MODELLING TONGUE MOVEMENTS

A DISSERTATION SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF MASTER OF SCIENCE
IN THE FACULTY OF SCIENCE AND ENGINEERING

2019

By
Sivananda Gorugantu
ID: 10152496
School of Computer Science

Contents

List of Tables

List of Figures

Abstract

MODELLING TONGUE
MOVEMENTS
Sivananda Gorugantu
ID: 10152496
A dissertation submitted to the University of Manchester
for the degree of Master of Science, 2019

Tongue lateralization refers to the movement of the sides of the tongue. This movement of the sides of the tongue is the rationale for the production of sounds. Keeping in mind the role of tongue in the production of sounds, a number of issues were highlighted during the proposal of this project. The proposer is interested in discussing issues such as tongue lateralization, symmetry and/or asymmetry of the tongue and the velocity of the movement of the tongue. To address these aforementioned issues, ultrasound tongue images (UTI) of a number of subjects have been collected beforehand as preliminary data. This project - ”Modelling Tongue Movements” - which can be treated as the first step in addressing the problem of tongue lateralization, also uses ultrasound tongue images (UTI) as the primary data. The initial step is to locate the tongue in ultrasound tongue images automatically.

”Modelling Tongue Movements” is a computer vision domain related project that aims to address the problem of locating the tongue in ultrasound tongue images (UTI). In order to achieve this, a statistical shape model (SSM) is built primarily. The obtained mean shape model is then used as a reference to identify and locate the exteriors of a tongue in UTIs. Active shape model (ASM) and random forest regression voting (RFRV) methods are used to carry out the process of searching (locating) tongue in ultrasound tongue images. The observed error rates of the respective methods and their overall performance provide a means to compare and evaluate the two methods. After

evaluation, it is observed that the random forest regression voting method performs better compared to the active shape model due to various factors. The methods, their results and the factors affecting them are addressed in this thesis.

Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to convey my deepest gratitude to my supervisor, Dr Tim Morris for his constant and excellent supervision and guidance. His support, advice and motivation played a major role in the successful completion of my dissertation and helped me in developing research skills.

I would also like to thank my parents for their sacrifice, aid and tremendous support. I thank my siblings and friends for their regular encouragement. Last but not least, I thank my colleagues for helping me improve throughout the course of my MSc.

Chapter 1

Introduction

1.1 Introduction

”Modelling tongue movements” is a computer vision domain related project, proposed by a lecturer in Linguistics and Quantitative Methods, School of Arts, Languages and Cultures, University of Manchester. The proposer is interested in the development of an automated analysis method of tongue lateralization based on ultrasound tongue imaging (UTI). Figure 1.1 is an example of an ultrasound tongue image (UTI).

Tongue lateralization refers to the movement of the sides of the tongue. This movement of the sides of the tongue is the rationale for the production of sounds. ”There are over 6500 spoken languages in the world” [3], various dialects for each language, various pronunciations of words and consequently, there are a lot of sounds. For simplicity, let’s consider only two languages, namely English and Welsh. One can visualize the shape of the tongue while producing the sound [l] in English or [ɬ] in Welsh (the ll-sound) as opposed to the shape of the tongue while producing the sound [o] in pronouncing the word ”go”. Now that the role of the tongue in producing sounds is established, it’s time to deliberate on tongue lateralization, the effect of symmetry or asymmetry of the tongue and the velocity of the movement of the tongue in producing sounds. The proposer is interested in investigating and learning the answers to the aforementioned research subjects. To address the above issues, ultrasound tongue

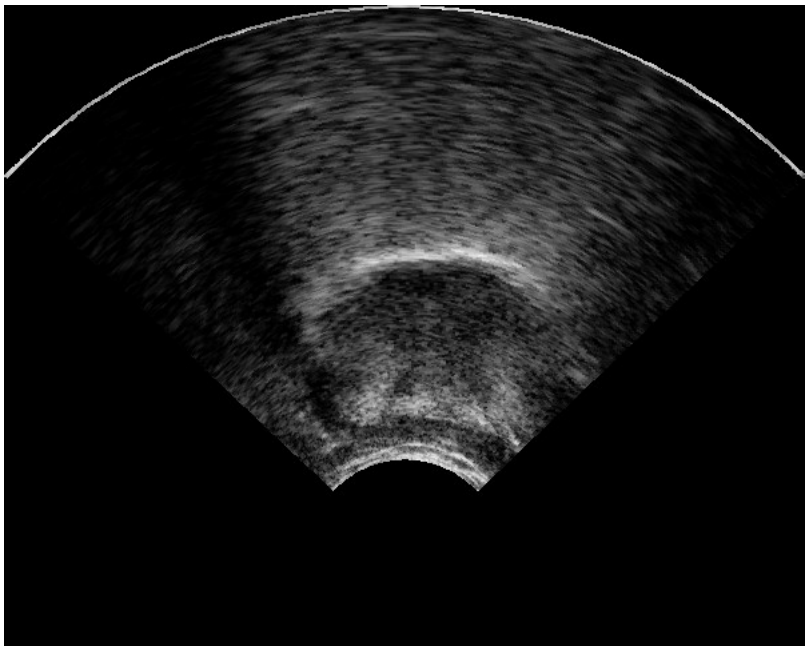


Figure 1.1: Example of an ultrasound tongue image (UTI).

images (UTI) of a number of subjects have been collected beforehand as preliminary data. This project "Modelling Tongue Movements" which can be treated as the first step in addressing the problem of tongue lateralization, also uses ultrasound tongue images (UTI) as the primary data. The initial step is to locate the tongue in ultrasound tongue images automatically.

1.2 Problem and Solution Approach

Visualising the shape of the tongue mentally is simpler. In contrast, determining the shape of the tongue and its outline in an ultrasound tongue image mathematically makes it a very challenging task. Therefore, the effect of symmetry and/or asymmetry of the tongue in producing sounds or the velocity of the movement of the tongue in producing sounds can only be determined after locating the tongue in ultrasound tongue images. Hence, this makes marking and locating the exteriors of the tongue in an ultrasound image the immediate requirement.

1.2.1 Problem

The shape of the tongue in an ultrasound tongue image depends on the speaker, the sound made by the subject and proper placement of transducer probe (a device that transmits ultrasound waves and captures their echoes). In this thesis, the sound made by the subject is considered to be the key factor for variation of the shape of the tongue. Different sounds generate different shapes. A mean shape is to be computed for each sound. Thereafter, the mean shape is to be used as an initial shape in locating the tongue in the inputted ultrasound tongue image or video.

One can spontaneously feel the change of shape of the tongue while making certain sounds. Say, for instance, the tongue needs to be placed in a particular shape to generate sounds like [a], [l], [o]. However, this change in shape needs to be identified and addressed using 2-dimensional ultrasound tongue images (UTI). The UTI of sounds [a], [e], [i], [l], [o], [t] and [u] are shown in Figure 1.2. Understanding the natural shape of the tongue for a particular sound will help in addressing the issue of tongue lateralization.

1.2.2 Solution Approach

The foundation for the research of tongue lateralization is locating the boundary of tongue in ultrasound tongue images. As specified in the above subsection, different sounds can be generated by changing the shape of the tongue. These shapes are to be identified and recorded. Previously, this process of identifying tongue in an ultrasound image was being carried out manually. The process of learning the natural shape of the tongue in UTI should be automated. A faithful way of tackling this problem is by using statistical shape models. A statistical shape model (SSM) of the tongue is built initially. This yields the mean shape of the tongue for that particular sound. This shape model is then used in two methods, namely, active shape modelling (ASM) and random forest regression voting (RFRV) to match the shape model in ultrasound tongue images.

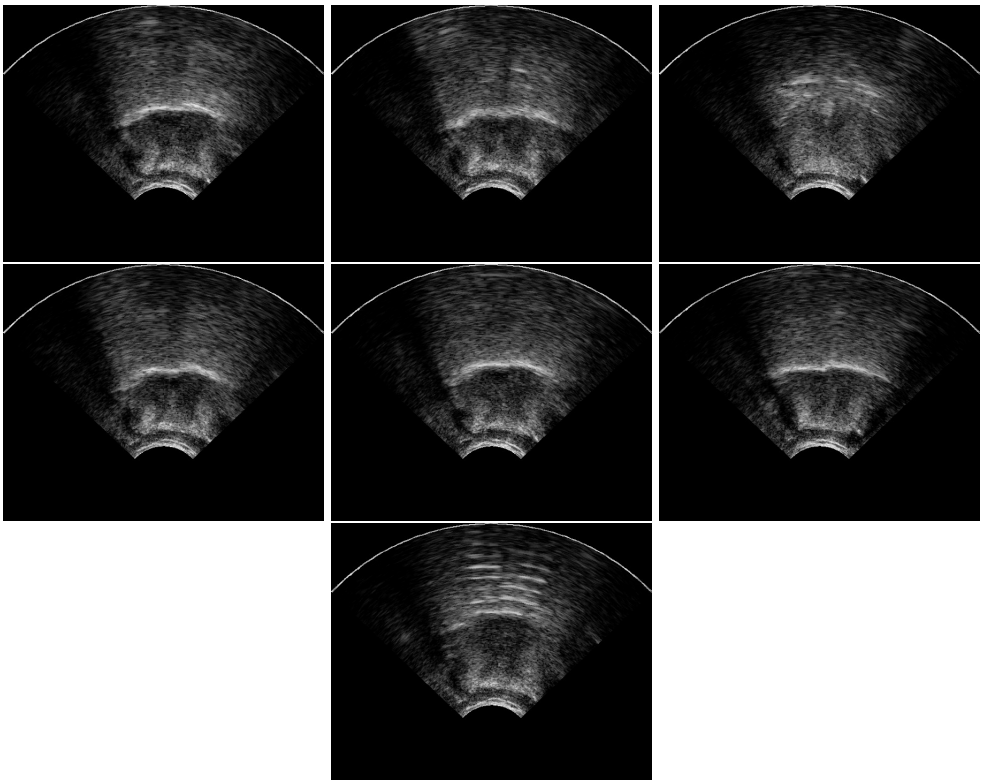


Figure 1.2: The ultrasound tongue images of sounds [a], [e], [i] respectively in first row, sounds [l], [o], [t] respectively in second row and sound [u] in third row.

1.3 Aims and Objectives

The aim of this project is to identify, locate and mark an outline of the exteriors (outer edge) of tongue in still images and in videos. This will firstly, assist in understanding the basic skeleton of the tongue. Secondly, it will give an insight of the structure of the tongue for individual sounds. This thesis will thus boost the understanding of the natural shape of the tongue both with respect to individual sounds and in general. Thereby, providing a basis to carry the work forward in the intended direction of comprehending tongue lateralization. In the pursuit of locating the tongue and identifying its shape automatically, an optimal method is also determined.

The objective of modelling tongue movements, that is, automatically identifying the tongue in ultrasound tongue images is achieved by following the seven individual

stages of this project. First, individual sounds are selected for annotating based on certain criteria (see section 3.2.1). Second, the images are annotated and the points are saved for building statistical shape models. Third, all the annotated shapes are aligned and the model is trained to obtain a mean shape of the tongue. Fourth, the mean shape and the variation along its principal axes are captured. Fifth, the aligned annotated points and the mean shape are further given as inputs for image searching. Two methods are used for image searching, namely, active shape models (ASM) and random forest regression voting (RFRV). Sixth, training the models using a number of annotated ultrasound images of tongue provided (9 training folds). Seventh, testing them on ultrasound images of tongue provided (1 test fold).

The above-mentioned objectives are accomplished by ensuring the following.

- All models must generate and mark an outline of the tongue's exteriors (outer edges) in the ultrasound tongue image.
- No model is over-fitting.
- No model generates offbeat results.
- Evaluating by comparing the results of the two methods and against the gold standard.

1.4 Motivation and Research Questions

The goal of this project is to build statistical shape models and train them to identify and mark the boundaries (tip and sides) of the tongue in ultrasound tongue images. Although the bigger objective of this research is to take a step towards comprehending tongue lateralization, this thesis focuses on spotting the tongue in the inputted ultrasound image. A number of challenges were encountered as the work progressed, that motivated to put forth a series of queries. This thesis investigates and addresses the

research questions. Having said that, it also states its vulnerabilities.

Keeping in mind the goal, the following research questions will be addressed in this thesis.

Research Questions:

- How can the shape of the tongue be modelled?
- What amount of variation is allowed for the deformable object (modelled tongue)?
- How can the tongue be identified in a given ultrasound tongue image (UTI)?
- Can an alternative method apart from the active shape model be used to achieve the goal?
- How do the results vary and what factors influence the results?

1.5 Dissertation Structure

The structure of this dissertation is as follows: the immediate chapter (that is, chapter 2) is about systematic literature review. A set of guidelines were followed in conducting the literature review. All the guidelines, the process involved and the narrowing down of literature to suit the topic at hand is explained seamlessly. The research methodology is outlined in chapter 3. Starting with modelling the tongue, use of statistical shape models for aligning modelled shapes and generating the mean shape. This is followed by explaining the use of active shape models and random forest regression voting approaches in image search (in locating tongue). The results obtained and the observations made are mentioned in the subsequent chapter (chapter 4). Further discussions about the thesis, in particular, the research questions are addressed in chapter 5. Chapter 6 marks the end of this thesis by summarising the work, concluding it and stating future work.

Chapter 2

Systematic Literature Review

A systematic literature review was conducted to acknowledge and learn the work of predecessors in this research topic. This thesis follows the guidelines proposed by Kitchenham and Charters (2007) [4] in conducting a systematic literature review. In this chapter, a rationale is provided as to why a systematic review was conducted in section 2.1. Section 2.2 talks about the search strategy that was formulated. This search strategy is adopted to find articles and documents for studying. Here, a list of search resources is also stated. The next section provides a detailed explanation of the study selection criteria and procedure. Both inclusion and exclusion study criteria are specified. Section 2.4 accommodates quality assessment procedure. The search result that yielded potential search documents for study and the manner in which it was carried out can be found in section 2.5. Subsequent sections are the dissemination of the literature review of this research topic.

