(Hz)	2272	2055	1905	1995

of F_2, however, remains 9.5% less than the audio signal. Figure 5.5 shows the normalized area profile along the VT at /i/ in our simulation, compared to the cine MRI data. Both the FE_{reg} and FE_{mesh} tongue models are able to capture the expected shape of the VT. Again, the noticeable mismatches happen at the areas influenced by the lips, velum, and epiglottis. Table 5.2 compares the formant frequencies of our simulations using both the FE_{reg} and FE_{mesh} for speaker B. Looking at F_2, we notice that while FE_{mesh} seems to decrease the gap between the simulation and the ground truth (audio and cine MRI), there is still a difference in values. This may be due to, (1) the discrepancy between the manual segmentations performed at the 1^{st} and 17^{th} time frame of the cine MRI, or (2) the arbitrary nature of centre-line extraction for 1D acoustic analysis. We will discuss the latter in more detail in the next section.

These quantitative results suggest that FE_{reg} and FE_{mesh} do not show an appreciable difference in acoustic performance for the simulation of the utterance /ə-gis/ using our source-filter based speech synthesizer (Doel and Ascher, 2008). Thus we conclude that the ArtiSynth generic tongue model (Buchaillard et al., 2009) provides sufficient resolution for modelling of this utterance at such a level of acoustic fidelity and cine MRI resolution.

5.2 Synthesis of Sustained Vowels

In Section 5.1, we coupled a state-of-the-art 1D acoustic synthesizer to subject-specific biomechanical models of the oropharynx, in order to generate sound in real-time. As shown in Subsection 5.1.3, the formant frequencies computed from the resultant acoustic signals are different from those of the audio signal. One hypothesis to explain this difference is that 1D acoustic analysis cannot capture the complex three dimensional shape of the VT, causing a discrepancy in the spectrum. In this section, we further investigate this hypothesis by running a series of experiments that aim to compare the results of 1D and 3D acoustic analysis. We hope such comparison assists in clarifying the limitations of our current models and/or speech synthesizer.

For the sake of comparison, we categorize the methods of acoustic analysis into two classes. The first is time-domain analysis, in which the wave equation is derived over time to generate sound. Finite Fourier transform (FFT) of the output sound signal is then divided by the FFT of the source signal (at the glottis) to yield the frequency spectrum of the filter (VT). We refer to the peaks of the spectrum as *formant* frequencies. Examples of the time-domain analysis are the acoustical models proposed by Doel and Ascher (2008) for 1D, and by Takemoto et al. (2014) for 3D analysis.

The second analysis is carried fully within the frequency domain. Since time is absent from the equations, the VT is not moving, i.e. the vowels are sustained. The VT transfer function (in frequency domain) is calculated by solving the frequency analogy of the wave equation. We refer to the peaks of the associated spectrum as *resonance* frequencies of the VT. Examples are the 1D and 3D models suggested by Aalto et al. (2014) and Kivelä (2015). Through the rest of this section, we explicitly use the terms *formant* and *resonance* to refer to the time-domain and frequency-domain analysis.

For our 3D analysis, we decided to follow Aalto et al. (2014) in calculating the Helmholtz resonances of our VT geometries using the 3D finite element method (FEM) analysis. The method is solely based on the shape of the VT and, does not require any glottal source excitation; thus, far fewer parameters need be adjusted, compared to a full 3D time-domain analysis, such as the one by Takemoto et al. (2014). Solving the equations directly in the Fourier domain simplifies the acoustical model by discarding any time-dependant damping or loss (such as effects of the vibration of the VT wall). Further simplification of the wave equations is also inevitable, in order to make mathematical derivations possible; however, the simulation runs close to real-time in comparison with time-domain 3D methods – that can take up to several hours for analysis of a single geometry (Takemoto et al., 2014).

5.2.1 Helmholtz Resonances

Assuming that the air flow in the VT is irrotational ($\nabla \times \mathbf{v} = 0$), we can define a velocity potential $\Phi(\mathbf{s}, t)$ with

$$\mathbf{v} = -\nabla\Phi \tag{5.6}$$

where we use vector $\mathbf{s}$ to denote the parameter of space in three dimensions. Solving the wave equation for either the pressure field ($\mathbf{P}$) or velocity field ($\mathbf{v}$) doesn't necessarily provide a simple answer for the other. Solving for Φ, on the other hand, yields $\mathbf{v}$ directly, from Equation 5.6, and $\mathbf{P}$, from the (linearized) Bernoulli equation for irrotational and unsteady flow:

$$\mathbf{P} = -\rho\frac{\partial\Phi}{\partial t} \tag{5.7}$$

The wave equation for the velocity potential is then summarized as follows:

$$\nabla^2\Phi = \frac{1}{c^2}\frac{\partial^2\Phi}{\partial t^2} \tag{5.8}$$

where ∇^2 is the Laplace operator. Making the velocity potential (in Equation 5.8) time-harmonic, by setting $\Phi_\lambda(\mathbf{s}, t) = \mathrm{Re}\{\Phi_\lambda(\mathbf{s})e^{i\lambda t}\}$, Equation 5.8 can be rewritten as

$$\nabla^2\Phi_\lambda(\mathbf{s}) + \frac{\lambda^2}{c^2}\Phi_\lambda(\mathbf{s}) = 0 \tag{5.9}$$

an eigenvalue problem that is, in terms of mathematics, considerably easier to solve than the wave equation. Using Equation 5.9, the vowel resonances are calculated by finding the eigenvalues (λ), and their corresponding velocity potential eigenfunction Φ_λ from the Helmholtz resonance problem:

$$c^2\nabla^2\Phi_\lambda + \lambda^2\Phi_\lambda = 0 \qquad \text{on } \Omega \tag{5.10a}$$

$$\Phi_\lambda = 0 \qquad \text{on } \Gamma_1 \tag{5.10b}$$

$$\alpha\lambda\Phi_\lambda + \frac{\partial\Phi_\lambda}{\partial\nu} = 0 \qquad \text{on } \Gamma_2 \tag{5.10c}$$

$$\lambda\Phi_\lambda + c\frac{\partial\Phi_\lambda}{\partial\nu} = 0 \qquad \text{on } \Gamma_3 \tag{5.10d}$$

where $\Omega \in \mathbb{R}^3$ is the air column volume, and $\partial\Omega$ is its surface – including the boundary at the mouth opening (Γ_1), the air-tissue interface (Γ_2) and a virtual plane above glottis (Γ_3). $\frac{\partial\Phi_\lambda}{\partial\nu}$ denotes the exterior normal derivative. Note that the Dirichlet boundary condition in Equation 5.10b simplifies the model by regarding the mouth boundary as an idealized open end of an acoustic tube, therefore neglecting lip radiation loss. The value of α regulates the energy dissipation through tissue walls, and the case $\alpha = 0$ corresponds with hard, reflecting boundaries. We calculate the numerical solution of Equation 5.10 with the Finite Element Method (FEM) using piecewise linear shape functions and approximately 10^5 tetrahedral elements. The imaginary parts of the eigenvalues (in the ascending order) yeild Helmholtz resonances ($\mathrm{H}_{R1}, \mathrm{H}_{R2}$, ...) of the VT air column in increasing order of frequency. We refer to Aalto et al. (2014) and Kivelä et al. (2013) for details of the implementation.

5.2.2 Webster Resonances

We also derive a 1D interpretation of Equation 5.8 in order to calculate what we will refer to as *Webster resonances* (W_R). This will allow investigation of acoustical differences between equations 5.3 and 5.10, independently from the dimensionality (1D vs. 3D) of the solution. To do so, we compute the averages of the 3D solution in tube cross-sections. If Φ is a solution to Equation 5.8, then such an average will be denoted by

$$\overline{\Phi}(s,t) = \frac{1}{A(s)}\int_{\Gamma(s)} \Phi \, \mathrm{d}A \tag{5.11}$$

where scalar s is the implicit parameter denoting points on the centre-line, and $\Gamma(s)$ is the normal cross-section at s. For the sake of simplicity, the area of this cross-section is considered to be circular – i.e., $A(s) = \pi R(s)^2$. The Webster equation is obtained after a lengthy derivation, as described by Lukkari and Malinen (2012):

$$(\frac{1}{c^2\Sigma(s)^2}\frac{\partial^2\Phi}{\partial t^2}) + \frac{2\pi\alpha W(s)}{A(S)}\frac{\partial\Phi}{\partial t} - \frac{1}{A(s)}\frac{\partial}{\partial s}(A\frac{\partial\Phi}{\partial s}) = 0 \tag{5.12}$$

Here Σ denotes the sound speed correction factor that depends on the curvature of the VT ($\kappa(s)$) as follows:

$$\Sigma(s) = (1 + \frac{1}{4}\eta^2(s))^{-1/2} \text{ with } \eta(s) = R(s)\kappa(s) \tag{5.13a}$$

$$W(s) = R(s)\sqrt{R'(s)^2 + (\kappa(s) - 1)^2} \tag{5.13b}$$

By setting $\Phi_\lambda(\mathbf{s}, t) = \text{Re } \Phi_\lambda(\mathbf{s})e^{i\lambda t}$ as before, the time-harmonic Webster equation can be written as follows:

$$(\frac{\lambda^2}{c^2}\frac{1}{\Sigma^2} + \lambda\frac{2\pi\alpha W}{A})\Phi_\lambda = \frac{1}{A}\frac{\partial}{\partial s}(A\frac{\partial\Phi_\lambda}{\partial s}) \qquad \text{on } [0, L] \tag{5.14a}$$

$$\lambda\Phi_\lambda - c\Phi_\lambda^{'} = 0 \qquad \text{at } s = 0 \tag{5.14b}$$

$$\Phi_\lambda = 0 \qquad \text{at } s = L \tag{5.14c}$$

Note that equations 5.3 and 5.12 both solve the so called Webster equation in 1D, but with different derivations: Equation 5.12 assumes that the centre-line is a smooth curve whose curvature affects the speed of sound in the VT, while Equation 5.3 assumes that the centre-line is piecewise linear. We refer to Kivelä (2015) for details of the implementation and parameter values.

