Computational Modeling of the Vocal Tract

Applications to Speech Production

SAEED DABBAGHCHIAN

Doctoral Thesis
Stockholm, Sweden 2018
Akademisk avhandling som med tillstånd av Kungl Tekniska högskolan framlägges
till offentlig granskning för avläggande av teknologie doktorsexamen i tal- och
musikkommunikation med inriktning mot talkommunikation fredagen den 7 december
2018 klockan 14.00 i D2, Kungl Tekniska högskolan, Långstedsvägen 5, Stockholm.

© Saeed Dabbaghchian, December 2018
In the name of Allah, the Most Gracious, the Most Merciful.

To those who devoted themselves to understand *Speech*.
# Contents

Abstract ix
Acknowledgements xi
Abbreviations xii

## 1 Introduction

1.1 Representing the vocal tract ................................. 1
1.2 The effects of vocal tract modeling dimensionality ....... 3
1.3 Biomechanical simulation of speech production .......... 4
1.4 Muscle activation pattern estimation .......................... 4
1.5 Vocal tract reconstruction from a biomechanical model .... 4
1.6 Coupled 3D biomechanical–acoustic simulations .......... 5
1.7 Applications .................................................. 6
1.8 Contributions .................................................. 7
1.9 Summary of the papers ....................................... 8
1.10 Relevant publications not included in the thesis ......... 10

## 2 Computational modeling

2.1 Scientific method cycle ....................................... 14
2.2 Applications .................................................. 14
2.3 Model choice ................................................... 16

## 3 Physiology of speech production

3.1 The anatomical directions and planes ....................... 19
3.2 The upper airway ............................................. 20
3.3 The nervous system .......................................... 24

## 4 Speech production models

4.1 Geometrical models .......................................... 28
4.2 Articulatory models ......................................... 31
4.3 Biomechanical models ....................................... 33
4.4 Speech production modeling in this thesis .................. 36
## Contents

5 Effects of geometry simplifications .................................................. 37
  5.1 The lips ................................................................. 37
  5.2 The vallecula and the piriform fossa ........................................ 39
  5.3 Slicing ................................................................. 39
  5.4 Shape of the cross-section .................................................. 39
  5.5 Bending ............................................................... 39
  5.6 Applications ......................................................... 40
  5.7 Implications ......................................................... 40

6 The biomechanical model ............................................................... 41
  6.1 Geometry of the anatomical structures ...................................... 41
  6.2 Muscle model ......................................................... 44
  6.3 Tongue muscles ....................................................... 45

7 Estimation of muscle activation patterns ........................................ 49
  7.1 Estimation based on articulation ........................................... 50
  7.2 Estimation based on acoustic ............................................. 53

8 Vocal tract reconstruction ............................................................. 57
  8.1 Boundary detection ..................................................... 59
  8.2 Surface reconstruction .................................................. 62

9 Acoustic modeling ........................................................................... 65
  9.1 The scalar wave equation .................................................... 65
  9.2 Three-dimensional wave equation .......................................... 68
  9.3 Moving boundaries ....................................................... 72

10 Sound synthesis ............................................................................. 75
  10.1 Vowel sounds ............................................................. 76
  10.2 Vowel-vowel sounds ....................................................... 77
  10.3 Influence of mechanical properties ...................................... 78

11 On the quantal theory ..................................................................... 81
  11.1 Quantal relationship ........................................................ 82
  11.2 Biomechanical constraints .................................................. 84
  11.3 Discussion .................................................................. 86

12 Conclusions and future directions .................................................. 87

Bibliography ......................................................................................... 91

Paper A Effects of lips geometry .......................................................... 107

Paper B Effects of geometry simplifications ........................................ 123
Note: The papers have not been included in the electronic version of the thesis.
Abstract

Human speech production is a complex process, involving neuromuscular control signals, the effects of articulators’ biomechanical properties and acoustic wave propagation in a vocal tract tube of intricate shape. Modeling these phenomena may play an important role in advancing our understanding of the involved mechanisms, and may also have future medical applications, e.g., guiding doctors in diagnosing, treatment planning, and surgery prediction of related disorders, ranging from oral cancer, cleft palate, obstructive sleep apnea, dysphagia, etc.

A more complete understanding requires models that are as truthful representations as possible of the phenomena. Due to the complexity of such modeling, simplifications have nevertheless been used extensively in speech production research: phonetic descriptors (such as the position and degree of the most constricted part of the vocal tract) are used as control signals, the articulators are represented as two-dimensional geometrical models, the vocal tract is considered as a smooth tube and plane wave propagation is assumed, etc.

This thesis aims at firstly investigating the consequences of such simplifications, and secondly at contributing to establishing unified modeling of the speech production process, by connecting three-dimensional biomechanical modeling of the upper airway with three-dimensional acoustic simulations.

The investigation on simplifying assumptions demonstrated the influence of vocal tract geometry features – such as shape representation, bending and lip shape – on its acoustic characteristics, and that the type of modeling – geometrical or biomechanical – affects the spatial trajectories of the articulators, as well as the transition of formant frequencies in the spectrogram.

The unification of biomechanical and acoustic modeling in three-dimensions allows to realistically control the acoustic output of dynamic sounds, such as vowel-vowel utterances, by contraction of relevant muscles. This moves and shapes the speech articulators that in turn define the vocal tract tube in which the wave propagation occurs. The main contribution of the thesis in this line of work is a novel and complex method that automatically reconstructs the shape of the vocal tract from the biomechanical model. This step is essential to link biomechanical and acoustic simulations, since the vocal tract, which anatomically is a cavity enclosed by different structures, is only implicitly defined in a biomechanical model constituted of several distinct articulators.

Keywords
vocal tract, upper airway, speech production, biomechanical model, acoustic model, vocal tract reconstruction
Sammanfattning

Människors talproduktion är en komplex process som innefattar neuromuskulära kontrollsignaler, effekter av talorganens biomekaniska egenskaper och den akustiska vägens utbredning i ett talrör med invecklad form. Genom att modellera dessa fenomen kan vår förståelse av de olika ingående mekanismerna förbättras. Det kan även bidra till framtida medicinska tillämpningar, som att vägleda läkare vid diagnosticering, behandlingsplanering och förutsägelser av effekterna av en operation, när det gäller relaterade medicinska problem, såsom munhålecancer, gomspalt, sömnapné, dysfagi etc.

För att förståelsen ska vara så fullständig som möjligt krävs att de ingående modellerna så nära som möjligt representerar de verkliga förhållandena. På grund av komplexiteten vid en sådan modellering, har förenklingar varit mycket vanliga inom talproduktionsforskning: fonetiska beskrivningar (såsom läget och storleken på talrörets minsta passage) används för styrning av modellen, talorganen beskrivs av tvådimensionella modeller, talröret anses vara en slät tub, plan vågutbredning antas, etc.

Den här avhandlingen syftar till att för det första undersöka vilka effekter sådana förenklingar har, och för det andra att möjliggöra en enhetlig modellering av hela talproduktionskedjan, genom att koppla samman tredimensionell biomekanisk modellering av talorganen med tredimensionell akustisk simuleringar.

Undersökningen av förenklingar visar att talrörets geometriska egenskaper – exempelvis hur formen representeras, om det modelleras som böjt eller rakt, och läpparnas form – påverkar den akustiska signalen, och att typen av modell – geometrisk eller biomekanisk – inverkar på såväl hur talorganen i modellen rör sig, som på som formanttransitionerna i spektrogram.

Sammanfogningen av tredimensionell biomekanisk och akustisk modellering möjliggör att på ett realistiskt sätt kontrollera dynamiska akustiska signaler, som exempelvis vokal-vokal-sekvenser, genom att aktivera de involverade musklerna. Detta förflytta och formar talorganen, vilka i sin tur definierar talrör, i vilket den akustiska vägen breder ut sig. Avhandlingens huvudsakliga bidrag på detta område är en ny och komplex metod som automatiskt återskapar talrören utifrån den biomekaniska modellen. Detta steg är nödvändigt för att koppla samman de biomekaniska och akustiska modellerna, eftersom talrören, som anatomiöst är tommrummet mellan olika strukturer, enbart är implicit definierat i en biomekanisk modell som består av flera olika delar.

Nyckelord
talrör, talorganen, talproduktion, biomekanisk modellering, akustisk modellering, talrörkonstruktion
Acknowledgments

This long journey towards getting my PhD was started 5 years ago, during Nowruz (Iranian new year’s day) holidays, when I was spending my vacation days in my parents’ house, and I did apply for a PhD position at the Department of Speech, Music and Hearing. Well, the application turned out to be successful, and I embarked on this journey. This journey would not come to an end without support of my supervisors, colleagues, ex-colleagues, friends, and family.

First of all, I would like to express my gratitude to Olov Engwall, for being an excellent mentor, and giving me an absolute freedom on this expedition. I also appreciate you being patient about my repeated phrase: “the paper is almost ready.”, and then waiting for another month before getting it ready to submit! Sten Ternström, for your generous attitude, valuable feedback and instructive discussions, proof-reading of my papers and the thesis, and certainly for running the phonetic course.

I also would like to give my thanks to all of my colleagues at TMH, Johan Sundberg, for answering my questions, detailed discussions, and providing cakes in vocal cafe! I have been inspired by your hard work even sometimes at weekends. Anders Friberg for our conversations about scientific and non-scientific issues while having ice cream together during the hot days of the summer! Andreas Selamitzis for having listening ears and conversations about almost everything from inter-religious matters to infinite element methods! Kalin Stefanov for being my companion on weekends so I would not feel alone at work! Bajibabu Bollepalli for cycling with me from the main campus to Lappis in dark and cold days of the autumn. Giampiero Salesi, for all lunch table conversations about publication practices, the past and future of AI and machine learning. David House for your course on “Scientific Writing” and your ever smiling face. I never have seen you otherwise! Johan Boye, for our conversations in front of the coffee machine! Martin Johansson, for helping me understand Swedish culture, and resolve daily life issues. Patrik Jonell for being my roommate, though for a short time. Mattias Bystedt for your help and encouragement in learning Swedish. My thanks also goes to Joakim Gustafsson, Jonas Beskow, Anders Elousson, Jens Edlund, Gabriel Skantze, Peter Nordqvist, Björn Granström, Rolf Carlson, André Pereira, Nils Axelsson, Erik Ekstedt, Per Fullgren, Gustav Henter, Dimosthenis Kontogiorgos, Zofia Malisz, Bo Schenkman, Bob Sturm, Eva Szekely, and our previous colleagues at TMH: José David Lopes,
Samer Al Moubayed, Sofia Strömbergsson, Simon Alexanderson, Anders Askenfelt, Jana Götze, Raveesh Meena. You all have been awesome!

During my studies, I worked with many wonderful people around the world, and I take the opportunity to thank them. Marc Arnela, and Oriol Guash, from The Universitat Ramon Llull in Barcelona, for your help to understand the acoustics, conducting acoustic simulations, and reviewing Chapter 9 of the thesis. The ArtiSynth team at the University of British Columbia, in particular John E. Lloyd and Ian Stavness for their technical help. Frank Guenther, from Boston University, for hosting me as a research visitor, Joseph Perkell for helping me to understand motor control theories, and members of the lab, Ayoub Daliri, Alfonso Nieta-Castanon, Andrés F. Salazar-Gómez, Barbara Holland, and Jason Tourville. Björn Lindblom, from Stockholm University, for your valuable feedback, and helping me to understand phonetic theories in depth. Pétur Helgason from Stockholm university for in-depth phonetic discussions. Ryan Keith Shosted for hosting me as research visitor at University of Illinois. Marissa Barlaz for working together on real-time MRI data. I would like to thank a number of colleagues for sharing data and code: Pierre Badin at Gipsa-lab for EMA data; Brad Story for the tube-talker synthesizer and Marc Tiede for EMG data.

I also would like to acknowledge all my previous mentors and colleagues, to name a few, Behbood Mashoufi, Ali Aghagolzadeh, Hossein Sameti, Bagher BabaAli, and all my friends.

Finally, my heartfelt appreciation goes out to my parents, dear mother and father, I kiss your hands, and I acknowledge that I would not reach so far without your support and your prayers for me! My wife, Masoumeh, I have been fortunate to have you as my companion through the whole journey. I cannot thank you enough nor express with words how grateful I truly am for all the support and care that you provided to me and Taha during these years. My son, Taha, for having joyful times playing together and forgetting about all my deadlines, bugs in the code, etc. My twin sister, Maryam, and my little sister, Monireh, whom I have missed a lot! My father-in-law who is no longer with us and my mother-in-law for their support. I am truly blessed to have such a great family!
Abbreviations

1D one-dimensional
2D two-dimensional
3D three-dimensional
CF closing and filling
CNS central nerve system
CT Computed Tomography
DIVA Directions Into Velocities of Articulators
EMA Electromagnetic articulography
EMG Electromyography
FEM Finite Element Method
GB geometrical boolean
GC growing circle
GG genioglossus
GGA genioglossus anterior
GGM genioglossus middle
GGP genioglossus posterior
GH geniohyoid
HFE high frequency energy
HG hyoglossus
IL  inferior longitudinal
LI  lower incisor
LoS  line of sight
MAP  muscle activation patterns
MH  mylohyoid
MRI  Magnetic resonance imaging
OOS  orbicularis oris superior
PG  palatoglossus
PNS  peripheral nerve system
rtMRI  real time magnetic resonance imaging
SG  styloglossus
SL  superior longitudinal
T  transversus
TB  tongue back
TM  tongue middle
TT  tongue tip
V  verticalis
Chapter 1

Introduction

“To ask the value of speech is like asking the value of life.”
— Alexander Graham Bell

This thesis is all about “speech”, an encrypted “sound”, that is produced and perceived by human beings, giving them the ability to communicate in the most natural way. Although speech production and perception are two pieces of the same puzzle, the effort in this thesis is devoted to the production side. We know that speech production requires coordinated function of three different systems, namely the nervous, muscular and respiratory. The nervous system conveys our thoughts to muscles that, in turn, move the articulators. The lungs generate the airflow that passes through the vocal folds, and is then modulated by the vocal tract, and radiated through the lips. However, this high level of explanation does not reveal details about the involved mechanisms. In order to understand the properties of “speech”, we need to create it. That is, to develop computational models that simulate the involved mechanisms. Towards this end, a three-dimensional (3D) biomechanical-acoustic model of the speech apparatus has been developed, and its application in speech synthesis and motor control studies is presented. This chapter elaborates the goals of the work, the utilized approaches, contributions, and the structure of the thesis.

1.1 Representing the vocal tract

The acoustic characteristics of speech is mainly formed by the shape of a cavity, namely the vocal tract. The vocal tract geometry may be acquired using medical imaging techniques such as Magnetic resonance imaging (MRI) or Computed Tomography (CT), and used for acoustic simulations. Such an approach is illustrated by Figure 1.1. Although acquiring the vocal tract geometry through direct imaging is absolutely essential for basic knowledge of anatomy and articulation, one may also obtain such a geometry by employing a computational articulatory/biomechanical
model. This has the advantage that alternative anatomical or articulatory configurations may be tested without acquiring new medical data. Furthermore, it may overcome some of the drawbacks of medical imaging. Acquiring medical data is time consuming, expensive, and there is a trade-off between temporal and spatial resolution. The latter causes difficulties in reconstructing the vocal tract geometry when the vocal tract motion needs to be captured, since the state-of-the-art real-time magnetic resonance imaging (rtMRI) does not provide enough spatio-temporal resolution. In a static configuration of the vocal tract, on the other hand, the image contrast at the air-tissue boundary may create problems for image segmentation algorithms, especially when there is a narrow constriction. Some structures such as the teeth may moreover not be captured by MRI. All these issues result in a significant amount of manual work to reconstruct the vocal tract geometry from medical data.

On the other hand, if using a computational model, it is possible to accurately simulate the motion of the structures and hence determine the vocal tract shape at each time step. The temporal resolution is then limited only by the time step of the simulation, the air-tissue contrast is high, and the reconstruction algorithm may be automated to a large extent. An example of an approach by utilizing a biomechanical model is illustrated by Figure 1.2. It is important to note that this approach is considered to be a complement rather than alternative to the direct imaging one (depicted in Figure 1.1). As shown in Figure 1.2, the proposed model has several
components, including biomechanical simulation, geometry reconstruction, acoustic simulation, estimation of muscle activation patterns (MAP).

In both approaches, usually, there are more than one approach to implement each component. One important concern in all components is that the spatial dimension significantly affects the computational cost, the difficulty of the problem to address, and the accuracy of the results. Although the 3D approaches, utilized in this work, fit well with reality, models with lower dimension have been proposed in previous works. Developing such a 3D model introduces several challenges, some of which are exclusively posed by increasing the dimension from one-dimensional (1D) to 3D in the acoustic simulation, or from two-dimensional (2D) to 3D in the biomechanical model.

Further discussion about the benefits of computational models and examples of their application in voice/speech research is given in Chapter 2. To develop a computational model of the vocal tract, some anatomical knowledge may be beneficial, and such information is presented in Chapter 3. Chapter 4 reviews speech production models available in the literature to position this work among others.

1.2 The effects of vocal tract modeling dimensionality

The vocal tract may be represented e.g., by its 3D geometry or the corresponding area function. The vocal tract has a complex geometry, but the area function representation considers the vocal tract as a symmetric tube, and provides a compact representation where only the cross-sectional area and the distance from the glottis are preserved. Determining the vocal tract area function is much simpler than reconstruction of the 3D geometry from medical images. Furthermore, when using the area function, the acoustic characteristics of the vocal tract may be determined using 1D acoustic simulation, while a 3D geometry requires a more sophisticated 3D approach. The increase of the complexity both for acquiring the vocal tract representation and acoustic simulations may lead to the question whether it is necessary to employ a 3D approach. Considering the theory of acoustic waves, and the typical dimensions of the vocal tract, it can be argued that the area function representation provides valid results up to frequencies around 4-5 kHz. However, there is no study to investigate how different features of the geometry (such as the shape of the lips, bending, etc.) influence the acoustic response. To answer this question, a systematic simplification of 3D geometries of the vocal tract was performed and the acoustic characteristics were determined at each simplification step, as part of this work. The simplification procedure offers several alternatives to the vocal tract representation to choose the right one that meets the accuracy and the complexity requirements for a given application. The simplification procedure and its consequences, which are summarized in Chapter 5, has been published in (Dabbaghian et al., 2015), (Arnela et al., 2016a, Paper A), and (Arnela et al., 2016b, Paper B).
1.3 Biomechanical simulation of speech production

The shape of the vocal tract is deformed by the articulators, and the motion of the articulators is a function of forces exerted by muscles, mechanical constraints, and mechanical properties of the articulators. This process is well represented by a biomechanical model. Although articulatory models can correctly represent the motion of the articulators, they can not account for their biomechanical properties, as for instance, how the tongue material, or the jaw mass contributes in speech production. When using a biomechanical model it is a natural choice to do so in 3D. Otherwise, it is not possible to represent the tongue’s transversus muscle, as 2D model includes only muscle fibers running in the sagittal direction. Further, volume preservation of the tongue, which is a result of the tongue’s muscular hydrostat structure, is implicitly handled in 3D, but not in a 2D model. The only reason that may justify using a 2D model is the computational cost, which is becoming less relevant at the present. An existing biomechanical model created in ArtiSynth, namely FRANK (Anderson et al., 2017), was adapted for the purpose of this work. Chapter 6 provides details about the original model, and modifications applied in this work.

