MRI-based 3D Finite Element Analysis and Acoustic Tube Modeling of Vocal Tract for /r/ Phoneme

Sai Hei Yeung
Advisor: Dr. Carol Espy-Wilson
MERIT 2005
Institute for Systems Research, University of Maryland, College Park

Abstract. The causal relationship between the geometry of the vocal tract and the speech sound produced can be studied using acoustic models of the human vocal tract. This project involves 3D finite element analysis (FEA) and acoustic tube modeling of the human vocal tract phonating the /r/ sound. This requires extraction of 3D tract geometry information from magnetic resonance images (MRI) of the tract using medical image processing techniques. Results obtained from the FEA simulation can identify a more accurate area funciton extraction method and the correct acoustic tube configuration, permitting important improvements to acoustic tube modeling for the liquid sound /r/. A better understanding of the vocal tract articulation for liquid sound /r/ would be beneficial to clinical applications as well as advancement in speech technology such as speech therapy, speech and speaker verification systems, and speech synthesis technology.

1 Introduction

1.1 Overview

In the past, research has shown a strong correlation between articulation of the vocal tract and the sound it produces. An effective way to study this geometry-sound relationship is by creating acoustic tube models of the human vocal tract. Significant progress has been achieved in producing vocal tract models for vowels (/a/-/u/), fricatives (/s/,/f/,/v/), and nasals (/m/,/n/). However, our articulatory knowledge of liquid sounds (/r/ and /l/) are still relatively limited. This is because of the non-uniqueness property of liquid phonemes; a speaker can use one of several articulatory strategies to produce the same acoustic profile of a liquid sound. This wide variety makes it difficult to formalize a general model of the respective vocal tract for phoneme /r/. However, despite the variety in articulation, the vocal tract generally involves a large frontal cavity often associated with the third frequency formant (F3) of the transfer function response, a key characteristic of the /r/ phoneme[1], and possibly of higher frequency formants. Knowledge of how to model this frontal cavity will provide important information in liquid phoneme acoustics modeling. Questions such as should the vocal tract be modeled as a single-tube or a central tube with a second side-branching tube and how should segmentation or area function calculations be performed along the tract path needs to be answered. Recent studies have suggested several possible strategies to model the vocal tract. These strategies and experiments will be elaborated in detail later in this section of the report.

This project involves three-dimensional acoustic tube modeling and finite-element-based harmonic analysis (FEHA) on the human vocal tract for sound /r/. In order to do so we must extract three-dimensional geometry information of the liquid sound vocal tract from mag-

[1] An exception to the F3-frontal cavity affiliation is described in [1]

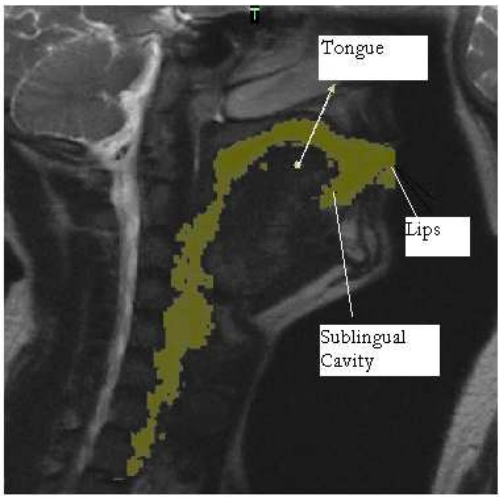


Figure 1: An outline of the vocal tract. Notice the divergence of path at the entrance of the frontal cavity.

netic resonance images(MRI)/footnotesrefer to section 2.2 for a more detailed description of MRI through medical image processing. By analyzing the FEHA simulation result, greater insight can be achieved on improving the acoustic modeling of sound /r/ by identifying the correct area function extraction method and acoustic tubes configuration. For example, the pressure flow distribution reveals directions of flow which can show whether the waves flow in one direction throughout the tube or diverges at the frontal cavity leading to the sublingual cavity (Figure 1), determining whether a single tube is sufficient to accurately model or a second side-branch tube is necessary. Iso-pressure contours produced by the flow diagram can guide the segmentation strategies to extract area function. The acoustic response can provide further information of the tract geometry such as formant-cavity relation or other patterns resulted from simulation.