2.1 Review Objective

The aim of the proposer is to build a collaboration to develop automated analysis methods of tongue lateralization based on Ultrasound Tongue Imaging (UTI) [5]. The movement of sides of the tongue is involved in the production of sounds such as English [l], or Welsh [ɬ] (the ll-sound). This project will provide foundations to answer questions about tongue lateralization. It will open up channels to methodologically

analysing and addressing issues such as the symmetry of the lateral movement of the tongue, how the lateral movement is affected by asymmetry, the speaker's dialect and other social factors.

In order to address these research questions, a more fundamental analysis must be carried out. A formal understanding of the working of ultrasound images is needed. It is also necessary to be aware of the flaws of using ultrasound imaging. Having sound knowledge of ultrasound image technology and its vulnerabilities helps in progressing to the next level of addressing research questions. Be that as it may, answering the next set of questions is more vital and poses a greater challenge.

The next stage puts forth numerous questions. Questions such as modelling the shape of the tongue, searching the modelled object in an image and the amount of allowable deformity of the object. A review of the generic computer vision algorithms is needed to answer these questions. The resulting search of reviewed generic algorithms is then refined so as to suit the requirements of this project. Furthermore, it is of great importance to look into the contributions of predecessors and the related, ongoing work in the field.

The rationale for the review is to have a clear and sound understanding of the background. This helps not only in understanding the problem more precisely but also increases the awareness of the susceptibilities. Therefore, allowing to formulate a plan to tackle them. Ultimately, this review assists in progressing the work on "modelling tongue movements". Thereby, addressing the diverse set of research questions.

2.2 Search Strategy

The very first and essential step in a review is formulating a search strategy. This can be achieved by first splitting the research question into finer and atomic units. Once

these terms are recognized, their synonyms, different spellings and/or alternate terms can be noted. These keywords are combined with Boolean ANDs and ORs to make sophisticated search strings [4].

2.2.1 Search Terms and Strings

The keywords and search items and/or their alternatives that are specific to this thesis are listed in table 2.1. These keywords are derived from the research questions. Two search strings were assembled using the search terms or keywords listed in table 2.1. The search strings are listed below.
Search strings:

1. (A1 OR A2 OR A3) AND (C1 OR C2 OR C3 OR C4 OR C5 OR C6)

2. (B1 OR B2) AND (C1 OR C2 OR C3 OR C4 OR C5 OR C6)

Table 2.1: Keywords and search terms.

A1. Tongue movements	B1. Deformable objects	C1. Computer vision
A2. Ultrasound tongue image	B2. Locating objects	C2. Statistical shape models
A3. Phonetics		C3. Active shape models
		C4. Random forest regression voting
		C5. Image search
		C6. Principal component analysis

2.2.2 Search Resources

Following is a list of digital libraries, online journals, search databases and research databases that were used as search resources for this project.

- IEEE Xplore Digital Library (https://ieeexplore.ieee.org/Xplore/home.jsp)

- ACM Digital Library (https://dl.acm.org/)

- ScienceDirect (https://www.sciencedirect.com/)

- SpringerLink (https://link.springer.com/)
- Web of Science (https://webofknowledge.com/)
- Google Scholar (for snowballing search)

2.3 Study Selection Criteria and Procedure

Once the study strategy is outlined, a selection criterion needs to be developed. The study selection criteria lay out the scheme of including search documents. Including all the resulting search documents is not practical. Hence, inclusion and exclusion criteria are stated. These protocols are necessary to determine the search documents that are to be included or excluded from the literature review. In other words, it helps in filtering relevant documents. The following subsections list the inclusion and exclusion protocols.

2.3.1 Inclusion Criteria

- The primary study directly addressing the research aim and objectives must be included.
- Include the study of active shape models.
- Include the study of random forest regression voting.
- The study must unequivocally address the deformability of the modelled objects.
- Include documents pertaining to ultrasound tongue imaging.
- Include documents where the study highlights the vulnerabilities of ultrasound images.
- Studies carried out in the industrial and academic domain can be considered.
- Full text of the search documents (articles or journals) must be available.

2.3.2 Exclusion Criteria

- Exclude documents that are not written in the English language.
- Documents that conduct a study of speech but are not related to the collaborated field of phonetics and computer vision.
- Documents that address image search, however, lack the concept of object deformability.
- Articles that are not related to any of the following:
    1. Ultrasound imaging.
    2. Statistical shape models.
    3. Model building and image search.
    4. Phonetics.

Once the documents were refined based on the inclusion and exclusion criteria, they were subjected to the next level of filtering. The relevance of the title, abstract and conclusion of the search document governed the rule of filtering. The filtered documents are finally considered for the purpose of the literature review.

2.4 Quality Assessment Procedure

A quality assessment checklist was prepared as shown in Table 2.2. Each resource was evaluated against this checklist. Resources that satisfied at least one question were considered.

2.5 Result of Search

The search strings formulated using the search terms (and keywords) in section 2.2.1 were used in the search resources and databases mentioned in section 2.2.2 to obtain

Table 2.2: Study quality assessment checklist.

Does the resource address research phonetics or ultrasound imaging?	Yes or No
Does the resource address modelling tongue movements?	Yes or No
Is the resource relevant to statistical shape modelling?	Yes or No
Does a resource address deformability of modelled shapes?	Yes or No
Does a resource address factors affecting image search?	Yes or No
Do the resource address issues that relate to the shape of the object?	Yes or No
Does the resource provide leads to any of the research questions?	Yes or No

search documents for reviewing. This process was carried out in April/May 2018. A total of 85,732 documents were obtained as a result of search string 1. Search string 2 yielded a sum total of 385,138 search documents. The number of documents obtained as a result of this search per individual search resource can be seen in Table 2.3. Keeping in mind the inclusion and exclusion study criteria, the search documents were refined down to 1,431 (for search string 1) and 6,461 (for search string 2). After applying quality assessment as the second stage of refinement and removing redundancies, 40 potential documents were yielded. The refinement results also considered the relevance of the document, that is, the title and the abstract of the document were studied for relevance and scrutinized the resulting search documents. The third and final stage of refinement using snowballing. The final number of documents that were studied for the review mount up to 44.

Table 2.3: Search results for search strings 1 and 2.
Total number of search documents found

Database	*Search string 1*	*search string 2*
IEEE Xplore Digital Library	43	1936
ACM Digital Library	39325	68937
ScienceDirect	13903	77259
SpringerLink	30506	233088
Web of Science	85	1828
Google Scholar	1870	2090
Total	85732	385138

2.6 Review of Ultrasound Imaging

An ultrasound (or ultrasonography) image is a 2-dimensional image of intensities of echoes of high-frequency sound waves. It is a similar technique as that of SONAR or echolocation technique used by bats [6]. Ultrasound tongue imaging also follows the same procedure. Before actually looking into the working of ultrasound imaging for capturing tongue movements, certain terminologies should be learnt. The terminologies and their definitions listed below are taken from [7].

Sonography: Graphical representation of a subject utilizing sound.

SONAR: The acronym stands for SOund NAvigation Ranging.

Echo: Reflected acoustic signal, received after hitting a tissue boundary.

Transducer: A device that converts energy from one form to another. In ultrasonics, this device is used in converting electrical energy to mechanical energy (transmitting acoustic waves) and mechanical energy to electrical (receiving reflected ultrasound waves).

Doppler Ultrasound: Ultrasound technique that uses the Doppler effect to detect the distance of tissue boundary from the origin based on the frequency of the reflected wave.

The transducer probe is placed under the subject's chin (see Figure 2.1). The probe transmits high-frequency sound waves (about 5 MHz). When the acoustic waves hit boundaries between tissues, sound waves are reflected back to the probe. The possible tissue boundaries are, the tongue - air, tongue - lower/upper gum and the tongue hard palate (superior boundary of the oral cavity). The probe then captures the reflected sound waves and broadcasts to the transducer to convert it to an electrical signal. The distances and intensities of echoes are computed by the machine and are then represented as a 2-dimensional ultrasound tongue image. In Figure 1.1, the brightest echoes in the middle of the image represent the exterior of the tongue. It can also be noted here that some sound waves carry forward without reflecting back to the probe after hitting a boundary. These waves might be reflected much later (possibly when they hit

Figure 2.1: Capturing ultrasound tongue images (UTI) by placing the transducer probe under the speaker's chin [1].

a stronger boundary) or they are reflected off some other tissue boundary and hence, a vast number of speckles can be observed around the tongue.

Hein and O'Brien Jr [8] concluded their review in 1993 about ultrasound techniques by stating that the Doppler techniques were used in the past to analyse the interaction of ultrasound with the tissue motion. However, in the recent times, time domain techniques are used for ultrasound imaging. A time-domain technique is computationally advanced and powerful than Doppler ultrasound. Ultrasound imaging has advanced even further. Improvement in ultrasonic techniques and equipment has led to greater quality and real-time test abilities. One of the most recent techniques, that offer great potential uses an ultrasonic array in the transducer. Bruce and Paul in [9], review "ultrasonic arrays for non-destructive evaluation". In this review, they comment on the state-of-the-art, the use of piezoelectric materials and the array geometries.

2.7 Literature Review of Phonetics

Image processing and computer vision collaborated with medical image analysis roughly two decades ago. Since then, a substantial amount of research has been carried out in this coupled field. Although the fusion with phonetics is fairly new, this collaboration undoubtedly resulted in the advancement of research in this budding field. Some of the recent and on-going works in this research area are reviewed here. Chollet et al. [10] give an overview of audio-visual speech processing. They also highlight few experiments that signify the relation of movement of articulators such as tongue and lips in recognizing words. This benefit hearing impaired in lip reading. Many others have published their work in this young collaboration. Another example is the work of Yuanyao and Qingqing [11]. Using localized active contour model-based methods, they successfully segmented lips in RGB images. This work enabled them in lip reading. This work is used as a foundation in human-computer interactions. Various such works are available and research in these fields is ongoing at this very moment. However, this thesis is restricted to only one organ of speech, namely, the tongue.

2.7.1 Role of Tongue in Phonetics

The tongue plays a major part in phonetics. It assists in producing various sounds. One can picture the shape of the tongue, the placement of the tongue and its positioning in producing sounds such as [l] in English or [t] in pronouncing the word ”tell”. By now, the role of the tongue in producing sounds is very well established. The proposer of this project is keen to build a collaboration to develop automated analysis methods of tongue lateralization based on ultrasound tongue images. Data analysing plays a key role in achieving this. Current analytical approaches are manual and labour intensive. Seeing Speech [1] is an online resource that was a collaborative effort of researchers from a number of Scottish Universities. It provides ”ultrasound tongue imaging (UTI) video of the speech, magnetic resonance imaging (MRI) video of the speech and 2D midsagittal head animations based on MRI and UTI data” [1].

2.7.2 PCA in Tongue Movements

Jackson and McGowan in [12] experimented with ”the use of principal component analysis in predicting midsagittal pharyngeal dimensions from x-ray images of anterior tongue positions”. The x-ray images were captured during the production of Swedish vowels by the subjects. Stone et al. [13] also used principal component analysis in their research. They were interested in learning tongue motion in subjects post glossectomy surgery. They studied the adaptive tongue motion of a tumour affected part of the tongue and the non-effected parts of the tongue.

Hueber et al. [14] used principal component analysis of pixel intensity values to extract eigentongue features. Their research was intended for the application of ”silent speech interface (SSI)”. In this study, Hueber et al. employed a global coding approach where the ultrasound images were projected onto a new feature space called the eigentongues. The results were fed to a neural network to learn the relation between the vocal tract of ultrasound images to the line spectrum frequencies. They evaluated their model with a tongue contour model and verified that the global coding approach is more efficient.

2.7.3 More Complex Methods

Translational deep belief networks (tDBNs) were used by Jeff Berry and Ian Fasel [15] to automatically extract the dynamics of tongue gestures from ultrasound images. Lisa and Ghassan [16] treat tongue tracking as a graph labelling model. They developed a semi-automatic approach in tracking tongue contours in two-dimensional ultrasound images. Yin and Xiaohu [17] dwell on the problems of 3-D reconstruction of the tongue as the movements of the tongue are subtle and swift. They propose ”a real-time visualization framework based on motion capture technique and constraint-driven elastic model” to address the underlined issues.

A review of different viewpoints of the definition of the role of a shape with respect to various academic disciplines is carried out by George Nagy and Naomi Nagy [18]. They also emphasize on collaborating image acquisition and shape analysis. By this, they propose advanced image processing methods for linguistics and articulatory phonetics. Lisa et al. [19] carried their work forward from representing and segmenting the tongue using a graph labelled model. They adopt a machine learning approach that eradicates the use of semi-automatic tongue segmentation approaches for analysis. They managed to construct velocity based and spatiotemporal gestural descriptors which encode tongue dynamics during the speech.

Samuel Silva and António Teixeira [20] address the issue of automatic tongue annotations in ultrasound images. This effort is to enable systematic analysis of surface electromyography (sEMG). This method intends to eventually address the issue of improving silent speech interfaces. Moisik et al. [21] used optical flow analysis to automate the process of finding the laryngeal state during Mandarin tone production. Their work signifies the role of extra glottal laryngeal mechanism in the production of higher or lower tones. This research, firstly, is in regard to the Mandarin language. Secondly, does not address the issue of tongue lateralization.

Plenty of literature is available for coupled fields of computer vision and phonetics or linguistics. Some of which are reviewed above. Information from this literature can be obtained in abundance. However, some of these methods are not extended for lateral tongue movements or some of them invariably, fail to address the issue of tongue lateralization. Therefore, it is inevitable to overlook this far-reaching literature and develop a method that can be extended to address the pivotal issue of tongue lateralization.