1.4 Muscle activation pattern estimation

One challenge in the biomechanical modeling approach in Figure 1.2 is that the control parameters, the muscle activation patterns, MAP, to a large extent, are not known for a given sound. Electromyography (EMG) is the commonly used method to measure the activation of muscles, but it is invasive and impractical for speech tasks, and EMG measurements of the tongue are subject to uncertainty because of the interwoven architecture of the muscles. Furthermore, the relation between EMG signal, which measures the electrical pulses, and the contraction of the muscles is not straightforward. In such a situation, an inverse modeling method provides an alternative way to estimate the MAP. The explanation of such an inverse method to estimate the MAP for a given articulation or sound (i.e formant frequencies) is given in Chapter 7. The MAP estimation from articulation data has been utilized in (Dabbaghchian et al., 2014, Paper D), (Dabbaghchian et al., 2016, Paper E), (Dabbaghchian et al., 2017), and (Dabbaghchian et al., 2018b, Paper F).

1.5 Vocal tract reconstruction from a biomechanical model

A biomechanical simulation of the speech production apparatus can not provide the vocal tract geometry directly. This statement might seem confusing, since the purpose of using the biomechanical simulation of the articulators has been stated to be to obtain the vocal tract geometry. However, in a biomechanical model, each physical structure, such as the tongue, the mandible, etc. is represented by a 3D geometry (either a volume or surface mesh), and the motion of the articulators
is determined by numerically solving the governing physical equations. The vocal tract itself is not a physical object, and it may therefore not be represented directly in a biomechanical model. Instead, it is represented indirectly, as a cavity, by the geometry of all surrounding structures. For acoustic simulations and for visualization purposes, an explicit geometry of the vocal tract is however needed. This poses a challenge in the biomechanical modeling approach depicted in Figure 1.2.

Obtaining the geometry of the cavity would be trivial if it were perfectly enclosed by all surrounding geometries. However, existing gaps and overlaps between the structures of the biomechanical model may cause holes in the boundary surface of the vocal tract cavity, making it unsuitable for acoustic simulations. These artifacts are unavoidable because of the methods used for developing biomechanical models. Furthermore, such gaps and overlaps may also appear again when the structures start to move. Another difficulty imposed by the acoustic simulation is that a computational domain (i.e. a volume mesh), in which the 3D wave equation is solved numerically, must satisfy certain requirements regarding the quality of the mesh elements, and be deformable to follow the motion of the vocal tract deformation. All these requirements lead to a very challenging problem.

To address this problem, several geometry reconstruction methods were developed to blend all the surrounding geometries of the cavity, and generate a deformable vocal tract geometry as an entity that satisfies the mesh quality requirements for acoustic simulations, despite existing artifacts. The reconstruction itself is an important contribution because it allows the linking of 3D biomechanical and acoustic models, thus offering a great potential for future applications of 3D voice production. The three different versions of the geometry reconstruction method that have been developed are summarized in Chapter 8. The first version was used in (Dabbaghchian et al., 2016, Paper E) for simulation of three cardinal vowels \([a, i, u]\). The second version, which is highly stable and more computationally efficient than the first version, has been used for both vowel and vowel-vowel sound synthesis (Dabbaghchian et al., 2018b, Paper F). The third version is significantly more complex and accurate than two others and capable of including sub-branches (such as piriform fossa, sublingual cavity, etc). The third version has been published in (Dabbaghchian et al., 2018a, Paper G).

### 1.6 Coupled 3D biomechanical–acoustic simulations

When it comes to acoustic simulations, there are different approaches implemented in 1D, 2D, and 3D. The 3D-based approaches are the natural choice to maintain fidelity to the modeled phenomena, but they are computationally expensive and introduce new challenges to deal with. Using a 1D acoustic model to simulate the sound waves in a 3D domain corresponds to the assumption of plane wave propagation. This means that such an acoustic model can not account for non-planar waves (i.e. high order modes), but it may still give accurate results when the wavelength is much larger than the dimension of the vocal tract along the
direction perpendicular to the propagation. Considering the typical vocal tract dimensions, this assumption is valid for frequencies below 4-5 kHz. However, a 3D approach replicates the high-frequency contents which may be perceptually important. Furthermore, in 1D approaches, the acoustic model needs to be adapted to different situations, e.g. when there is a large area discontinuity, or if sub-branches such as the piriform fossa, the sublingual cavity, the nasal tract, etc. are to be included. In some other situations, it may not even be possible to utilize a 1D approach since the nature of the problem requires non-planar wave propagation. One such example is the study of the acoustic interaction between the left and the right piriform fossa. On the other hand, with a 3D approach, all these situations are addressed intrinsically (Chapter 4 expands further on the importance of using 3D modeling). In the research within this thesis, 3D acoustic simulations predominate and a 1D approach was utilized only when it was not possible to utilize a 3D approach, such as in estimation of MAP from acoustic data, which involves an iterative procedure (see Chapter 7), or the investigation of the contact role in motor control of speech (see Chapter 11).

1.7 Applications

A 3D biomechanical-acoustic model may have various applications, such as the detailed study of realistic speech sequences, in particular those involving dynamic sounds like vowel-vowel utterances or syllables. As mentioned above, such sequences are difficult to study with MRI, due to its limitations. The outcome of this work also represents a further step in the ambitious field of creating a virtual human physiology, which is expected to play a predominant role in patient-specific modeling, medical surgery (e.g. in glossectomy), and treatments in a not so distant future.

Synthesis of vowel and vowel-vowel sounds is presented in Chapter 10. Vowel-vowel sounds were utilized using both approaches: direct imaging and biomechanical modeling. In the direct imaging approach, the simplified geometry of vowel [a] was linearly interpolated into the geometry of vowel [i], and an [ai] sound was synthesized. This work, published in (Arnela et al., 2017, Paper C), confirms that the use of adequately simplified geometries may result in significant decrease in the complexity of the problem while keeping the accuracy high enough. Using the biomechanical modeling approach, six vowels and three vowel-vowel utterances were synthesized (Dabbaghchian et al., 2018b, Paper F). As another application, the role of contact between the tongue and other structures in producing vowel sounds was investigated (Dabbaghchian & Engwall, 2017). The results are reported in Chapter 11.
1.8 Contributions

The contributions of this thesis are grouped based on the approach (direct imaging or biomechanical modeling) as outlined below. All 3D acoustic simulations have been conducted in La Salle, Universitat Ramon Llull in Barcelona, Spain, by Marc Arnela and Oriol Guash.

A. Simulations based on direct imaging

The use of the geometry simplification procedure resulted in the following contributions, compared to the state-of-the-art:

- Determining the influence of the lips on the acoustic characteristics (Paper A)
- Determining the influence of the vocal tract geometry features, such as bending, shape irregularities etc. (Paper B)
- Geometrical interpolation to synthesize vowel-vowel sequences (Paper C), demonstrating how static MRI data may be used for dynamic sounds

B. Biomechanical modeling

Most, though not all, of the contributions are linked to the geometry reconstruction method, which allows for an unprecedented coupling between a 3D biomechanical model and 3D acoustics.

- 3D reconstruction of the vocal tract geometry
  - first approach, excluding sub-branches, which illustrates the concept (Paper E)
  - second approach, stable and computationally efficient, excluding sub-branches, applicable to synthesis of vowel-vowel sequences (Paper F)
  - third approach, accurate and detailed reconstruction, including sub-branches, demonstrating the possibility to reconstruct realistic and more complex shapes (Paper G)
- Linking 3D biomechanical and acoustic models
  - synthesizing the cardinal vowels, allowing to corroborate and refine the biomechanical model (Paper E)
  - synthesizing vowel-vowel utterances, demonstrating that synthesizing the transition with muscle activation patterns leads to different acoustic output than pure geometrical interpolation (Paper F)
  - obtaining more accurate synthesis results in low and high frequencies compared to standard simulations using the plane wave assumption (Paper E, Paper F, Paper G)
• Estimation of muscle activation patterns
  – from Electromagnetic articulography (EMA) for a vowel-consonant-vowel utterance, demonstrating that possible MAP may be identified directly from articulation data (Paper D)
  – from artificially generated EMA-like data, illustrating that the biomechanical model is able to recreate movements typically observed in EMA data, such as loops (Paper F)
  – from acoustic data, i.e. formant frequencies, as a proof-of-concept of acoustic-to-muscle-activation inversion (Chapter 7)
• Providing insights
  – on the delay between the contraction of muscles and sound onset (Paper D)
  – on the influence of the articulators mechanical properties on their trajectory and the spectrogram of the generated sound (Paper F)
  – on the role of interdental space in generating a dip in the acoustic transfer function of vowel [i] (Paper G)
  – on the quantal relation between muscle activation and acoustic spaces and role of the contact (Chapter 11).

1.9 Summary of the papers

• Paper A
In this work, the influence of the lips’ geometry on the acoustic characteristics of the vocal tract was investigated. SD and MA discussed how to simplify the geometries. SD created all the simplified geometries and MA performed the acoustic simulations. MA wrote the paper with inputs from SD and other co-authors.

• Paper B
In a similar way as Paper A, this work analyzes the vocal tract geometry features, including bending, cross-sectional shape, and number of cross-sections.
SD designed the simplification procedure in collaboration with MA and OE. SD implemented the method, and generated all simplified geometries. MA conducted the acoustic simulations and wrote the paper with inputs from SD, on the simplification procedure, and other co-authors.

- **Paper C**

As an application of the geometry simplification method, a sequence of vocal tract geometries was generated by interpolating between two static ones. The mixed wave equation was solved to generate the [ai] sound. SD adapted the simplification method for this problem by considering the maximum deformation that cross-sections may tolerate before the distortion of mesh elements. MA developed the interpolation method and conducted the acoustic simulations. MA wrote the paper with inputs from SD, and comments from the remaining co-authors.

- **Paper D**

The focus in this paper is the estimation of the tongue muscles for a given EMA trajectory. This work only considers the biomechanical model of the tongue. IN run the simulations with help of OE. SD analyzed the results and wrote the paper with inputs from OE.

- **Paper E**

The first version of the geometry reconstruction was presented in this paper (see Chapter 8 for variations of the method). It was used to link an adapted biomechanical model with acoustic simulations. The MAP were estimated using EMA data, the vocal tract geometry was obtained for the cardinal vowels [α, i, u], and the corresponding sounds were synthesized. SD proposed and developed the reconstruction method, and performed the MAP estimation. MA conducted the acoustic simulations and sound synthesis. SD analyzed the results with help from OE. SD wrote the paper with inputs from MA, and comments from the remaining co-authors.
Paper F

The proposed approach in Paper E for geometry reconstruction was significantly revised to increase its stability and reduce the complexity. The proposed approach is highly stable and efficient, and it can be used to generate a time sequence of geometries for a deforming vocal tract. SD proposed the geometry reconstruction approach, and performed the MAP estimation. MA conducted the acoustic simulations and sound synthesis. SD wrote the paper with inputs from MA on the section describing acoustic simulations. OE and OG provided comments on the manuscript. Part of this work has been presented in Dabbaghchian et al. (2017).

Paper G

The proposed approach for geometry reconstruction in both Paper E and Paper F excludes the sub-branches. Another method was therefore developed for accurate and detailed reconstruction of complex vocal tract geometries. As a proof of concept three vowels were simulated, and the acoustic characteristics of the sub-branches were observed. SD proposed the geometry reconstruction approach and generated detailed geometries for the vowels, a lateral approximant, and an example of dynamic geometries. MA conducted the acoustic simulations. SD wrote the paper, except the section on acoustic simulations, written by MA and OG. OE guided the writing; MA and OG provided comments.

1.10 Relevant publications not included in the thesis


Chapter 2

Computational modeling

“What I cannot create, I do not understand.”
— Richard P. Feynman

In the late 1980s, the "in-silico" expression was invented, contrasting with “in-vitro” and “in-vivo”\(^1\), to refer to a study using computer simulation. Since that date, progress in numerical methods and high performance computing promise a new era of computational modeling in science. Nowadays, many ambitious projects have been launched, including Virtual Physiological Human (Viceconti & Hunter, 2016) to simulate the human physiology, Human Brain Project, euHeart for patient specific modeling of the heart and cardiovascular diseases, SimVascular for cardiovascular simulations, ArtiSynth for biomechanical modeling of human anatomical structures involved in speech and swallowing, and EUNISON for human voice simulation, As Brodland (2015) stated: “Indeed, computational modelling is transitioning into mainstream science in much the same way that statistics did many years ago.”

A computational model is a mathematical expression of a phenomenon implemented in computer code. To make such a model, one needs to comprehend a phenomenon carefully before developing its model. This raises the question why there is a need for a computational model as the phenomenon under study is already known. It may be argued that developed computational models will reflect our knowledge in the best case, and not any more. Indeed, this is true. However, this does not mean that computational models are useless. In this chapter, the role of the computational modeling in the scientific method is explained, and some applications of the computational models are presented.

\(^{1}\)In-vitro (in Latin means “within the glass”) refers to an experiment which is conducted outside of a living organism but in a controlled environment. In-vivo (in Latin means within the living) refers to an experiment which is conducted in a living organism.
2.1 Scientific method cycle

A scientific method to study a phenomenon involves two phases, namely induction and deduction. The induction includes measurements, reasoning, and hypothesis proposal. In the deduction phase, the hypothesis is tested and if the predicted results are not consistent with the empirical data, then the hypothesis is improved, usually by conducting new laboratory experiments, so that the predicted results get closer to the evidence. This process is known as the scientific method cycle as illustrated by Figure 2.1. Even if a hypothesis can explain all observations, it cannot be proven. A computational model can only be utilized in the deductive phase, and it cannot replace the laboratory experiments which provide empirical data. Thus a computational model is considered as a complementary tool that facilitates the deductive phase.

2.2 Applications

A. Numerical solution

Most of the time, there are no analytical solutions for the problem under study. For example, propagation of acoustic waves in a medium is expressed by the acoustic wave equations. The solution of this equation can be found analytically only in a few simple domains. For a complex domain, e.g., the vocal tract, a numerical method needs to be employed. Another example, is the movement of the tongue in response to the forces caused by the contraction of muscles. The relation between force and acceleration can be represented well by the Newton’s second equation, but it is not possible to solve this equation analytically in this case. In such cases, computational modeling can contribute with insights through numerical solutions.
B. Sensitivity analysis

A computational model offers an inexpensive solution for sensitivity analysis, as an important step in characterizing a phenomenon, which usually provides insights and may lead to new hypotheses. For instance, analysis of the formant frequencies in a simple vocal tract model revealed that for the most common vowels in the world languages, [a,i,u], the second formant has the least sensitivity to the constriction location (Fant, 1971). Along the same lines, Stevens (1989, 2000) hypothesized that the relation between the speech articulation and acoustic has a quantal nature, known as the quantal theory. In another model-based study, Perkell (1996) showed that the quantal relation does not hold for the constriction area. Last but not least, Buchaillard et al. (2009) studied how the formant frequencies may change in response to changes of muscles’ contraction. These examples demonstrate how computational modeling can lead to new theoretical insights.

C. Exploring the limits

Computational models are well suited to explore the limits of a phenomenon. As an example, using a tube model of the vocal tract, Carre (2004, 2009) showed that the shape of the vowel diagram (trapezoid) is a pure result of the acoustic characteristics of the vocal tract.

D. Unusual situations

Let us assume that we would like to investigate the separate influence of the lips’ on the acoustic response. This cannot be done in a laboratory experiment, since the subject’s lips cannot be removed, but it can on the other hand be achieved using a computational model (Arnela et al., 2016a). In another example, the vocal tract was simplified to investigate the acoustic consequences (Arnela et al., 2016b). The possibility to test a system in an unusual situation may also benefit the medical field, e.g. to predict a surgery outcome (Zhou et al., 2013a; Takatsu et al., 2017), or patient specific modeling (Neal & Kerckhoffs, 2010; Saha et al., 2016).

E. Examining a hypothesis

Even if a computational model may not be used to prove a hypothesis, it may be utilized to falsify it. A computational model further makes it possible to examine if some certain conditions are sufficient for a given phenomenon to happen. However, one should be more cautious when discussing necessary conditions. As an example, contraction of the genioglossus muscle suffices to create a constriction in the oral cavity, but it is not necessary, since the positioning of the jaw may compensate for a lack of genioglossus contraction.
F. Knowledge integration

A computational model allows one to integrate knowledge from different sources and disciplines, and check the consistency among them. For example, considering the vocal tract as a tube, which can be deformed in an arbitrary way may lead to a different conclusion than considering that such a deformation is constrained by the tongue. Considering the innervation of the tongue further helps in defining the control units of the tongue. Yet another example is the loop-like trajectories observed in EMA. This may be attributed to the complexity of control parameters (i.e. motor commands), but it has shown that mechanical properties of the tongue may be responsible for such trajectories (Payan & Perrier, 1997; Dabbaghchian et al., 2018b). In this case, considering only the articulatory information might be misleading since it suggests that the control parameters need to be complex to generate such loop-like trajectories.

G. Documentation

A computational model can represent the knowledge about a phenomenon in the best quantitative way, and the results can be reproduced. This makes it an ideal tool for documenting and conveying the knowledge. It may also serve as an educational tool for students.

H. Inverse modeling

When there is no direct way to observe (or measure) a phenomenon, an inverse model may be utilized. An inverse model estimates the parameters of the forward model for an indirect given observation. This introduces two different challenges: finding a model of the phenomenon that is consistent with the observations, and dealing with non-uniqueness of the solution (Tarantola, 2005). In speech modeling, inverse models may be utilized to estimate articulatory parameters (see, e.g., Toutios et al., 2011), or to estimate the contraction of muscles in a biomechanical model (see, e.g., Zandipour et al., 2004; Dabbaghchian et al., 2014; Harandi et al., 2017).

2.3 Model choice

Model scale and order are two determining factors of model complexity. The choice of the scale and order of the model depend on the research question, required accuracy and available resources.

A. Scale

Scale of the model refers to the size of the basic structure in the model, and can range from molecules, cells, tissue, organ, to the whole body. Although, multiscale
models may offer a maximum knowledge integrity, they are extremely complex, and usually different physics are involved at different scales. Additionally, coupling of different physics in a multiphysics model is a challenge. In this work, the human speech production is considered at the organ functionality level which involves biomechanics and acoustics.