The 1D approach suffers from ambiguity in the area functions, due to the lack of a uniquely definable centre-line for the 3D VT surface meshes. Different centre-lines have different acoustic lengths. This is a source for formant error not directly related to the Webster model itself, but the geometry processing it requires. In an effort to better compare the Webster and Helmholtz resonances, we linearly scale the length of the VT centre-line so that the

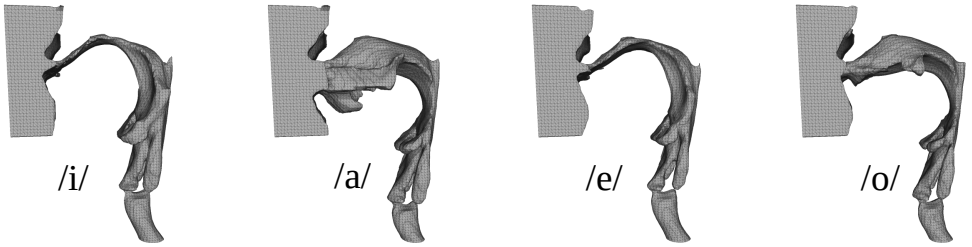


Figure 5.6. VT geometries extracted from MRI data (Aalto et al., 2013).

three lowest Webster resonances coincide in average with the corresponding Helmholtz resonances from the same VT configuration. We use only the three lowest resonances because they are purely longitudinal and, hence, accounted for by the Webster model. After scaling, the acoustic length is corrected based on Helmholtz model. The results (S_R) are expected to take away the systematic discrepancy between resonances from 1D and 3D acoustics.

5.2.3 Results and Discussion

For the simulations in this section, we use static MRI images acquired with a Siemens Magnetom Avanto 1.5 T scanner. A 12-element Head Matrix Coil, and a 4-element Neck Matrix Coil, allow for the Generalize Autocalibrating Partially Parallel Acquisition (GRAPPA) acceleration technique. One speaker, a 26-year-old male, was imaged while he uttered four sustained Finnish vowels. The MRI data covers the vocal and nasal tracts, from the lips and nostrils to the beginning of the trachea, in 44 sagittal slices, with an in-plane resolution of 1.9mm. Figure 5.6 shows the VT surface geometries extracted from MRI data using an automatized segmentation method (Aalto et al., 2013). Note the greater detail captured in these geometries compared to those extracted from dynamic MRI (in Section 5.1), especially below the epiglottis, at the posterior pharyngeal side branches, and between the teeth.

Figure 5.7 shows the first two formant/resonance frequencies, computed for the four Finnish vowels. Webster formants (W_F) are calculated by solving Equation 5.3, as suggested by Doel and Ascher (2008). Helmholtz (H_R) and Webster resonances (W_R) are obtained from equations 5.10 and 5.14, respec-

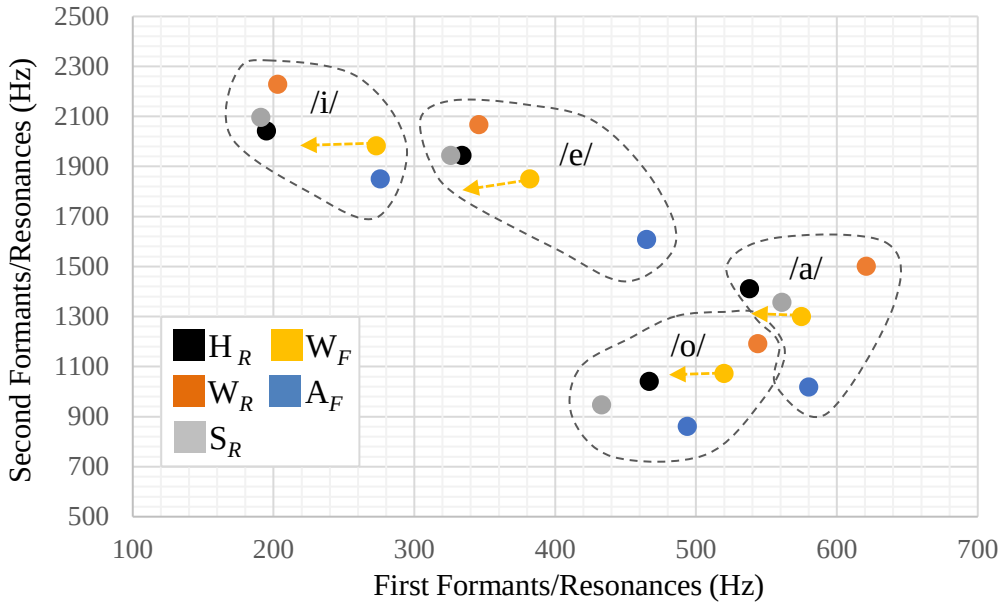


Figure 5.7. Simulation results for first and second formant/resonance frequencies for different vowels: Helmholtz resonances (H_R), Webster resonances (W_R) and their scaled version (S_R), Webster formants (W_F) and formants from audio signal (A_F).

tively (Aalto et al., 2014). S_R denotes the scaled version of W_R, as described in Subsection 5.2.2. The figure also includes the formant frequencies (A_F) computed from audio signals recorded in an anechoic chamber (Aalto et al., 2014). The values are averaged over 10 repetitions of each vowel utterance.

As we can see in Figure 5.7, the resonance values (H_R, W_R and S_R) lie close together for vowels /i/ and /e/, with S_R being closer to H_R, as expected. For vowels /o/ and /a/ there is more difference in the first resonances of H_R and W_R; For /o/, although S_R lies closer to H_R, its first resonance is surprisingly low. For all of the vowels in Figure 5.7, the second formant of the audio is less than the computed results. This finding contradicts our results in Table 5.1, where second Webster formants from cine MRI of /i/ are consistently lower than the values from the audio signals. The vowel /i/ is expected to be very sensitive to glottal end position, which, in turn, suggests the significance of adequate MRI resolution and accurate geometry processing for its spectral analysis.

Interestingly, the Webster formants (W_F) remain closer to the audio formants (A_F) than any of the resonances in the case of /i/, /e/, and /a/. For /o/ the distance to the A_F is almost equal for W_F and H_R, with both having similar values for the second formant/resonance; however, the first H_R is lower, and the first W_F is higher, than the first A_F.

The time-domain Webster analysis (Doel and Ascher, 2008) accounts for the VT wall-vibration phenomenon that is missing in the resonance analysis. This is done by substituting $A(x,t)$, from equation 5.3, with $A(x,t) + C(x,t)y(x,t)$: where $C(x,t)$ is the slow-varying circumference and $y(x,t)$ is the wall displacement governed by a damped mass-spring system. Setting $y(x,t)$ to zero, the Webster formants move along the arrows in Figure 5.7, reducing in their first formants. This moves the W_F closer to the H_R as both acoustical models now ignore the wall vibration. Meanwhile, W_F moves away from the audio formants in the case of /i/, /e/, and /a/. The distance between W_R and W_F remains large, despite the fact that both acoustical models solve the Webster equation. The results imply that 3D Helmholtz analysis is more realistic than its 1D Webster version, as expected.

Overall, our experiments suggest that the time-domain interpretation of acoustic equations provides more realistic results – even if it requires reducing from 3D to 1D. This may be partially due to the fact that time-domain analysis allows for more complexity in the acoustical model such as inclusion of lip radiation and wall loss. Certainly unknown parameters always remain (such as those involved in glottal flow, coupling between fluid mechanics and acoustical analysis, etc.), which are estimated indirectly, based on observed behaviour in simulations.

It should be noted that our experiments are solely based on data from a single speaker. A larger database – inclusive of more speakers from different genders and languages – is needed in order to confirm the validity/generality of our findings.

5.3 Conclusions

In this chapter, we have investigated the acoustics of speech production. In Section 5.1 we enable acoustical synthesis of the vowel phonemes, based on

data-driven simulations of our speaker-specific biomechanical models. We model the VT as an air-tight skin mesh that deforms, along with the articulators. We then use a standard acoustic synthesizer (Doel and Ascher, 2008) that generate sounds, based on a 1D representation of VT geometries. Comparison of the formant frequencies of generated sounds, with those from the audio signals and cine MRI data, demonstrates discrepancies. We speculate that low contrast and resolution of dynamic MR images contribute to these results. In addition, our deformed VT often suffers from a mismatch with images in the regions around the epiglottis, soft-palate, and lips (whose biomechanical models were not included).