2.8 Review of Shape Models

Loosvelt et al. [22] propose a biomechanical model for tracking tongue in ultrasound tongue images. With the help of experiments, they prove that this model is efficient and also works in cases where the tongue movement is abrupt. Anastasios Roussos et al. [23] propose the usage of active appearance model (AAM) in tracking tongue movements in ultrasound tongue images. Nevertheless, this thesis aims to develop a shape model to represent tongue based on statistical shape modelling and employ active shape models and investigate the effectiveness of random forest regression voting in tracking tongue contour in ultrasound tongue images.

This thesis focuses on building a model for the tongue. As seen in chapter 1, the tongue in the ultrasound tongue image (see Figure 1.1) is the brightest echo in the middle region of the ultrasound image. Therefore, it can perhaps be treated as a rigid model with constrained flexibility. Recognizing rigid objects in images using model-based [24][25] approaches are well known. The challenge is to identify objects that can deform. A number of approaches were developed to address the issue of recognizing deformable objects in images. These flexible objects are governed by a number of control parameters to control the modelled shape of the object.

Fundamental components such as a line, circle or arc can be used in building flexible models with a slight variation in their shape, scale and orientation. Such hand-crafted models were used by Yuille et al. [26] to recognize features of a face. The best fit is found by minimizing an energy function by varying the parameters. P. Lipson et al. [27] extract features from medical images (CT images) using a similar model. A. Hill and C. J. Taylor ”describe the application of genetic algorithms in model-based image interpretation” [28] by identifying the left ventricle in echocardiograms.

"A generalised Hough transform approach is used in recognizing articulated objects consisting of rigid parts connected by rotary or prismatic joints" [29] an articulated model scheme used by Beinglass and Wolfson. Although this approach is far better than rigid object recognition, in that, it is capable of recognizing objects with some degree of freedom. The shape of the tongue can't be modelled using rotary or prismatic joints. Therefore, this method cannot be extended in modelling and recognizing tongues.

Snakes or active contour models described by Kass et al.[30] are also used in mapping flexible contour models to image features. However, a number of parameters are needed to be set by the user. An efficient method cannot be determined to set parameters such as the initial size and position of the contour. Assuming the initial position of the contour is set and the parameters are specified such that the contour is attracted towards the features of interest in the image, numerous issues come to light. Firstly, it is important to keep in mind that the contour is superimposed on an ultrasound image. Secondly, the following question needs to be addressed. Should the contour be attracted towards the bright echoes or darker region? The above two issues remind that there are a lot of speckles in the ultrasound tongue image and therefore, snakes or active contours are bound to fail in recognizing tongue in ultrasound tongue images. Hence, this is also not a sensible approach for modelling tongues.

Hinton et al.[31] worked on recognizing hand-printed digits by having a number of control points that store their "home" location and are moved away from their "home" location when there are deformations. Fourier series shape models as proposed by Scott [32] is a method of modelling shapes based on equation 2.1. Different shapes are formed by varying parameters: a_n, b_n, ϕ_n, ψ_n.

$$x = x_0 + \sum_n a_n \sin(n\theta + \phi_n)$$
$$y = y_0 + \sum_n b_n \sin(n\theta + \psi_n) \tag{2.1}$$

Staib and Duncan [33] take it forward by using them to interpret medical images. Bozma and Duncan [34] use this approach to model organs in medical images.

Statistical models - a set of boundary points (landmark points) that define the shape of an object to be modelled, were used by Goodall in [35], Grenander et al. [36] also use this approach to represent the shape of hands. Colin Goodall describes a model-based Procrustes approach. This approach is valuable in analysing shapes that are translated, rotated or of variable scales. Goodall presents beautifully, the approach of Procrustes analysis in shape modelling. His work presents the details of mathematical derivations a trace of how he arrived at modelling shapes. A similar approach is used in this thesis in building the shape model of the tongue.

Active shape models proposed by Cootes et al. [2] overcomes the drawback of all the earlier models by not compromising the robustness of models by restricting its variation. Therefore, the deformation of the model is only permitted in ways consistent to training data. In addition to providing a robust approach in modelling shape models by imposing constraints on the variability of the shape of the object, Cootes et al. propose a robust method of searching the mean model in images. This method superimposes the mean shape model on the image. The edges along the normal of each landmark point are inspected for their respective edge strengths. This is an efficient method of determining object boundaries. The mean shape, however, is not moved directly to the point of strongest edge. It is projected onto a lower dimensional space and reconstructed in the original image space. This efficient and robust approach of modelling and searching is studied and implemented in modelling tongue movements.

A number of shape model matching methods such as active shape models [2], active appearance models [37], pictorial structures, constrained local models [38] and so on have been used for object recognition in images. A preview of active shape models is seen in the previous paragraph. Active appearance models, besides addressing the constraints on shape variability, take a step further and also consider the variation in texture in modelling shapes. This approach, although a valuable literature, has a possibility of failing when it encounters ultrasound tongue images with discrete echoes. That is, the captured texture of such models might add overwhelming variability to the overall texture model. Thereby, disturbing the mean shape model. ”Pictorial structure models for object recognition” [39] is also a robust method. Pedro and Daniel [39] propose a method of modelling shape objects by identifying and dividing the deformable object into a number of individual parts. These distinctive deformable objects are modelled separately. The object to be recognized, therefore, is a collection of such modelled parts with spring-like connections between them. However, this is not a desirable approach for modelling tongue objects.

Various regression based matching methods such as boosted regression [40][41], random forest regression [42], shape regression machine [43] and so on are also available to estimate and match the shape and pose of an object. However, regression-based voting methods such as generalized Hough transform [44], implicit shape model [45], Hough forests [46], kernel SVM based regressors [47] and many more voting based approaches are proven to be more effective for identifying and locating shapes in images.

T. F. Cootes et al. ”use random forest regression, together with a global shape model” [48] to obtain faster and accurate results. This method uses the shape model built using statistical shape models. The optimal position of each point in an image is determined by the votes cast by a regressor. This approach is proved to be more

accurate than boosted regression or discriminative based methods which were trained on similar or identical data. The model is based on two principal concepts. Firstly, a constrained local model approach is used for matching points that were yielded from a mean shape model obtained from statistical shape modelling. Secondly, these matched points are evaluated over a region of interest (ROI). Voting of these points with the help of random forest regressors takes place in a grid in the specified region of interest around the point. In this thesis, the random forest regression voting [48] approach is studied and investigated in detail along with active shape models [2], as an alternative. This literature review acted as the foundation for this thesis and contributed to addressing the research questions.

As a result of the literature review, this thesis emulates the following implementation and evaluation plan. Tongue shape models are built using statistical shape modelling approach. Thereafter, these models are used as an initial estimation of the tongue in searching ultrasound images. Two methods, namely, active shape models and random forest regression voting approaches are implemented to support image search. Subsequently, the models built using the two approaches are evaluated against each other.

Chapter 3

Research Methodology

The aim of this thesis was to build models that are capable of tracking tongue contour in ultrasound images. To achieve the set goal, this thesis adhered to the research methodology elucidated in this chapter. The structure of this chapter is as follows: section 3.1 explains the process of data acquisition. The methodology practised is described in section 3.2. In this section, the design opted to represent the shape of the tongue is illustrated. The role of active shape models and random forest regression voting in image search and their implementation are made clear. Section 3.3 puts forth the tools used for development and the challenges encountered are stated in the subsequent section.

3.1 Data Acquisition

The primary data that is necessary to work on this project is the ultrasound tongue images (UTI). A number of speakers contributed to extracting ultrasound tongue images. The ultrasound images were captured while the speakers were producing various sounds. Sounds such as [a], [e], [l], [t], [i], [o] and [u]. Sample ultrasound images captured while producing these sounds can be seen in Figure 1.2. The transducer probe is placed under subjects' chin as they generate sounds. The probe transmits high-frequency sound waves which are reflected back on hitting the boundary between

tissues. These mechanical sound waves are converted to electrical signals and captured as two-dimensional ultrasound images. The data was collected by a lecturer in Linguistics and Quantitative Methods, School of Arts, Languages and Cultures, University of Manchester. The data provided for this thesis is secondary data and is fully anonymised.

3.2 Methodology

Tracking tongue contour in ultrasound images needs a shape model that represents the tongue. The represented mean shape model acts as the primary data for searching unseen ultrasound tongue images. To successfully accomplish this objective, a literature of great depth was explored as seen in chapter 2. The literature obtained from the archive was narrowed down to three most essential concepts. The statistical shape models (SSM) for building a mean shape model of the tongue. The active shape models [2] (ASM) for searching and fitting the mean shape model in the ultrasound tongue image. The random forest regression voting [48] approach to do the same and act as an alternative to active shape modelling.

3.2.1 Shape Design

The very first step in building shape models is annotating the object to be modelled. Accordingly, annotating boundary of the tongue in ultrasound tongue images is the preliminary task. In order to annotate the tongue, certain criteria have to be followed. Looking at Figure 1.2, it is evident that locating tongue in ultrasound images representing sounds [i] and [u] is impractical. Even an expert would fail to identify the boundaries of the tongue in such images. Therefore, the first criterion is to eliminate such data. Proceeding with the remaining data after discarding data that represents sounds [i] and [u], a number of ultrasound images were identified where the tongue was indistinguishable. That is, the images were extremely noisy and the tongue was

barely noticeable. Figure 3.1 is one such example. Hence, the second criterion: eliminate images which are extremely noisy. These two criteria were followed in annotating tongue images.

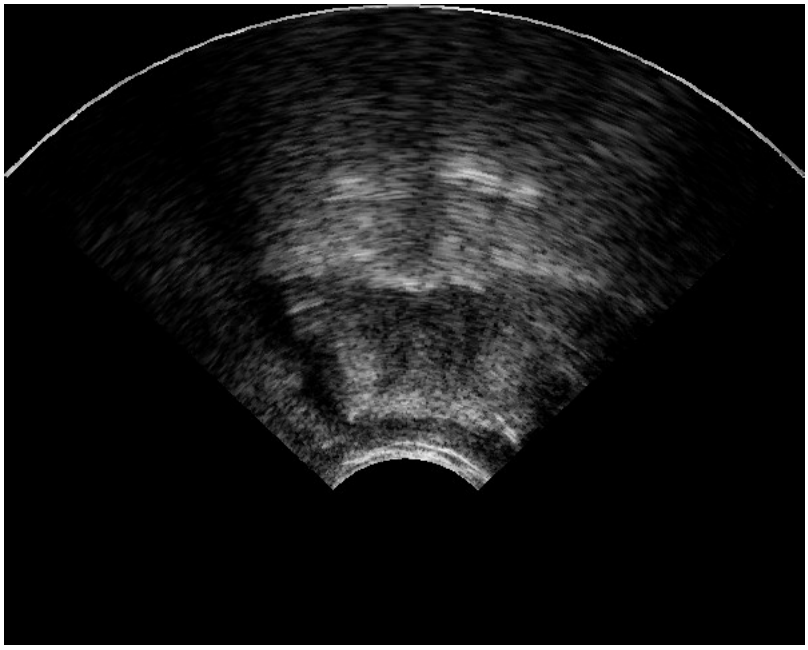


Figure 3.1: Example of an extremely noisy ultrasound tongue image where the tongue is barely noticeable.

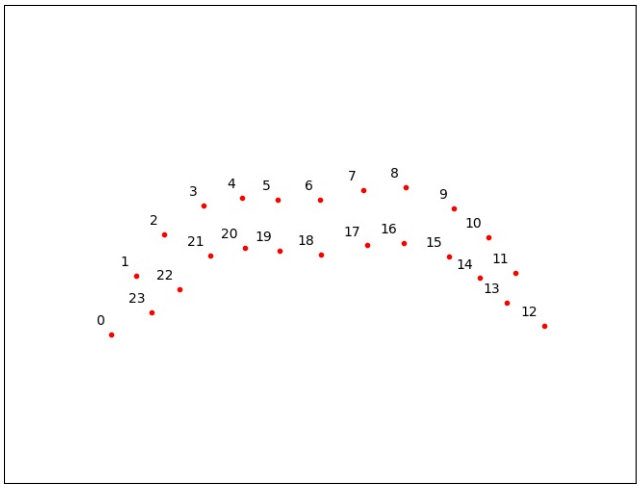


Figure 3.2: Labelled tongue shape.

Bookstein [49] talks about four principles in addressing the shape variations in a morphometric study. The first and the second principles talk about the landmark locations and shape coordinates. Bookstein marked the landmark points to study and compare the shape variation with numerous other factors. Cootes et al. [2] simplified the explanation of identifying landmark points by categorising them into three classes. The first class of landmark points are the regions that pose great significance to a particular application. For instance, vertices of triangles. The second category is the landmark points that are not application specific. The third class of points are a mixture of both. For instance, equidistant points plotted between two significant landmark points. A similar approach is adopted in annotating tongue images in this thesis.

Every annotated point represents the boundary of the tongue. Each point has its significance. The design is selected in a way that sufficient landmark points represent a substantial part of tongue boundary. Figure 3.2 displays a sample of labelled tongue shape model. Points 0 and 12 represent the endpoints of the tongue. The centres of the upper and lower boundaries of the tongue are represented by points 6 and 18. Points 3, 9, 15 and 21 are the control points representing any significant change or curve on the tongue. The remaining points are placed equidistant between the above-specified control points. There are two equidistant points between any two control points. These points represent the boundary of the tongue. An example of an annotated ultrasound tongue image is shown in Figure 3.3.

The described set of instructions are to be followed closely. If the landmark points are not placed in the region it represents or the region is misjudged by the annotator, a significant amount of variation is added to the mean shape. This will later play a vital role in influencing the identification of tongue while searching ultrasound tongue images. Keeping in mind, the constraints and the guidelines, the 24-point annotation method was adopted in this thesis.