B. Order

The model order refers to the degree of approximation in which the model mimics the reality. As an example, the vocal tract may be approximated by its area function, assuming plane wave propagation, or by its 3D shape, in which case such an assumption is not needed.
Chapter 3

Physiology of speech production

“The human body is the most complex system ever created. The more we learn about it, the more appreciation we have about what a rich system it is.”

— Bill Gates

Production of voice and speech requires coordinated function of the nervous, muscular and respiratory systems. The nervous system convey commands to motor neurons. These neurons cause muscle fibers to contract, to exert forces, and to move the speech articulators. There are approximately 100 muscles that are involved in speech production. The involved organs range from the brain, which controls the process, over the ribcage, lungs and the trachea involved in producing and transmitting the airflow, to the upper organs that form the vocal tract, including the larynx, pharynx, velum, tongue, etc.¹

This chapter briefly reviews the anatomy and physiology of selected speech organs and the terminology used in the remainder of this thesis, as a basic understanding of the physiology is required when addressing biomechanical modeling of speech production. More details about the anatomy and physiology can be found in Zemlin (2011); Gick et al. (2012); Kandel et al. (2012); Drake et al. (2014); Guenther (2016).

3.1 The anatomical directions and planes

The standard terms for relative anatomical directions and planes that are used throughout this thesis are illustrated in Figure 3.1a and Figure 3.1b, respectively. The directions are anterior-posterior, superior-inferior, and lateral-medial, and the reference planes are sagittal, coronal, and transverse.

¹By considering auditory and somatosensory feedback, the ears and somatosensory receptors are also part of the speech production system. However, these organs are not discussed in this chapter.
Chapter 3

3.2 The upper airway

The upper airway plays a vital role in sound production, breathing, and swallowing. The upper airway is not a structure in itself, but consists of several small and large cavities, and is usually divided into several regions, namely the laryngopharynx, oropharynx, oral cavity and the nasal cavity. Figure 3.2 illustrates the anatomy of the structures that form to the upper airway.

Figure 3.1: Anatomical terms defining relative directions and reference planes (head image by M. Arnela).

Figure 3.2: Anatomy of the upper airway.


**A. The larynx**

The larynx, depicted in Figure 3.3, is important in speech production primarily since it houses the vocal folds. The inferior of the larynx is situated just above where the trachea\(^2\) and the esophagus\(^3\) merge to form the pharynx tract. The superior part of the larynx is attached to the hyoid bone\(^4\), and its outer surface is covered mainly with the cricoid and the thyroid cartilages. During swallowing, the epiglottis acts as a lid to close the trachea, preventing the bolus to enter the trachea. A membrane, namely the thyrohyoid membrane, attaches the superior border of the thyroid to the inferior border of the hyoid bone. When air is emerging from the trachea, it passes the vocal folds. When the folds are close, the airflow causes them to vibrate\(^5\), generating the sound source for voiced sounds. When the folds are further apart, as in voiceless sounds, they do not vibrate. The vocal folds and the slit-like opening between them is named glottis, which is usually used as a reference point to measure the vocal tract length. The piriform fossa (or sinus piriformis) are two small, pear-shaped (hence the name) cavities on either side of the larynx, which are bounded laterally by the aryepiglottic fold and the thyroid cartilage, and posteriorly by the pharynx wall. The piriform fossa are of interest in speech production research, as they affect the generated sound. The larynx structure may move up or down during speech, thus altering the vocal tract length and consequently the resonance frequencies.

---

\(^2\)The trachea is a windpipe, which belongs to the respiratory system.
\(^3\)The esophagus is a food pipe, which belongs to the digestive system.
\(^4\)The hyoid bone is U-shaped and is the only in the body that is not attached to another bone.
\(^5\)The typical vibration frequency of the vocal folds is 100 Hz for males and 200 Hz for females.
B. The tongue

The posterior part of the tongue root is attached to the hyoid bone and the anterior part is attached to the mandible as shown in both Figure 3.2 and Figure 3.4. The tongue is a muscular hydrostat structure, i.e., it mainly consists of muscles without skeleton support; it is incompressible and preserves its volume. The tongue is the most active articulator and plays a crucial role in speech production and swallowing.

Tongue muscles have a complex interweaved structure. Some muscles originate outside the tongue, namely the extrinsic muscles genioglossus (GG), hyoglossus (HG), styloglossus (SG), palatoglossus (PG), geniohyoid (GH), and mylohyoid (MH). The origin and insertion points of the extrinsic muscles are summarized in Table 3.1. The other tongue muscles are running inside the tongue, namely the intrinsic muscles superior longitudinal (SL), inferior longitudinal (IL), verticalis (V), and transversus (T) as shown in Figure 3.4. In general, extrinsic muscles cause larger displacement/deformation than the intrinsic muscles.

![Figure 3.4: Anatomy of the tongue and its muscles. Reprinted from Stavness (2010) with original source Drake et al. (2014), © (2014), with permission from Elsevier.](image-url)
Table 3.1: Origin and insertion points of the tongue extrinsic muscles.

<table>
<thead>
<tr>
<th>muscle</th>
<th>origin</th>
<th>insertion</th>
<th>muscle</th>
<th>origin</th>
<th>insertion</th>
</tr>
</thead>
<tbody>
<tr>
<td>GG</td>
<td>mandible</td>
<td>tongue</td>
<td>PG</td>
<td>soft palate</td>
<td>tongue</td>
</tr>
<tr>
<td>HG</td>
<td>hyoid bone</td>
<td>tongue</td>
<td>GH</td>
<td>mandible</td>
<td>hyoid bone</td>
</tr>
<tr>
<td>SG</td>
<td>styloid process</td>
<td>tongue</td>
<td>MH</td>
<td>mandible</td>
<td>hyoid bone</td>
</tr>
</tbody>
</table>

Based on the gross anatomy of the tongue muscles and their innervation (see, e.g., Gick et al., 2012), all tongue muscles except PG are innervated by the hypoglossal (XII) nerve. However, their detailed innervation inside the tongue is still under investigation (see, e.g., Mu & Sanders, 2010). These studies are crucial in order to understand the functional units of the tongue control. For instance, it has been very common in most of the previous studies to divide the GG into three compartments, namely posterior, middle, and anterior (Dang & Honda, 2002; Stone et al., 2004; Gérard et al., 2006; Buchaillard et al., 2009; Anderson et al., 2017). However, neuroanatomy studies of the hypoglossal (XII) suggests that the GG muscle consists of two functional parts, namely horizontal and oblique (Mu & Sanders, 2010; Sanders & Mu, 2013). A similar controversy exists for the SG muscle regarding whether it consists of one or two functional units (Takano & Honda, 2007).

In general, contraction of the tongue muscles have direct and indirect consequences. A direct consequence is any displacement or deformation that can be attributed to a muscle’s fiber direction. For example, contraction of the HG muscle moves the tongue in the inferior-posterior direction. An indirect consequence is any deformation that can be attributed to the tongue volume conservation property. That is, if one part of the tongue is compressed, because of muscle forces, some other parts needs to be released so that the volume remains constant. Although direct consequences are mainly attributed to fiber directions, indirect consequences of a muscle contraction depend not only on other muscles, but also the surrounding structures. For example, the tongue can not be released in a region in which it is surrounded by hard structures such as bones and cartilages.

The tongue muscles are usually contracted as agonist-antagonist pairs. An antagonist muscle produces the motion that is in the opposite direction of an agonist muscle. For example, HG may be considered as an antagonist to SG. Muscles may also act as synergists. A synergist muscle may not contribute to the motion but its contraction causes stability in the motion by preventing an unwanted motion.

C. The airway cavities

The upper airway is indeed a complex cavity, and the motion of the articulators may reform its shape. The pharynx cavity, which begins just above the larynx and continues towards the soft palate, is mainly affected by the motion of the tongue. Small cavities, namely the piriform fossa and the vallecula, are attached
to the pharynx cavity, but they may be connected or closed off, depending on the articulation. The piriform fossa (see section A.), are attached at the lateral sides to the most inferior part of the pharynx. The vallecula is a cavity between the posterior part of the tongue and the anterior side of the epiglottis, and if it is formed depends on the relative positions of the tongue and epiglottis: the vallecula only exists when the tongue root and the epiglottis are apart, in which case the vallecula is attached to the anterior of the pharynx cavity. The upper part of the pharynx cavity branches into the nasal and oral cavities. The nasal cavity has a static shape and plays a role in speech production when the soft palate is lowered. This opens the passage to the nasal cavity, and causes nasalization of the sound as additional resonances and anti-resonances are formed in this cavity. When the soft palate is raised, the passage is closed and the nasal cavity is decoupled from the airway and does thus not affect the sound. The shape of the oral cavity is governed by the movements of the tongue, jaw and lips. The left and right interdental spaces (space between the upper and lower teeth), and the sublingual cavity (the cavity between the tongue tip and mouth floor) are three small cavities in the oral region that affect the acoustics.

3.3 The nervous system

The nervous system is divided into the central nerve system (CNS) and the peripheral nerve system (PNS). The CNS consists of the brain and the spinal cord, and the PNS includes all of the nerves that branch out from the CNS to reach all receptors and motor neurons within the body. The PNS conveys motor commands from CNS to motor neurons and also collects the sensory information from receptors.

Different areas of the brain in the cerebral cortex and subcortical structures are involved in speech production and perception as shown in Figure 3.5. It is known that aphasia\(^6\) is caused by injury to Broca’s or Wernicke’s areas. The other areas of the cerebral cortex that are involved in speech production/perception are the auditory, somatosensory, and motor cortex. Furthermore, cerebellum, basal ganglia, and thalamus are some of the subcortical areas playing a role in speech (see Guenther, 2016, for a detailed explanation of the involved areas in speech).

The speech articulators are innervated by 6 cranial nerves\(^7\) as shown in Figure 3.6a. These nerves are either exclusively motor, sensory or mixed (see Gick et al., 2012, for more details).

A peripheral motor nerve is connected to a motor neuron that acts as an interface between the PNS and muscle fibers as shown in Figure 3.6b. Firing of the motor neurons contracts the muscle fibers (see Kandel et al., 2012, for more details). The detailed innervation of the nerves and motor neuron types are still under investigation (Mu & Sanders, 2010).

---

\(^6\)Aphasia is an impairment of language that affects the production or comprehension of speech, as well as the ability to read or write.

\(^7\)Cranial nerves are 12 pairs of the peripheral nerves that emerge from the brainstem.
Figure 3.5: Cortical and subcortical areas of the brain involved in speech (Guenther, 2016), © 2016 by the Massachusetts Institute of Technology, published by the MIT Press, reprinted with permission.

Figure 3.6: (a) Innervation of the organs involved in speech, (b) A motor unit and its connection to muscle fibers.
Chapter 4

Speech production models

“A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant.”
— Manfred Eigen

The source-filter theory (Fant, 1960) considers the vocal tract as a filter that forms the speech acoustics of the source signal (produced by the vocal folds or by turbulence at a constriction). The characteristics of such a filter are determined by the vocal tract shape formed by the speech articulators (c.f. Chapter chapter 3), which, in turn, are controlled by the contraction of muscles. Developing a physics-based computational model of the whole process is an ambitious goal that requires a multiphysics approach, and a tremendous effort. Computational cost and validation of the model would be other concerns for such a model. Usually, parts of the whole process are modeled using simplifying assumptions about other parts. Based on the literature review presented in this chapter, the speech production modeling approaches may be divided into the four categories geometrical, articulatory, biomechanical, and neuromuscular, which, in that order, address the modeling with increasing levels of detail and fidelity towards human speech production, as shown in Figure 4.1. Each category incorporates different physical phenomena and the amount of effort to develop the model, and the resulting computational cost may therefore vary substantially. Usually, the demand increases significantly from the apex to the base of the pyramid. In this chapter, a literature review of vocal tract models is presented. The geometrical approach considers the vocal tract as a tube or an area function, whereas both articulatory and biomechanical models consider the vocal tract as a cavity, rather than an entity, which is formed by the position and shape of the speech articulators. There are two main differences between articulatory and biomechanical models. In a biomechanical model, the mechanical properties of the articulators directly influence the speech acoustics, whereas an articulatory model only considers the kinematics of the articulators, signifying that only the time-changing vocal tract geometry is considered. Another difference is
that, in an articulatory model, high level phonetic descriptors control the model, whereas a biomechanical model is controlled by the contraction of the muscles.

### 4.1 Geometrical models

In early attempts, the vocal tract was modeled as a tube with varying cross-sections. Although a single uniform tube may characterize the neutral schwa vowel ([ə]), a tube with variable cross-sectional area is required to represent other sounds (e.g. [a, i, u]). Figure 4.2 shows tube models with different number of cross-sections, which are able to represent the basic acoustic characteristics, i.e., the resonance frequencies, of the vocal tract. Using X-ray imaging, Fant (1960) proposed a four-tube model for vowels and a three-tube model for consonants. Later on, this proposal was extended to a tube with $n$ cross-sections. Such a tube is usually represented by the area function (Baer et al., 1991; Story et al., 1996), which describes the

![Figure 4.2: Tube model of the vocal tract: (a) uniform tube, (b) two-tube model, (c) three-tube model](image-url)
cross-sectional area of the vocal tract against the distance from the glottis (c.f. Figure 4.3a). The emergence of magnetic resonance imaging (MRI), which is considered to be safe for the subjects’ health, became an important tool, from the early 1990’s, for more extensive studies of the vocal tract’s 3D geometry. Further improvement of the MRI technology allows one to capture high resolution images and hence detailed 3D reconstruction of the vocal tract shape (Aalto et al., 2014), as shown in Figure 4.3b). Both the area function and 3D geometries may be used as input to acoustic simulation.

Different methods have been proposed for the acoustic analysis of the area function including transmission line analogy (Stevens et al., 1953; Fant, 1960), time domain methods (Ishizaka & Flanagan, 1972), and wave-reflecting analogy (Story, 1995). Using the area function simplifies the acoustic analysis, since it assumes plane wave propagation. In this case, the geometrical shape of the cross-sections is not important and the acoustic characteristics are only governed by the areas of the cross-sections (and in particular the location of smaller cross-sectional areas, related to the place of articulation). However, this assumption limits the accuracy of the results up to a certain frequency. The maximum valid frequency depends on the maximum diameter of the tube, \(d\), and is approximately determined by \(c/2d\) (\(c\) being the speed of sound), which is the frequency at which the first transverse mode appears (Takemoto et al., 2006). Based on the area functions used in his study, with a maximum cross-sectional area of \(6 \text{ cm}^2\), the acoustic analysis was considered valid for the frequencies below \(6 \text{ kHz}\). Indeed this limit could be much lower than \(6 \text{ kHz}\) because of two reasons. Firstly, based on the analysis of the vocal tract geometries in (Aalto et al., 2014), the maximum cross-sectional area of vowel [\(a\)] is \(\approx 13 \text{ cm}^2\), which limits the validity of the result to frequencies below \(4 \text{ kHz}\). Secondly, the maximum diameter is underestimated in the tube model by assuming
circular cross-sections. The same analysis showed that the maximum diameter for the vowel [i] is \( \approx 6 \) cm. This corresponds to a maximum valid frequency of 3 kHz. Work presented in this thesis (Arnela et al., 2016b, Paper B) has further shown that not only the cross-sectional shape, but also bending of the vocal tract, and number of cross-sections influence the acoustic transfer function. Even in the low frequency range, deviations of up to 14% (considering \( F_4 \) of vowel [i], see Table 1 and also Figure 7 of (Arnela et al., 2016b, Paper B) were reported. In another study, Blandin et al. (2015) showed that the high order modes may appear in frequencies as low as 4-5 kHz.

Using the area function and the plane wave assumption consequently affect the accuracy of the acoustic analysis in primarily the high frequency range, but also for low frequencies, though the deviations are less significant at lower frequencies. On the other hand, 3D simulations require solving the wave equation by employing complex numerical methods, such as the finite element method (Motoki, 2002), finite difference (Takemoto et al., 2010), boundary element method (Kagawa et al., 1992), or waveguide (Mullen et al., 2006). These methods introduce extra problems and requirements on numerical stability, quality of the mesh elements, and increase the computational cost. Thus one may prefer to work with area function rather than 3D geometry arguing that such deviations are unimportant and perceptually insignificant. However, as Monson et al. (2014a) stated, even if the frequencies above 5 kHz have traditionally been believed to be unnecessary, they have never been shown to be perceptually insignificant. In contrast, there are some studies highlighting the importance of high frequency energy (HFE) for speech naturalness, intelligibility etc. Moore & Tan (2003) reported that the naturalness score of the speech signal decreased significantly when the cut-off frequency changed from 10.9 kHz to 7 kHz. In another study, it was shown that in some conditions, speech ineligibility increased when the lowpass filter cutoff frequency was increased from 5 kHz to 7 kHz (Moore et al., 2010). Best et al. (2005) showed that speech localization was improved when keeping the frequency contents up to 16 kHz. Using computer models, it has also shown that the quality of synthesized signals are improved by tuning a 2D vocal tract model (Arnela & Guasch, 2017), and the accuracy of the formants increase by using high order models (Gully et al., 2018). It is further observed that the propagation of a higher order mode can strongly modify the speech directivity, i.e., the directional acoustic characteristics of the speech (Blandin et al., 2016).

In addition to the above evidence about the importance of HFE, simulation of the wave propagation using a 3D geometry of the vocal tract provides a unified approach. That is, in traditional 1D approaches, the proposed acoustic model is adapted to address different phenomena and to correct formant values. For example, Kang & Ji (2008) proposed a modification for large area discontinuities, which have been observed in the vocal tract geometry. Another modification is needed to include side-branches. Other modifications have been proposed to account for the acoustic influence of the piriform fossa and the nasal tract (Dang & Honda, 1996; Dang et al., 2016). However, with 3D approaches, these phenomena are addressed
intrinsically. The side-branches of the vocal tract, such as the piriform fossa, vallecula, and the interdental and sublingual spaces, can easily be included in the acoustic simulation, which is not the case when using the area function. Takemoto et al. (2013) showed that using a 3D geometry of the vocal tract facilitates modeling of the acoustic interaction between the left and right piriform fossa, which was not possible to do with the transmission line analogy approach. Furthermore, to address the computational cost, Dabbaghchian et al. (2015) proposed a simplification procedure that limits the computational cost by reducing the geometrical detail in a manner that maintains the acoustic properties. The acoustic influence of different simplifications is described in (Arnela et al., 2016b). Other alternatives are adapted techniques utilizing a two-dimensional representation of the vocal tract (Arnela & Guasch, 2014).