In Section 5.2 we look at the sustained vowels in order to experiment with a 3D, frequency-based analysis of the VT (Aalto et al., 2014). We mainly seek to investigate if such analysis (performed in 3D) would yield more accurate results than the 1D Doel-Ascher synthesizer. This time, we use the VT surface geometries extracted from the high resolution static MRI of one Finnish speaker uttering four sustained vowel phonemes. Our results suggest the advantage of time-domain over the frequency-domain analysis, as the Webster formants (Doel and Ascher, 2008) lie closer to the formants of audio signals than the Helmholtz resonances (Aalto et al., 2014).

We believe in effective coupling of biomechanical and acoustical models of the oropharynx, in order to provide a better understanding of speech production. A system that is capable of providing a mapping from activation units to articulatory movements and sound, would greatly assist with speech rehabilitation planning. To accomplish this, several challenges on each side should be addressed. The speaker-specific biomechanical models should include the epiglottis, velum, and face, to produce accurate VT deformations. In addition, the accuracy of the models (and their simulations) depends highly on the resolution and contrast of dynamic medical images. On the acoustical side, 3D time-domain analysis of the VT is still impractical, but greatly anticipated, for real-time simulations. In addition, biomechanical and acoustical methods should move from generic parameter-tuning to meet the needs of speaker-specific models. This necessitates further advances in data measurements from the oropharyngeal structures. An example of this is audio and image recording from the subglottal area, in order to enable synthesis of consonant phonemes. A high-resolution MRI depiction of muscle fibers could also be used to adjust the tongue model – especially for the pathological anatomy.

Chapter 6

Conclusions

This dissertation develops methods for subject-specific modeling and simulation of the oropharynx, with direct applications to speech research. Driven by a clinical need for better understanding of speech biomechanics, we enable investigation of inter- and intra-speaker variability in speech production, based on medical imaging data. To do so, we incorporate, develop and evaluate methods that address several challenges to data processing, mesh generation, data-driven simulation, and acoustical modeling.

In Chapter 3 we first design a real-time, click-based expert-interaction scheme for a mesh-to-image registration method, in order to delineate the tongue surface from volumetric MR images. To accomplish this, we use a shape matching algorithm, driven by gradients of intensity profiles, in the direction normal to surface mesh. Confined to the segmented surface, we then generate a hexahedral-dominant mesh that bears the desired spatial resolution and mesh quality. Using the Mesh-Match-and-Repair registration method, we augment our mesh with biomechanical information embedded in a standard, generic tongue model, while permitting adjustments to muscle fiber definition. We finally couple our FE tongue models to rigid-body models of the mandible, maxilla and hyoid, and successfully perform forward simulations that comply with the literature.

In Chapter 4 we perform data-driven simulations of our subject-specific models in order to investigate inter- and intra-speaker variability in muscle activation patterns. We derive our inverse simulations using tissue trajectories extracted from tagged- and cine-MRI data (subject to damping and regularization constraints that compensate for muscle redundancy and system instability). This is the first time that such comprehensive, quantified tissue-point motion has been used to drive an oropharyngeal, biomechanical model. Our simulations lead to a novel interpretation of the data itself, by distinguishing between the active and passive muscle shortenings, and identifying

the co-contraction of antagonist muscles where no regional shortening is observed.

Next, we generate sound by coupling our biomechanical models with a 1D standard acoustic synthesizer. A vocal tract skin mesh translates the motion of the articulators to deformations of an airtight airway. Such deformations update the extracted centre-line and area profiles, and alter the synthesized sound accordingly. Using the recorded audio as our ground truth, we extend our experiments to 3D methods, and identify the sources of inaccuracy in computed formant frequencies. Our attempts show promise in linking acoustical and biomechanical models for speech research.

6.1 Concluding Remarks

By moving from generic to speaker-specific models, we fill a gap that exists between the oropharyngeal medical images and the biomechanical models of articulatory motion. We demonstrate that the current technology for super-resolution reconstruction of cine- and tagged-MRI provides adequate information for building and data-driven simulation of speaker-specific oropharyngeal models. We carry such transition (from data to individualized models) by facilitating MRI processing and FE meshing. In particular, we demonstrate significant improvement in time efficiency using our tongue segmentation method, and allow adjustability in mesh configurations of our FE tongue models.

Our inverse simulations prove beneficial in investigating motor control variability across speakers. Looking at the laminal and apical /s/ in four healthy English speakers, we indicate dominant use of the genioglossus muscle, with its posterior portion being more active than the anterior when /s/ follows /i/. The reverse is true when /(s)/ precedes /u/. We also find the anterior digastric muscle, as well as the inferior-lateral pterygoid muscle, to be the key active muscles of the jaw and hyoid during /s/. Our results do not indicate any substantial difference between activation patterns in apical and laminal /s/ types, confirming the hypothesis that the associated variations in tongue shape are more attributable to palate constraint. In addition, we are able to verify the theories of motor control that suggest multiple functionally-distinct segments throughout key muscles of the tongue. Our results show

variety in activation mainly through the genioglossus, but less through verticalis and transverses muscles. Such speaker-specific simulations provide abundant opportunities for testing, modifying, and validating the theories of speech strategy across the population.

Finally, we introduce a platform to generate sound, based on the predicted biomechanics. The accurate and realistic synthesis of speech phonemes proves challenging, and falls outside of the scope of this study; nevertheless, our attempts set the stage for a unified framework, in which the articulators are driven (based on the medical data) to generate a particular sound. We show that synthesized vowels suffer from ambiguity in vocal tract representation for 1D acoustical models, and require higher MRI resolution for 3D methods. The results also show sensitivity to the generic – and often conventional – parameter-tuning performed for acoustical models, in order to generate a realistic sound.

6.2 Future Work

I suggest potential improvements and future work at the end of each chapter, but now would like to highlight a few directions from a global perspective.

To improve our subject-specific modeling and simulation framework, I recommend including models of other oropharyngeal organs, such as the velum, uvula, epiglottis and lips (upon which the shape of the vocal tract depends). Due to the small size of these organs, such modeling would require MR images of higher resolution and contrast. For larger scale models, such as the face, I recommend implementing FE registration methods that work directly with MRI images (rather than surface meshes) to eliminate the burden of segmentation.

Our modeling and simulation framework would certainly benefit from a more comprehensive set of medical data. CT images remove the complexity of bone segmentation; jaw optical tracking can increase the reliability of inverse simulations. Other biomechanical measurements, such as maximum jaw exertion force, could help with tuning each speaker-specific model. A more in-depth post processing of tagged MRI data could enable tracking of each individual muscle in the tongue. The results might be used for validation of inverse

simulations, or be incorporated into a muscle-based (as opposed to a point-based) inverse simulation scheme, ensuring a more meaningful averaging of tagged MRI tracking data. In addition, a higher resolution MRI is (deemed to be) necessary for enabling 3D acoustical analysis.

The vocal tract skin mesh, which is used to link our biomechanical models to their acoustical counterparts, could benefit from methods that allow changes in the topology of the airway (such as shortening or discontinuity). This feature might come in handy where the biomechanical models of teeth and face are present, and could prove essential for modeling highly deformed or, at times, isolated air cavities in the mouth.

Lastly, I recommend that a generic tongue model incorporate a higher resolution representation of muscle fibers, using fibers extracted from digitization of the cadaver tissue. Such a model, when adapted to speaker data, would require faster computation methods. Experiments with mesh-less elastic models, instead of FE, could reduce simulation time; however, some technical difficulties (such as ambiguity in definition of a unique attachment site) need to be addressed accordingly. It would also be beneficial to compare the accuracy of such simulations against FE models and image data in speech production.

Bibliography

Aalto D, et al.2014. Large scale data acquisition of simultaneous MRI and speech. J Appl Acoust. 83:64–75.

Aalto D, et al. 2013. Algorithmic Surface Extraction from MRI Data-Modelling the Human Vocal Tract. Proceeding of 6th International Joint Conference on Biomedical Engineering Systems and Technologies; Barcelona, Spain.

Aalto D, et al.2012. How far are vowel formants from computed vocal tract resonances? arXiv:1208.5963.

Alliez P, Cohen-Steiner D, Yvinec M, Desbrun M. 2005. Variational tetrahedral meshing. ACM Trans Graph. 24(3): 617–625.

Alvey C, Orphanidou C, Coleman J, McIntyre A, Golding S, Kochanski G. 2008. Image quality in non-gated versus gated reconstruction of tongue motion using magnetic resonance imaging: a comparison using automated image processing. Int J Comput Assist Radiol Surg. 3(5):457–464.

Anderson FC, Pandy MG. 2001. Static and dynamic optimization solutions for gait are practically equivalent. J Biomech. 34(2): 153–161.

Arnela M, Guasch O. 2014. Three-dimensional behavior in the numerical generation of vowels using tuned two-dimensional vocal tracts. Proceeding of 7th Forum Acousticum; Krakw, Poland.

Badin P, Bailly G, Reveret L, Baciu M, Segebarth C, Savariaux C. 2002. Three-dimensional linear articulatory modelling of tongue, lips and face, based on MRI and video images. J Phonetics. 30(3):533–553.

Baer T, Alfonso PJ, Honda K . 1988. Electromyography of the tongue muscles during vowels in /apup/ environment. Ann Bull RILP. 22:7-19.