Mimics [2], a medical computer-aided design (CAD) software, is used to extract the three-dimensional geometry from the MRI data. The three-dimensional model produced from the MIMICS software may be very complex, requiring excessive computation and memory usage. Determining an efficient way to minimize the amount of computations and memory

[2]Medical Image Processing Software, Materialise, Inc.

is desirable. Magics[3] is used to make corrections and improvements to the extracted model for optimal computation performance. Finally, Femlab[4] is used to implement the FEHM simulation.

1.2 Acoustic Source-Filter Model

The production of speech sounds can be modeled as a linear time-invariant source-filter model(Figure 2) often encountered in the realm of systems and signals processing[2]. In this model, the periodic compressions and rarefactions of air columns, formed by vocal folds (or vocal cords) vibration of air from the lungs, serves as a source. The vocal tract geometry, consisting of the pharyngeal cavity and oral cavity, serves as a filter (Figure 3). The source is processed through the filter to yield speech sounds. A typical transfer function spectrum, a property of the tract geometry, contains several peaks (or poles) that represents the formants (F1, F2, etc.), each one corresponding to a mode of vibration. A unique characteristic of the /r/ phoneme is the low frequency locale of the third formant(F3), making it very close to the second formant(F2).

1.3 Previous Vocal Tract Models

In past research, the /r/ phoneme vocal tract is modeled as a single tube with an expanded volume in the frontal cavity to account for the sublingual cavity extension (Figure 4, left). An alternative model suggests the sublingual cavity extension to be modeled as a side-branch to the main tube instead of a volume expansion to the frontal cavity (Figure 4, middle). It has been shown that both models result in negligible difference in the locations of the first three formants [3]. However, there is considerable differences for the positions of higher formants.

[3] `.stl` correction software ,Materialise, Inc.
[4] Multiphysics Finite-Element Simulation Software, COMSOL, Inc.

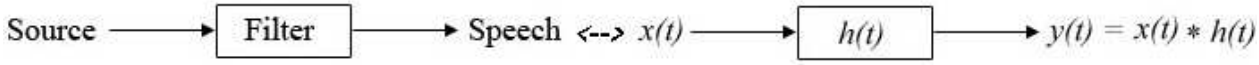


Figure 2: LTI source-filter and vocal tract model equivalence

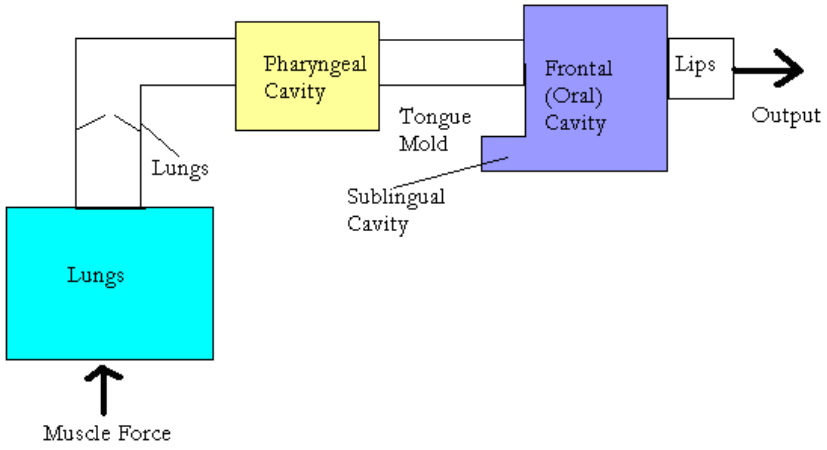


Figure 3: Detailed source-vocal tract model. The muscle force pushed through the lungs and vocal cords serve as the source while the filter composes of pharyngeal and frontal cavities.

Recent research [4] has suggested another model where the side-branch length is extended from the sublingual cavity to the top of the frontal cavity (Figure 4, right). The rationale for the side-branch extension is because there may be air trapped at the bottom of the frontal cavity, regardless of the sublingual cavity presence. The acoustic response for the first three formants show little difference from previous models, however, the higher formants show a larger difference from the single-tube model than the central model with sublingual side branch.