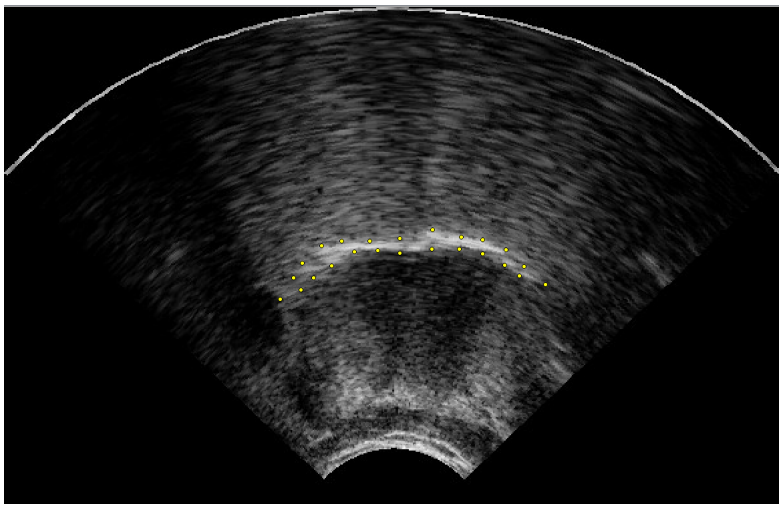


Figure 3.3: Example of an annotated ultrasound tongue image.

3.2.2 Mean Shape Model

A statistical shape model approach is used in computing the mean shape model of the tongue. The landmark points obtained after annotating tongue images hold the information of the coordinates of the points in the image frame. The training data has to be aligned before computing the mean shape. The logic behind computing the mean shape is to compare the equivalent points in the training set. These points can only be compared when they are aligned with respect to a set of reference points. Also, there is a possibility of each shape varying in scale or rotated at an angle different from others. Cootes et al. [2] use a modified Procrustes method [50] in aligning annotated resistor, hand and heart ventricle shapes. Their method is extended in aligning tongue shapes in this thesis.

Every annotated tongue shape is represented using the equation (3.1), where, x_{i0}, y_{i0} are the coordinates of the annotated point 0 of the i^{th} training shape. Every training shape is scaled by a factor s, translated by (t_{xj}, t_{yj}) and rotated by an angle θ in order to map one shape onto another so as to minimise the objective function in equation (3.2).

$$\mathbf{x}_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, ..., x_{in-1}, y_{in-1})^T \tag{3.1}$$

$$E_j = (\mathbf{x}_i - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j)^T \mathbf{W} (\mathbf{x}_i - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j) \tag{3.2}$$

where,

$$M(s,\theta) \begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} & - & (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} & + & (s\cos\theta)y_{jk} \end{pmatrix}, \tag{3.3}$$

$$\mathbf{t}_j = (t_{xj}, t_{yj}, ..., t_{xj}, t_{yj})^T \quad \textit{and}$$

$\mathbf{W}$ - a diagonal matrix of weights.

Following are the steps and mathematics involved in aligning the training data:

1. Consider every annotated shape and compute the distance between each point R_{kl}. Where k and l are two points in the shape and R is the distance between them.

2. Compute the variance $V_{R_{kl}}$ among the distances over the entire training data. That is, the variance in distance between point 0 and point 1, point 0 and point 2 and so on, in the training set.

3. Choose a weight for each point based on their respective variance. One method of computing weights is given in equation (3.4).

$$w_k = \left(\sum_{l=0}^{n-1} V_{R_{kl}} \right) \tag{3.4}$$

4. Compute weight matrix W using equation (3.5).

$$W = \sum_k w_k \tag{3.5}$$

5. Consider the very first shape as reference shape and compute X_1 and Y_1 by substituting $i = 1$ in equation (3.6).

$$\begin{aligned} X_i &= \sum_{k=0}^{n-1} w_k x_{ik} \\ Y_i &= \sum_{k=0}^{n-1} w_k y_{ik} \end{aligned} \tag{3.6}$$

6. For every other shape in the training data,

   6.1. Compute parameters X_2 and Y_2 by substituting $i = 2$ in equation (3.6). Also compute Z, C_1 and C_2 using equation (3.7).

$$\begin{aligned} Z &= \sum_{k=0}^{n-1} w_k (x_{2k}^2 + y_{2k}^2) \\ C_1 &= \sum_{k=0}^{n-1} w_k (x_{1k} x_{2k} + y_{1k} y_{2k}) \\ C_2 &= \sum_{k=0}^{n-1} w_k (y_{1k} x_{2k} - x_{1k} y_{2k}) \end{aligned} \tag{3.7}$$

   6.2. Solve equation (3.8) using standard matrix methods to obtain a_x, a_y, t_x, t_y. where, $a_x = s\cos\theta$ and $a_y = s\sin\theta$.

$$\begin{bmatrix} X_2 & -Y_2 & W & 0 \\ Y_2 & X_2 & 0 & W \\ Z & 0 & X_2 & Y_2 \\ 0 & Z & -Y_2 & X_2 \end{bmatrix} \begin{bmatrix} a_x \\ a_y \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} X_1 \\ Y_1 \\ C_1 \\ C_2 \end{bmatrix} \tag{3.8}$$

   6.3. Store the resulting (mapped) shape as the aligned, scaled, translated and rotated shape of $\mathbf{x}_2$ for further computation.

7. Compute the mean shape $\overline{\mathbf{x}}$.

The idea here is to consider the very first shape as the reference shape to align the data for the first iteration. Once an initial mean shape is computed, this will then be used

as the reference shape for subsequent iterations. The computed mean shape must be normalized. This is achieved by aligning the mean shape with the first shape in the training data. In other words, the orientation, origin and the scale of the obtained mean must be aligned with the reference shape. Once the mean shape is normalized, the training set must be re-aligned with the mean shape. The process of aligning the training set with the mean shape is repeated until convergence. The convergence criteria opted for this process is shown in equation (3.9). That is, the difference between the deviation of the current mean to the new mean in two consecutive iterations must be less than 0.03.

$$\begin{aligned} diff_i &= \sum(\overline{\mathbf{x}}_{current} - \overline{\mathbf{x}}_{new}) \\ convergence &: |diff_i - diff_{i+1}| \leq 0.03 \end{aligned} \tag{3.9}$$

3.2.3 Capturing the Variation of Aligned Shapes

The next stage of implementation is to capture the variation of the aligned shapes. This is a decisive stage as a decision is made on the allowable deformability of the mean shape model by analysing the variation of all shapes over the aligned training data set. To accomplish this, the principal axes of the 2n-dimensional (n = 24) data are obtained. These principal axes are identified by applying principal component analysis [51] over 48-dimensional training shapes.

The number of dimensions is 48 as each shape has 24 landmark points and each point is represented by its x and y coordinates in the image frame. The data for each data-point is arranged as shown in equation (3.1). The procedure is explained below:

1. Compute the mean. (equation 3.10)

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{3.10}$$

2. Compute covariance matrix S.

$$S = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \tag{3.11}$$

3. Solve the eigenvalue problem.

$$SP_i = \lambda_i P_i \tag{3.12}$$

where, λ_i is the i^{th} eigenvalue and P_i is the corresponding eigenvector.

4. Arrange the eigenvectors in the decreasing order of their corresponding eigenvalues. That is, $\lambda_1 > ..\lambda_i > ..\lambda_n$.

5. Apply proportion of variance (PoV) for extracting the first m (< 48) principal components contributing to 95% of the variation.

$$PoV = \frac{\sum_{i=1}^{m}\lambda_i}{\sum_{i=1}^{2n}\lambda_i} \tag{3.13}$$

3.2.4 Image Search Using Active Shape Model

The primary method implemented for searching and locating the tongue in unseen ultrasound tongue images is the active shape model method. This uses the mean shape model built using statistical shape modelling approach explained in section 3.2.2. The mean shape model that is mapped onto the ultrasound image is represented by the equation (3.14). Cootes et al. [2] explain the mathematics involved in locating the modelled object in an unseen image. This approach is extended in searching and identifying tongue in ultrasound images.

$$\begin{aligned} \mathbf{X} &= M(s,\theta)[\mathbf{x}] + \mathbf{X}_c \\ \mathbf{X}_c &= (X_c, Y_c, ..., X_c, Y_c)^T \end{aligned} \tag{3.14}$$

where, $\mathbf{X}_c$ represents the centroid of the mean shape model.

The algorithm design is as follows (see appendix for derivations and explanation).

1. Determine the normal at each point of the model. This is accomplished by first computing the slope of the line connecting two consecutive points in the shape using the equation of *slope of a line* in (3.15). The slope of the normal is then computed using the equation specified by *slope of normal* in (3.15). The obtained slope of the normal and the reference point at which a normal is intended to be computed are substituted back in the equation of a line to extract ± 10 points relative to the reference point along the normal.

$$\begin{aligned} slope\ of\ normal, \qquad & m_2 = \frac{-1}{m_1} \\ slope\ of\ a\ line, \qquad & m_1 = \frac{y_2 - y_1}{x_2 - x_1} \end{aligned} \tag{3.15}$$

2. Find the strongest edge along the normal. The candidate boundary positions are represented as $d\mathbf{X}$ as shown in equation (3.16). Figure 3.4 shows a sample of recommended movement along the normal of the boundary based on strong edge detection.

$$d\mathbf{X} = (dX_0, dY_0, ..., dX_{n-1}, dY_{n-1})^T \tag{3.16}$$

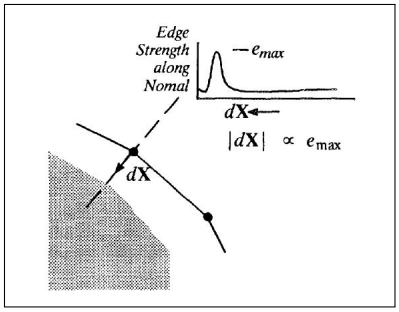


Figure 3.4: Indicating strong edge along the normal [2].

3. The pose parameters (s, θ, X_c, Y_c) are computed relative to the suggested new movement of the model boundary $\mathbf{X} + d\mathbf{X}$.

4. The computed pose parameters influence the change required to the mean shape model $d\mathbf{x}$.

$$\begin{aligned} d\mathbf{x} &= M((s(1+ds))^{-1}, -(\theta + d\theta))[y] - \mathbf{x} \\ where, \qquad y &= M(s,\theta)[\mathbf{x}] + d\mathbf{X} - d\mathbf{X}_c \end{aligned} \tag{3.17}$$

5. The suggested new shape of the model $\mathbf{x} + d\mathbf{x}$ needs to fit in the allowable shape domain to keep the flexibility and shape of the model in check. Therefore, the new shape is projected onto the model parameter space. The reconstruction equation (3.18) can be approximated to (3.19). Subtracting (3.18) from (3.19) and the fact that $\mathbf{P}^T = \mathbf{P}^{-1}$ simplifies the projection equation to (3.20)

$$\mathbf{x} = \overline{\mathbf{x}} + \mathbf{P}\mathbf{b} \tag{3.18}$$

$$\mathbf{x} + d\mathbf{x} \approx \overline{\mathbf{x}} + \mathbf{P}(\mathbf{b} + d\mathbf{b}) \tag{3.19}$$

$$d\mathbf{b} = \mathbf{P}^T d\mathbf{x} \tag{3.20}$$

6. The computed model space parameters $\mathbf{b}$, indicate the lower dimensional projection of the new tongue shape model.

7. If the model space parameters $\mathbf{b}$ are not in the defined allowable shape domain, they are rescaled using equation (3.21).

$$b_k = b_k \cdot \frac{D_{max}}{D_m} \tag{3.21}$$

8. A contour in the image space is reconstructed using the equation (3.18).

9. The pose parameters are then updated using the following equations.

$$\begin{aligned} X_c &= X_c + w_t dX_c \\ Y_c &= Y_c + w_t dY_c \\ \theta &= \theta + w_\theta d\theta \\ s &= s(1 + w_s ds) \\ \mathbf{b} &= \mathbf{b} + \mathbf{W}_b d\mathbf{b} \end{aligned} \tag{3.22}$$

10. This iterative process is repeated until convergence. The convergence criterion is specified in equation (3.23). That is, the absolute sum of difference of the current and new tongue shape models must be less than 0.1.

$$convergence: \qquad \left|\sum(\mathbf{x} - \mathbf{x}_{new})\right| < 0.1 \tag{3.23}$$

3.2.5 Image Search Using Random Forest Regression Voting

The random forest regression voting approach proposed by Cootes et al. [48] rely on two principal concepts. Namely, constrained local models [38] and voting based on random forest regressors. As stated earlier, the mean shape model built using the statistical shape model approach is the initial stage for the random forest regression voting method of searching images. The model represents each shape model using the equation (3.24). This equation represents the shape model generated by reconstructing the mean shape model represented by $\overline{\mathbf{x}}_i$, using only the m opted modes of variation that is subjected to a global transformation T.

$$\mathbf{x}_i = T(\overline{\mathbf{x}}_i + \mathbf{P}_i\mathbf{b}; m) \tag{3.24}$$

The aim is to minimize the objective function (3.25) by reasonably selecting $\mathbf{x}_i$ and their respective $\mathbf{b}$ and m parameters. A noteworthy assumption is that the pose parameters of the model in the image space and that in the model parameter space

are equally likely. The equation (3.25) can be simplified to equation (3.26) once the quality of fit $C_i(\mathbf{x}_i)$ (equation (3.27)) is computed. The shape constraints, that is, the allowable shape flexibility is kept in check by the first term of the objective function (3.26) and the image matching information is concealed in the second term.

$$Z_p = -\log p(\mathbf{b}, m|I) = -\log p(\mathbf{b}) - \alpha \sum_{i=1}^{N} \log p(\mathbf{x}_i|I) \tag{3.25}$$

$$Z_p = -\log p(\mathbf{b}) + \alpha \sum_{i=1}^{N} C_i(\mathbf{x}_i) \tag{3.26}$$

$$C_i(\mathbf{x}_i) = -\log p_i(\mathbf{x}_i|I) \tag{3.27}$$

A general purpose optimiser [38] is used to optimise the objective function (3.26). The optimiser starts with the initial radius and searches for the best fit for a given point. It then narrows the radius down and the process is repeated until convergence (maximum number of iterations or the minimum radius).