Baghdadi L, Steinman DA, Ladak HM. 2005. Template-based finite-element mesh generation from medical images. Comput Meth Programs Biomed. 77(1):11–21

Bai Y, Han X, Prince JL. 2004. Super-resolution reconstruction of MR brain images. Proceedings of 38th Annual Conference on Information Sciences and Systems; Princeton, New Jersey, USA.

Benade AH, Jansson EV. 1974. On Plane and Spherical Waves in Horns with Nonuniform Flare I. Theory of Radiation, Resonance Frequencies, and Mode Conversion. ACTA ACUST UNITED AC. 31(2): 80–98.

Birkholz P, Jackl D, Krger BJ. 2007. Simulation of losses due to turbulence in the time-varying vocal system. IEEE Trans Audio Speech Language Process. 15(4): 1218–1226.

Birkholz P. 2013. Modeling Consonant-Vowel Coarticulation for Articulatory Speech Synthesis. PLoS ONE 8(4): e60603. doi:10.1371/journal.-pone.0060603

Blemker SS, Pinsky PM, Delp SL. 2005. A 3D model of muscle reveals the causes of nonuniform strains in the biceps brachii. J Biomech. 38(4):657-665.

Boersma P, Weenink D. 2005. Praat: doing phonetics by computer (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org.

Bresch E, Narayanan S. 2009. Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance images. IEEE Trans Med Imag. 28(3):323–338.

Bresson X, Vandergheynst P, Thiran JP. 2006. A variational model for object segmentation using boundary information and shape prior driven by the Mumford-Shah functional. Int J Comput Vis. 68(2):145–162.

Buchaillard S, Perrier P, Payan Y. 2009. A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning. J Acoust Soc Am. 126:2033–2051.

Bucki M, Lobos C, Payan Y. 2010. A fast and robust patient specific finite element mesh registration technique: application to 60 clinical cases. Med Image Anal. 13(3):303-317.

Bucki M, Lobos C, Payan Y, Hitschfeld N. 2011. Jacobian-based repair method for finite element meshes after registration. Eng Comput. 27(3):285–297.

Cachier P, Ayache N. 2001. Regularization in image non-rigid registration: I. Trade-off between smoothness and intensity similarity. Technical report, INRIA.

Chen M, Lu W, Chen Q, Ruchala KJ, Olivera GH. 2008. A simple fixed-point approach to invert a deformation field. Med Phys. 35(1): 81–88.

Chuanjun C, Zhiyuan Z, Shaopu G, Xinquan J, Zhihong Z. 2002. Speech after partial glossectomy: a comparison between reconstruction and non-reconstruction patients. J Oral Maxillofac Surg. 60(4):404–407.

Cootes TF, Taylor CJ, Cooper DH, Graham J. 1995. Active shape models: Their training and application. Comput Vis Image Understand. 61(1):38–59.

Couteau B, Payan Y, Lavallee S. 2000. The mesh-matching algorithm: an automatic 3D mesh generator for finite element structures. J Biomech. 33(8):1005–1009.

Cveticanin L. 2012. Review on mathematical and mechanical models of the vocal cord. J Appl Math. 2012: 1–18.

Dang J, Honda K. 2004. Construction and control of a physiological articulatory model. J Acoust Soc Am. 115:853-870.

Dart S. 1991. Articulatory and acoustic properties of apical and laminal articulations. UCLA Working Papers in Phonetics 79.

Doel K van den, Vogt F , English RE, Fels S. 2006. Towards Articulatory Speech Synthesis with a Dynamic 3D Finite Element Tongue Model. Proceeding of the 7th Intentional Seminar on Speech Production; Ubatuba, Brazil.

Doel K van den, Ascher UM. 2008. Real-time numerical solution of Webster's equation on a non-uniform grid. IEEE Trans Audio Speech Lang Processing. 16:1163–1172.

Drake R, Vogl AW, Mitchell AW. 2010. *Gray's anatomy for students.* Churchill Livingstone, Elsevier Inc., Philadelphia, PA, 2nd edition.

Dubuisson MP, Jain AK. 1994. A modified Hausdorff distance for object matching. Proceedings of the 12th IEEE International Conference on Pattern Recognition; Jerusalem, Israel.

Engwall O. 2003. Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. Speech Comm. 41(2):303–329.

Engwall O. 2003. A revisit to the Application of MRI to the Analysis of Speech Production-Testing our assumptions. Proceedins of 6th International Seminar on Speech Production; Sydney, Australia.

Erdemir A, McLean S, Herzog W, van den Bogert AJ. 2007. Model-based estimation of muscle forces exerted during movements. Clin Biomech. 22(2):131-154.

Eryildirim A, Berger MO. 2011. A guided approach for automatic segmentation and modeling of the vocal tract in MRI images. Proceedings of 19th European Signal Processing Conference; Barcelona, Spain.

Fang Q, Fujita S, Lu X, Dang J. 2009. A model-based investigation of activations of the tongue muscles in vowel production. Acoust Sci Tech. 30(4):277–287.

Fant G. 1960. Acoustic theory of speech production. The Hague, Netherlands: Mouton.

Flanagan JL, Landgraf LL. 1968. Self-oscillating source for vocal-tract synthesizers. IEEE Trans Audio Electroacoust. 16(1): 57–64.

Freedman D, Zhang T. 2005. Interactive graph cut based segmentation with shape priors. Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005; San Diego, USA.

Freitag L, Plassmann P. 2000. Local optimization-based simplicial mesh untangling and improvement. Int J Numer Meth. Eng 49(12):109125.

Foulonneau A, Charbonnier P, Heitz F. 2009. Multi-reference shape priors for active contours. Int J Comput Vis. 81(1):68–81.

Fowler CA, Saltzman E. 1993. Coordination and coarticulation in speech production. Lang speech. 36(2-3): 171–195.

Fujii N, et al.. 2011. Evaluation of swallowing using 320-detector-row multislice CT. Part I: single-and multiphase volume scanning for three-dimensional morphological and kinematic analysis. Dysphagia. 26(2): 99–107.

Gaige TA, Benner T, Wang R, Wedeen VJ, Gilbert RJ. 2007. Three dimensional myoarchitecture of the human tongue determined in vivo by diffusion tensor imaging with tractography. J Magn Reson Imaging. 26(3):654–661.

Gentil M, Gay T. 1986. Neuromuscular specialization of the mandibular motor system: speech versus non-speech movements. Speech Commun. 5(1):69–82.

George PL, Borouchaki H, Laug P. 2002. An efficient algorithm for 3D adaptive meshing. Adv Eng Softw. 33(7): 377–387.

Gerard JM, Wilhelms-Tricarico R, Perrier P, Payan Y. 2006. A 3D dynamical biomechanical tongue model to study speech motor control. Recent Res Develop Biomech. 1:49–64.

Gilles B, Pai D. 2008. Fast musculoskeletal registration based on shape matching. Proceedings of the 11th International Conference on Medical Image Computing and Computer Assisted Intervention. New York, USA.

Glupker L, Kula K, Parks E, Babler W, Stewart K, Ghoneima A. 2015. Three-dimensional computed tomography analysis of airway volume changes between open and closed jaw positions. Am J Orthod Dentofacial Orthop. 147(4): 426–434

Grady, L. 2006. Random walks for image segmentation. IEEE Trans Pattern Anal Mach Intell. 28(11):1768–1783.

Grosland NM, Bafna R, Magnotta VA. 2009. Automated hexahedral meshing of anatomic structures using deformable registration. Comput Method Biomech. 12(1): 35–43.

Hannam AG, Stavness I, Lloyd JE, Fels S. 2008. A dynamic model of jaw and hyoid biomechanics during chewing. J Biomech. 41(5):1069-1076.

Heimann T, Münzing S, Meinzer HP, Wolf I. 2007. A shape-guided deformable model with evolutionary algorithm initialization for 3D soft tissue segmentation. Proceedings of 20th International Conference on Information Processing in Medical Imaging. Kerkrade, The Netherlands.

Heimann T, Meinzer HP. 2009. Statistical shape models for 3D medical image segmentation: A review. Med Image Anal. 13(4):543–563.

Hillenbrand J, Getty LA, Clark MJ, Wheeler K. 1995. Acoustic characteristics of American English vowels. J Acoust Soc Am. 97(5): 3099-3111.

Hillenbrand J. 2003. American English: Southern Michigan. J Int Phon Assoc. 33: 121–126.

Ho AK, et al. 2014. 3D dynamic visualization of swallowing from multislice computed tomography. ACM SIGGRAPH 2014 Posters; Vancouver, Canada.

Hughes TJR. 2000. The finite element method: linear static and dynamic finite element analysis. Dover Publications.

Iltis PW, Frahm J, Voit D, Joseph AA, Schoonderwaldt E, Altenmüller E. 2015. High-speed real-time magnetic resonance imaging of fast tongue movements in elite horn players. Quant Imaging Med Surg. 5(3): 374–381.

Inamoto Y, et al.. 2015. Anatomy of the larynx and pharynx: effects of age, gender and height revealed by multidetector computed tomography. J Oral Rehabil.

Ishizaka K, Flanigan JL. 1972. Synthesis of voiced sounds from a two-mass model of the vocal cords. J Bell Syst Tech. 51: 1233–1268.

Jacewicz E, Fox R A. 2013. Cross-dialectal differences in dynamic formant patterns in American English vowels. In Vowel inherent spectral change. Springer Berlin Heidelberg; pp. 177–198.