The experiments in this project will simulate pressure flow through realistic three-dimensional model instead of two-dimensional simplified models used in earlier experiments. Results will give greater insight as to where to separate the main tube from the side branch or whether a side branch is necessary at all. Furthermore, acoustic flow fields of the three-dimensional model can reveal a more accurate way to segment the tube for area function extraction. In

previous models, the segmentation is generally done perpendicular to the axis of the central tube.

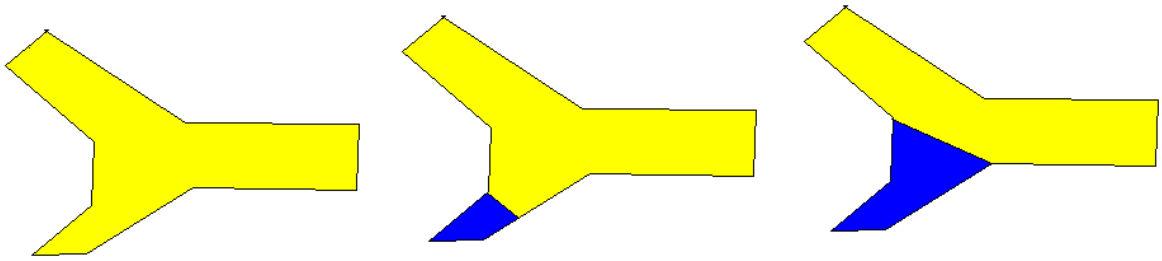


Figure 4: Single-tube model with waves propogating from input(left) to output(right)(*Left*). Central Tube with sublingual cavity as a side-branch (*Middle*). Central Tube with an extended side-branch.(*Right*)[4]

1.4 Applications

A better understanding of the vocal tract for liquid sounds would be beneficial to clinical applications as well as advancement of speech technology. Clinical applications include diagnosis and treatment for articulatory and phonological disorders involving the speech of liquid sounds and speech therapy using biofeedback of articulatory and acoustic data. An application of this research to speech technology is improvement on current speech recognition systems, in which the liquid sounds /r/ and /l/ are frequently sources of error. Other applications of this research are related to speech synthesis and speaker verification system.

2 Methodology

2.1 Image Processing and Three-Dimensional Geometry Extraction

The tool of choice for the image processing component of this project is Mimics. This software's core functionality is processing and editing of scanned 2D medical images. Mimics provides a large range of image editing tools which include performing Boolean operations

(and, or, xor, etc.) on images, morphology operations such as eroding, dilating, and contracting of pixel sets, manual pixel-by-pixel manipulation on images, and image registration features for combining multiple sets of images. Mimics also allows exportation of the resulting geometry model in `.stl` format, the input format necessary to perform the subsequent harmonic analysis of the geometry. Finally, Mimics is also very user-friendly and provides a comprehensive but summary user manual.

2.2 The Subject and Data information

The subject is a middle-sized, age 18 male. The data consists of MRI images from three different directions of the subject: axial, coronal, and midsagittal. The inter-spacing thickness between images range from $3mm$ in some sets to $5mm$ in others. The resolution of the image is $256x256$ pixels with a pixel size of $.938mmx.938mm$.

2.3 MRI Image Importation

Magnetic Resonance Imaging (MRI) is a technology that can detect the presence of protons that are conveniently present in human tissue and process it in image form [5]. Human tissue is concentrated with hydrogen atoms, which when placed under a magnetic field set at correct frequency, the dipole moment of the atoms begin to precess at a rate proportional to the strength of the field. At proper frequency, the atom will exert energy onto the magnetic coil, providing a signal of its activity. Proper modification of the field by superimposing a magnetic gradient makes it possible to associate the signal given off by the atom with the spatial coordinates of the signal origin. The intensity of the activity is depicted by a level of intensity in grayscale spatially mapped onto an image.

The available MRI data of the subject includes three separate sets of images in the coronal view. There is one set for the front cavity and another for the back cavity. The former

set spans from the lips to highest point of the vocal tract and has a span of $50mm$ in 11 images. The latter set continues from the first set and ends the back vertical wall of the vocal tract, also with a span of $50mm$ in 11 images. Both sets have slices that are 5mm in thickness. The remaining set (Figure 5) consists of 12 images from the front cavity to the entrance of the back cavity, spanning a distance of $33mm$. Each slice has a thickness of 3mm in this set.