4.2 Articulatory models

In a geometrical model, it is not possible to study the role of the speech articulators since the vocal tract is considered as an entity. An articulatory model, as the next level in the modeling pyramid, allows an analysis of the movement of the articulators. Early articulatory models were developed based on the midsagittal plane (Liljencrants, 1971; Mermelstein, 1973; Coker, 1976; Maeda, 1982; Kröger et al., 1995). These models are useful for studies of coarticulation (Öhman, 1967), bite-block restriction (Gay et al., 1981), motor control (Lindblom et al., 1977; Saltzman & Munhall, 1989), or articulatory compensation (Maeda, 1990). Two-dimensional articulatory models are based on the assumption that the vocal tract shape is symmetric on either side of the midsagittal plane. This assumption

Figure 4.4: Examples of articulatory models and control variables: (a) Reprinted with permission from Toutios et al., 2011 © 2011 Acoustical Society of America, (b) Reprinted from Engwall, 2003 © (2003), with permission from Elsevier, (c) Birkholz, 2013.
may be safe to some extent, but it certainly can not be generalized to all vocal tract shapes, an in particular not to articulations such as the lateral approximant [l]. It is also not possible to represent the side-branches such as vallecula, or piriform fossa as these cavities are located on the lateral sides. Simulation of abnormal speech, e.g. in glossectomy patients (Zhou et al., 2013a,b; Ha et al., 2016; Takatsu et al., 2017), is another case that a 2D model can not address. In response to these limitations, three-dimensional articulatory models were developed (Badin et al., 1998; Engwall, 2003; Birkholz et al., 2006).

Figure 4.4 shows three different articulatory models and their control parameters. In an articulatory model, the control parameters are usually defined by analyzing articulation data captured with measurement techniques such as EMA or MRI. The parameters usually correspond with high level phonetic descriptors such as jaw height, tongue height, etc. However, the parameter set is not uniquely known for a given articulation and needs to be determined empirically. Engwall (2003) utilized a combinatorial search to adjust the parameters. Alternatively, an inverse problem may be solved, so that the desired articulation is obtained with a plausible combination of parameters. Since the relation between articulation and parameters is one-to-many, an optimization approach may be utilized to minimize a cost function in the inversion problem. Toutios et al. (2011) utilized an optimization-based approach to estimate the control parameters of an articulatory model for a given trajectory of flesh-points. Such trajectories were collected using EMA. In another work, Birkholz (2013) used MRI to adjust the parameters manually for the best possible match between the MRI contours and the model.

Although an articulatory model predicts the movement of the articulators, it can not be used directly for acoustic simulations. That is, the articulatory model simulates the kinematics of the articulators while the acoustic study requires the vocal tract shape. As already discussed, an acoustic simulation uses either the area function or the geometry of the vocal tract. In order to perform an acoustic simulation based on an articulatory model, an extra step is required to create the appropriate representation of the vocal tract. In the case of 2D models, the $\alpha\beta$ method estimates the area function $A(x)$ from the midsagittal dimensions $d(x)$, using one multiplication $\alpha$ and one exponential factor $\beta$ as $A(x) = \alpha d(x)^\beta$, where $x$ is the distance from the glottis (Heinz, 1965; Perrier et al., 1992). The model requires training data to learn the relationship between the variables, e.g., using images captured in perpendicular slices to the vocal tract midline when a subject articulates sustained vowels. The accuracy of the approach varies along the vocal tract regions; it is sensitive to training data (see Soquet et al., 2002, for a review), and the estimation is valid only for the particular subject for which the variables have been determined. Still, the method provides a method for area function estimation from a 2D model.

In the case of 3D models, it is appropriate to calculate the area function in 3D rather than using the $\alpha\beta$ method. This is straightforward when the geometry (i.e.

---

1See Paper D for a description of measurements with Electromagnetic Articulography, EMA
boundary surface) of the vocal tract is available (Badin et al., 1998). However, for more complex articulatory models that consider the vocal tract as a cavity bounded by articulators without explicit representation of the vocal tract, the task becomes difficult. Two methods have been proposed to address this problem. In the first method, a set of cutting planes intersect the geometries of the articulators. Then, the area of a closed contour is calculated on each plane (Buchaillard et al., 2009). In the second approach, an initial 3D shape of the vocal tract in rest position is constructed. Then, each vertex of the shape is attached to one or more articulators so that the movement of articulators deforms the vocal tract geometry and the area function is calculated by intersecting the deformed geometry by cutting planes (Anderson et al., 2017). A similar idea has been proposed by Birkholz et al. (2006).

4.3 Biomechanical models

In an articulatory model, each articulator is represented only by its 2D midsagittal contour or outer 3D shape, without specifying its mechanical properties such as mass, or stiffness. Consequently, an articulatory model can replicate only the kinematics of the articulators, while failing to account for dynamic properties. As some properties of the produced speech can be attributed to the mechanical properties of the speech apparatus, this is a simplification. For instance, it has been shown that loop-like trajectories observed in EMA data (Mooshammer et al., 1995) may be the result of mechanical properties (Dabbaghchian, 2018; Payan & Perrier, 1997), and that mechanical inertia may account for coarticulation (Recasens et al., 1997; Baum & Waldstein, 1991) rather than explicitly planned motor commands. As another example, Perkell (1996) hypothesized that mechanical saturation effects (Fujimura & Kakita, 1979) account for the acoustic stability of vowels [a, i] against variation of motor commands.

Figure 4.5: Examples of biomechanical models: (a) in 2D (reprinted with permission from Perrier et al., 2003 © 2003 Acoustical Society of America, and (b) in 3D (Dabbaghchian et al., 2017).
In a biomechanical model, such as the ones shown in Figure 4.5, mechanical properties are assigned to the speech articulators, which leads to more realistic modeling, but introduces challenges regarding how the mechanical properties of the articulators may be determined. Bony structures are usually modeled as rigid bodies. Among the soft and deformable structures, the tongue has the most complex structure, being a muscular hydrostat. It is either modeled as a hyper-elastic material, or is considered as a constitutive model. In the later case, parameters of the constitutive model are determined by conducting ex/in vivo experiments (Hermant et al., 2017).

In addition to the mechanical properties, a biomechanical model comprises muscles that contract to generate the required force of the articulation. To include a muscle model, knowledge of the muscle anatomy and their innervation is required, and these are not yet fully understood. This has resulted in variations and controversy in the literature. For example, the genioglossus muscle is usually assumed to be comprised of three parts: posterior, middle, and anterior (Dang & Honda, 2002; Takano & Honda, 2007; Buchaillard et al., 2009; Fang et al., 2009; Wu et al., 2014). However, in other studies Mu & Sanders (2010); Sanders & Mu (2013) it was divided into horizontal and oblique parts, based on the innervation analysis of five adult human tongues. As another example, the styloglossus has been considered as muscle comprised of three parts (Takano & Honda, 2007), two parts (Fang et al., 2009), or as a single muscle (Dang & Honda, 2002; Buchaillard et al., 2009; Wu et al., 2014). Addressing the innervation and functional units of the tongue muscles, as an important still open question, requires conducting more experimental in/ex vivo studies (Mu & Sanders, 2010). Using high-resolution 3D MRI to visualize the anatomy of muscles, and exploring the variations across subjects may help to refine biomechanical models (Stone et al., 2016). Diffusion tensor imaging is another emerging technology which may reveal more aspects of the muscles anatomy and their physiology (Shinagawa et al., 2008; Murano et al., 2010).

Muscle force generation is usually simulated using a Hill-type model (Hill, 1938). The generated force consist of active and passive forces, and the active force is a function of muscle cross-sectional area, current length and its shortening rate. As an extension of the Hill model, continuum-based muscle models have also been developed (Blemker et al., 2005). Other alternatives are the equilibrium point hypothesis (λ-model) proposed by Feldman (1986) or its extensions (Nazari et al., 2013).

In a biomechanical model, contraction of muscles, as the input, controls the articulators. However, these inputs are to large extent unknown for a given articulation or sound. EMG can in general measure the activity of muscles, but it is invasive and impractical in speech production task. EMG measurements of speech have hence been reported in few studies. Mac Neilage & Sholes (1964) used surface EMG to measure 13 locations on the tongue surface of a single subject while articulating [pvp] utterances. In another study, Baer et al. (1988) measured the extrinsic tongue muscles GG HG, SG, and MI, GH, and a muscle of the lips, namely orbicularis oris superior (OOS) while producing [spvp] utterances. EMG
measures the electrical signal of motor neurons, which may not easily be attributed to the contraction of the muscle. Moreover, as tongue muscle fibers are interwoven in a complex manner, EMG measurements of individual muscle activation risk having low confidence. Other alternatives have been proposed to determine the control parameters including combinatorial searching (Maeda & Honda, 1994; Dang & Honda, 2004; Buchaillard et al., 2009), iterative search (Wu et al., 2014), or inverse methods (Stavness et al., 2012), or using MRI to measure the change in the muscle length Takano & Honda (2007). In general, determining the muscle activation patterns for a given speech sequence is both computationally expensive and challenging in terms of determining physiologically correct combinations, as there is a many-to-one mapping between muscle activation patterns and articulatory output.

Early biomechanical models were developed in 2D in the midsagittal plane (Perkell, 1974; Payan & Perrier, 1997; Sanguineti et al., 1998). Figure 4.5a depicts an example of such a model presented by Payan & Perrier (1997). In such a model, it is not possible to represent structures in the sagittal direction, such as the tongue’s transversus muscle, and the verticalis muscle has to be represented in the same plane as the genioglossus, which is anatomically incorrect. In addition, volume preservation of the tongue, which is a result of the tongue’s muscular hydrostat structure, is not easily implemented in a 2D model. This has lead to the development of 3D biomechanical models (Takemoto, 2001; Dang & Honda, 2001, 2004; Fang et al., 2009; Buchaillard et al., 2009; Anderson et al., 2017), which are more truthful representations of the anatomy. An example of a 3D biomechanical model is depicted in Figure 4.5b (Dabbaghchian et al., 2017).

In the same way as the articulatory model, a biomechanical model does not simulate the acoustic waves and it must be linked with an acoustic model. Payan & Perrier (1997) used the \(\alpha\beta\) method to estimate the area function in a 2D biomechanical model. Buchaillard et al. (2009) calculated the area function directly from the 3D geometries (Buchaillard et al., 2009). In another work, (Stavness et al., 2014) proposed a method, namely "skinning", in which a 3D skin mesh, i.e. surface mesh, is created of the vocal tract, and each vertex of the skin is assigned to one or more articulators. When an articulator moves, its assigned vertices belonging to the vocal tract skin are also moved. This provides a way to deform the skin mesh by the movement of the articulators. The skin mesh may then be used to calculate the area function, or alternatively, it may potentially be used for 3D acoustic simulations, but it would not be physically correct to consider the vocal tract as a structure of its own. Experiments performed in this thesis moreover revealed that large deformation leads to elongated triangles and self-intersections of the skin mesh, which makes it useless for numerical acoustic simulation.

To the author’s knowledge, in all previous models, the articulation has therefore been linked to acoustics using the area function, and acoustic simulations have usually been limited to the calculation of two to three formant frequencies, without any generation of sound.

With a 3D biomechanical model it is natural to perform the acoustic simulations in a 3D domain, to take full advantage of the details of the biomechanical model.
However, this introduces several challenges. Firstly, the acoustic model must simulate the wave propagation in a 3D domain with moving boundaries. Secondly, such an acoustic simulation requires a 3D deformable, air-tight and non-self-intersecting, geometry of the vocal tract at each time step. Construction of such a geometry is not a trivial task and requires complex computational geometry algorithms, since the vocal tract is a cavity, created by the void between the different articulatory structures in the model.

### 4.4 Speech production modeling in this thesis

Using a three-dimensional biomechanical model and a physics-based approach for speech synthesis allows for high fidelity in modeling of human speech production: model control is performed using muscle activation patterns corresponding to human speech control, the resulting vocal tract shape closely corresponds to the geometry of the human vocal tract and the properties of the acoustic wave propagation are modeled. However, the computational cost increases significantly compared to other speech production models, and several challenges are introduced. The question whether it is essential to use such a complex model for speech production may thus be asked. The emergence of a new era in scientific computation, with the aim of simulating the entire human physiology (Hunter et al., 2010; Magnenat-Thalmann et al., 2014; Viceconti & Hunter, 2016), and detailed patient-specific modeling (Neal & Kerekhoffs, 2010), nevertheless justifies the use of biomechanical modeling in health care and speech production modeling. Such a model may be used in prediction of oral surgery outcomes, i.e., to examine how speech production may be altered after surgery, e.g., in glossectomy patients (Zhou et al., 2013b; Takatsu et al., 2017). It may also be used in speech motor control studies (Perrier et al., 2003; Zandipour et al., 2004).

Considering the arguments above, a 3D biomechanical model, further described in chapter 6, is used for speech production modeling in this thesis.

Although significant efforts have been made in both 3D biomechanical and 3D acoustic modeling, little attention has been paid to combined 3D biomechanical-acoustic models. This thesis proposes a method to create the vocal tract geometry at each time step in a biomechanical simulation, constraining the created geometry to be suitable for 3D acoustic modeling. The method itself is an important contribution because it allows us to link biomechanics and acoustics in a 3D biomechanical-acoustic model. The potentials of the combined 3D biomechanical-acoustic model have further been demonstrated through the synthesis of six vowels [a, o, e, i, u] and three vowel-vowel utterances [ui, ou, ui]. The influence of the mechanical properties of the articulators was further investigated by comparisons of spectrograms of a vowel-vowel utterance generated based on a biomechanical model with those generated with a geometrical method that only considers the kinematics of the vocal tract.
Chapter 5

Effects of geometry simplifications

“Everything should be made as simple as possible, but not simpler.”

— Albert Einstein

The previous chapter considered speech production models with different levels of detail. It is natural that the more a model is simplified, the more it departs from the phenomena that it should model. However, the actual consequences of the simplifications are often not fully explored. This chapter therefore assesses how the acoustic output is affected by a number of common simplifications.

The area function and the 3D geometry are two common representations of the vocal tract. Acoustic theory suggests that either representation may lead to a similar acoustic response up to frequencies around 4-5 kHz (considering the typical dimensions of the vocal tract). However, no numerical study has thoroughly investigated how different features of the geometry (such as cross-sectional outline, lip shape, bending, etc.) influence the acoustic response. A simplification procedure was developed and employed to answer this question. 3D geometries of the vocal tract were simplified systematically, and at each step, the acoustic characteristics were determined to reveal how different features of the geometry contributes to the acoustic characteristics. Furthermore, the simplification procedure generates several alternative vocal tract representation with a trade-off between complexity and accuracy. This chapter summarizes the simplification procedure (Dabbaghchian et al., 2015), and the acoustic analysis of the simplified geometries that is the topics of (Arnela et al., 2016a, Paper A), and (Arnela et al., 2016b, Paper B). Some applications, in which the simplified geometries may be used, are also introduced.

5.1 The lips

The acoustic waves, after traveling along the vocal tract, are radiated to the free-field space through the lips. Although the characteristics of this radiation may
be influence by the lips shape, it is common to terminate the vocal tract with a vertical cross-section, ignoring the shape of the lips. This may facilitate the acoustic simulations, but the effects on the radiation characteristics need to be examined in order to assess if it is a valid simplification. To this end, the vocal tract geometry of three vowels [a, i, u] (Aalto et al., 2014) were chosen, and the lips termination was replaced with a vertical cross-section as shown in Figure 5.1. Conducting Finite Element Method (FEM) simulations, and comparing the transfer functions for vocal tract geometries including and excluding the lip shape, the following conclusions were obtained: Removing the lips had the largest impact on vowel [a], and the entire frequency range was affected. Among the formant frequencies, the maximum deviation of $\approx 12\%$ was observed for the second formant, which may be perceptually relevant. Formant bandwidths were increased by removing the lips, $\approx 19\%$ for $B_2$ and $\approx 49\%$ for $B_5$. Furthermore, a level increase of up to $7 \, dB$ in high frequencies was observed, and it has been shown that such a difference in level is perceptually detectable (Monson et al., 2011, 2014a,b). A smaller impact was observed for vowel [i], and it was almost negligible for vowel [u].

Figure 5.1: Geometry simplification of vowel [i]
5.2 The vallecula and the piriform fossa

Removing these small cavities causes some dips in the frequency range $4 - 6 \, kHz$ to disappear. This result is consistent with previous studies (Takemoto et al., 2010, 2013; Vampola et al., 2015).

5.3 Slicing

To further simplify the 3D geometry, it is sliced into a set of cross-sections. A specific arrangement of the planes, namely a semi-polar grid is used, since it suits the vocal tract morphology (Dabbaghchian et al., 2015). That is, the slicing planes are approximately perpendicular to the vocal tract midline, which makes the cross-sections adequate for plane wave modeling. The use of slicing causes an upward shift of the formant frequencies compared to acoustic response from the full 3D geometry. The deviation is negatively correlated with the number of cross-sections, with larger deviations as the number of cross-sections decreases. This is consistent for all three examined vowels $[a, i, u]$.

5.4 Shape of the cross-section

The cross-sections of the vocal tract have irregular shapes. Traditionally, when using tube-models or the area function, it is however assumed that the cross-sections can be considered to be elliptical or, more frequently, circular. In order to investigate the influence of these assumptions, the shapes of the vocal tract cross-sections were converted into equivalent elliptical and circular shapes with the same area. In the case of the ellipse, the major axis was set to be equal to the lateral dimension of the original shape, and the minor was adjusted to maintain the original cross-sectional area. The transfer function is, as expected, not sensitive to the cross-sectional shape in the low frequency region, since plane wave propagation dominates. However, some deviations were observed for vowel $[a]$, which may be attributed to the large area discontinuities (Kang & Ji, 2008). In contrast, in the high frequency range, the transfer function is highly sensitive to the cross-sectional shape, and larger deviations due to shape simplification were observed. In general, the acoustic transfer function of the original vocal tract shape is better approximated by using elliptical shapes than circular ones.

5.5 Bending

In all examined vowels, straightening a geometry causes a downward shift for some formants regardless of the cross-sectional shape, and most of the antiresonances disappeared. This was particularly striking for the vowel $[a]$, which presents many high order modes. The use of circular cross-sections results in radial symmetry and smaller lateral dimension, which may prevent the onset of high order modes.
Furthermore, straightening of elliptical and circular cross-sections generates centric shapes while the bent configuration is essentially acentric (see Blandin et al., 2015, to compare centric and acentric).

5.6 Applications

Using a 3D multimodal approach (see, e.g. Blandin et al., 2015) may speed up the acoustic simulations in comparison with FEM-based methods. However, it may not be applicable for a complex vocal tract geometry, whereas straightened geometries make it possible to apply the multimodal approach (Arnela et al., 2016b).

A deformable vocal tract geometry, which is essential for acoustic simulations of dynamic sounds, is another application of simplified geometries. That is, an interpolation approach can be utilized to deform the vocal tract geometry of one articulation into that of another. Interpolation between the original complex geometries is not trivial and poses several challenges such as the possible occurrence of self-intersection of the surface mesh. Using the simplified geometries decreases the complexity of the interpolation and self-intersections can be prevented. Figure 5.2 shows an example of such an interpolation. This application has been presented in (Arnela et al., 2017, Paper C).