B. Joe. 2008. Shape measures for quadrilaterals, pyramids, wedges, and hexahedra. Technical Report. Retrieved from http://members.shaw.ca/bjoe.

Kawakami S, et al.. 2012. Mechanomyographic activity in the human lateral pterygoid muscle during mandibular movement. J Neurosci Meth. 203(1):157–162.

Kelly KL, Lochbaum CC. 1962. Speech Synthesis. Proceeding of Fourth International Congress on Acoustics; Copenhagen, Denmark.

Kerwin W, Osman N, Prince J. 2000. Image processing and analysis in tagged cardiac MRI. In: Bankman I, editor. Handbook of Medical Imaging, chapter 24. Academic Press; p. 375–391.

Keyak JH, Meagher JM, Skinner HB, Mote CDJ. 1990. Automated three-dimensional finite element modelling of bone: a new method. J Biomed Eng. 12(5):389–397.

Kim YC, Hayes CE, Narayanan SS, Nayak KS. 2011. Novel 16channel receive coil array for accelerated upper airway MRI at 3 Tesla. J Magn Reson. 65(6):1711–1717.

Kim YC, Proctor MI, Narayanan SS, Nayak KS. 2011. Visualization of Vocal Tract Shape Using Interleaved Real-Time MRI of Multiple Scan Planes. Proceeding of 12th Annual Conference of the International Speech Communication Association; Florence, Italy.

Kitamura, et al. 2005. Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner. Acoust sci technol. 26(5):465–468.

Kivelä A. 2015. Acoustics of the vocal tract: MR image segmentation for modelling, Master's thesis, Aalto University School of Science.

Kivelä A, Kuortti J, Malinen J. 2013. Resonances and mode shapes of the human vocal tract during vowel production. Proceedings of 26th nordic seminar on computational mechanics; Oslo, Norway.

Knupp PM. 2000. Achieving finite element mesh quality via optimization of the Jacobian Matrix norm and associated quantities. Int J Numer Meth Eng. 48(8):1165–1185.

Ladefoged P. 2001. Vowels and consonants. Phonetica 58:211-212.

Larkman DJ, Nunes RG. 2007. Parallel magnetic resonance imaging. Phys Med Biol. 52(7):R15.

Lee J, Woo J, Xing F, Murano EZ, Stone M, Prince JL. 2013. Semi-automatic segmentation of the tongue for 3D motion analysis with dynamic MRI. Proceedings of the 10th IEEE International Symposium on Biomedical Imaging; San Francisco, USA.

Leventon ME, Grimson WEL, Faugeras O. 2000. Statistical shape influence in geodesic active contours. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2000; Hilton Head, USA.

Liu X, Abd-Elmoniem K, Stone M, Murano E, Zhuo J, Gullapalli R, Prince JL. 2012.Incompressible Deformation Estimation Algorithm (IDEA) from Tagged MR Images. IEEE Trans. Med. Imaging 31(2): 326-340.

Livesu M, Sheffer A, Vining N, Tarini M. 2015. Practical hex-mesh optimization via edge-cone rectification. ACM Trans Graphic. 34(4): 141.

Lloyd JE, Stavness I, Fels S. 2012. ARTISYNTH: A fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In: Payan Y, editor. Soft Tissue Biomechanical Modeling for Computer Assisted Surgery. Springer Berlin Heidelberg; p. 355–394.

Lobos C. 2012. A set of mixed-elements patterns for domain boundary approximation in hexahedral meshes. Stud Health Technol Inform. 184:268–272.

Lobos C, Bucki M, Hitschfeld N, Payan Y. 2007. Mixed-element mesh for an intra-operative modeling of the brain tumor extraction. Proceedings of 16th International Meshing Roundtable; Seattle, USA.

Luboz V, Chabanas M, Swider P, Payan Y. 2005. Orbital and maxillofacial computer aided surgery: patient-specific finite element models to predict surgical outcomes. Comput Methods Biomech Biomed Eng. 8(4):259–265.

Lukkari T, Malinen J. 2012. Webster's equation with curvature and dissipation. arXiv preprint arXiv:1204.4075.

Ma SYL, Whittle T, Descallar J, Murray GM, Darendeliler M A, Cistulli P, Dalci O. 2013. Association between resting jaw muscle electromyographic activity and mandibular advancement splint outcome in patients with obstructive sleep apnea. Am J Orthod Dentofacial Orthop. 144(3): 357–367.

Mac Neilage PF, Sholes GN. 1964. An electromyographic study of the tongue during vowel production. J SPEECH LANG HEAR R. 7(3): 209–232.

Martins JAC, Pires EB, Salvado R, Dinis PB. 1998. A numerical model of passive and active behavior of skeletal muscles. COMPUT METHOD APPL M. 151(3):419–433.

Mijailovich SM, Stojanovic B, Kojic M, Liang A, Wedeen VJ, Gilbert RJ. 2010. Derivation of a finite-element model of lingual deformation during swallowing from the mechanics of mesoscale myofiber tracts obtained by MRI. J Appl Phys. 109(5):1500–1514.

Miyawaki O, Hirose H, Ushijima T, Sawashima M. 1975. A preliminary report on the electromyographic study of the activity of lingual muscles. Ann Bull RILP. 9(91):406.

Montagnat J, Delingette H, Scapel N, Ayache N. 2000. Representation, shape, topology and evolution of deformable surfaces: Application to 3D medical image segmentation. Technical Report, INRIA.

Mory B, Somphone O, Prevost R, Ardon R. 2012. Real-Time 3d image segmentation by user-constrained template deformation. Proceedings of the 15th International Conference on Medical Image Computing and Computer Assisted Intervention; Nice, France.

Mu L, Sanders I. 2000. Neuromuscular specializations of the pharyngeal dilator muscles: II. Compartmentalization of the canine genioglossus muscle. Anat Rec. 260(3):308–325.

Muller M, Heidelberger B, Teschner M, Gross M. 2005 July. Meshless deformations based on shape matching. In ACM Trans Graph. 24(3):471–478.

Murano EZ, Shinagawa H, Zhuo J, Gullapalli RP, Ord RA, Prince JL, Stone M. 2010. Application of diffusion tensor imaging after glossectomy. Otolaryngol Head Neck Surg. 143(2):304–306.

Murray GM. 2012. The lateral pterygoid muscle: function and dysfunction. Semin Orthod. 18(1):44–50.

Narayanan S et al.2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). J Acoust Soc Am. 136(3): 1307–1311.

Nealen A, Mller M, Keiser R, Boxerman E, Carlson M. 2006. Physically based deformable models in computer graphics. In Comput Graph Forum. 25(4):809–836.

NessAiver MS, Stone M, Parthasarathy V, Kahana Y, Paritsky A. 2006. Recording high quality speech during tagged cineMRI studies using a fiber optic microphone. J Magn Reson. 23(1):92–97.

Ohala JJ. 1993. Coarticulation and phonology. Lang Speech. 36(2-3): 155–170.

O'Kusky JR, Norman MG. 1995. Sudden infant death syndrome: increased number of synapses in the hypoglossal nucleus. J Neuropath Exp Neur. 54: 627–634.

Osman NF, McVeigh ER, Prince JL. 2000. Imaging Heart Motion Using Harmonic Phase MRI. IEEE Trans Med Imaging. 19(3): 186–202.

Peled S, Yeshurun Y. 2001. Superresolution in MRI: application to human white matter fiber tract visualization by diffusion tensor imaging. Magn Reson Med. 45(1):29–35.

Peng T, Kerrien E, Berger MO. 2010. A shape-based framework to segmentation of tongue contours from MRI data. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing 2010; Dallas, USA.

Perrier P, Payan Y, Zandipour M, Perkell J. 2003. Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. J Acoust Soc Am. 114(3):1582–1599.

Peterson GE, Barney HL. 1952. Control methods used in a study of the vowels. J Acoust Soc Am. 24(2):175–184.

Plenge E, Poot DHJ, Bernsen M, Kotek G, Houston G, Wielopolski P, Weerd LVD, Niessen WJ, Meijering E. 2012. Super-resolution in MRI: better images faster?. Proceedings of SPIE Confetrence on Medical Imaging: Image Processing 2012. San Diego, USA.

Reichard R, Stone M, Woo J, Murano E, Prince J. 2012. Motion of apical and laminal /s/ in normal and post-glossectomy speakers. Acoustical Society of America, 3346.

Roberts TJ, Gabaldn AM. 2008. Interpreting muscle function from EMG: lessons learned from direct measurements of muscle force. Integr Comp Biol. 48(2): 312–320.

Saddi KA, Rousson M, Chefdhotel C, Cheriet F. 2007. Global-to-local shape matching for liver segmentation in CT imaging. Proceedings of the 10th International Conference on Medical Image Computing and Computer Assisted Intervention; Brisbane, Australia.

Sánchez CA, Stavness I, Lloyd JE, Fels S. 2013. Forward dynamics tracking simulation of coupled multibody and finite element models: Application to the tongue and jaw. Proceedings of the 11th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering; Salt Lake City, USA.

Sánchez CA, Lloyd JE, Fels S, Abolmaesumi P. 2014. Embedding digitized fibre fields in finite element models of muscles. Comput Methods Biomech Biomed Eng: Imaging Vis. 2(4):223–236.