Random forest regressors are used for voting the position of each point in a selected region of interest. Voting with random forest regressors is very well explained by Cootes et al. in [48]. This approach is extended in this thesis for searching the tongue in ultrasound images. Haar-like features [52] are used for training, to benefit the process of sampling features around a point and evaluating them to predict the best position of a given point.

3.3 Tools for Development

A tool to annotate points on 2-dimensional images was taken from [53]. The points annotated were stored with a .pts file extension and were later used in building shape models. For a greater part of this project, python programming language was used. The implementation of statistical shape model in constructing mean shape model of the tongue, capturing the variation of the mean shape model, implementation of active

shape models and plotting numerous graphs for evaluation was done using python programming language. Various libraries such as matplotlib for plotting graphs, OpenCV for drawing tongue contours on ultrasound tongue images and many more were included. Constrained Local Model (hclm_build_model), best fit optimiser and other tools were used in training models, building trees based on random forest regression voting and for testing them. BoneFinder software [54] was used for visualising the results (the identified and located tongue in the inputted ultrasound image) obtained after training the gdss model using random forest regression voting approach.

3.4 Challenges Encountered

The challenges encountered during the development of the adopted methodology and the methods implemented to overcome them are listed below.

1. Annotating images was time-consuming. Firstly, some amount of thought and time was dedicated to analysing and finalizing the design of the shape of the tongue model. Once the design was finalized and approved, annotating the data consumed a considerable amount of time. Annotating sufficient examples was a necessity to train and build robust models.

2. Unwanted noise/speckles in the ultrasound images. The greatest challenge by far in this project was to address the issue of eradicating the unwanted noise/speckles from the images. The initial approach to tackle this problem was to use filters to smooth the image. This did not produce the desired results. The next approach was to use despeckling methods. Hence, a median filter was implemented. It undoubtedly performed better than the smoothing filter. However, the result was still not satisfactory. The next approach was to use a simple rank filter that replaces the value of the pixel with the highest value in the opted window. This method showed a lot of promise and exhibited tremendous improvements. Therefore, every ultrasound tongue image is subjected to filtering using this rank filter prior to model fitting.

3. Allowable shape domain. Another challenge encountered in implementing active shape model was establishing the allowable shape deformation of the modelled tongue. keeping in mind the set conditions of the proportion of variance *PoV* in identifying the first m principal components that capture 95% of the variation of the shape model, two methods suggested by Cootes et al. in [2] were extended in this thesis. The first method is to allow the variation (the value of the parameter **b**) of the obtained m principal components to vary between 3 standard deviations as shown in equation (3.28). The second method is to compute the parameter **b** using the equation (3.29). However, it is bound to the condition that the Mahalanobis distance (D_m) is less than a set maximum (D_{max}).

$$-3\sqrt{\lambda_k} \leq b_k \leq 3\sqrt{\lambda_k} \tag{3.28}$$

$$D_m^2 = \sum_{k=1}^{t} \frac{b_k^2}{\lambda_k} \leq D_{max}^2 \tag{3.29}$$

Both methods however returned bizarre results in model fitting. By trial and error, a method was designed that combined both formulae. Firstly, if the computed model parameter b was within 1 standard deviation, the shape was unaltered and the iteration continued. If, on the other hand, the model parameter deviated to a great degree, the second condition was checked. That is, the Mahalanobis distance must be less than a set maximum. Only if the model parameter fails to satisfy both conditions, the shape parameter is altered and scaled accordingly using the equation (3.21).

Chapter 4

Results

The aim of this project was to develop a collaborated model primarily to address the research questions (section 1.4). Also, assuming that the results and the answers obtained after evaluation of this thesis will establish a firm foundation in discussing the research questions put forth by the proposer. For experimentation, a total of 2900 ultrasound tongue images were captured. The data acquisition procedure is explained in section 3.1. The data was fragmented based on the sound made by the speaker while capturing it. The collected data was filtered using the set criteria (see section 3.2.1). After refinement, a total of 1177 ultrasound images were selected. These images represent three sounds, namely, [a], [l] and [o]. A total of four models are built. One for each individual sound and one representing the combination of all sounds. A total of 368 ultrasound images representing sound [a], 402 images of sound [l] and 407 images of sound [o] were annotated.

In the following sections, the results obtained after conducting various experiments are analysed. Firstly, the mean shape model of the tongue for each individual sound and one for the combined model is analysed, the variation obtained for these models is shown in the section 4.2. Section 4.3 sheds light on the image search results obtained after training the models. The drafted evaluation plan is comprehended and their results are examined in section 4.4.

4.1 The Mean Shape Model

The annotated ultrasound tongue images were subjected to statistical shape modelling as explained in section 3.2.2. The mean shape models obtained after training the statistical shape models using the annotated ultrasound tongue images are shown in Figure 4.1. It is evident from these figures that the mean shape of the tongue alters depending on the sound. Therefore, it is assumed that the individual models will achieve good results in comparison to the universal model. The assumption is based on the fact that the annotated points used to train these individual models belong to their respective sounds. On the other hand, the universal model is trained on a combined pool of annotated points of all the three individual sounds. These mean shapes are superimposed on the ultrasound tongue image that needs to be searched. That is, these are the initial shapes that their respective models use as a reference to search images for tongues.

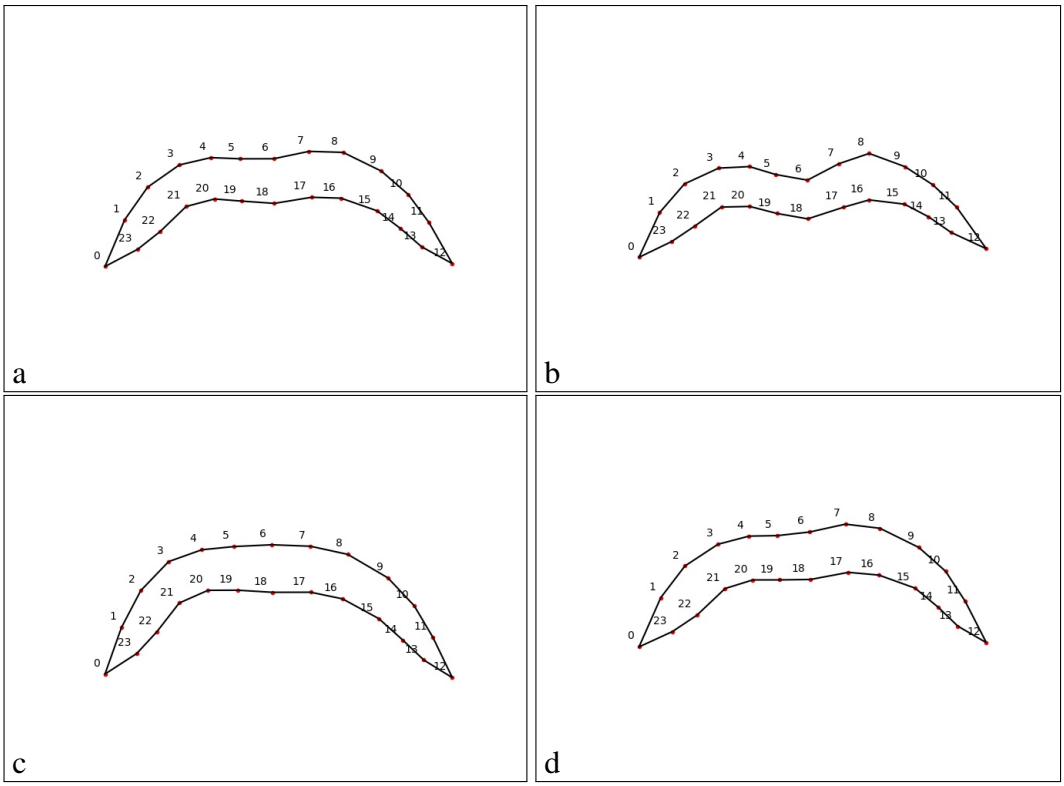


Figure 4.1: Mean shape models of: (a) universal model, (b) model representing sound [a], (c) model representing sound [l] and (d) model representing sound [o].

4.2 Captured Variation

The mean shape models and their respective aligned data were subjected to principal component analysis. Section 3.2.3 explains the methodology in detail. The requirement was to capture 95% of the variation. In order to achieve this, a proportion of variance *PoV* method was implemented.

The universal model. Table 4.1 shows the percentage of variation each principal component contributes to successfully capturing 95% variation in constructing the mean shape model for the universal model. From this table, it can be seen that the first three principal components are the major contributors to the variation of the mean shape model. Therefore, the variation across these three principal components is shown in Figure 4.2. To capture these variations, the mean shape is projected onto the model parameter space and reconstructed back in the image space. In projecting the shape to the model parameter space, only the parameter b_k representing the principal component is varied across one standard deviation. All the other parameters are kept constant at zero.

Figure 4.2 shows the variation across the first three principal components. The first principal component (parameter b_1) is responsible for the variation in the movement of the endpoints of the tongue. This variation results in capturing the tongue with different scales. The second principal component (parameter b_2), is responsible for the variation of control points that mark the curves of the tongue. From this figure, it can also be noted that the third principal component captures the variation of the central area of the tongue. It is important to understand that these tongue representations are the reconstructions of the tongue model from the model parameter space and are not the annotated tongue shapes. The variation depicted shows the possibility of encountering a similar tongue shape in UTI images.

The model representing the sound [a]. Table 4.2 shows the percentage of variation each principal component contributes in capturing 95% variation in constructing the mean shape model for the model representing the sound [a]. The first principal component captures 40.4% of the variation. This principal component is responsible for capturing the variation indicating different scales of the tongue model. Also, it can be seen in Figure 4.3, that it is responsible for the movement of the endpoints of the tongue. The second principal component, similar to the previous model, is responsible for capturing the variation in the parts representing the curve of the tongue. Although it captures the variation of the curves similar to the universal model, it can be observed that the variation is catastrophic. The third principal component captures the variation of the central area of the tongue. An interesting observation is that this principal component along with capturing the variation of the central area of the tongue also captures the variation in the left-part of the tongue indicating asymmetry.

The model representing the sound [l]. The percentage of variation contributed by each principal component can be seen in Table 4.3. The third principal component captures the variation in the central area of the tongue. However, the first principal component captures the variation in the left curve and the lower hemisphere of the tongue. The second principal component captures the variation of how the tongue scales. The variations captured are shown in Figure 4.4.

The model representing the sound [o]. The individual percentage of variation of each principal component is shown in Table 4.4 and the captured variation is shown in Figure 4.5. The variation captured by the principal components of this model are similar to that of the universal model. Although, the variation across third principal axes is negligible.

Table 4.1: Percentage of variation contributed by each principal component for universal model.

Principal Component	Eigenvalue	$\frac{\lambda_i}{\lambda_T} \times 100\%$
1	λ_1	30.6%
2	λ_2	24.1%
3	λ_3	14.2%
4	λ_4	9.2%
5	λ_5	3.9%
6	λ_6	2.0%
7	λ_7	1.9%
8	λ_8	1.3%
9	λ_9	1.2%
10	λ_{10}	1.2%
11	λ_{11}	1.1%
12	λ_{12}	1.0%
13	λ_{13}	0.9%
14	λ_{14}	0.9%
15	λ_{15}	0.7%
16	λ_{16}	0.6%

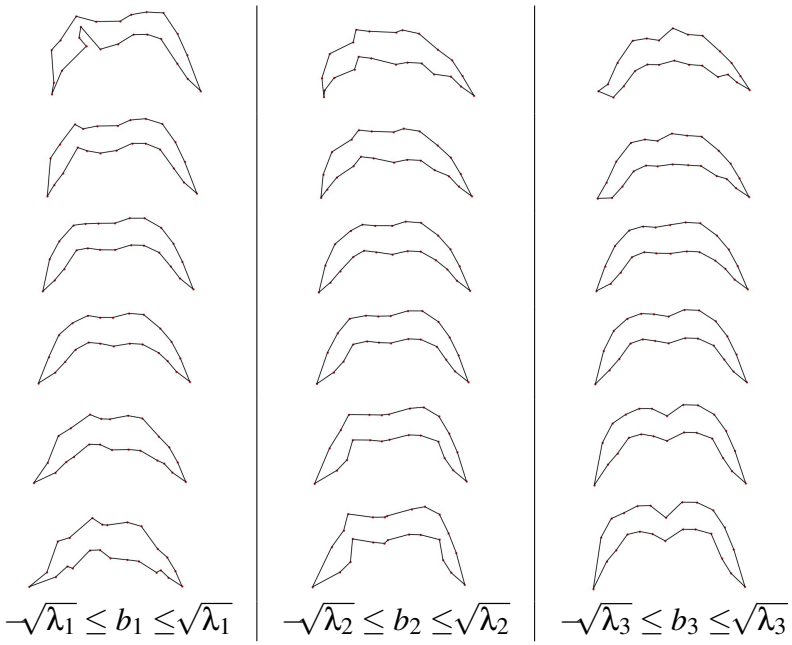


Figure 4.2: Variation across the first three principal components of the universal model.

Table 4.2: Percentage of variation contributed by each principal component for model representing sound [a].