Furthermore, the simplification procedure and acoustic analysis of simplified geometries provided the basis for geometry reconstruction from a biomechanical model that is explained in Chapter 8.

5.7 Implications

This chapter has shown that although simplified representations of the vocal tract may be valid under some circumstances, all simplifications have an impact on the acoustics. Furthermore, simplified models may be efficient for simulations, but it comes at the cost of decreasing understanding of the actual mechanisms; such as anatomical constraints, muscular properties and control mechanisms; behind the speech production process. It may therefore be beneficial to consider a more complex biomechanical model, as outlined in Chapter 6.

Figure 5.2: The linear geometrical interpolation from vowel [a] to vowel [i].
Chapter 6

The biomechanical model

“Essentially, all models are wrong, but some are useful.”

Box & Draper (1987)

The biomechanical constraints of the speech apparatus play an essential role in shaping speech acoustics, and it is thus important to consider such constraints for high-fidelity modeling of speech production. A 3D biomechanical model of the upper airways in ArtiSynth (a biomechanical modeling platform, see Lloyd et al., 2012) was adapted to be used in this work. The adapted model is a combination of two available models in ArtiSynth, namely BadinJawHyoidTongue and FRANK, merged with some essential modifications. The BadinJawHyoidTongue model was used as the base model, but the missing structures were imported from FRANK. Two variations of the model adapted for the work in this thesis are currently available through the ArtiSynth repository, namely EunisonGenericModel and VocalTractCavity. The original models and their adaptation are described in this chapter.

6.1 Geometry of the anatomical structures

As explained in (Anderson et al., 2017), the geometry of different anatomical structures in FRANK model come from different sources (see Table 6.1). This causes some mismatch between different structures, since their size and anatomical shape may not fully correspond, and thus some modifications are needed for speech production modeling. Some of the geometries were modified in the adapted model in order to get adequate simulation results, to decrease the complexity, or to increase the stability of the simulations. The criteria for adequate simulation results were the ability of the model to achieve typical formant frequencies of the three cardinal vowels [a, i, u] (Dabbaghchian et al., 2016, Paper E), and that the shape of the vocal tract for the same vowels was reasonable in comparison with available articulation data, e.g., from MRI.
Table 6.1: Structures in the biomechanical model (Anderson et al., 2017)

<table>
<thead>
<tr>
<th>Structure</th>
<th>Source</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tongue</td>
<td>CT, MRI, cryo, female, 59 years old</td>
<td>Gérard et al. (2006)</td>
</tr>
<tr>
<td>Jaw, hyoid pharynx, skull</td>
<td>CT, male, 35 years old</td>
<td>Stavness et al. (2006)</td>
</tr>
<tr>
<td>Soft palate</td>
<td>MRI, male</td>
<td>Chen et al. (2012)</td>
</tr>
<tr>
<td>Larynx</td>
<td>CT, MRI, cryo, male, 38 years old</td>
<td>Moisik &amp; Gick (2017)</td>
</tr>
<tr>
<td>Lips (face)</td>
<td>CT, female</td>
<td>Nazari et al. (2010)</td>
</tr>
</tbody>
</table>

A. Pharynx and larynx

In FRANK, the inferior part of the pharynx geometry (in the laryngopharynx region) is oblique. As a result, a large cavity is formed in this region even when the tongue is in the most posterior position, and thus no proper constriction can be formed when producing the vowel [a]. Furthermore, compared to available data (Dang & Honda, 1997; Takemoto et al., 2010; Aalto et al., 2014) the piriform fossa become relatively large. In order to adjust for this, the lower part of the pharynx was modified to be more vertical, as shown in Figure 6.1. This modification causes the pharynx to partly overlap with the larynx in the model, and the larynx structure was therefore translated 7 mm towards the anterior to decrease the overlap. It was not possible to totally avoid overlap, as the size of the larynx structure is large relative to the other anatomical structures. The thyrohyoid membrane, which is a thin layer between the thyroid and hyoid bones, was created and added to the biomechanical model, as shown in Figure 6.1c, in order to be able to form a full vocal tract shape from the biomechanical model. The epiglottis was left out of the model, since we lack data on how to control it systematically.

![Figure 6.1](image)

Figure 6.1: Lateral view of the pharynx geometry in FRANK(a), and in the adapted model (b), (c) the thyrohyoid membrane and its attachment to the hyoid and thyroid bones.
Figure 6.2: Superior view of the velopharyngeal port, (a) when it is open, (b) when it is closed by the soft palate.

Figure 6.3: Posterior view of the oral cavity, (a) in FRANK without the face geometry, (b) in the adapted model with skin mesh to cover the gaps.

B. Soft palate

The soft palate geometry is the same as in FRANK, except that the uvula was slightly altered to avoid its collision with the tongue when producing a high-back vowel. In the neutral position, the velopharyngeal port is open (Figure 6.2a), allowing the air to enter the nasal cavity. For the simulations in this thesis, the port was always closed, by activating the muscles of the soft palate (Figure 6.2b).

C. Face and skin mesh

To decrease the complexity of the numerical simulations, the face mesh in FRANK was removed, only keeping the lips. Removing the face mesh signifies that the adapted model can not control rounding of the lips, it lacks the relevant muscles. The upper lip is attached to the maxilla and is static, while the lower lip is attached
to the mandible and follows its movement, which means that the lip opening may be altered, but not the rounding. This is naturally a caveat for the simulations of rounded vowels, but it was necessary to achieve stability in the simulations.

An internal surface of the face mesh was further cropped and modified to cover the gaps of the cheeks and the ramus of the mandible, as shown in Figure 6.3.

D. Mechanical properties

Table 6.2 summarizes the density of the structures and their attachments (Anderson et al., 2017). In the adapted model, the tongue, lower lip, mandible, and hyoid bone are dynamic while the other structures are static. The contact is modeled between the tongue and the maxilla, the mandible, the pharynx, and the soft palate. The tongue is controlled by the eleven muscles (see section 6.3). Two muscle groups control the jaw opening/closing. The jaw muscles in FRANK were adjusted using the work by Peck et al. (2000).

6.2 Muscle model

A muscle in ArtiSynth can be represented either as a volumetric finite element or as a spring-like point-to-point model. The latter is computationally efficient and more stable and the biomechanical model used in this work therefore utilizes the point-to-point muscle model. The equation to be used for the force calculation may vary depending on the material property of the point-to-point muscle. However,

Table 6.2: Properties and attachments of the structures in FRANK (Anderson et al., 2017) and the adapted model.

<table>
<thead>
<tr>
<th>Name</th>
<th>$\rho$ (kg/m$^3$)</th>
<th>Attachments</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxilla</td>
<td>1900</td>
<td>jaw, face, soft palate</td>
</tr>
<tr>
<td>jaw</td>
<td>3470</td>
<td>maxilla, face, tongue</td>
</tr>
<tr>
<td>hyoid</td>
<td>1900</td>
<td>tongue, larynx</td>
</tr>
<tr>
<td>thyroid</td>
<td>1900</td>
<td>larynx, pharynx</td>
</tr>
<tr>
<td>tongue$^2$</td>
<td>1040</td>
<td>hyoid, jaw, pharynx</td>
</tr>
<tr>
<td>soft palate$^2$</td>
<td>1040</td>
<td>maxilla, pharynx</td>
</tr>
<tr>
<td>pharynx$^3$</td>
<td>1040</td>
<td>thyroid, tongue, soft palate</td>
</tr>
</tbody>
</table>

$^1$Mooney-Rivlin material with $c_{01}, c_{11}, c_{02} = 0$ Pa, and bulk modulus $K = 10 \times c_{10}$, $(c_{10}, c_{20})$ increasing linearly from $(1037, 486)$ Pa at no activation to $(10370, 4860)$ Pa at full activation (Gerard et al., 2005).

$^2$In FRANK, the soft plate material is defined by Young’s modulus $E = 500$ Pa and Poisson ratio $\nu = 0.4995$. In the adapted model, it is treated as a rigid body.

$^3$In FRANK, the larynx material is defined by Mooney-Rivlin with $c_{01} = 2500, c_{20} = 1175, c_{10}, c_{11}, c_{02} = 0$ Pa, and the bulk modulus $K = 10 \times c_{01}$. In the adapted model, it is treated as a rigid body.
in general, the muscle force consists of passive \( f_p \) and active \( f_a \) components. Passive forces depend only on the muscle length, whereas the active forces are a function of length, \( l \), and its derivative, \( \dot{l} \). The muscle net force is then:

\[
f(e, l, \dot{l}) = f_p(l) + e f_a(l, \dot{l}) \tag{6.1}
\]

where \( e \) is the excitation level of the muscle in the range \([0, 1]\). Two different materials have been used in the model, namely ConstantAxialMuscle for the tongue muscles, and PeckAxialMuscle for the jaw muscles. ConstantAxialMuscle is implemented as:

\[
f = (e + k) f_{\text{max}} + d \times \dot{l} \tag{6.2}
\]

where \( k \) is a coefficient representing the passive fraction of the force, \( d \) is the damping factor, \( \dot{l} \) is the first derivative of the muscle length, \( f_{\text{max}} \) is the maximum force that can be generated by the muscle, and \( e \) is the excitation level in the range \([0, 1]\). In this work, the relationship has been simplified by setting \( k = d = 0 \), and thus only using \( e \) as the variable input. \( f_{\text{max}} \) was determined and then distributed among the different constitutive fibers based on the proportional volume of the surrounding elements (Buchallard et al., 2009). The force for PeckAxialMuscle material, used for the jaw muscles, is calculated as:

\[
f = (a \times e + p \times k) f_{\text{max}} + d \times \dot{l} \tag{6.3}
\]

where \( a \) and \( p \) are related to the active and the passive portion of the force.

\[
a = \begin{cases} 
0.5(l + \cos(2\pi l_n)), & 0.5 < l_n < 1.5 \\
0, & \text{otherwise}
\end{cases} \tag{6.4}
\]

\( l_n \) is the normalized length of the muscle and computed as:

\[
l_n = \frac{l - rl_0}{l_0 - rl_0} \tag{6.5}
\]

where \( l \) is the current length, \( l_0 \) is the length at rest, and \( r \) is the ratio of the muscle tendon. The passive portion, \( p \), is calculated using the following equation:

\[
p = \begin{cases} 
0, & l < l_m \\
\frac{l - l_0}{l_m - l_0}, & l_0 < l < l_m \\
1, & l > l_m
\end{cases} \tag{6.6}
\]

### 6.3 Tongue muscles

The tongue is the only deformable body in the adapted model; all other structures are treated as rigid. The tongue is controlled by eleven muscles, namely genioglossus posterior (GGP), genioglossus middle (GGM), genioglossus anterior (GGA), SG, HG, MH, GH, V, T, IL, SL. The fiber directions of these muscles are illustrated in Figure 6.4 and Figure 6.5, which indicate the tongue muscles, their fibers, and their response in positioning and deforming the tongue when activated independently.
Figure 6.4: Tongue muscles, their fibers, and their response in positioning and deforming the tongue when activated independently.
Figure 6.5: Tongue muscles, their fibers, and their response in positioning and deforming the tongue when activated independently.
Chapter 7

Estimation of muscle activation patterns

In a biomechanical model, MAP control the articulators, and hence the generated sound. However, MAP are, to a large extent, still unknown for any given articulation, as they are difficult to measure. EMG may in general be used to determine the activation of muscles, by measuring the electric potential generated by the muscle cells, when these are activated. However, for speech tasks, EMG is impractical for several reasons, as already mentioned in the Introduction: the technique relies on inserting needles into the muscles that are to be measured, which firstly would quite obviously interfere with the subject’s speech production, and is secondly vitiated by the uncertainty of measuring the intended muscle, because of the complex, interwoven structure of tongue muscles. Furthermore, the relation between the electrical pulses of the EMG signal and the contraction of muscles is not straightforward in itself, which introduces additional uncertainty.

EMG has nevertheless been used in two studies of speech production, according to the author’s knowledge. The first study is the work by MacNeilage & Sholes (1964) in which a surface EMG was employed to measure the tongue activation during production of [pvp] utterances. The second work was conducted by Baer et al. (1988) who employed EMG to measure extrinsic tongue muscles during production of [avvp] utterances. EMG can more readily be used to measure the tongue muscles for other tasks than speech production, if the task does not involve tongue motion (see, e.g., Sauerland & Mitchell, 1975; Pittman & Bailey, 2009).

Alternatively, the MAP may instead be estimated computationally for a given articulation or acoustic response, using an inverse method. Inverse methods can be explained by assuming a nonlinear function $f$ that maps input $x$ to output $y$ so that $y = f(x)$, and defining the problem of finding the inverse mapping so
that $x = f^{-1}(y)$. In general, two issues need to be addressed, namely the non-linearity of the forward function $f$, and the redundancy (i.e., the infinite number of solutions for the inverse mapping). In this case, a combinatorial search of different MAP combinations is performed, and the most suitable combination, given some criterion, is selected.

This chapter reviews two approaches utilized to estimate the MAP for a given articulation in section 7.1, and for a given acoustic in section 7.2. MAP estimation for a given articulation has been utilized in (Dabbaghchian et al., 2016, Paper E), and (Dabbaghchian et al., 2018b, Paper F). The work of estimating the MAP for a given acoustic output were inspired by the Directions Into Velocities of Articulators (DIVA) model (Nieto-Castanon et al., 2005; Guenther, 2016). The results on MAP estimation from acoustics presented in this chapter are preliminary and have not been published before.

### 7.1 Estimation based on articulation

In this method, MAP are estimated for a reference articulation so that feeding the estimated values to the biomechanical model generates an articulation similar to the reference. An already available inverse method in ArtiSynth (Anderson et al., 2017; Stavness et al., 2012) was employed. The method is based on the governing equations of the biomechanical model, which reads

\[
\dot{M}u = f_a(q, u, a) + f_p(q, u),
\]

\[
f_a(q, u, a) = a\Lambda(q, u),
\]

with variables standing for $M$: the mass matrix, $q$: positions, $u$: velocity, $\dot{u}$: time-derivative of the velocity, $a$: MAP, $f_a$: active forces, $f_p$: passive forces, $\Lambda$: nonlinear function relating position and velocity. To determine the MAP, $a$, a quadratic program, defined in (7.2), is solved to minimize the velocity tracking error at each time step,

\[
\arg\min_a (\|v - Ha\|^2 + \alpha\|a\|^2 + \beta\|\dot{a}\|^2),
\]

\[
\text{subject to } 0 \leq a \leq 1.
\]

The matrix $H$ represents all the biomechanical characteristics. The $l^2$-norm regularization coefficient, $\alpha$, solves the redundancy of the solution space and the damping factor, $\beta$, secures smooth transition of the solution by penalizing sudden jumps. We should note that the utilized approach leads to a suboptimal solution, since it considers only the current time step. Furthermore, it can not account for contact constraints since this requires the introduction of unilateral constraints in the equations, which leads to a mathematical programming problem with complementarity constraints.
A. EMA data

EMA data is used as the reference articulation data that is needed by the inverse method. Figure 7.1 depicts an example of EMA data for the utterance [aii] (Youssef, 2011). To use the EMA data as the reference articulation, a set of markers are positioned on the biomechanical model representing the EMA sensors as shown in Figure 7.2, with one marker attached to the lower incisor (LI), and three on the tongue surface: tongue tip (TT), tongue middle (TM), and tongue back (TB). Trajectories of the markers in the biomechanical model can be prescribed directly from EMA data, and this direct approach is used in Paper D (Dabbaghchian et al., 2014).

Figure 7.1: An example of EMA data for the utterance [aii].

Figure 7.2: Positions of four markers attached to the biomechanical model on the lower incisor (LI), tongue tip (TT), tongue middle (TM), and tongue back (TB).
B. EMA-like trajectories

For several reasons, the direct use of EMA data is however problematic: 1) the anatomical differences between the subject for whom the EMA data was collected and the one on which the biomechanical model was developed, including differences in tongue and palatal shapes, 2) uncertainties in positioning the sensors in the biomechanical model at the same anatomical landmarks as in the EMA data collection, 3) the possible inability of the tongue model to follow the complex trajectories that may occur in the EMA data.

As an alternative, an artificial EMA-like trajectories may therefore be generated by fitting a Bézier curve to a set of three points: \( \{P_{\text{org}}, P_{\text{des}}, P_{\text{mid}}\} \) as shown in Figure 7.3. \( P_{\text{mid}} \) is calculated using the following equations:

\[
\vec{P}_{\text{mid}} = \vec{P}_{\text{org}} + 0.5 \times \vec{P}_{\text{des}} + 0.25 \times d \times [\cos \theta \quad \sin \theta].
\]

This approach was used in F. Five artificial trajectories were generated to replicate the motion of EMA sensors from the neutral position to target positions, corresponding the five vowels \([a,e,i,u] \) (where \([u] \) is the unrounded vowel corresponding to \([u] \)). Using these trajectories as reference, the inverse method estimates the MAP, and feeding the estimated values to the forward model should generate similar trajectories as the references. The artificially designed trajectories and their corresponding forward model output is depicted in Figure 7.4. The model outputs reasonably follow the desired trajectories and reach the target positions. In the case of vowels, reaching the target position is more critical than replicating the trajectory. The latter is essential for dynamic sounds, such as vowel-vowel utterances, covered in Paper F (Dabbaghchian et al., 2018b).

Figure 7.3: An example of artificial EMA-like trajectory with its velocity profile.
7.2 Estimation based on acoustic

Estimation of MAP based on acoustic data may be favored over the corresponding articulation data, not only because of the availability of acoustic data, but also because of its similarity with the development of speech motor skills during childhood. Such an approach has been implemented in DIVA, a computer model that simulates the acquisition of speech motor skills. In DIVA, the vocal tract model is a 2D articulatory model represented by the midsagittal outline of the vocal tract. The method presented here has been inspired by the DIVA model, but estimates MAP for a given acoustic data in the 3D biomechanical model. The involved mathematics is reviewed and some examples are presented.