Scherer RC, Titze IR, Curtis JF. 1983. Pressureflow relationships in two models of the larynx having rectangular glottal shapes. J Acoust Soc Am. 73(2): 668–676.

Sharp GC, Lee SW, Wehe DK. 2002. ICP registration using invariant features. IEEE Trans Pattern Anal 24(1):90–102.

Shepherd J. 2007. Topologic and Geometric Constraint-Based Hexahedral Mesh Generation. Doctoral Dissertation, University of Utah.

Sherif MH, Gregor RJ, Liu LM, Roy RR, Hager CL. 1983. Correlation of myoelectric activity and muscle force during selected cat treadmill locomotion. J Biomech. 16(9):691–701

Shimada Y, Nishimoto H, Kochiyama T, Fujimoto I, Mano H, Masaki S, Murase K. 2012. A technique to reduce motion artifact for externally triggered cine-MRI (EC-MRI) based on detecting the onset of the articulated word with spectral analysis. Magn Reson Med Sci. 11(4):273–282.

Sifakis E, Neverov I, Fedkiw R. 2005. Automatic determination of facial muscle activations from sparse motion capture marker data. ACM Trans Graph. 24(3)–417-425.

Sigal IA, Hardisty MR, Whyne CM. 2008. Mesh-morphing algorithms for specimen-specific finite element modeling. J Biomech. 41(7):1381–1389.

Silva MP, Ambrsio JA. 2004. Sensitivity of the results produced by the inverse dynamic analysis of a human stride to perturbed input data. Gait Posture. 19(1): 35–49.

Slaughter K, Li H, Sokoloff AJ. 2005. Neuromuscular Organization of the Superior Longitudinalis Muscle in the Human Tongue. Cells Tissues Organs. 181:51–64.

Sokoloff AJ, Deacon TW. 1992. Musculotopic organization of the hypoglossal nucleus in the cynomolgus monkey, Macaca fascicularis. J Comp Neurol 324: 81-93.

Sondhi MM, Schroeter J. 1987. A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans Acoust Speech Sinal Process. 35(7): 955–967.

Stavness I. 2010. Byte your tongue: A computational model of human mandibular-lingual biomechanics for biomedical applications. Doctoral Dissertation, University of British Columbia.

Stavness I, Lloyd JE, Payan Y, Fels S. 2011. Coupled hard-soft tissue simulation with contact and constraints applied to jaw-tongue-hyoid dynamics. Int J Numer Method Biomed Eng. 27(3):367–390.

Stavness I, Lloyd JE, Fels S. Automatic Prediction of Tongue Muscle Activations using a Finite Element Model. J Biomech. 45(16):2841–2848.

Stavness I, Nazari MA, Flynn C, Perrier P, Payan Y, Lloyd JE, Fels S. 2014a. Coupled Biomechanical Modeling of the Face, Jaw, Skull, Tongue, and Hyoid Bone. In: Magnenat-Thalmann N, Ratib O, Choi HF, editors. 3D Multiscale Physiological Human. Springer London. p. 253–274.

Stavness I et al.2014b. Unified Skinning of Rigid and Deformable Models for Anatomical Simulations. Proceeding of ACM SIGGRAPH Asia; Shenzhen, China.

Stone M, Epstein MA, IskarousK. 2004. Functional segments in tongue movement. Clin Linguist Phons. 18(6):507–521.

Stone M, Rizk S, Woo J, Murano EZ, Chen H, Prince JL. 2012. Frequency of apical and laminal /s/ in normal and post-glossectomy patients. J Med Speech Lang Pathol. 20(4):106–111.

Story BH, Titze IR. 1995. Voice simulation with a bodycover model of the vocal folds. J Acoust Soc Am. 97(2): 1249–1260.

Somphone O, Mory B, Makram-Ebeid S, Cohen L. 2008. Prior-Based Piecewise-Smooth Segmentation by Template Competitive Deformation Using Partitions of Unity. Proceedings of the 10th European Conference on Computer Vision; Marseille, France.

Sotiras A, Christos D, Paragios N. 2012. Deformable Medical Image Registration: A Survey. Technical Report, INRIA.

Takano S, Honda K. 2007. An MRI analysis of the extrinsic tongue muscles during vowel production. Speech Comm 49(1): 49–58.

Takemoto, H. 2001. Morphological analyses of the human tongue musculature for three-dimensional modeling. J Speech Lang Hear R 44(1):95–107.

Takemoto H, Mokhtari P, Kitamura T. 2014. Comparison of vocal tract transfer functions calculated using one-dimensional and three-dimensional acoustic simulation methods. Proceeding of 15th Annual Conference of the International Speech Communication Association; Singapore, Singapore.

Teo JCM, Chui CK, Wang ZL, Ong SH, Yan CH, Wang SC, Wong HK, Teoh SH. 2007. Heterogeneous meshing and biomechanical modeling of human spine. Med Eng Phys. 29(2): 277–290.

Titze IR. 1988. The physics of smallamplitude oscillation of the vocal folds. J Acoust Soc Am. 83(4): 1536–1552.

Top A, Hamarneh G, Abugharbieh R. 2011. Active learning for interactive 3d image segmentation. Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention; Toronto, Canada.

Tsai A, Yezzi AJ, Wells W, Tempany C, Tucker D, Fan A, Grimson WE, Willsky A. 2003. A shape-based approach to the segmentation of medical imagery using level sets. IEEE Trans Med Imag. 22(2):137–154.

Tuller B, Harris KS, Gross B. 1981. Electromyographic study of the jaw muscles during speech. J Phon. 9: 175–188.

Uecker M, Zhang S, Voit D, Karaus A, Merboldt KD, Frahm J. 2010. Real-time MRI at a resolution of 20 ms. NMR Biomed. 23(8): 986–994.

Välimäki V, Karjalainen M. 1994. Improving the kelly-lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques. Proceeding of the 3rd International Conference on Spoken Language Processing; Yokohama, Japan.

Vasconcelos MJ, Ventura SM, Freitas DR, Tavares, JMR. 2012. Inter-speaker speech variability assessment using statistical deformable models from 3.0 Tesla magnetic resonance images. P I MECH ENG G-J AER. 226(3): 185–196.

Ventura SR, Freitas DR, Tavares JMR. 2009. Application of MRI and biomedical engineering in speech production study. Comput Methods Biomech Biomed Eng. 12(6): 671–681.

Ventura SR, Freitas DR, Tavares, JMR. 2011. Toward dynamic magnetic resonance imaging of the vocal tract during speech production. J Voice. 25(4):511–518.

Ventura SR, Freitas DR, Ramos IMA, Tavares, JMR. 2013. Morphologic differences in the vocal tract resonance cavities of voice professionals: an MRI-based study. J Voice. 27(2):132–140.

Vercauteren T, Pennec X, Perchant A, Ayache N. 2009. Diffeomorphic demons: Efficient non-parametric image registration. Neuroimage. 45(1): 6-1–72.

Wand M. 2015. Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Modeling. KIT Scientific Publishing.

Wolánski W, Gzik-Zroska B, Kawlewska E, Gzik M, Larysz D, Dzielicki J, Rudnik A. 2015. Preoperative Planning of Surgical Treatment with the Use of 3D Visualization and Finite Element Method. In Developments in Medical Image Processing and Computational Vision. Springer International Publishing; pp. 139–163.

Woo J, Murano EZ, Stone M, Prince JL. 2012. Reconstruction of High Resolution Tongue Volumes from MRI. IEEE Trans Biomed Eng. 6(1): 1–25.

Woo J, Lee J, Murano, EZ, Xing F, Al-Talib M, Stone M, Prince JL. 2015. A high-resolution atlas and statistical model of the vocal tract from structural MRI. Comput Methods Biomech Biomed Eng: Imaging Vis. 3(1):47–60.

Xing F, Ye C, Woo J, Stone M, Prince J. 2015. Relating speech production to tongue muscle compressions using tagged and high-resolution magnetic resonance imaging. Proceeding of SPIE Medical Imaging; Munich, Germany.

Xing F, Woo J, Murano EZ, Lee J, Stone M, Pronce JL. 2013. 3d Tongue Motion from Tagged and Cine MR Images. Proceeding of the 16th International Conference on Medical Image Computing and Computer-Assisted Intervention; Nagoya, Japan.

Yoshida K, Takada K, Adachi S, Sakuda M. 1982. Clinical Science EMG Approach to Assessing Tongue Activity Using Miniature Surface Electrodes. J Dent Res. 61(10): 1148–1152.

Yushkevich PA, Piven J, Hazlett HC, Smith RG, Ho S, Gee JC, Gerig, G. 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage, 31(3): 1116–1128.

Zajac FE. 1988. Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control. Crit Rev Biomed Eng. 17(4): 359–411.

Zhang Y, Bajaj C, Xu G. 2005. Surface smoothing and quality improvement of quadrilateral/hexahedral meshes with geometric flow. Proceedings of the 14th International Meshing Roundtable; San Diego, USA.

Appendix A

Oropharyngeal Muscles

A.1 Tongue Muscles

The human tongue is unique in the body. The structure that resembles it most closely is the heart, since both are composed entirely of soft-tissue; but, the tongue typically moves at 10 times the rate of the heart. Tongue muscle architecture is considerably more complex than heart muscle, which allows the tongue to be extremely versatile: chewing requires that it throws food onto the teeth; some languages utilize it in the pronunciation of over 150 phonetic sounds; and every time we inhale the genioglossus posterior contracts in order to keep the pharynx open.