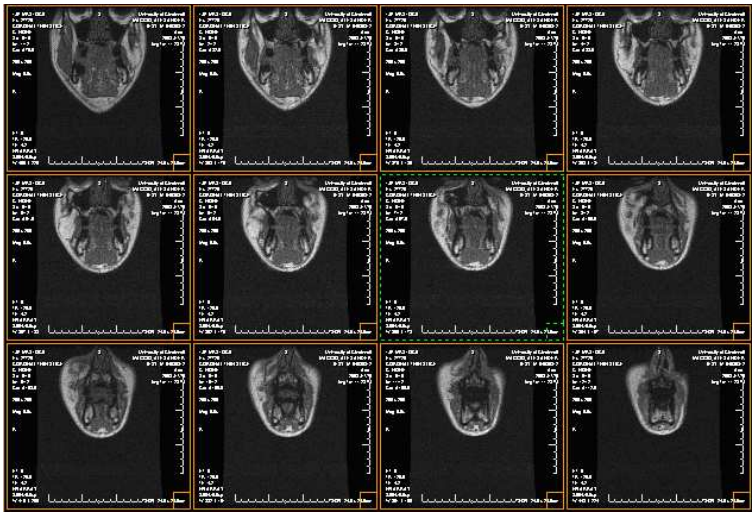


Figure 5: A set of thin-slice MRI-images in coronal view with 3mm thickness. This set of images trace the vocal tract from the entrance of the frontal cavity to the oral output.

The orientation and the relative positions of each set must be matched before importation. These information is contained within the `.dicom`[5] files and can be extracted using the MATLAB[6] command `dicominfo`. Each image must be set to the correct orientation and then placed into the correct position of the resulting composite set. This can be manually manipulated using either MATLAB's image-processing module or Mimics. This project employs both methods. The resulting composite set contains 34 images. When the parameters of the `.dicom` files are corrected, the image set is imported into Mimics. Mimics also re-

[5] `.dicom` stands for Digital Imaging and Communications in Medicine, and is widely used as a way to view medical images because of its capability to store information of the subject and image properties [5].

[6] MATLAB is a high-level language used for scientific simulation and computation.

quires specifications of the corresponding directions on the images which are obtained in the `..dicom` files.

2.4 Thresholding

The geometry is extracted using the overlapping connections between masks formed on the MRI images. A mask is essentially a set of planes with filled-in pixels, each plane corresponding to each image of the set. The process of producing the first mask requires thresholding, a filtering process to produce a desired mask. The MRI images are in grayscale in which each pixel is a transitional intensity within the extremes of black and white, which corresponds to minimum and maximum intensities respectively. The intensity of the pixel is proportional to the attenuation encountered or in other words, the intensity of the magnetic resonance [5]. The higher the concentration of protons, the more intense the signal becomes. Thresholding permits the setting of minimum and maximum of these intensities to filter out the intensities that are not within the set range (Figure 6). Since the geometry of the vocal tract is desired, a setting that permits low intensities of air while filtering high intensities of tissue is chosen.

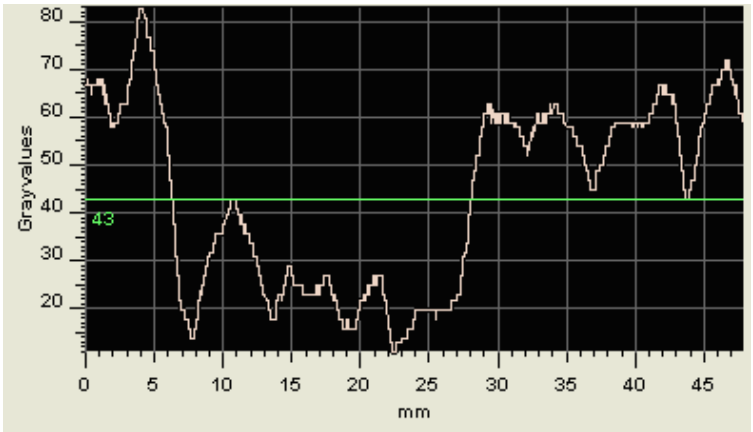


Figure 6: Grayscale intensity levels over the length of a profile line. The higher intensities represent presence of tissue while the lower intensities represent air space or proton deficiency.