A. Mathematical formulation

Given the problem formulated above of finding the inverse mapping so that 

\[ x = f^{-1}(y), \]

the issue of non-linearity of \( f \) can be addressed by using the first-order approximation

\[ f(x + \Delta x) \approx f(x) + \frac{\partial f(x)}{\partial x} \Delta x = y + \Delta y, \]  

(7.4)

and

\[ \Delta y = \frac{\partial f(x)}{\partial x} \Delta x. \]  

(7.5)

The equation in the matrix form, assuming \( X \in R^n \) and \( Y \in R^m \), reads

\[ \Delta Y = J \Delta X, \]  

(7.6)
where \( J \) is the Jacobian matrix defined as
\[
J = \begin{pmatrix}
\frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\
\frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\
\cdots & \cdots & \cdots & \cdots \\
\frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n}
\end{pmatrix}_{M \times N}.
\]

(7.7)

The problem of interest is to find a solution of \( \Delta X \) for a given \( \Delta Y \) so that (7.6) is satisfied. In general, \( J \) is a rectangular matrix and thus not invertible. However, using the generalized inverse, the solution can be written as
\[
\Delta X = J^+ \Delta Y + (I - J^+ J)v,
\]
where \( J^+ \) is the generalized inverse of \( J \) satisfying the following condition,
\[
JJ^+ J = J.
\]

(7.9)

The second term in (7.8), \((I - J^+ J)v\), is known as the null space where \( v \in \mathbb{R}^n \) is an arbitrary vector, \( I \) is the identity matrix and the null projector matrix \( N \) is defined as
\[
N = (I - J^+ J).
\]

(7.10)

Using (7.10) and (7.9) it can be easily shown that:
\[
JNv = 0
\]

(7.11)

Indeed, the null space consists of all vectors with basis of \( Nv \) that do not affect the \( \Delta Y \) if they are added to the basic solution \( J^+ \Delta Y \). Using the null space projector matrix, (7.8) can be rewritten as
\[
\Delta X = J^+ \Delta Y + Nv.
\]

(7.12)

Assuming that \( J \) has full row rank, then \( J^+ \) is calculated as
\[
J^+ = J^T (JJ^T)^{-1}.
\]

(7.13)

This can be verified easily by entering it into (7.9).

Now, entering (7.13) in (7.12) and (7.10),
\[
\Delta X = J^T (JJ^T)^{-1} \Delta Y + Nv,
\]

(7.14)

\[
N = (I - J^T (JJ^T)^{-1} J).
\]

(7.15)
To improve the convergence, some variations of this method have been proposed including Levenberg method, which adds a damping or regularization term,

\[ \Delta X = J^T(JJ^T + \lambda I)^{-1} \Delta Y + Nv, \quad (7.16) \]
\[ N = (I - J^T(JJ^T + \lambda I)^{-1}J), \quad (7.17) \]

where \( \lambda \) is the regularization factor.

Another modification is the Levenberg-Marquardt method which scales the additional regularization term based on the gradient. Then the equation reads

\[ \Delta X = J^T(JJ^T + \lambda I)^{-1} \Delta Y + Nv, \quad (7.18) \]
\[ N = (I - J^T(JJ^T + \lambda \text{diag}(JJ^T))^{-1}J), \quad (7.19) \]

with \( \text{diag} \) keeping the diagonal elements and setting the other matrix elements to zero.

B. Examples

The method was applied to estimate MAP for formants corresponding to the cardinal vowels [a,i,u]. As the estimation approach is iterative, necessitating many simulations of the acoustic output, the 3D approach for acoustic simulations used in the remainder of this thesis is not practical and a 1D acoustic model, namely VTAR (Zhou et al., 2004; Espy-Wilson et al., 2007) is instead utilized to compute the formant frequencies. In order to overcome the complexity of the inverse modeling, only the tongue is dynamic in this simulations, and the jaw is fixed.

Figure 7.5 depicts the desired formant values and the model output. In general, the estimator is able to find the MAP that moves the articulators in the right direction to increase or decrease the formants. However, some difficulty to reach the target formants of vowel [i] can be observed. One possible explanation is that the tongue can not form a proper constriction because of the static jaw.

![Figure 7.5: Examples of desired formant values (blue line) and the corresponding model output (red line) corresponding to, (a) vowel [a], (b) vowel [i], (c) vowel [u].](image-url)
C. Discussion

Although the formant trajectories can be tracked by the DIVA model, only static vowels, with constant formants, are presented here. The reason for this is that the use of a 3D biomechanical model instead of the 2D articulatory model in DIVA poses some challenges for formant tracking. One of the challenges is the time delay caused by the mechanical inertia. In an articulatory model, the response of the forward model to a change in a control variable is immediately available while this is not the case in a biomechanical model. The second challenge is the complexity of the input-output relationship. The relation between the control variables and the formants in the DIVA model is approximately monolithic, making it easy to search the space in the inverse method. In a biomechanical model, however, the relation between MAP and formants is extremely complex because of the nonlinearity involved in coactivation of the muscles, and the indirect consequences of a muscle’s contraction. The third challenge is that some changes in the vocal tract 3D shape may affect the formants significantly. This was observed for the vowel [i] in which the motion of the tongue tip causes the sublingual cavity to be included or excluded as part of the vocal tract. This affects the effective length of the vocal tract and hence the formants.
Chapter 8

Vocal tract reconstruction

The previous chapters have described the biomechanical model and how it may be controlled with MAP. In order to use the model for acoustic simulations, the geometry of the vocal tract needs to be determined and since the vocal tract is a cavity rather than a physical object, it is not represented as an entity in a biomechanical model. Instead, its geometry is defined by the surrounding anatomical structures, i.e. the maxilla, the tongue, the pharynx, etc. Such an implicit representation of the geometry is suitable neither for visualization of the vocal tract shape nor for acoustic/fluid simulation. Acquiring an explicit vocal tract geometry requires addressing problems of the biomechanical model, including gaps between structures and non-aligned model boundaries. Some gaps are the result of missing structures, and adding the missing structures to the model will fix the problem, as described in chapter 6. However, non-aligned boundaries are inevitable since they are caused by the model development process that involves segmentation and registration of medical images. Mesh resolution and simulation of the contact between structures also result in similar artifacts. As an airtight geometry with regular elements (not elongated or squeezed), and without any self-intersection is needed for the acoustic/fluid simulation, a dedicated procedure is required in order to reconstruct the vocal tract geometry from the biomechanical model with structures that are not perfectly aligned. This may in fact be an important challenge, in particular when aiming for automatic methods.

This chapter addresses the problem of reconstructing the vocal tract geometry by blending the non-aligned geometries of the surrounding structures. Although the method was primarily developed to address the geometry reconstruction from a biomechanical model, it is applicable to any 3D articulatory model consisting of separate parts. It may also contribute to similar problems that arise in the geometry and image processing fields. The geometry reconstruction method has three vari-
lations, published in (Dabbaghchian et al., 2016, Paper E), (Dabbaghchian, 2018),
(Dabbaghchian et al., 2018a, Paper F), (Dabbaghchian et al., 2018b, Paper G).

Figure 8.1 illustrates the block diagram of the geometry reconstruction with an example. In the first step, to decrease the complexity of the problem, 3D geometries are converted to cross-sections. A set of planes, named slicing planes (Figure 8.1a), intersect all anatomical structures in the model (Figure 8.1b). The special arrangement of the slicing planes is chosen to be roughly perpendicular to the midline of the vocal tract. Figure 8.1c depicts four examples of the cross-sections belonging to 3D geometries.

Figure 8.1: Reconstruction of the vocal tract geometry.
Vocal tract reconstruction

to different regions from the laryngopharynx to the oral cavity. In the second step, the boundaries of the vocal tract on each cross-section is detected as shown by some examples in Figure 8.1d. Extracted boundaries are transformed back into 3D in the third step (Figure 8.1e). In the last step, a surface is reconstructed from the cross-section. Figure 8.1f shows an example of such a surface. Among these four steps, step one and three are using standard methods of the computational geometry and thus will not be discussed further. Step two, the boundary detection, and step four, surface reconstruction, are elaborated in sections 8.1, and 8.2 respectively.

8.1 Boundary detection

Figure 8.2 illustrates the boundary detection with a simple example in which a region is enclosed by two polygons. In this case, the boundary of the enclosed region is implicitly, but uniquely, defined by the polygons and a boundary detection, such as geometrical boolean (GB) operations, can convert such an implicit representation to an explicit one. The general problem is that the GB operations works well when the region is fully enclosed, which requires the boundaries of two polygons to be perfectly aligned. If there is a tiny gap, in the order of the computational precision, between non-aligned boundaries, GB fails to extract any boundary just because the region is not closed. Non-aligned boundaries are defined as boundaries with gaps or/and overlaps between them, as illustrated in the upper row of Figure 8.3. The lower row of Figure 8.3 illustrates unsuccessful boundary extraction of GB for problematic enclosed regions, including non-detection, tail-like corners or sliver polygons. Three different methods, of different detail and complexity, are proposed to address this problem.

Figure 8.2: An example of an enclosed region and its extracted boundary using the GB method.

Figure 8.3: Examples of non-aligned boundaries with gaps and overlaps. Applying the GB method results in $\varnothing$ (no boundary is identified), boundaries with tail-like corners, and with sliver polygons.
A. First approach: closing and filling

Since the GB method fails when there is a gap, a preprocessing method, named closing and filling (CF), may identify and close/fill the gaps before applying GB. The closing technique adds polygons to close the region of interest, as shown by the example in Figure 8.4. The location of the gap can be identified by searching for the smallest non-zero distance between the polygons of two non-aligned boundaries. This technique is useful for simple gap artifacts. In some cases, the combination of gaps and overlaps create a more complex situation. Then, another technique, filling, is useful to avoid forming of sliver polygons. When two boundaries intersect each other in more than one point, it is sufficient to identify the intersection points and modify the polygons of one of the boundaries to be aligned with those of the other one. Figure 8.5 shows an example of filling.

The question is whether these methods are applicable in more complex situations with several polygons involved. In fact, as long as the gaps are identified correctly, the closing and filling methods work well. However, it is not always trivial to identify gaps correctly. Gap identification is the most complex part of the reconstruction method, but using anatomical knowledge helps to develop an hierarchical decision making algorithm when searching for gaps. This method does not intrinsically limit the boundary detection to one polygon (as an enclosed region may consists of several polygons), the current implementation only considers the main region of interest to avoid additional complexity. As a result, this method cannot reconstruct sub-branches (piriform fossa, sub-lingual cavity) when applied to the vocal tract problem. This method was utilized in (Dabbaghchian et al., 2016, Paper E) to reconstruct the vocal tract geometry of three corner vowels [a,i,u].

Figure 8.4: Closing the gap between two polygons by using two small black rectangles that results in an enclosed region.

Figure 8.5: Filling the gap between two polygons prevents the generation of sliver polygons.
B. Second approach: line of sight

The complexity of the gap searching algorithm substantially increases as the number of polygons increases, making the closing and filling algorithm computationally inefficient. It is also sensitive to the orientation of the slicing planes. The algorithm needs to be aware of all possible gaps and overlaps between structures, and it must hence be adapted if a new structure is introduced or the orientation of the slicing planes is changed. This drawback may prevent the extension of the method to time-varying geometries. The second variation of the extraction method is line of sight (LoS), which is an efficient and stable approach. The idea of LoS method is relatively simple, and is illustrated by an example in Figure 8.6. In this method, there is a viewpoint, approximately in the center of the region, and the boundary of the region consists of all the points on the surrounding structures that can be seen from the viewpoint. The accuracy of the boundary is directly determined by the number of rays emanating from the viewpoint; the larger the number, the more details are captured. One advantage of the LoS method is that slivers and tail-like corners are avoided naturally, as shown in Figure 8.7. However, the method fails to detect regions behind a bulge. Such a situation is very rare when the main cavity (excluding subbranches) of the vocal tract is considered. This method was used to reconstruct the vocal tract geometry both for static and time-varying geometries presented in (Dabbaghchian et al., 2017), (Dabbaghchian et al., 2018b, Paper F).

![Figure 8.6](image1)

Figure 8.6: Boundary of a region is detected using the LoS method.

![Figure 8.7](image2)

Figure 8.7: Examples of non-aligned boundaries with gaps and overlaps. Applying the LoS method detects the boundary of the region of interest despite the existing gaps, and avoids generation of tail-like corners or sliver polygons.
Chapter 8

Figure 8.8: Examples of enclosed regions with complex shapes.

Figure 8.9: Illustration of the GC method, (a) with one circle, (b) with an example with concave region in which the boundary is detected by using three circles.

C. Third approach: Growing circles

Both previous methods fail to reconstruct the sub-branches of the vocal tract. To include the sub-branches, the boundary detection method requires dealing with many complex situations, as depicted in Figure 8.8. The growing circle (GC) method has been developed for such complex cases. The basic idea of the method is illustrated by Figure 8.9. A circle at the center (labeled 1) starts to grow by increasing its radius (labeled 2), but its growth stops whenever it touches a boundary (labeled 3). The method with one circle is similar to the LoS method, and it cannot deal with complex situations. However, the idea can be easily extended to use more circles as shown by the example in Figure 8.9(b), where three circles can detect the boundary of the region. Each circle detects part of the region and the whole region is determined by the union of partial regions. Boundary detection based on the GC method can identify boundaries very accurately. The details of the method have been described in (Dabbaghchian, 2018), and (Dabbaghchian et al., 2018a, Paper G).

8.2 Surface reconstruction

Once the boundaries of the vocal tract on each cross-section are determined, its surface can be reconstructed. Two methods have been developed for surface reconstruction. The first method is useful for tube-like geometries, and it cannot reconstruct the vocal tract with sub-branches. In this method, each polygon is resampled to have the same number of vertices and then the corresponding vertices are connected to create a triangulated surface, as shown in Figure 8.10. The resampling process not only generates the same number of vertices but also considers the orientation of the vertices (clock-wise or counter-clock-wise). It also supports
Vocal tract reconstruction

Figure 8.10: An example of using the first method for surface reconstruction.

vertex constraints, which helps avoiding self-intersections and elongated triangles. One may, e.g., define a constraint for some of the vertices to lie on a specific line. This surface reconstruction method can be combined with any boundary detection method that produces one polygon for each cross-section, and it was utilized in (Dabbaghchian et al., 2016, Paper E) in combination with the CF method and in (Dabbaghchian et al., 2017), (Dabbaghchian et al., 2018b, Paper F) in combination with the LoS method.

The second method takes into account sub-branches, and can deal with significant changes from one cross-section to the next. Briefly explained, each cross-section may have an arbitrary number of polygons, and some polygons may even have holes. In such a situation, two problems need to be addressed: finding the correspondence between the polygons of different cross-sections, and triangulation of the corresponding polygons (Bajaj et al., 1996; Zou et al., 2015). To find the correspondence, three consecutive cross-sections are analyzed, and the polygons of each cross-section is partitioned into three parts, one part each to be used for connecting to the previous and next cross-section, and one part belonging to the cross-section itself. After partitioning, a Delaunay triangulation creates a surface between the corresponding partitions. Figure 8.11 illustrates these steps with an example. In Figure 8.11, three consecutive cross-sections have been marked with $C_{i-1}$, $C_i$, and $C_{i+1}$. The partitioning step, based on the $C_i$ as the middle cross-section, generates three set of polygons $C_i^p$, $C_i^r$, and $C_i^n$. In the next step, the polygons are paired and surfaces are reconstructed. $C_i^r$ does not have any pair, and is triangulated on the cross-section without making any surface with the previous or the next cross-section. Full explanation and details of the method can be found in (Dabbaghchian et al., 2018b, Paper G), where the this surface reconstruction method has been used in connection with the GC method (for boundary detection) to reconstruct the vocal tract geometry with sub-branches.
Chapter 8

Figure 8.11: An example of using the second method for surface reconstruction. $C_{i-1}$, $C_i$, and $C_{i+1}$ are three consecutive cross-sections. Two steps of the method, namely the partitioning and the triangulation are illustrated.
Chapter 9

Acoustic modeling

Having obtained the geometry of the vocal tract, using the methods outlined in chapter 8, one may use it for acoustic simulations, as described in the following two chapters. The present chapter deals with the fundamentals of acoustic modeling, whereas chapter 10 describes the acoustic synthesis experiments performed within the scope of this thesis. The physical phenomena involved in the generation of vowel sounds, in which the vocal tract is relatively open, may be defined by the acoustic wave equation. Except in a few simple cases, there is no closed form solution to the wave equation, and finding the solution requires employing numerical methods. Using the scalar wave equation, the basic acoustic characteristics of a uniform tube is derived in section 9.1. Section 9.2 provides details of the FEM utilized to solve the 3D wave equation with specified boundary conditions for static vocal tract geometries. The numerical approach to address dynamic geometries, with moving boundaries, is described in section 9.3.

9.1 The scalar wave equation

The scalar wave equation considers the wave propagation only in one dimension. This can be considered as plane wave propagation in 3D, and is a reasonable approximation under some conditions. In this section, the scalar wave equation is used to derive basic acoustic characteristics of a uniform tube. Starting with the equation

$$\frac{\partial^2 p(x, t)}{\partial t^2} = c^2 \frac{\partial^2 p(x, t)}{\partial x^2}, \quad (9.1)$$

where $c$ is the speed of the wave in the media ($\approx 350 \text{ m/s}$ in the case of air), and $x$ is a scalar, the general solution is expressed as

$$p(x, t) = f(x + ct) + g(x - ct), \quad (9.2)$$
where $f$ and $g$ are arbitrary well-behaved functions, i.e., that the second order derivative exists. This shows that $f$ and $g$ are respectively traveling in the $-x$ and $+x$ directions with the speed of the wave ($c$) in the media. The general solution can be validated by entering (9.2) in (9.1) and using the chain rule,

$$\frac{\partial^2 p(x,t)}{\partial t^2} = \frac{\partial}{\partial t} \frac{\partial}{\partial t} [f(x + ct) + g(x - ct)]$$

$$= \frac{\partial}{\partial t} (cf'(x + ct) - cg'(x - ct))$$

$$= c^2 f''(x + ct) + c^2 g''(x + ct)$$

$$= c^2 p''(x, t). \quad (9.3)$$

Similarly,

$$\frac{\partial^2 p(x,t)}{\partial x^2} = \frac{\partial}{\partial x} \frac{\partial}{\partial x} [f(x + ct) + g(x - ct)]$$

$$= \frac{\partial}{\partial x} (f'(x + ct) + g'(x - ct))$$

$$= f''(x + ct) + g''(x + ct)$$

$$= p''(x, t). \quad (9.4)$$

Comparison of (9.3) and (9.4) proves that (9.2) satisfies the wave equation.

More specific solutions can be found by imposing initial values and boundary conditions. In the following, examples of a uniform tube with different boundary conditions are analyzed to derive their resonance frequencies.