The tongue is a perfect muscular hydrostat, meaning it has no skeleton and no sack. It is composed entirely of soft tissue. There are two definitive features associated with muscular hydrostats: they have orthogonal muscle orientation, and they are volume preserving. Compression in one location means expansion in another, and motion is accompanied by deformation.

The tongue consists of 100 laminae (alternating layers of orthogonal fibers), enabling complex deformations. The hypoglossal nerve (CN XII) supplies motor fibres to all the muscles of the tongue – except the palatoglossus muscle, which is innervated by the vagus nerve (CN X). These nerves arise from the hypoglossal nucleus (of the caudal brain stem), which includes about 13,000 motoneurons (O'Kusky JR and Norman, 1995). It is possible that every one of these has independent control, potentially enabling very complex deformations. The complex muscle fiber direction is compatible with localized innervation of muscle fibers.The implication lies in localized control, not control of muscle compartments.

Previously, extrinsic muscles of the tongue were thought to move the tongue

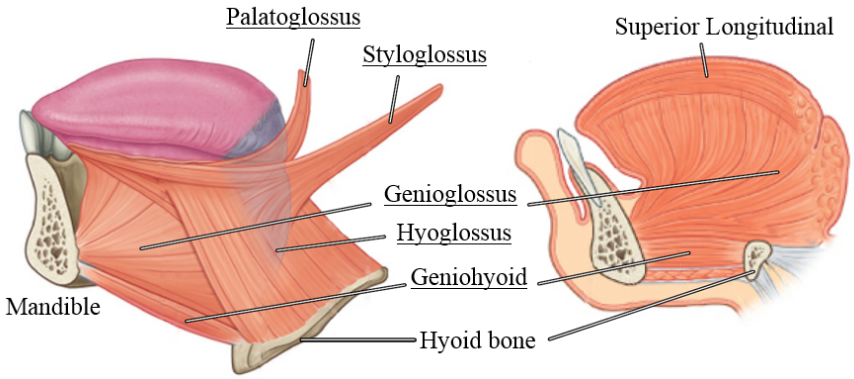


Figure A.1. Lateral and sagittal cross-section views of the tongue, denoting extrinsic muscles (underlined). ©Elsevier, Drake et al. (2010), adapted with permission.

as a rigid body, while intrinsic muscles would fine tune tongue shape into a minimal deformation. Today we know that both extrinsic and intrinsic muscles participate in the movement and deformation of the tongue.

A.1.1 Extrinsic Muscles

The extrinsic muscles of the tongue originate on the bones outside the tongue, and insert into the tongue surface. Extrinsic muscles, as shown in Figure A.1, are named for origin and insertion: genioglossus (GG), hyoglossus (HG), styloglossus (STY) and palatoglossus (PG). The suffix *glossus* refers to the tongue. Fibers are located very laterally, back to front (SG, HG, PG), or very medially, front to back (GG), and mostly interdigitate with the intrinsic muscles.

The genioglossus (GG) originates from the mandibular symphysis, and inserts into the mid-line tongue surface, excluding the tip, in a fan shape. Here, the prefix *genio* refers to the greek word for *chin*. The genioglossus anterior (GGA) depresses the anterior tongue; and the genioglossus posterior (GGP) pulls the posterior of the tongue forward. GG is the major muscle responsible for protruding the tongue.

The hyoglossus (HG) originates from the hyoid bone (hence the prefix *hyo*), and inserts into the posterior tongue, interdigitating with the transverse, and

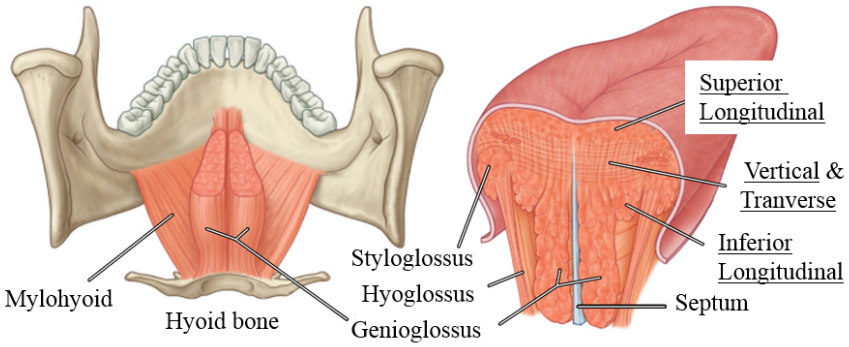


Figure A.2. Posterior views of tongue muscles with horizontal and vertical cut-away, denoting intrinsic tongue muscles (underlined). ©Elsevier, Drake et al. (2010), adapted with permission.

along the lateral margin to the tip, inserting along its length. It retracts and depresses the tongue.

The styloglossus (STY) arises from the styloid process of temporal bone (hence the prefix *stylo*), and like the HG, inserts into the posterior tongue, interdigitating with the transverse muscle, and along the lateral margin to the tip, inserting along its length. It retracts and elevates the tongue, pulling it up and back.

The palatoglossus (PG) originates from the oral surface of the soft palate (hence the prefix *palato*), and spreads into the lateral tongue, where some of its fibers interdigitate with the transverse muscle. The PG elevates the posterior tongue, during swallowing, in order to close the oropharyngeal isthmus.

A.1.2 Intrinsic Muscles

The intrinsic muscles of the tongue originate and insert within the soft tissue of the tongue. Fibers course lengthwise or crosswise, and interdigitate mostly with other muscles. Intrinsic muscles are named for direction of their fibers: superior longitudinal(is) (SL), inferior longitudinal(is) (IL), vertical(is) (VERT) and transvers(e/us) (TRANS). Figure A.2 shows the intrinsic muscles in mid-sagittal and mid-coronal cross-cuts of the tongue.

The superior longitudinal (SL) originates at the tongue tip, and inserts into the posterior surface, above the hyoid. Its activation shortens the tongue while elevating and curling the tip.

The inferior longitudinal (IL) originates from the tongue blade, and inserts into the posterior surface, above the hyoid and below the SL. Its activation shortens the tongue while depressing and curling the tip.

The verticalis (VERT) arises at the upper surface of tongue, and inserts into the upper surface of the IL and ventral surface of the tongue. Its activation flattens and widens the tongue; it also protrudes the tongue if activated alongside with the TRANS.

The transversus (TRANS) originates from the median septum, and inserts into the upper lateral surface of the tongue. It narrows and lengthens the tongue.

A.2 Jaw and Hyoid Muscles

The human jaw (the bone structure at the entrance of the mouth) is articulated by the motion of its lower section (mandible), while its upper section (maxilla) is mostly fixed into the skull. The hyoid bone is a small, u-shaped bone distantly anchored to the skull; it supports tongue motion by providing attachments to the HG as well as mouth floor muscles. The muscles of the jaw and hyoid insert into (or originate from) the mandible. Figure A.3 identifies the attachment sites of each muscle on the mandible.

The jaw muscles (the masseter, temporalis, medial pterygoid and lateral pterygoid) are shown in Figure A.4. They are all inverted by the mandibular division (V3) of the fifth cranial nerve. The lateral pterygoid is the only muscle from the four to open the jaw, while the bilateral activation of the others results in jaw closing.

A.2.1 Jaw Closers

The medial pterygoid (MP) is a thick quadrilateral, originating from above the medial surface of the lateral pterygoid plate (deep head), as well as

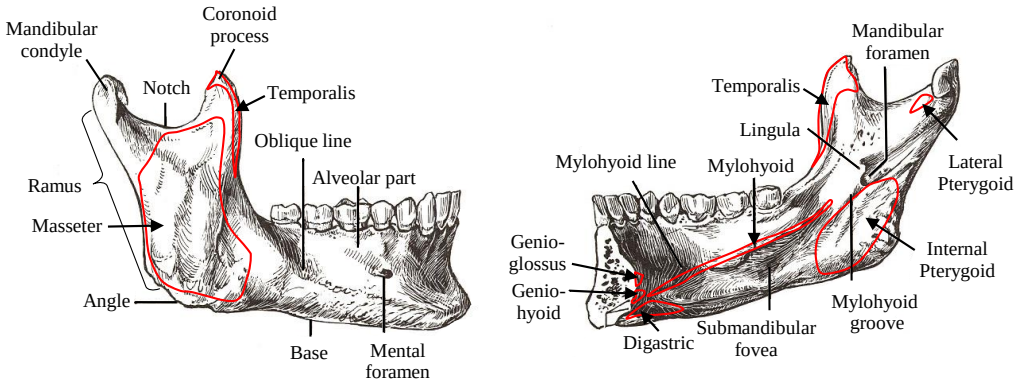


Figure A.3. The insertion sites of jaw, and hyoid muscles, on the mandible. Public Domain. Adapted from Health, Medicine and Anatomy Reference Pictures, 2013.

the maxillary tuberosity, and the pyramidal process of the palatine bone (superficial head). Passing downward, lateral and posterior, the fibers insert into the internal surface of the ramus, and down to the angle of the mandible. When activated, the MP muscle elevates the mandible, and closes the jaw.