2.5 Region Growing and Isolation of the Vocal Tract

The region growing feature permits the isolation of the vocal tract region from the outer region included in the original mask due to its low intensity. Region growing exploits the connectivity of the pixels within each image as well as overlapping of regions between images to isolate the geometry desired. In order to isolate the vocal tract, any connections between it and the outer environment, either within each image or between each image, must be disconnected. This is done by completely erasing all of the pixels of planes that are between the desired and undesired regions. The mask planes to be erased include a plane between the upper tract wall and the upper boundary of the skull from the axial view window(Figure 7) and a plane on each of the sides between vocal tract and outer environment from the midsagittal view window(Figure 7). Special treatment is given to the first plane of the coronal direction, where all of the pixels except for the desired region between the lips are erased. When it is ensured that the vocal tract is isolated from its outer environment, Mimics' region growing function produces a new mask consisting only pixels of the vocal tract region. Few iterations of trial and error may be necessary for complete isolation.

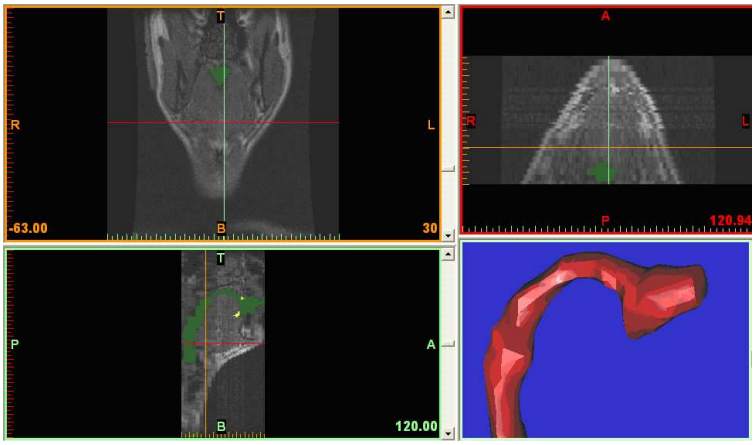


Figure 7: Display of Mimics environment. The upper-left is the coronal view window. Upper-right is the axial view window. Lower-left is the midsagittal view window. Lower-right is the 3D geometry window.

2.6 Morphology Operations

A major problem with the MRI images is presence of noise and lack of boundary definition stemming from artifacts in the subject's vocal tract or errors from the MRI scan itself. The noise present in the data presents a problem when creating the model because it can lead to a coarse surface and small holes on the geometry that may cause errors in the harmonic analysis simulation. Mimics' morphology operations can correct these problems. The first operation is using the closing function to close in any tiny holes introduced as artifacts. The second operation is using erode and dilate functions to eliminate jaggy edges on mask boundaries and then dilating it to compensate for the pixels taken away from eroding. The resulting mask is noise-free but not necessarily accurate to the human vocal tract yet. Errors sourced from the estimation of the boundary, though small, needs to be taken account and teeth-compensation needs to be performed.

2.7 Teeth Compensation

One of the drawbacks in using MR images to model the vocal tract is its inability to detect the teeth. The problem arises because teeth lack concentration of protons, the main source of attenuation for magnetic resonance to indicate presence of tissue or other substance. Therefore, in order to accurately represent the frontal cavity of the vocal tract, compensation for the teeth absence must be done. Outlines of the gum as well as known information about teeth positions in the mouth are used to determine the locations of the teeth on the MR images. The pixels for the teeth are then manually added. The errors introduced from the estimation of teeth position cannot be disregarded. However, any disparities between estimated and actual geometry should be small enough to render it insignificant.

2.8 Exportation to .STL Format

The edited and completed mask is exported as a `.stl` file. The `.stl` or stereolithography format is an ASCII[7] or binary file often used for prototyping in manufacturing. The file consists of a list of sets of three vertices representing triangular surfaces. The data is then used to describe a computer generated solid model. The resulting `.stl` model is then used as the input to perform the harmonic analysis in FEMLAB, the scientific analysis tool used in this project.