A. Closed-closed tube

As the first example, a uniform tube of length $L$, and closed at both ends, is analyzed. This imposes the boundary conditions

$$p(0,t) = 0, \quad (9.5a)$$

$$p(L,t) = 0. \quad (9.5b)$$

Substituting (9.5a) in (9.2),

$$p(0,t) = f(ct) + g(-ct) = 0 \Rightarrow f(ct) = -g(-ct).$$
Since this equation should be valid for all \( t \), a new variable \( z \) is defined as \( z = ct \) which results in

\[ f(z) = -g(-z). \tag{9.6} \]

Now, the other boundary condition as expressed in (9.5b) can be applied,

\[ p(L, t) = f(L + ct) + g(L - ct) = 0 \Rightarrow f(L + ct) = -g(L - ct). \]

By defining a new variable \( z \) as \( z = L + ct \),

\[ f(z) = -g(2L - z), \tag{9.7} \]

and considering equations (9.6), and (9.7),

\[ g(z) = g(z + 2L). \]

This means that \( g \) is a periodic function with period \( 2L \), and this implies \( f \) to be periodic as well with the same period. Consequently, \( p \) will be periodic

\[ p(x) = p(x + 2L). \]

This shows that by imposing the boundary conditions expressed in (9.5), traveling waves (\( f \) and \( g \)) are not independent anymore, and both are periodic with period \( 2L \). In other words, they form a standing waveform with wavelength of \( 2L \). The resonance frequencies of the closed-closed tube can hence be determined as

\[ f_n = \frac{n c}{\lambda}, \forall n \in (1, 2, 3, ...) \]

and considering \( \lambda = 2L \),

\[ f_n = \frac{n c}{2L}, \forall n \in (1, 2, 3, ...). \tag{9.8} \]

**B. Closed-open tube**

The example when one end of the tube is open can be analyzed subject to the boundary conditions

\[ p(0, t) = 0, \tag{9.9a} \]

\[ \frac{\partial p}{\partial x}(L, t) = 0. \tag{9.9b} \]

Since the tube is closed at \( x = 0 \), applying the boundary condition of (9.9a) results in the same equation as (9.6), and taking the derivative gives us

\[ f'(z) = g'(-z). \tag{9.10} \]
To apply the other boundary condition expressed in (9.9b), the first order spatial derivative is calculated as
\[
\frac{\partial p(x,t)}{\partial x} = f'(x+ct)+g'(x-ct),
\]
now applying (9.9b),
\[
\frac{\partial p}{\partial x}(L,t) = f'(L+ct)+g'(L-ct) = 0 \Rightarrow f'(L+ct) = -g'(L-ct),
\]
with a variable change as \( z = L+ct \),
\[
f'(z) = -g'(2L-z), \tag{9.11}
\]
and considering equations (9.10), and (9.11),
\[
g(z) = -g(z+2L).
\]
So, in this case, \( f \) and \( g \) are odd functions with period \( 4L \), thus their sum \( (p) \) will be an odd periodic function with period \( 4L \). Being an odd function, the even harmonics will be zero, and thus the resonance frequencies of the close-open tube can be determined as:
\[
f_n = (2n-1) \frac{c}{\lambda}, \quad \forall n \in (1, 2, 3, \ldots),
\]
considering \( \lambda = 4L \),
\[
f_n = (2n-1) \frac{c}{4L}, \quad \forall n \in (1, 2, 3, \ldots). \tag{9.12}
\]
The resonance frequencies of an open-open tube are the same as the closed-closed tube, and can be derived in the same way. It is important to note that the resonance frequencies calculated in equations (9.8) and (9.12) are only subject to the boundary conditions and independent of initial values.

### 9.2 Three-dimensional wave equation

Finding the solution of the wave equation with a complex computational domain requires employing a numerical approach. This section describes the Finite Element Method, FEM, to solve the wave equation in a 3D domain, which is expressed as
\[
c^2 \nabla^2 p = \frac{\partial^2 p}{\partial t^2}, \quad \text{in } \Omega, \tag{9.13}
\]
with
\[
\nabla^2 p = \frac{\partial^2 p}{\partial x^2} = \frac{\partial^2 p}{\partial x_1^2} + \frac{\partial^2 p}{\partial x_2^2} + \frac{\partial^2 p}{\partial x_3^2}.
\]
To get a discrete equation in the time domain, the partial time derivative in (9.13) is approximated using finite differences as
\[ \frac{\partial^2 p}{\partial t^2} \approx \frac{p - 2p^1 + p^0}{\Delta t^2} \]
where \( p^1 \) and \( p^0 \) express the value of the solution at one and two time steps back, respectively, and \( \Delta t \) defines the time step. Doing so, the wave equation reads
\[ c^2 \Delta t^2 \nabla^2 p = p - 2p^1 + p^0, \]
multiplying by a test function and integration over the domain \( \int_\Omega \), we get
\[ \int_\Omega c^2 \Delta t^2 \nabla^2 p v d\mathbf{x} = \int_\Omega \left( p - 2p^1 + p^0 \right) v d\mathbf{x}, \]  
(9.14)
and using Green’s first identity for integral by parts, the left hand side of (9.14) can be written as
\[ \int_\Omega \nabla^2 p v d\mathbf{x} = -\int_\Omega \nabla p \nabla v d\mathbf{x} + \int_{\partial\Omega} (\nabla p \cdot \mathbf{n}) v d\mathbf{s}, \]  
(9.15)
where \( \partial\Omega \) describes the surface of the domain, and \( \nabla p \cdot \mathbf{n} \) expresses the gradient of the solution in the normal direction \( \mathbf{n} \) of the surface. Entering (9.15) in (9.14) and rearranging the equation, the variational formulation reads as
\[ \int_\Omega (pv + c^2 \Delta t^2 \nabla p \nabla v) d\mathbf{x} = \int_\Omega (2p^1 - p^0) v d\mathbf{x} + c^2 \Delta t^2 \int_{\partial\Omega} (\nabla p \cdot \mathbf{n}) v d\mathbf{s}. \]  
(9.16)
To further proceed, we need to specify the boundary conditions. Two different examples of boundary conditions are provided.

A. Boundaries without radiation and with lossless walls

Figure 9.1a depicts the vocal tract geometry and its computational domain. The domain is split into three parts subject to different boundary conditions. The mouth exit, \( \Gamma_M \), is closed by a baffle, assuming that acoustic pressure is zero on the baffle. This simple boundary condition cannot account for the acoustic radiation from the lips. The vocal tract walls, \( \Gamma_W \), which are assumed to be lossless, and the glottis, \( \Gamma_G \), has a Gaussian inflow. The boundary conditions in this case reads
\[ \nabla p \cdot \mathbf{n} = g \quad \text{on } \Gamma_G, \]  
(9.17a)
\[ \nabla p \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_W, \]  
(9.17b)
\[ p = 0 \quad \text{on } \Gamma_M. \]  
(9.17c)
In this example \( \partial\Omega = \{\Gamma_G, \Gamma_W, \Gamma_M\} \), so the integral over the domain surface in (9.16) is expanded as
\[ \int_{\partial\Omega} (\nabla p \cdot \mathbf{n}) v d\mathbf{s} = \int_{\Gamma_G} (\nabla p \cdot \mathbf{n}) v d\mathbf{s} + \int_{\Gamma_W} (\nabla p \cdot \mathbf{n}) v d\mathbf{s} + \int_{\Gamma_M} (\nabla p \cdot \mathbf{n}) v d\mathbf{s}, \]  
(9.18)
and applying the boundary conditions specified in (9.17) yields to
\[
\int_{\partial \Omega} (\nabla p \cdot n) vds = \int_{\Gamma_G} g vds + \int_{\Gamma_W} 0 vds + \int_{\Gamma_M} 0 vds = \int_{\Gamma_G} g vds, \tag{9.19}
\]
substituting (9.19) in the variational formulation (9.16) gives
\[
\int_{\Omega} (pv + c^2 \Delta t^2 \nabla p \nabla v) dx = \int_{\Omega} (2p^1 - p^0)v dx + c^2 \Delta t^2 \int_{\Gamma_G} g v ds. \tag{9.20}
\]
Equation (9.20), derived by applying the boundary conditions in the variational formulation, is converted to a system of equations by substituting different test functions and the solution is calculated by solving the system of equations.

**B. Boundaries with radiation and wall loss**

The boundary conditions of (9.17) can not account for the lips’ radiation and the loss at the vocal tract walls. To address these issues, a different computational domain, shown in Figure 9.1b, is used with the boundary conditions

\[
\nabla p \cdot n = g \quad \text{on } \Gamma_G, \tag{9.21a}
\]
\[
\nabla p \cdot n = p/Z_w \quad \text{on } \Gamma_W, \tag{9.21b}
\]
\[
\nabla p \cdot n = 0 \quad \text{on } \Gamma_H, \tag{9.21c}
\]
\[
\nabla p \cdot n = p/Z_0 \quad \text{on } \Gamma_\infty, \tag{9.21d}
\]

where (9.21b) is imposed to replicate the loss of the vocal tract walls, while equations (9.21c), and (9.21d) are used to account for radiation. The variational formulation can be derived in the same way as in the previous example.
C. Transfer function

The vocal tract acoustic transfer function, $H(f)$, is defined as

$$H(f) = \frac{P_o(f)}{Q_i(f)},$$

(9.22)

where $P_o(f)$ and $Q_i(f)$ are the Fourier transform of the acoustic pressure at the mouth exit ($p_o(t)$ in Figure 9.1b) and input volume velocity at the glottis. To determine the transfer function, a Gaussian pulse is imposed at the glottal cross-section, $\Gamma_G$, and the acoustic pressure $p_o(t)$ is captured. Takemoto et al. (2010) proposed the Gaussian inflow as

$$g(n) = \exp \left[ - \left( \frac{n\Delta t - T_{gp}}{0.29T_{gp}} \right)^2 \right],$$

(9.23)

where $T_{gp} = 64.6 \mu s$ and $\Delta t$ is the time step of the simulation. Figure 9.2 depicts the Gaussian pulse in the time (a) and frequency (b) domains. As expected, its spectrum is similar to a white noise, with 3 dB change in the whole frequency range $[0 10]$ kHz. Figure 9.2c depicts an example of the transfer function.

Figure 9.2: (a) Gaussian pulse used as the boundary condition ($g$) at the glottal cross-section ($\Gamma_G$), and (b) its spectrum, (c) an example of the acoustic transfer function.
D. Sound synthesis

To generate the corresponding sound of the vocal tract geometry, a train of glottal pulses, shown in Figure 9.2, is convolved with the inverse Fourier transform of the transfer function. That is,

\[ h(t) = F^{-1}\{H(f)\}, \]
\[ s(t) = h(t) \ast g(t). \]

Figure 9.3 shows an example of the glottal pulse train and the synthesized sound.

9.3 Moving boundaries

To deal with moving boundaries, the wave equation is expressed in ALE mixed form,

\[ \frac{1}{\rho c^2} \partial_t p - \frac{1}{\rho c^2} u_{\text{dom}} \cdot \nabla p + \nabla \cdot u = 0, \quad (9.24a) \]
\[ \rho \partial_t u - \rho u_{\text{dom}} \cdot \nabla u + \nabla p = 0, \quad (9.24b) \]

where \( c \) stands for the speed of sound, \( \rho \) for the air density, \( u_{\text{dom}} \) for the velocity of the domain (i.e. the vocal tract), and \( \partial_t \) for the first partial time derivative.

The ALE mixed wave equation (9.24) may be supplemented with the following boundary and initial conditions

\[ u \cdot n = g(t) \quad \text{on } \Gamma_G, t > 0, \quad (9.25a) \]
\[ u \cdot n = p/Z_w \quad \text{on } \Gamma_W, t > 0, \quad (9.25b) \]
\[ p = 0 \quad \text{on } \Gamma_M, t > 0, \quad (9.25c) \]
\[ p = 0, u = 0 \quad \text{in } \Omega, t = 0. \quad (9.25d) \]
Since the acoustic transfer function changes over time, it is not possible to use the convolution method for sound synthesis. Instead, the acoustic pressure $p_o(t)$ is captured during the simulations. Although this method can also be used in the case of static geometries, it is computationally inefficient, since it requires the simulations to run for the whole period.

In contrast to static geometries, the computational domain changes according to the vocal tract deformation, i.e. $u_{dom}$. This adds an extra complexity to the problem. That is, the geometry reconstruction presented in Chapter 8 addresses the deformation of the vocal tract walls. Yet, one needs to deform the computational domain (i.e., the volume mesh) based on the motion of the walls (i.e., the surface mesh). A Laplacian equation for the node displacement $w$ may be solved numerically to compute the coordination of the inner vertices of the domain. Such an equation reads

$$\nabla^2 w^n = 0 \quad \text{in } \Omega, \ t = t^n, \quad (9.26a)$$

$$w^n = x^n_w - x^{n-1}_w \quad \text{on } \Gamma_W, \ t = t^n, \quad (9.26b)$$

$$w^n \cdot n = 0 \quad \text{on } \Gamma_G, \ t = t^n, \quad (9.26c)$$

$$w^n \cdot n = 0 \quad \text{on } \Gamma_M, \ t = t^n, \quad (9.26d)$$

where $x_w$ refers to the coordinates of those vertices on $\Gamma_W$, and the superscript $n$ denotes the time step. Equation (9.26b) prescribes the motion of the boundary vertices while (9.26c) and (9.26d) set the displacement to zero in the normal direction $n$ (pointing outwards) to avoid an artificial lengthening of the vocal tract. Solving this equation smoothly translates the motion of the boundary vertices to the inner ones.

Deformation of a computational domain is one of the challenges of mesh-based approaches. Although the general solution is remeshing, it is computationally expensive not only because it needs a new volume mesh to be generated but also that the solution of the old mesh should be projected onto the new mesh. Solving the Laplacian equation (9.26a) helps to avoid the involved complexities of the remeshing. However, it requires the computational domain to be simple. For example, the computational domain of Figure 9.1b may not be addressed by this approach. It may also not be suitable for geometries with sub-branches, if the deformation of the sub-branch may change the topology of the domain. Such examples of sub-branches are the sublingual cavity, interdental space, or the vallecula, which may appear or disappear due to the movement of the tongue.

All 3D acoustic simulations were conducted in La Salle, Universitat Ramon Llull in Barcelona, Spain, by Marc Arnela and Oriol Guasch. For further details, see Arnela (2014); Arnela & Guasch (2013); Guasch et al. (2016).
Chapter 10

Sound synthesis

Physics-based speech synthesis aims to produce natural sounds. Such approach requires the vocal tract geometry. Such a geometry, as already described in Chapter 1, may be acquired either by direct imaging or utilizing a biomechanical model. The current MRI technology allows for accurate reconstruction of the static vocal tract geometry (Aalto et al., 2014). However, it is not practical for the dynamic vocal tract, since rtMRI does not provide enough spatio-temporal resolution (see Fu et al., 2017; Lingala et al., 2017, for the most recent achievements in rtMRI). To simulate dynamic sounds, two approaches have been proposed in this thesis work. The first approach is to interpolate between two static geometries to simulate the dynamic vocal tract (Arnela et al., 2017, Paper C), (Gully et al., 2018). The other alternative is to utilize the biomechanical model and reconstruct the vocal tract geometry at each time step of the simulations. The difference between the two alternatives is that the former approach employs a linear interpolation, which may not account for the mechanical constraints of the articulators, whereas the latter incorporates them to the extent that they are included in the model. A key question regarding the validity of the two approaches is hence whether such constraints are important, and if they influence the synthesized sound.

This chapter explains the synthesis of static vowels\(^1\) and vowel-vowel sequences, as an application of the developed 3D biomechanical-acoustic model. Then the influence of mechanical properties of the articulators is investigated by analyzing the motion of the jaw and the tongue, and comparing the spectrograms of the same sound sequence generated with geometrical interpolation and with the biomechanical model. The details of the study are presented in (Dabbaghchian et al., 2018b, Paper F).

\(^1\)Also known as sustained vowels or point vowels.
10.1 Vowel sounds

Using the 3D biomechanical-acoustic model, the six vowels [a, o, e, i, u] were synthesized. As explained in Chapter 7, an artificial trajectory is designed for each vowel to move markers on articulatory fleshpoints to the target positions and the corresponding MAP are determined through inverse modeling. The MAP are then fed to the biomechanical model (see Chapter 6), causing the articulators to move to the target positions. After reaching the equilibrium, the geometry reconstruction method, presented in Chapter 8, generates the geometry of the vocal tract which is then used for the acoustic simulations detailed in Chapter 9. Figure 10.1 depicts the articulation of the six vowels and the resulting formant frequencies and bandwidths are reported in Table 10.1.

Table 10.1: First five formant frequencies and bandwidths for vowels.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Formant frequencies (Hz)</th>
<th>Formant bandwidths (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F_1$</td>
<td>$F_2$</td>
</tr>
<tr>
<td>[a]</td>
<td>664</td>
<td>1134</td>
</tr>
<tr>
<td>[o]</td>
<td>521</td>
<td>1666</td>
</tr>
<tr>
<td>[e]</td>
<td>513</td>
<td>1763</td>
</tr>
<tr>
<td>[ɛ]</td>
<td>463</td>
<td>1916</td>
</tr>
<tr>
<td>[i]</td>
<td>300</td>
<td>2116</td>
</tr>
<tr>
<td>[u]</td>
<td>386</td>
<td>1362</td>
</tr>
</tbody>
</table>
10.2 Vowel-vowel sounds

The MAP of vowels are interpolated linearly to generate the MAP for vowel-vowel sounds. For example, if $M_1$ and $M_2$ are the MAP for vowels $V_1$ and $V_2$ respectively, then linear interpolation in the muscle activation space from $M_1$ to $M_2$ is used as the MAP for the utterance $V_1V_2$ (see Paper F, or Dabbaghchian et al., 2017). Feeding the interpolated MAP, moves the articulators, and the vocal tract geometry is reconstructed at each time step, which is then used in the acoustic solver. The tree utterances [ai], [au], and [ui] were synthesized and their spectrograms are depicted in Figure 10.2.

![Spectrograms of three vowel-vowel utterances](image)

Figure 10.2: Spectrogram of three vowel-vowel utterances synthesized using the 3D biomechanical-acoustic model (colorbar is in dB, image by M. Arnela and O. Guasch).
10.3 Influence of mechanical properties

The advantage of using the biomechanical model is that the influence of mechanical properties on the synthesized sound is simulated, which is not the case in articulatory models. Two observations were made regarding such an influence.

Analyzing the motion of the markers attached to the jaw and the tongue, loop-like trajectories were observed as shown in Figure 10.3. A similar observation has been reported using a 2D biomechanical model (Perrier et al., 2003). This observation is important since it suggests that such loop-like trajectories observed in EMA data (Mooshammer et al., 1995) may be the result of mechanical properties of the articulators, rather than an active control mechanism.

To analyze the influence of the mechanical properties on the spectrogram, the utterance \([\text{ai}]\) was synthesized using both the linear geometrical interpolation method (Arnela et al., 2017, Paper C) and the MAP interpolation in the biomechanical model. In the first case, linear geometrical interpolation is performed between the two static geometries of the vowels ([a] and [i]). One may use the geometries of these vowels generated in section 10.1, but in order to compare the results with the second method (interpolation in the MAP space), a specific method was employed: the same time-sequence as for the reconstructed geometries in section 10.2 is used, but the geometries between \(t_1 = 120\) ms and \(t_2 = 230\) ms were replaced with the ones that are generated by linear geometrical interpolation between the geometries at time instants \(t_1\) and \(t_2\). In other words, the geometries of both methods are identical before \(t_1\) and after \(t_2\), and the difference between the geometries during the time interval \([t_1,t_2]\) is due to mechanical constraints of the articulators being reflected in the geometries of the MAP-interpolation method. Synthesizing the spectrogram of the utterance \([\text{ai}]\) revealed how mechanical constraints of the articulators influence the spectrogram, as shown in Figure 10.4.