The masseter is a strong mastication muscle that parallels the medial pterygoid. The superficial head of the masseter (SM) originates from the zygomatic process of the maxilla, and the zygomatic arch; it inserts into the angle and ramus of the mandible. The deep head of the masseter (DM) arises from the lower border and medial surface of the zygomatic arch. Its fibers pass downward and forward to insert into the upper half of the ramus. At its insertion, the masseter joins the MP to form a common sling, allowing for powerful jaw elevation. The temporalis closes the jaw and pulls the mandible back. It arises from the temporal fossa and the deep part of temporal fascia, and inserts within the coronoid process of the mandible. If the entire muscle contracts, the main action is to elevate the mandible and raise the lower jaw; however, the contraction of the posterior, by itself, retracts the mandible.

A.2.2 Jaw Openers

The lateral pterygoid is the only jaw muscle that protrudes the mandible, hence opens the jaw. Its superior head (SP) arises from the sphenoid bone, and inserts onto the capsule of the temporomandibular joint; its inferior head

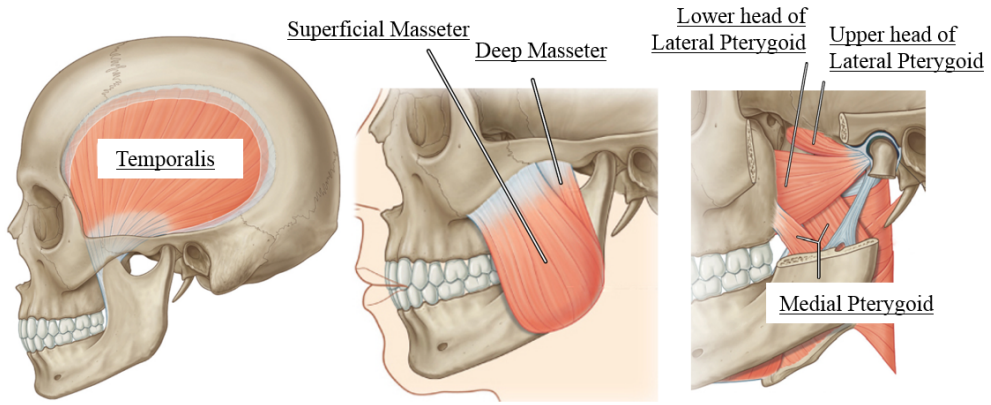


Figure A.4. Illustration of the jaw muscles (underlined) inserting into the mandible ©Elsevier, Drake et al. (2010), adapted with permission.

(IP) arises from the lateral pterygoid plate, and inserts onto the condyloid process.

Other jaw openers include the muscles of the mouth floor, as illustrated in Figure A.5. The geniohyoid (GH) originates from the back of the mandibular symphysis and inserts on the anterior surface of the hyoid body. It depresses the mandible, and opens the jaw (if the hyoid is kept in position), or it pulls the tongue body forward and up by elevating the hyoid (if the hyoid is not stabilized).

The mylohyoid (MH) is a flat, triangular muscle arising from the mandibular and inserting to the hyoid. Here, the prefix *mylo* refers to the Greek word for *molar*. The MH forms the muscular floor of the oral cavity and has similar action to the GH.

Figure A.5 also shows some of the suprahyoid muscles. The digastric is a narrow muscle that includes two bellies. The anterior belly of the digastric (AD) arises from the lower border of the mandible, and closes to the symphysis. It opens the jaw and pulls the tongue body forward and up, if the hyoid bone is not stabilized. The posterior belly of the digastric (PD) originates at one end from the mastoid process of the temporal bone; at the other end, it forms a tendon that attaches to the hyoid. The digastric muscle shows action similar to the GH.

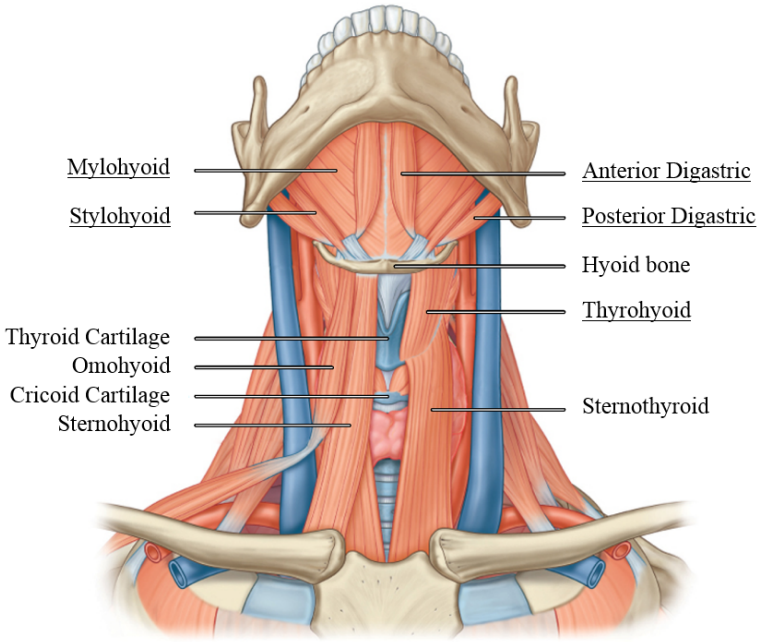


Figure A.5. Frontal view of the submandibular and neck muscles, denoting muscles of the mouth floor (underlined). ©Elsevier, Drake et al. (2010), adapted with permission.

THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC)

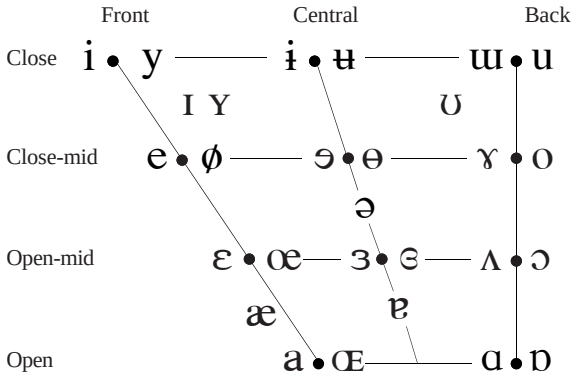
	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

CONSONANTS (NON-PULMONIC)

Clicks		Voiced implosives		Ejectives	
ʘ	Bilabial	ɓ	Bilabial	’	Examples:
|	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
!	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
‖	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

VOWELS



Where symbols appear in pairs, the one to the right represents a rounded vowel.

OTHER SYMBOLS

ʍ	Voiceless labial-velar fricative	ɕ ʑ	Alveolo-palatal fricatives
w	Voiced labial-velar approximant	ɺ	Voiced alveolar lateral flap
ɥ	Voiced labial-palatal approximant	ʢ	Simultaneous ʃ and x
ʜ	Voiceless epiglottal fricative		
ʡ	Voiced epiglottal fricative		
ʡ	Epiglottal plosive		

Affricates and double articulations can be represented by two symbols joined by a tie bar if necessary. k͡p t͡s

SUPRASEGMENTALS

ˈ	Primary stress
ˌ	Secondary stress ˌfoʊnəˈtɪʃən
ː	Long eː
ˑ	Half-long eˑ
˘	Extra-short ĕ
|	Minor (foot) group
‖	Major (intonation) group
.	Syllable break ɹi.ækt
‿	Linking (absence of a break)

TONES AND WORD ACCENTS

LEVEL		CONTOUR	
e̋ or ˥	Extra high	ě or ˩˥	Rising
é ˦	High	ê ˥˩	Falling
ē ˧	Mid	e᷄ ˦˥	High rising
è ˨	Low	e᷅ ˩˧	Low rising
ȅ ˩	Extra low	e᷈ ˥˦˥	Rising-falling
↓	Downstep	↗	Global rise
↑	Upstep	↘	Global fall

DIACRITICS Diacritics may be placed above a symbol with a descender, e.g. n̥ŋ̊

◌̥	Voiceless	n̥ d̥	◌̤	Breathy voiced	b̤ a̤
◌̬	Voiced	s̬ t̬	◌̰	Creaky voiced	b̰ a̰
◌ʰ	Aspirated	tʰ dʰ	◌̼	Linguolabial	t̼ d̼
◌̹	More rounded	ɔ̹	◌ʷ	Labialized	tʷ dʷ
◌̜	Less rounded	ɔ̜	◌ʲ	Palatalized	tʲ dʲ
◌̟	Advanced	u̟	◌ˠ	Velarized	tˠ dˠ
◌̠	Retracted	e̠	◌ˤ	Pharyngealized	tˤ dˤ
◌̈	Centralized	ë	◌̴	Velarized or pharyngealized	ɫ
◌̽	Mid-centralized	e̽	◌̝	Raised	e̝ (ɹ̝ = voiced alveolar fricative)
◌̩	Syllabic	n̩	◌̞	Lowered	e̞ (β̞ = voiced bilabial approximant)
◌̯	Non-syllabic	e̯	◌̪	Advanced Tongue Root	e̪
◌˞	Rhoticity	ɚ a˞	◌̺	Retracted Tongue Root	e̺
◌̃	Nasalized	ẽ	◌̻	Apical	t̻ d̻
◌ⁿ	Nasal release	dⁿ	◌̺	Laminal	t̺ d̺
◌ˡ	Lateral release	dˡ	◌̚	No audible release	d̚