2.9 Improving the Back Cavity of the Vocal Tract

The frame-to-frame change on images of the coronal direction around the back cavity is very large partly due to the large distance between images of $5mm$, but mainly due to the rapid descent of the path towards the back. The solution to improve the accuracy of the back cavity is to combine the data obtained from the coronal view with that of the axial view. Therefore, the thresholding, region growing, and editing process are done to the relevant parts of the set of images in the axial direction. Because of the difficulty in combining the data within Mimics, the appropriate solution is to correctly combine the data together in the `.stl` domain. Magics provides `.stl` geometry manipulation tools to translate, cut, merge and connect different geometric bodies. Using these tools, the back cavity geometry produced from the axial image set is merged with the front cavity of the geometry produced from the coronal image set. The resulting geometry is much more accurate (Figure 8).

2.10 Correction and Optimization of .STL file

Before inputting into FEMLAB for harmonic analysis, alterations to the `.stl` file are necessary. The resulting `.stl` file contains errors that prevents acceptance in FEMLAB such as

[7] ASCII (*American Standard Code for Information Interchange*) is a character-encoding standard set used in computers.

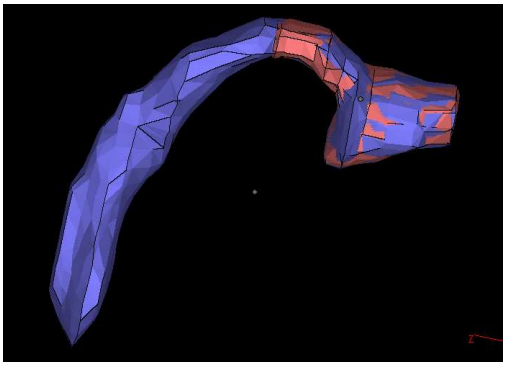


Figure 8: Vocal tract geometry after merging coronal and axial data sets.

intersecting triangle planes, extremely short or long triangles, overlapping planes or double shells, and violation of vertex to vertex rule. Furthermore, the calculation time and memory use are also worth consideration. Calculation for one set of parameter may take from minutes to hours. Thus, calculation for a range of frequencies may take from hours to days to perform. Memory usage also depends on the complexity of the `.stl` data. Therefore, manipulation of the `.stl` for optimal performance is desirable.

The tool to complete the tasks is Magics, a software directly compatible with Mimics. Magics allows detection of bad edges, double shells, or other errors that may prevent the subsequent FEMLAB processing. Manual manipulation is then required to fix the problems indicated. Magics also provides tools to smooth the surface of the `.stl` and reduce the number of triangles of the model. The proper level of smoothing and triangle reduction is chosen to yield optimal performance for the following harmonic analysis while still retaining the accurate geometry (Figure 9).

2.11 Harmonic Simulation in FEMLAB's Acoustics Module

The finished `.stl` file is then imported into FEMLAB's acoustics module. The geometry needs to be forced into solid form and meshed in preparation for the simulation. The mesh

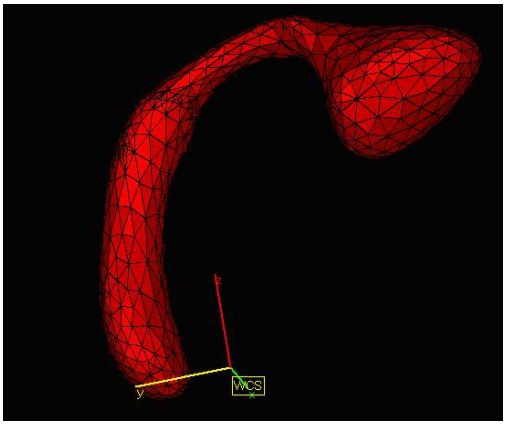


Figure 9: `.stl` geometry after smoothing and optimization operations.

on the surface of the solid creates a template for the boundaries of each triangle, allowing discretization of the solid. The resulting mesh contains 64349 elements with 98976 degrees of freedom. Boundary conditions are set on the finished mesh and then a sinusoidal input is forced into the entrance of the vocal tract to begin the simulation. The simulation is described by Helmholtz Equation(2.1). On a Pentium 4 2.8GHz computer with 2G RAM, the simulation took 70 seconds for each frequency. The simulation is performed at frequencies from 0Hz and 6000Hz. Two pressure flow distributions at 500Hz and 3000Hz, representing a low and high frequency respectively, are analyzed in the next section.

Helmoltz Equation:

$$\nabla(-\frac{1}{\rho_0}\nabla(p)) - \frac{\omega^2 p}{\rho_0 c_s^2} = 0 \tag{2.1}$$

Where $\omega = 2\pi f$ is the angular frequency, ρ_0 is the fluid density, and c_s is the speed of sound.