Principal Component	Eigenvalue	$\frac{\lambda_i}{\lambda_T} \times 100\%$
1	λ_1	40.4%
2	λ_2	21.1%
3	λ_3	15.3%
4	λ_4	3.8%
5	λ_5	2.4%
6	λ_6	2.2%
7	λ_7	1.7%
8	λ_8	1.4%
9	λ_9	1.3%
10	λ_{10}	1.1%
11	λ_{11}	1.1%
12	λ_{12}	1.0%
13	λ_{13}	0.9%
14	λ_{14}	0.8%
15	λ_{15}	0.8%

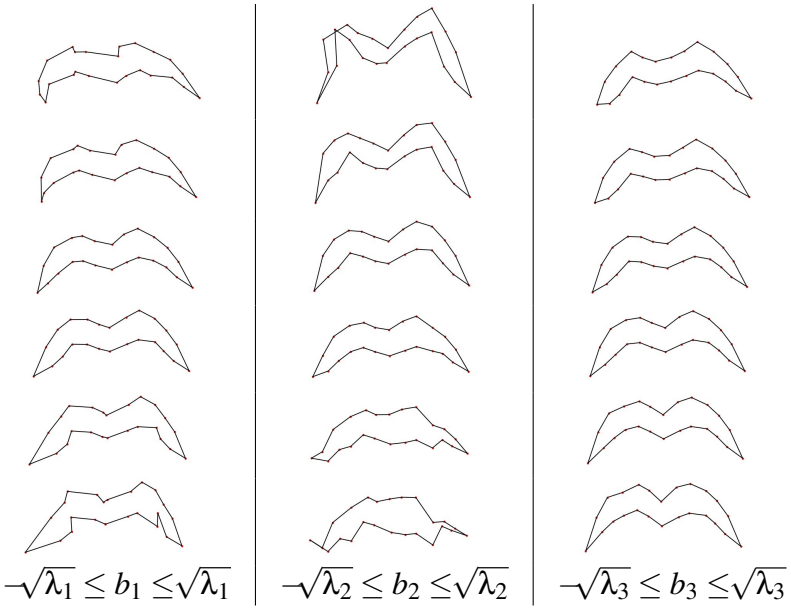


Figure 4.3: Variation across the first three principal components of the model representing sound [a].

4.3 Image Search

The aligned and trained tongue models aid the process of modelling a mean shape model that can be used by the object search methods as a starting point. Also, the

Table 4.3: Percentage of variation contributed by each principal component for model representing sound [l].

Principal Component	Eigenvalue	$\frac{\lambda_i}{\lambda_T} \times 100\%$
1	λ_1	32.5%
2	λ_2	26.1%
3	λ_3	13.3%
4	λ_4	7.8%
5	λ_5	4.4%
6	λ_6	2.0%
7	λ_7	1.8%
8	λ_8	1.4%
9	λ_9	1.3%
10	λ_{10}	1.2%
11	λ_{11}	1.0%
12	λ_{12}	0.9%
13	λ_{13}	0.8%
14	λ_{14}	0.6%

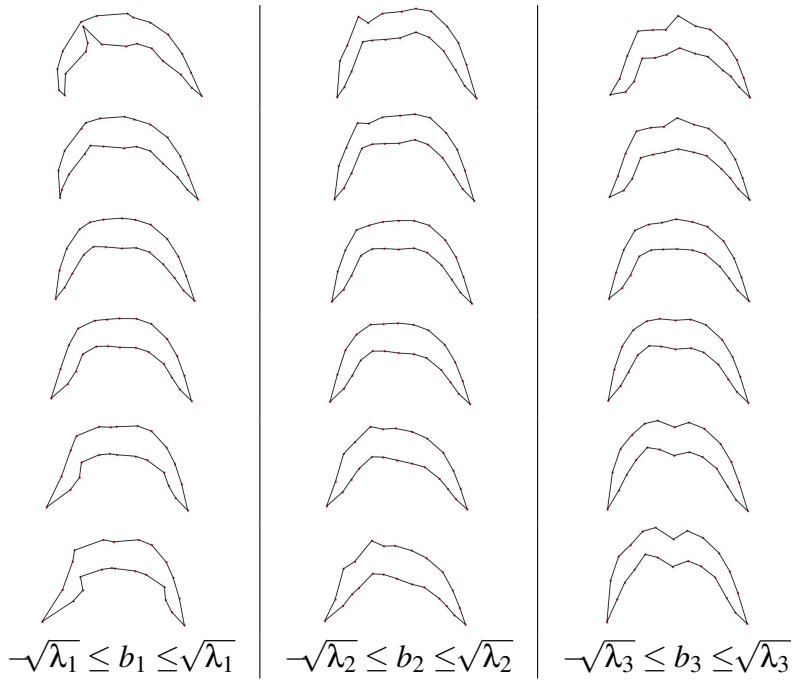


Figure 4.4: Variation across the first three principal components of the model representing sound [l].

Table 4.4: Percentage of variation contributed by each principal component for model representing sound [o].

Principal Component	Eigenvalue	$\frac{\lambda_i}{\lambda_T} \times 100\%$
1	λ_1	32.3%
2	λ_2	17.9%
3	λ_3	11.2%
4	λ_4	6.6%
5	λ_5	5.8%
6	λ_6	3.6%
7	λ_7	2.7%
8	λ_8	2.1%
9	λ_9	1.9%
10	λ_{10}	1.7%
11	λ_{11}	1.6%
12	λ_{12}	1.5%
13	λ_{13}	1.3%
14	λ_{14}	1.2%
15	λ_{15}	1.2%
16	λ_{16}	1.1%
17	λ_{17}	0.9%
18	λ_{18}	0.7%

variation captured is used by these methods in confining the shape of the model to preserve its meaning. Two image search methods are used as explained in the section 3.2. Namely, the active shape model approach and the random forest regression approach.

4.3.1 Active Shape Models

The methodology and implementation details of the active shape model image search algorithm are given in section 3.2.4. In the following paragraphs, the results obtained for each model are analysed.

Figure 4.6 shows a couple of search results in which, the universal model identifies the tongue contour in the UTI. The result obtained, that is, the exterior of the tongue contour accomplishes the task of locating the tongue. At the same time, manages to

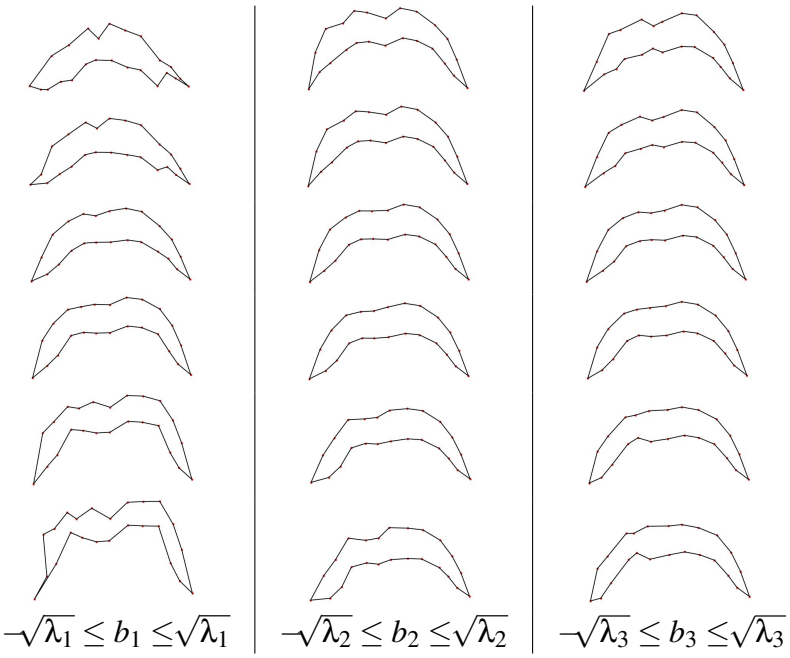


Figure 4.5: Variation across the first three principal components of the model representing sound [o].

satisfy the set criteria of model deformability. On the other hand, when the model encounters images such as the ones shown in Figure 4.7, it performs poorly. The reason for its poor performance can be due to the fact that the strong edge detected might not represent the boundary of the contour as shown in Figure 4.7(a). In other words, the boundary point of the object being searched might not be a strong edge. Another reason for the failure of this is due to the shape constraints. Figure 4.7(b) is a good example. The shape of the tongue in the UTI is not similar to the modelled mean shape and this shape is also not captured in just one standard deviation of the variation of the model. Therefore, in such images, this model tends to fail.

Figure 4.8 and 4.9 show the results obtained by the model representing the sound [a]. The model achieves to successfully label the tongue contour in the UTI as shown in Figure 4.8. Here, it can be seen that the model successfully captures the required variability. Figure 4.8 (a) and (b) have a different tongue shape. However, the model manages to identify both shapes. On the other hand, this model performs poorly for

UTI images shown in Figure 4.9. It is genuinely difficult to identify the tongue in such images. Having said that, the model tries to capture the shape and locate the tongue contour as can be seen in the Figure 4.9(b).

The results obtained by the model representing the shape [l] are shown in Figure 4.10 and 4.11. The model successfully manages to locate the tongue in UTI where the shape of the tongue is similar to the mean shape model with minor variations as seen in Figure 4.10. The model undoubtedly fails to precisely draw the tongue contour in images where even an expert (human) cannot identify. This can be seen in the Figure 4.11 (a). The model does not perform expectedly when the variation of the shape of the tongue in UTI is not within the defined limits as seen in the Figure 4.11 (b).

Figure 4.12 shows the satisfactory results obtained by the model representing the sound [o]. The model's performance is poor in UTI images shown in Figure 4.13 due to the discussed reasons.

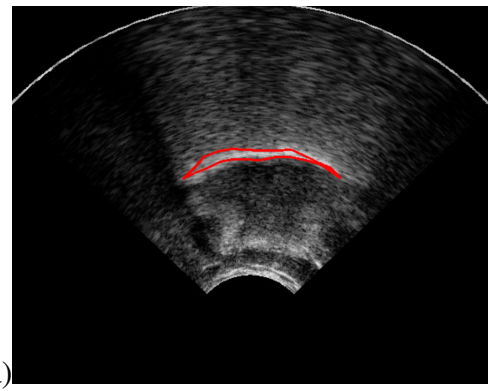
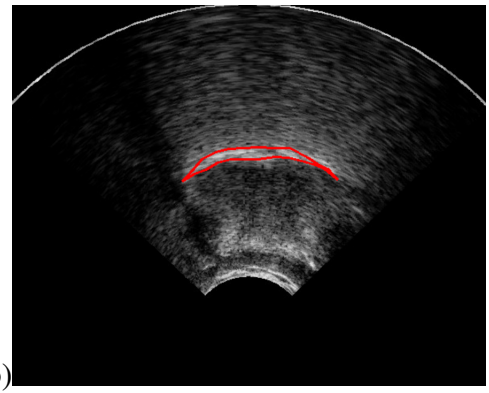
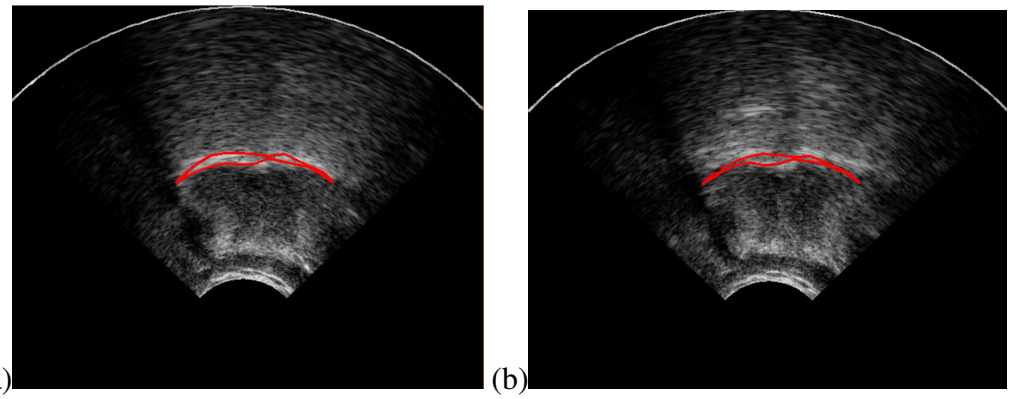
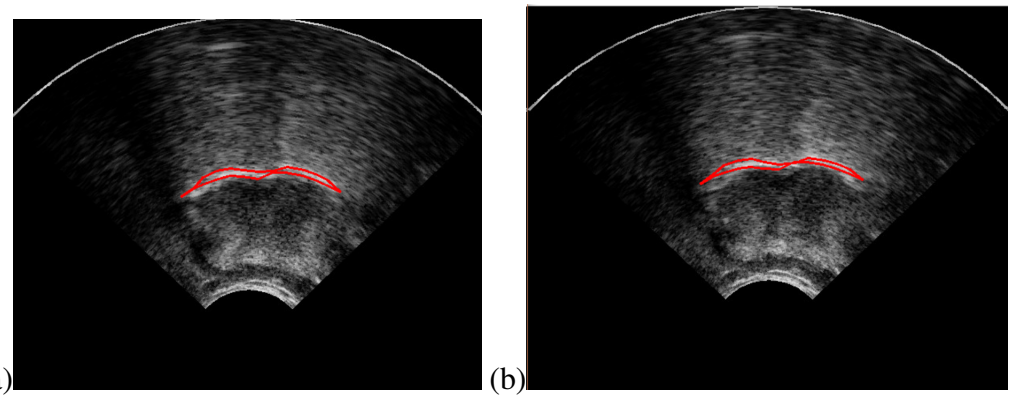
(a)  (b) 

Figure 4.6: Satisfactory search results of the tongue in UTI representing the universal model using ASM.