As expected, the largest differences are produced between 120 ms and 230 ms (the deviations at high frequencies at \(t > 250\) ms are due to numerical errors caused by the very low pressure levels in this frequency region). The differences between

![Figure 10.3: Loop-like trajectories observed in biomechanical simulation of the utterances.](image)
spectrograms, which can be observed in all frequencies, are less likely to change the perception of the vowel category in a vowel-vowel utterance. However, such differences may still be perceptually relevant. That is, the slope of the second formant, $F_2$, which is the most affected one (see Figure 10.4c), is known to be an important perceptual cue for consonants in CV utterances (Sussman & Shore, 1996; Lindblom & Sussman, 2012). A perceptual test with more variations of utterances is beneficial to fully understand the perceptual influence of the mechanical properties of the articulators. Furthermore, in a biomechanical model, coarticulation is implicitly addressed because of the mechanical inertia. That is, the speech rate affects the target formants and transitions in a nonlinear way, whereas, in a geometrical interpolation, changing the timing of the interpolation only scales the time axis, and not the target formants.
Chapter 11

On the quantal theory

The high frequency of the three cardinal vowels \([a, i, u]\) in the world’s languages (see Table 11.1), and significantly lower frequency of other vowels raises the question of the reason of selectivity in the vowel inventories. Two different hypothesis have been proposed, namely the dispersion theory (Liljencrants & Lindblom, 1972; Lindblom, 1986) and the quantal theory (Stevens, 1989; Perkell, 1996). Both theories predict the preference of the cardinal vowels, though for different reasons.

According to the dispersion theory, the preferred vowels are the ones, among all possible produced by the vocal tract, which have maximum perceptual discrimination. That is, if a language uses only three vowels, the three cardinal vowels are preferred, since they can be produced with the largest possible discrimination.

The quantal theory, on the other hand, argues that the preferred vowels are the ones that can be the most reliably produced. In other words, the vowels for which the acoustic output has the smallest possible sensitivity are preferred. It should be noted that the dispersion theory considers both production and perception capabilities, whereas the quantal theory only considers production capabilities.

Table 11.1: Statistics of vowels in the world’s languages (451 languages, Ref: UCLA Phonological Segment Inventory Database)

<table>
<thead>
<tr>
<th>IPA</th>
<th>Description</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>high front unrounded</td>
<td>87.14</td>
</tr>
<tr>
<td>a</td>
<td>low central unrounded</td>
<td>86.92</td>
</tr>
<tr>
<td>u</td>
<td>high back rounded</td>
<td>81.82</td>
</tr>
<tr>
<td>e</td>
<td>mid front unrounded</td>
<td>41.24</td>
</tr>
<tr>
<td>o</td>
<td>mid back rounded</td>
<td>40.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of vowels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min 2</td>
</tr>
<tr>
<td>Max 31</td>
</tr>
<tr>
<td>Typical 5</td>
</tr>
</tbody>
</table>
Using the biomechanical model and the vocal tract geometry extraction method, described in the preceding chapters, this chapter provides some insights into the quantal theory, the role of biomechanical constraints in producing vowel sounds, and its connection to speech motor control. The formant frequencies computed in this chapter uses a 1D acoustic software, namely VTAR (Espy-Wilson et al., 2007). This study has been presented in 7th international conference on speech motor control (Dabbaghchian & Engwall, 2017).

11.1 Quantal relationship

Figure 11.1a depicts an extreme quantal mapping between an input and an output space. There are regions of stability, $S_1$ and $S_2$, where the change in the input does not affect the output significantly. However, in a transient region, $T$, the output is highly sensitive to changes of the input. Using tube models of the vocal tract, Stevens (1989) showed that the articulation-acoustic mapping is quantal and that cardinal vowels are located in stable regions. The investigated articulatory parameter in his work was the constriction location. The work has been reproduced for two-tube and three-tube models in the following sections.

A. Two-tube model

In a tube-tube model, shown in Figure 11.1b, the total length of the tube $L_1 + L_2$ is constant, and is set to 16 cm. The cross-sectional areas are set to $A_1 = 0.5 \text{ cm}^2$ and $A_2 = 3 \text{ cm}^2$. Then the constriction location $L_1$ was manipulated and the formant frequencies were determined. The result is presented in the left panel of Figure 11.2. When $L_1 = 8 \text{ cm}$, the derivative has a minimum, and the lower formants, i.e. $F_1$ and $F_2$, change slowly in an interval of 7–9 cm, which is considered as the region of stability. This region, shown with a dashed box, corresponds to the vowel [a] articulation. In contrast to the constriction location $L_1$, there is no stability region regarding the constriction area (Perkell, 1996). In another simulation, the influence of the constriction area was investigated by setting the total length of the tube to

![Figure 11.1](image-url)
16 cm, the constriction location to \( L_1 = 8 \text{ cm} \), and \( A_2 = 3 \text{ cm}^2 \). The right panel of Figure 11.2 depicts the result, which shows no stable region.

**B. Three-tube model**

Similar simulations were conducted for the three-tube model shown in Figure 11.1c. To analyze the sensitivity to the constriction location \( L_1 \), the variables were set to \( A_1 = 3 \text{ cm}^2 \), \( A_2 = 0.5 \text{ cm}^2 \), \( A_3 = 3 \text{ cm}^2 \), \( L_2 = 2 \text{ cm} \), and the \( L_3 \) value depends on \( L_1 \), considering the total length of 16 cm. Figure 11.3 depicts the results, where the most stable region for the location of the constriction has been marked with a dashed box between \( L_1 = 8.5 \text{ cm} \) and \( 10.5 \text{ cm} \). To check the sensitivity of the constriction area, the constriction location was set to \( L_1 = 9.5 \text{ cm} \), and the constriction area, \( A_2 \), was changed as shown in the right panel of Figure 11.3.
11.2 Biomechanical constraints

To investigate the role of physiological constraints in the production of cardinal vowels, a study was conducted using the 3D biomechanical model. The aim of the study was to explore the sensitivity of the formants to changes in MAP, while considering biomechanical constraints. According to the quantal theory, the cardinal vowels are preferred because they can be articulated with some variations in the articulatory space and still maintaining the acoustic output stable. However, as (Perkell, 1996) showed and it has been reproduced in this study, there is no such an stability regarding the constriction area. Perkell (1996) proposed that the combination of two mappings, motor commands to articulation and articulation to acoustics result in a quantal mapping. By analyzing x-ray microbeam data and using a 2D biomechanical model, he hypothesized that a physiologically-relevant mechanism, i.e. contact, contributes to the the control of the constriction area and causes the production of relatively stable constrictions, despite the variability of motor commands, when producing vowels [a] and [i].

Using the 3D biomechanical model, the production of low-back, [a], and high-front, [i], vowels were investigated, considering contact or not in the biomechanical model. To produce these vowels an exciter is defined as the combination of four muscles:

\[
e = k_1(0.1 \times GG_p + 0.03 \times SL) + k_2(0.1 \times HG + 0.1 \times GG_A)\]

(11.1)

where only \(k_1 \neq 0\) or \(k_2 \neq 0\). When \(k_1 \neq 0\) and \(k_2 = 0\), the exciter combines the activation of the posterior GG and the SL which moves the tongue in the high-front direction, corresponding to a movement towards the [i] articulation. Choosing the second term \(k_1 = 0\) and \(k_2 \neq 0\), combines the HG and the anterior GG, which cause low-back movement of the tongue, corresponding to a movement towards the articulation of [a].

The formants' sensitivity was analyzed with respect to \(k_1\) and \(k_2\) under two different conditions. In the first condition, the contact between the tongue and other structures were not considered, and the tongue may penetrate the other structures. This is not realistic, but nevertheless useful in order to investigate the role of contact. In the second condition, contact was considered.

A. Low-back vowel

In this case, a target excitation was set in (11.1) by choosing \(k_1 = 0\), and \(k_2\) was linearly increased from zero until reaching the target value \(k_2 = 7\) after 200 ms. Figure 11.4 depicts the first and the second formants for both conditions, with and without contact. As \(k_2\) becomes larger, the sensitivity decreases in both conditions. However, the stable region is wider when the contact is considered.
Figure 11.4: In the 3D biomechanical model: formants sensitivity to muscle exciter coefficient $k_2$.

B. High-front vowel

The same simulation was conducted by choosing $k_2 = 0$ and a target $k_1 = 10$ in (11.1), giving the result depicted in Figure 11.5. No difference was found when simulating contact.

Figure 11.5: In the 3D biomechanical model: formants sensitivity to muscle exciter coefficient $k_1$. 
11.3 Discussion

When both \( k_1 \) and \( k_2 \) are zero, there is no active muscle force and the formants are highly sensitive. As moving to the corners by increasing \( k_1 \) or \( k_2 \), the sensitivity decreases. There are at least two reasons for such a stability. Firstly, as the activation of the muscle increases, the tongue becomes stiffer and thus it gets more difficult to move it further. Secondly, as the tongue reaches the corners, the contact between the tongue and other structures may prevent the tongue to move further.

The observations in this study suggest that tongue stiffening, in addition to contact, is also important and may contribute in forming the stable regions. For the low-back vowel, the influence of the contact seems to be more important than the stiffening. For the high-front vowel, it seems that the tongue stiffening contributes to forming a stable region.

In general, the presented results are preliminary observations that need to be investigated further. Simulating the physics of the contact is not trivial. The contact simulations presented here detects the penetration nodes and move them to discard the penetration. A more realistic situation would be to simulate tongue deformation based on the contact force. Furthermore, in the above simulations, the jaw position was fixed, which may influence the observation for the high-front vowel. Analysis of empirical data on pressure sensitive palatography may help to reach a concrete conclusion about the role of contact in producing this vowel (see, Tiede et al., 2003; Ono et al., 2004; Jeannin et al., 2008; Yano et al., 2012; Sardini et al., 2013, for works in this area).
Chapter 12

Conclusions and future directions

When concluding on the work presented in this thesis, the main message is that a full understanding of “speech”, as any other thing, requires that we create it. One way to do so is to translate our knowledge into computer code, thus creating a computational model. However, such a computational model not only necessitates modeling of each block involved in the process, but also linking them in a chain, and thus the overall accuracy is limited by the lowest accuracy of the individual blocks.

This thesis has argued for the benefits of using modeling of acoustics and articulation that is as truthful as possible to human speech production, i.e., that both a 3D acoustic and a 3D biomechanical model should be used. The arguments are summarized below.

Speech acoustics, as one of the blocks in the chain, predicts the propagation of sound waves within and outside the vocal tract. Although sound waves propagate in 3D, in the earlier works on acoustics, it was simplified as a 1D propagation phenomenon. This simplification was necessary due to technical limitations, but the models could nevertheless contribute to our understanding of the fundamentals of speech production. However, employing of 3D acoustic simulations offers several advantages that justifies paying extra effort and computational resources: 1) In 1D approaches, the model needs to be tweaked for different situations e.g., if there is an abrupt change in adjacent cross-sectional areas, or if sub-branches are to be included. Utilizing a 3D approach address these issues intrinsically. 2) Considering the typical dimensions of the vocal tract, it is argued that 1D approaches produce results with adequate accuracy for low frequencies, i.e. frequencies up to 4-5 kHz. However, as was shown by Arnela et al. (2016b, Paper B), some deviations may be observed also in the lower portion of the spectrum when comparing 1D and 3D approaches. Even if such deviations are less likely to affect the perceived vowel category, it is not clear if they are insignificant in applications such as personalized
speech production. 3) 1D approaches can not simulate the non-planar wave propagation that is relevant e.g., in the oral cavity when producing low-back vowels (Dabbaghchian et al., 2018a, Paper G), or the interaction between the right and left piriform fossa (Takemoto et al., 2013). This affects the dips (in both low and high frequencies) and the whole region of the high frequencies, which is perceptually relevant (see, Monson et al., 2014a, for a review). 4) Sound synthesis using 3D approaches may result in more natural and intelligible speech in comparison with the corresponding 1D approaches (Gully et al., 2018). In conclusion, it seems reasonable to employ 3D acoustic simulations, unless the computational cost is insurmountable.

The acoustic model requires a computational domain, i.e. the vocal tract geometry, for conducting the simulations. Such a geometry may be provided either by direct imaging of the speech articulators or by a model. In this thesis, a biomechanical model has been used, and it has the advantage that different configurations can be tested without the need to acquire new medical images. It also offers high temporal resolution, only limited by the simulation time step, which is as an advantage in comparison with rtMRI, since its spatio-temporal resolution is not high enough. Earlier biomechanical models were developed in 2D representing the midsagittal cross-section, despite the fact that the articulators are 3D structures. As demonstrated in this thesis, accurate modeling of the anatomical parts and their mechanical properties can not be implemented in lower dimensions than 3D. Thus, 3D biomechanical models should be utilized if the computational cost is not a concern.

As already mentioned, the overall accuracy of the model is limited by the lowest accuracy of individual blocks in the chain. This means that linking a 2D biomechanical model with a 3D acoustic model does not yield realistic results, since the 2D model discards the properties of the vocal tract cross-sectional shape when estimating the area function from the midsagittal outline. In the same way, linking a 3D biomechanical model with a 1D acoustic model is sub-optimal, since the complex geometry of the vocal tract needs to be converted to area function representation, in which only the cross-sectional area and the distance from the glottis are preserved. The cavity reconstruction method proposed in (Dabbaghchian et al., 2018a, Paper G) allows linking 3D biomechanical and acoustic models to accurately model anatomical structures and the acoustic transfer function. It hence provides a time-varying vocal tract geometry with high spatio-temporal resolution.

The reconstruction method may also be beneficial in developing and refining 3D biomechanical models, since it provides an accessible visualization of the vocal tract shape. For instance, a small resonance followed by an antiresonance was observed between 1 kHz and 2 kHz in producing the vowel [i]. The visual investigation of the vocal tract geometry suggests that a lateral channel between the upper and lower teeth, parallel to the oral cavity, generates this pair of resonance and antiresonance. Further, biomechanical modeling may be used for modeling of speech production phenomena and for understanding of the mechanical properties of the articulators. Using the coupled biomechanical-acoustic model, the six vowels and three
vowel-vowel utterances that were synthesized reproduce the loop-like trajectories that have been observed in EMA data. Since this was achieved by linear interpolation of MAP, one may hypothesize that such trajectories are the result of mechanical properties, rather than an active control mechanism. Analysis of the spectrograms of vowel-vowel utterances also revealed the influence of the mechanical properties of the articulators in generating vowel sounds. The most extreme deviation in the spectrogram was observed for $F_2$ during the sequence [ai]. Even if such a deviation is less likely to affect the perception of the vowel category, it may be important in a vowel-consonant-vowel utterance where the slope of $F_2$ plays an essential role in the perception of the consonant (Sussman & Shore, 1996; Lindblom & Sussman, 2012; Berry & Weismer, 2013).

The examples above are merely a few of the potential insights that may be gained by exploring the established link between a 3D biomechanical model and 3D acoustics. We next consider some suggestions for future work.

The geometry simplification proposed in (Arnela et al., 2016a, Paper A), and (Arnela et al., 2016b, Paper B) provides a quantitative comparison between 1D and 3D approaches. However, a perceptual test between 1D and 3D approaches, and also between different simplified geometries would be beneficial to investigate how important these differences are. Such a perceptual test however requires more variations of sounds.

The geometry reconstruction developed in this work allows reconstruction of intricate vocal tract shapes. However, to get the full advantage of this method, the acoustic model should be extended for simulations of consonants. Then one may use the model, e.g., to investigate the role of the sublingual cavity in distinguishing between [f], and [s] (Perkell et al., 2004). As another extension of the acoustic model, implementation of remeshing allows one to include subbranches in simulating dynamic sounds. This may be useful for studies similar to the one reported by Honda et al. (2010). In that study, some possible extensions to the quantal theory were suggested. That is, when articulating [ai], detaching the interdental spaces from the airway causes some inference between vowel formants. A similar model-based study may be utilized to investigate the role of the sublingual cavity in utterance [ai].

Although the larynx structure is available in the biomechanical model, it was treated as a static structure that forms the larynx tube of the airway. As a future direction, a more unified simulation may be to use the model to generate the glottal pulse instead of prescribing it as a boundary condition in the acoustic model (Degirmenci et al., 2017). However, this by itself is a large research area that requires solving Navier-Stokes equation (see, e.g., Zhao et al., 2002; Mihaescu et al., 2010) and addressing the fluid-structure interaction (see, e.g., Zhang, 2016; Jiang et al., 2017).

Another direction would be to extend the verification of the model. The most important challenge that limits verification of such a model is that the involved mechanisms are mainly invisible to the naked eye. The acoustic signal is the most accessible way to verify such a model. At the articulation level, the model can be verified using MRI, or rtMRI. However, when it comes to the muscular level, there
is a severe shortage in how to collect experimental data for verification. Emerging technologies such as diffusion tensor imaging (Shinagawa et al., 2008; Murano et al., 2010) may overcome this difficulty.

While such techniques are still under development, inverse models may be utilized even though there is no simple way of verifying the results. The main reason is the many-to-one mappings from muscle activation to articulation space, and from articulation to acoustic space. The variability across subjects is another challenge for verification. Developing the entire biomechanical model based on single subject may decrease the uncertainty involved in using the inverse model. Furthermore, to constrain the inverse model further, one may use MRI or rtMRI, instead of the four EMA markers, in order to increase the amount of geometrical input data to the inversion. However, this requires a refined tongue mesh with more flexible deformation in order to be able to follow the articulation data.

In general, when uncertainty is involved, sensitivity analysis is beneficial. That is, the estimated MAP are changed systematically, and the simulations is repeated. If small changes in muscle space do not change the articulation and acoustic significantly, it shows that the estimated MAP with a good margin results in desired articulation or acoustics. This requires developing a systematic approach to change MAP and to conduct a large amount of simulations. Due to the complexity of the presented model, it requires enormous computational power, unless one could develop an efficient approach in manipulating the MAP to limit the number of simulations.

The last but not least direction would be to extend the work presented in Chapter 7, and to fully establish the link between biomechanical model and a neural control model such as DIVA (Guenther, 2016). Then one could e.g., systematically explore how neural control influences acoustics and vice versa, how the perception of speech acoustics contributes to optimizing neural control and MAP. In such a model, where iterative acoustic simulations are required, the 3D acoustic model may be replaced by a 1D model to facilitate the simulations.

Finally, as the production of “speech” is difficult to understand, as it involves many different aspects of physiology and physics, it requires a multidisciplinary effort. The achievements within different disciplines are like pieces of a puzzle. One may not see the whole picture unless putting the pieces together and improvements in one piece leads to a sharper full picture.
Bibliography


Bibliography