3 Results

3.1 $500Hz$ Source

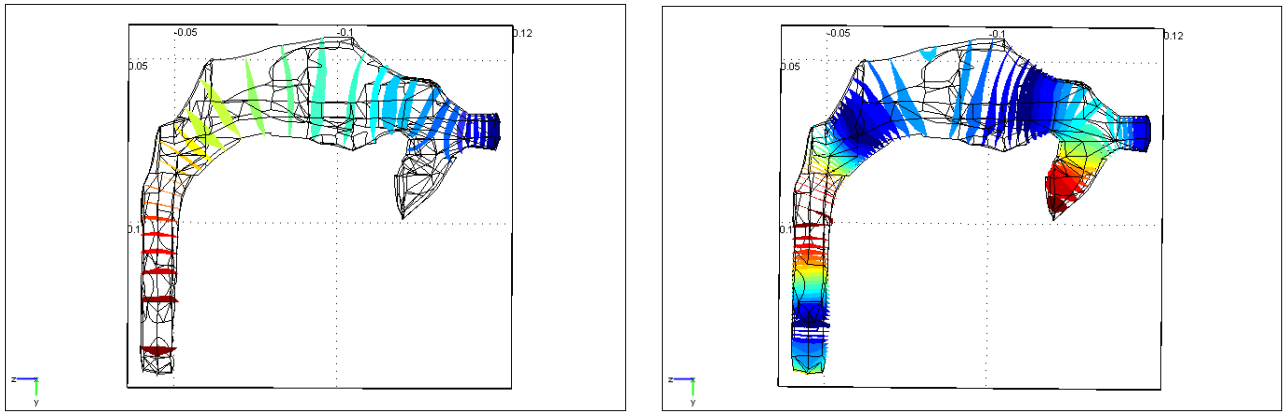
The pressure flow field at $500Hz$ (Figure 10a) shows uniform wave propagation from the posterior to anterior, where the iso-pressure contours are for the most part orthogonal to the midsagittal axis. The variation in intensity is also shown, with highest pressure at the tract input (color red) and lowest pressure at the output (color blue), hinting the proximity of the fundamental resonance frequency of the first formant, as depicted in the acoustic frequency response (Figure 11).

3.2 $3000Hz$ Source

At $3000Hz$, the flow field exhibits a non-uniform propagation that is unlike that observed at $500Hz$ (Figure 10b). The angles iso-pressure contours form with the midsagittal axis varies throughout the tract length. Furthermore, periodicity in intensity variation is observed throughout the tract, with peak intensities at the ends (input and output) as well as in the middle, hinting the proximity of the fourth fundamental resonance. The acoustic frequency response further confirms this with the location of the fourth formant a little above $3000Hz$ (Figure 11).

3.3 Acoustic Frequency Response

Peaks (or poles), representing the formants, are shown in the acoustic response (Figure 11). The locations of the first and fourth formants correspond nicely to the fundamental modes predicted by the pressure flow fields. A zero is observed between F5 and F6.



(a) (b)

Figure 10: Pressure-flow distribution field at (a)500Hz. (b) 3000Hz.

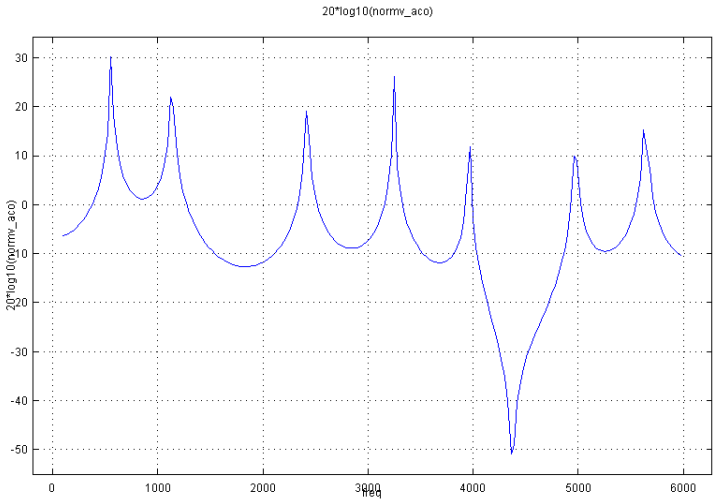


Figure 11: Acoustic frequency response of the harmonic analysis.