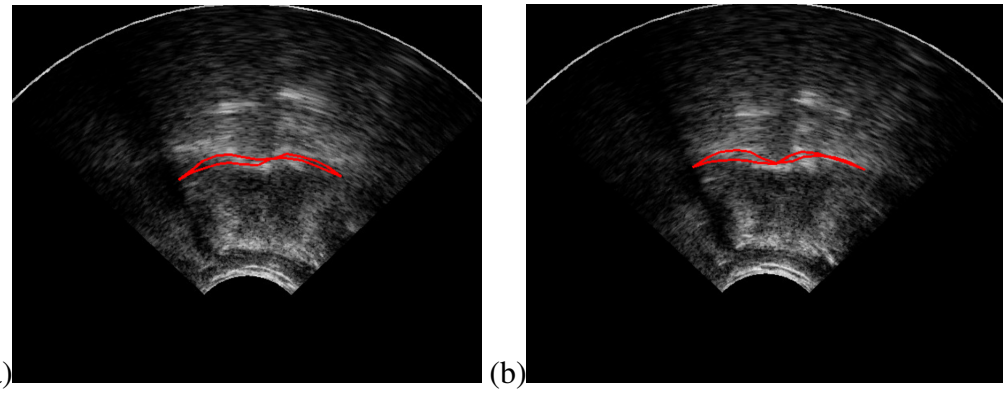
(a) (b)

Figure 4.7: Poor search results of the tongue in UTI representing the universal model using ASM.



(a) (b)

Figure 4.8: Satisfactory search results of the tongue in UTI representing the sound [a] using ASM.



(a) (b)

Figure 4.9: Poor search results of the tongue in UTI representing the sound [a] using ASM.

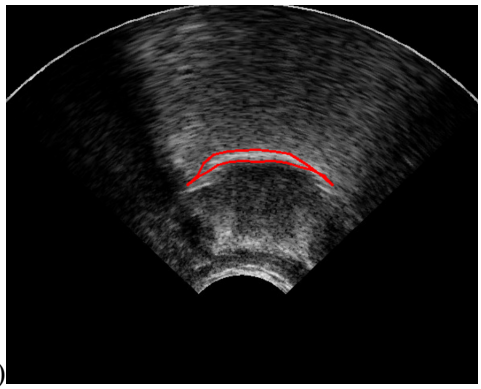
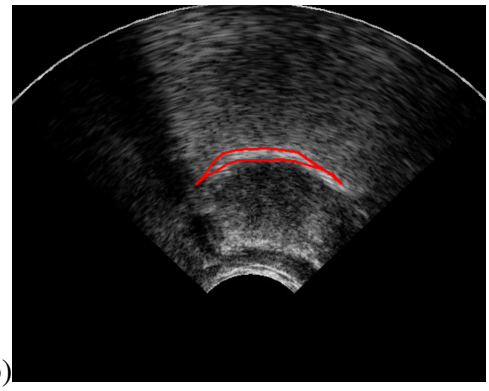
(a)  (b) 

Figure 4.10: Satisfactory search results of the tongue in UTI representing the sound [l] using ASM.

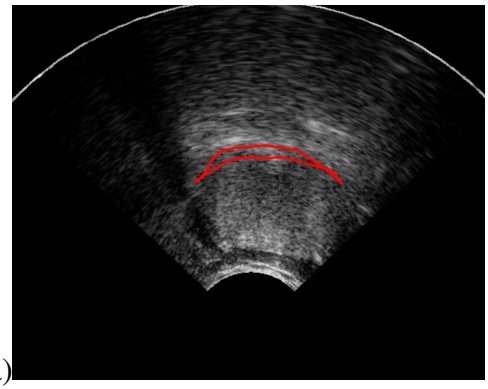
(a)  (b) 

Figure 4.11: Poor search results of the tongue in UTI representing the sound [l] using ASM.

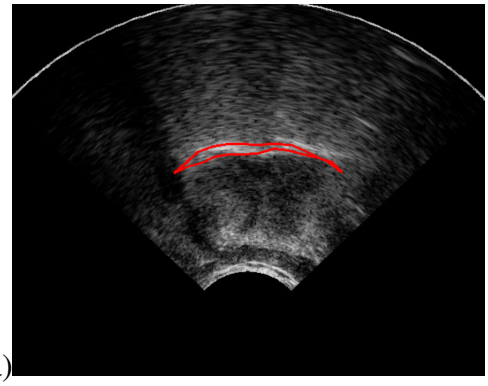
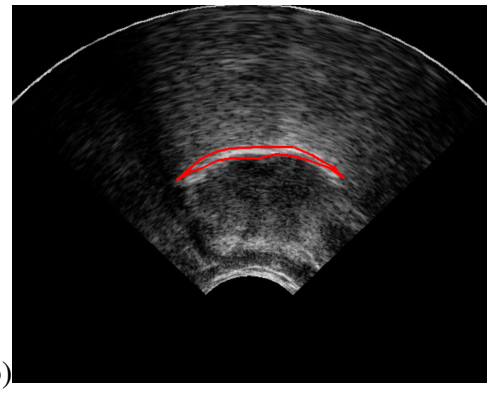
(a)  (b) 

Figure 4.12: Satisfactory search results of the tongue in UTI representing the sound [o] using ASM.

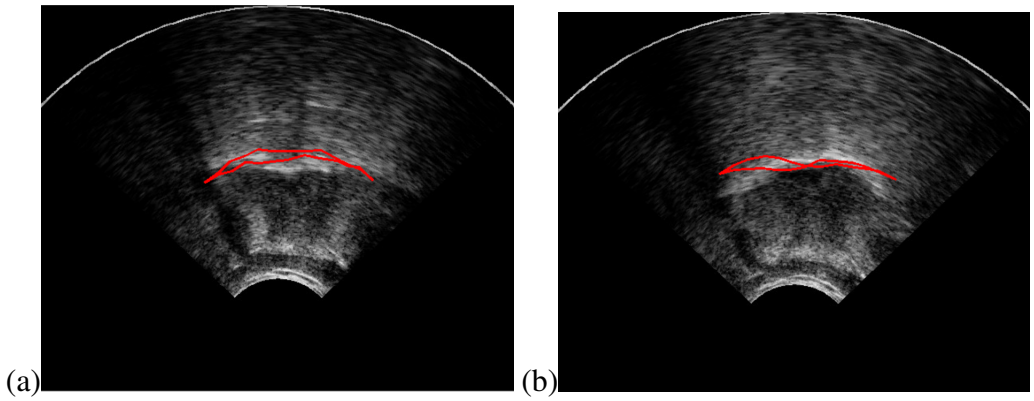


Figure 4.13: Poor search results of the tongue in UTI representing the sound [o] using ASM.

4.3.2 Random Forest Regression Voting

Section 3.2.5 elaborately explains the methodology of random forest regression voting approach in image search. A constrained local model built with the ability of regression voting to select an optimal position of a point in the modelled shape is used for searching UTI images. A single stage model with a maximum of 16 modes that captures 95% proportion of variance and a frame width of 15 pixels around each model point is used. The results obtained using all four models are analysed below.

Figure 4.14 shows the result of the universal model. The located shape in these images can further be refined. However, if the shape is subjected to refinement, the model does not guarantee a sensible tongue shape model. Although this method is robust and more promising, it still fails to locate the tongue in UTI images such as the ones shown in Figure 4.15.

Figure 4.16 shows how good this method is for a model representing the sound [a]. It precisely locates the tongue in UTI. However, it does not perform this well in certain images, such as the one shown in Figure 4.17. Nevertheless, it is not a bad approximation.

Figure 4.18 shows close to perfection results of model representing sound [l]. This model, however, performs not so well in UTI images where the identification of tongue is extremely difficult even to the naked eye. A sample of this is shown in Figure 4.19.

The model representing sound [o] yields good results as shown in Figure 4.20. Nonetheless, it fails to identify the tongue in images such as the ones shown in Figure 4.21. This might just be a result of the set parameters. There is a possibility that the model might perform well with different parameter settings.

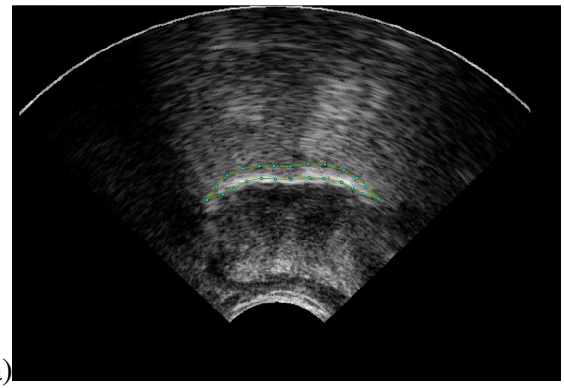
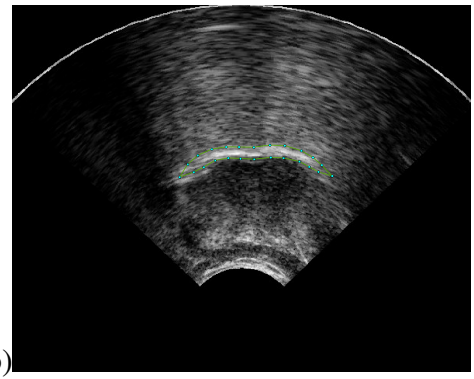
(a)  (b) 

Figure 4.14: Satisfactory search results of the tongue in UTI representing the universal model using RFRV.

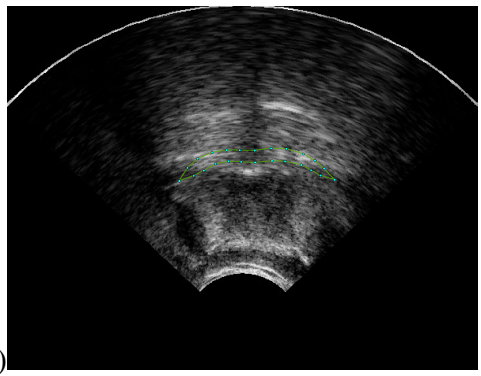
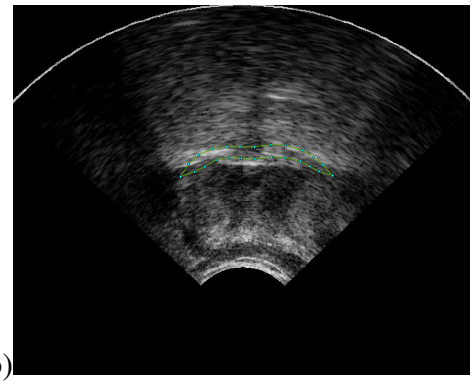
(a)  (b) 

Figure 4.15: Poor search results of the tongue in UTI representing the universal model using RFRV.

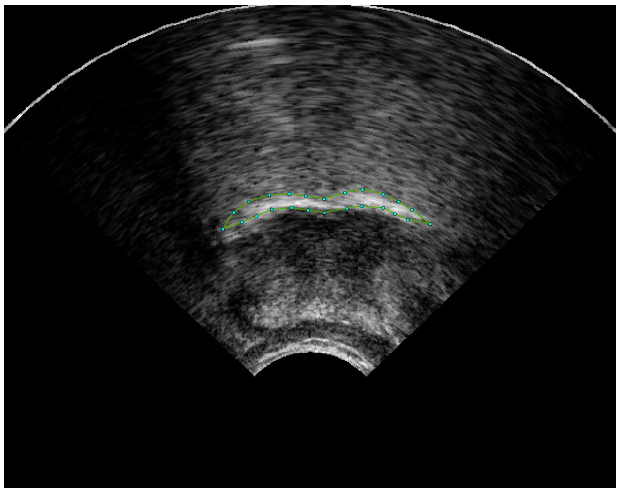


Figure 4.16: Satisfactory search result of the tongue in UTI representing the sound [a] using RFRV.

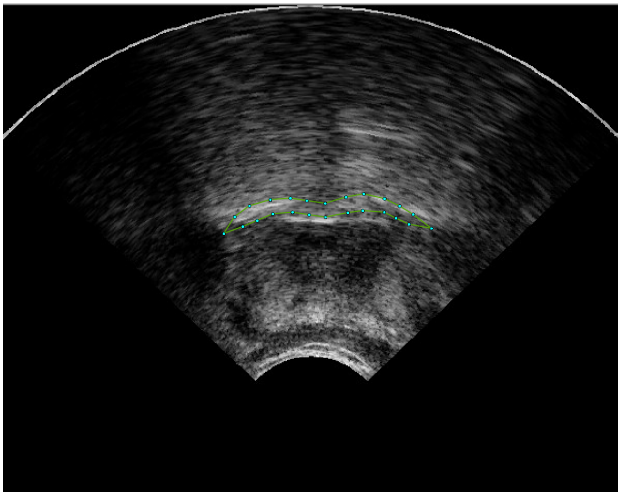
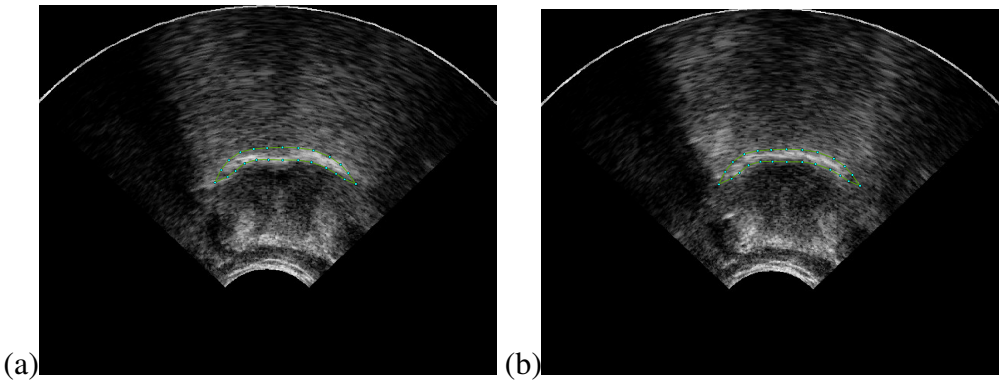


Figure 4.17: Poor search result of the tongue in UTI representing the sound [a] using RFRV.