4 Discussion & Conclusions

Several conclusions can be made by the flow field of the 500Hz source simulation. The uniform direction of the propagation from input to output suggests that the one-tube model is sufficient to model the tract under low frequency conditions, confirming results of previous experiments described in [3] and [4]. The orthogonality between the iso-pressure contours and the midsagittal axis suggests segmentation of the vocal tract for area function extraction should be done perpendicular to the midsagittal axis along the tract.

The results under 3000Hz source show non-uniform wave propagation along the tract, particularly observable at the frontal cavity entrance, with the wave diverging to two branches. One branch leads to the output while the other towards the sublingual cavity. The divergence of flow paths justifies the use of a central branch with a side branch model for higher frequencies, as concluded from the results of [4]. Also observable from the flow field is the node where the divergence begins, located near the tip of the tongue, indicating the position for the main branch-side branch separation.

Previous simple-tube model based simulations used perpendicular segmentation with respect to the midsagittal axis to extract the area functions [4],[3]. The variation in angles between the iso-pressure contour planes and the midsagittal axis suggests invalidity of these methods. The segmentation should be done in parallel with the iso-pressure contours to preserve uniform pressure intensity for each measurement of the area function.

An obvious discrepancy is observed between the F2 and F3 locations on the acoustic response and the proximity of F2 and F3 that is characteristic of the /r/ phoneme. The inaccuracy of this may be attributed to the coarse quality of the mesh. However, the inaccuracy of the formant locations should have no effect on the validity of results obtained from the pressure flow distributions.

A notable characteristic of the acoustic response is the zero generated between F5 and F6. The source of the zero can be attributed to the sublingual cavity, where it absorbs all of the energy at the anti-resonance frequency of the zero.

Future work should concentrate on improving the quality of the mesh. Furthermore, to

validate the accuracy of the three-dimensional model, a comparison analysis with the response from a real sound source should be performed. Improvements on the geometry model may also improve the acoustic response accuracy. Higher resolution MRI or CT images with finer thickness can accomplish this. If MRI is used, a more precise way to compensate for the missing teeth can also affect the overall accuracy. Methods for teeth compensation are suggested in [5].

5 Summary

The core of this project involves the extraction of a three-dimensional model of the human vocal tract for the /r/ phoneme. The extraction procedure, involving numerous steps such as tissue-air thresholding, morphology operations, and optimization, is outlined in this report. The geometry data is then analyzed through finite-element based harmonic analysis to yield pressure flow field data and acoustic frequency response.

The results confirmed previous conclusions from experiments based on two-dimensional simple-tube models. That is, the single tube model is valid for low frequencies, but inaccurate for higher frequencies, which require a side-branch appended to the central branch for an accurate model. The oblique angles of the iso-pressure contour planes with respect to the midsagittal axis suggests a more accurate way for segmentation of the vocal tract used for area function extraction. That is, to segment it parallel with the contour planes instead of exclusively perpendicular with the midsagittal axis used in the current simple-tube model.

6 Acknowledgements

Thanks to Xinhui Zhou for contributing his time and energy to teach and guide me with generous patience throughout the project.

Thanks to the National Science Foundation(NSF) and University of Maryland for supporting the MERIT projects.

References

[1] Zhaoyan Zhang, S. Boyce, C. Espy-Wilson, and M. Tiede. Acoustic strategies for production of american english retroflex /r/. *Proceedings of the International Congress of Phonetic Sciences*, 2003.

[2] R. Kent and C. Read. *Acoustics: Analysis of Speech*. Thomson Learning, second edition, 2002.

[3] C. Espy-Wilson, S. Boyce, N. Jackson, S. Narayanan, and S. Alwan. Acoustic modeling of american english /r/. *J.Acoust. Soc. Am.*, pages 343–356, 2000.

[4] Zhaoyan Zhang, Carol Espy-Wilson, S. Boyce, and Mark Tiede. Modeling of the front cavity and sublingual space in american english rhotic sounds. *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference*, 1:893 – 896, 2005.

[5] Mark Tiede. An mri-based morphological approach to vocal tract area function estimation. *ATR Technical Report TR-H-XXX.*, 2000.