4.4 Evaluation

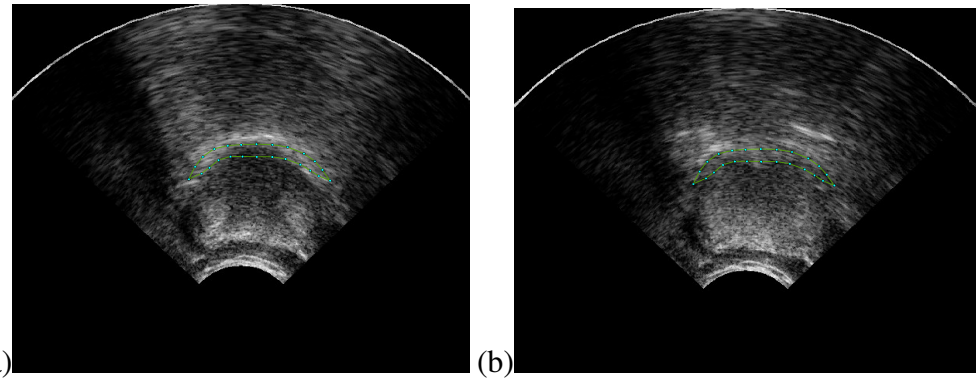
The testing and evaluation procedure followed in this thesis is as follows:

1. *Testing active shape model:* A figure of merit is obtained by training the models and testing them against the gold standard. Also, training and testing all four models using 10-fold cross validation.



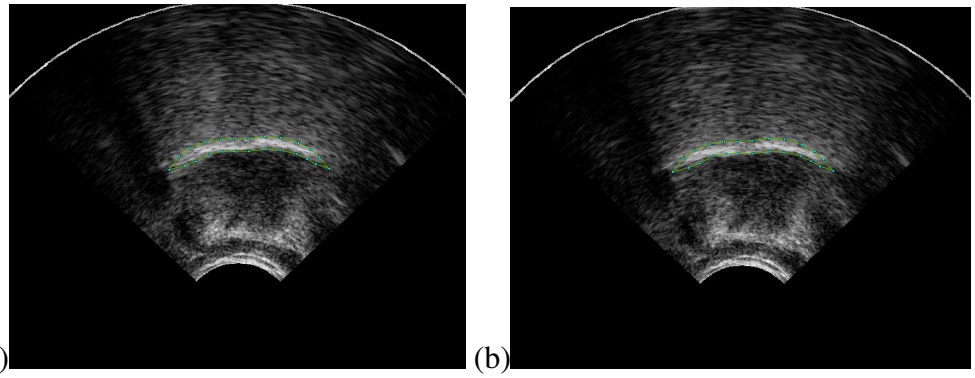
(a) (b)

Figure 4.18: Satisfactory search results of the tongue in UTI representing the sound [l] using RFRV.



(a) (b)

Figure 4.19: Poor search results of the tongue in UTI representing the sound [l] using RFRV.



(a) (b)

Figure 4.20: Satisfactory search results of the tongue in UTI representing the sound [o] using RFRV.

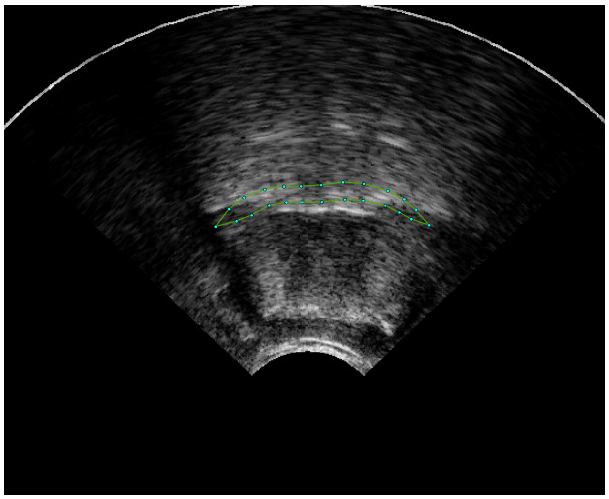


Figure 4.21: Poor search result of the tongue in UTI representing the sound [o] using RFRV.

2. *Testing random forest regression voting:* Training and testing models by 10-fold cross validation. Also, reviewing the performance of models based on cumulative distributive frequencies (CDFs).

3. *Evaluating the project:* The project is evaluated by comparing the performance of the above two methods. In that, a comparison of root mean squared error (RMSE) of each model.

4.4.1 Testing ASM

To test and evaluate the active shape model method, two independent strategies were implemented. A figure of merit was obtained by using the equation (4.1). Here, the average of the sum of Euclidean distances of the identified tongue points in UTI and the gold standard (annotated points) is computed for each annotated image. This method is also repeated with the mean shape model of the respective models.

$$fom = \frac{\sum_{j=0}^{n-1}\sqrt{(x_{ij}-X_{ij})^2+(y_{ij}-Y_{ij})^2}}{n} \tag{4.1}$$

where, x_{ij}, y_{ij} represent a predicted point j of shape i and
X_{ij}, Y_{ij} represent a point j of shape i, that is, gold standard.

$$dist = pixels \times size_{pixel} \tag{4.2}$$

Table 4.5 shows the figure of merit computed against the gold standard. The error values are computed both in terms of the number of pixels and the distance in millimetres (mm). The distance in mm is calculated by using the equation (4.2). From this table, it is evident that the model representing the sound [o] has the least distance error of 1.81 mm with a standard deviation of 0.5 and as a result, it is the best model. Table 4.6 shows the figure of merit computed against the respective mean shapes. This table also signifies that the best model is the one representing the sound [o] with an error of 0.14 ($\pm$ 0.11) mm.

The second strategy implemented for testing and evaluating models with active shape model is the 10-fold cross validation. Each model is trained 10 times. Every time, it is trained on 9 folds and tested on the remaining fold. A root mean squared error as shown in equation (4.3) is computed for each test data and averaged over the testing fold. This averaged error represents the overall error of the test fold. Error bars of each model over all the folds are plotted.

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n-1} (\mathbf{x}_{pred} - \mathbf{x}_{original})^2}{n}} \tag{4.3}$$

4.4.2 Testing RFRV

To test and evaluate the models using random forest regression voting, 10-fold cross validation strategy is used. The training data is fragmented into 10 folds. Every model is trained on 9 folds and tested on the remaining fold. The process is repeated for all

folds. The RMSE is computed and error bars are plotted. Also, a cumulative distributive frequency (CDF) is computed for each fold. The CDFs of each fold are plotted against point to point errors. A measure of performance is noted over all the folds. Figure 4.26 shows the relative performance graphs of all the four models. It can be observed that the performance of all the folds of the universal model and the model representing the sound [o] are consistent. The performance of the model representing the sound [a] on the other hand, shows the CDFs of folds scattered and probably also affecting the overall error.

4.4.3 Comparison

The 10-fold cross validation strategy was enforced on models with active shape model and models with random forest regression voting. The root mean squared error of each fold along with its standard deviation is plotted for each model. Figure 4.22 shows error bars of the universal model with ASM and RFRV. It can be observed that both methods follow similar trends. However, random forest regression voting has an upper hand. In that, the mean errors are relatively low and also, the deviation of the error rate is high only for one-fold (fold 5). On the other hand, ASM has three folds with relatively high deviation.

Figure 4.23 shows a plot of error bars of the model representing the sound [a] using both ASM and RFRV. These plots can create some ambiguity. In that, they look very similar. However, the error rate of the ASM method is less than that of the RFRV. Therefore, making ASM the winner. This can be justified by highlighting two points. First, although two folds have very high deviation, the deviation of the rest of the folds is minimal and the error rates are consistent. Second, the error rates of RFRV are inconsistent and the error of the fifth-fold is very high with a very high deviation.

The error bars of ASM and RFRV of the model representing the sound [l] are plotted in Figure 4.24. The unequivocal winner here is RFRV. The ASM has high error rates and the deviation of one-fold is very high. On the contrary, RFRV has low and consistent error rates. Also, their deviation is very minimal.

The error bars of the model representing the sound [o] are plotted in Figure 4.25. The comparison of this model is tough. Both ASM and RFRV methods have similar error rates and both display similar trends. The deviation of the folds of both methods are not very high and the error rates are also consistent. However, having a close look at both the plots, it is observed that the RFRV outperforms ASM.

The 10-fold cross validation errors of each individual model are computed by averaging the RMSE of each fold of the model. The averaged errors are listed in Table 4.7. These values can be addressed in two ways. First, comparing the 10-fold cross validation errors of all the models with respect to each method. That is, comparing the cross validation error of all models with respect to ASM, the model representing the sound [o] undoubtedly performs better than others. It has an error rate of 9.21 ($\pm$0.59) pixels. A low error rate and a standard deviation of 0.59 indicating its consistency, makes it obvious. Having said that, the model representing the sound [a] also does well with an error rate of 9.34 ($\pm$1.17) pixels. Finally, it can be said that the main contributor to a high error rate of the universal model is the training set that represents the sound [l] with an error rate of 10.66 ($\pm$1.08) pixels. Now, looking at RFRV, the model that performs the best is the model representing the sound [l] with an error rate of 8.55 ($\pm$0.71) pixels. Nevertheless, all the other models have a relatively low error rates and perform equally well.

The second way of addressing the 10-fold cross validation error rates is by comparing the two methods. The comparison of ASM and RFRV results in RFRV outperforming ASM. In that, RFRV has consistent error rates. It has better error rates than ASM

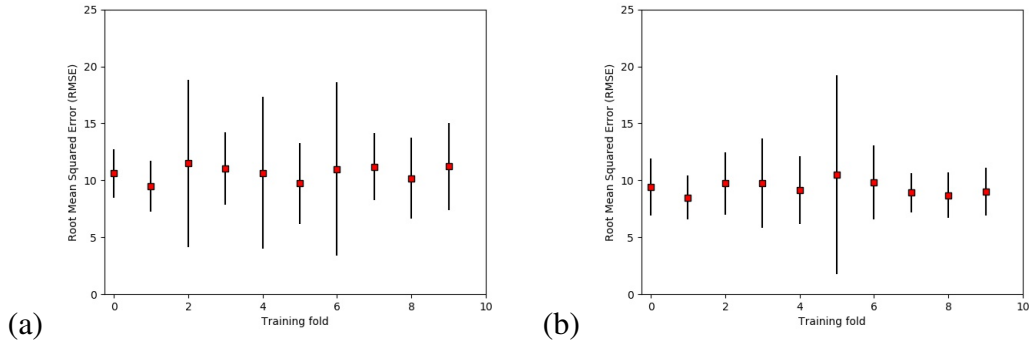
Table 4.5: Figure of Merit (FoM) computed against gold standard.

Sounds	Minimum		Maximum		Mean		Standard Deviation	
	pixels	mm	pixels	mm	pixels	mm	pixels	mm
All	5.02	0.81	108.81	17.52	13.25	2.13	5.97	0.96
[a]	5.38	0.86	105.57	16.99	11.57	1.86	7.48	1.20
[l]	7.10	1.14	106.96	17.22	13.19	2.12	5.76	0.93
[o]	5.51	0.88	22.46	3.61	11.26	1.81	3.13	0.50

Table 4.6: Figure of Merit (FoM) computed against mean model.

Sounds	Minimum		Maximum		Mean		Standard Deviation	
	pixels	mm	pixels	mm	pixels	mm	pixels	mm
All	0.53	0.09	8.24	1.33	1.30	0.21	0.96	0.15
[a]	0.86	0.14	6.01	0.97	1.56	0.25	0.73	0.12
[l]	0.48	0.08	8.54	1.37	1.38	0.22	1.08	0.18
[o]	0.36	0.06	6.00	0.97	0.88	0.14	0.70	0.11

for three models. Namely, the universal mode, the models representing the sounds [l] and [o]. On the other hand, error rates of ASM are inconsistent and it manages to outperform RFRV only using the model representing the sound [a]. The margin, however, is minimal. Nevertheless, the bottom line is that the RFRV outperforms ASM.



(a) (b)

Figure 4.22: Root mean squared error (RMSE) of each fold of the universal model using: (a) ASM, (b) RFRV.

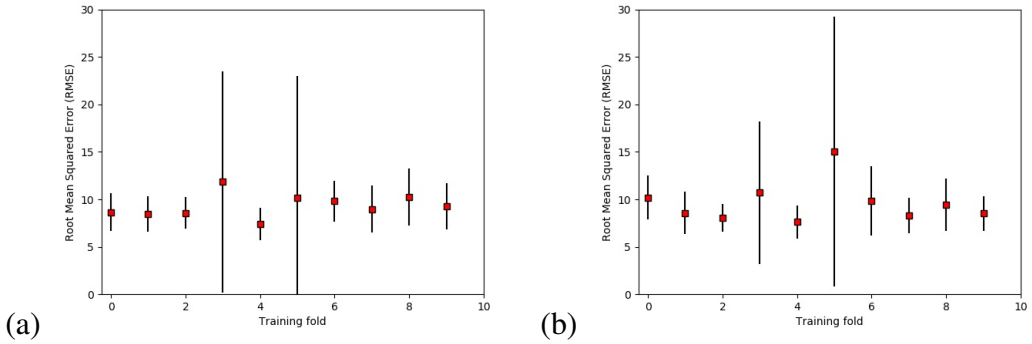


Figure 4.23: Root mean squared error (RMSE) of each fold of model representing sound [a] using: (a) ASM, (b) RFRV.

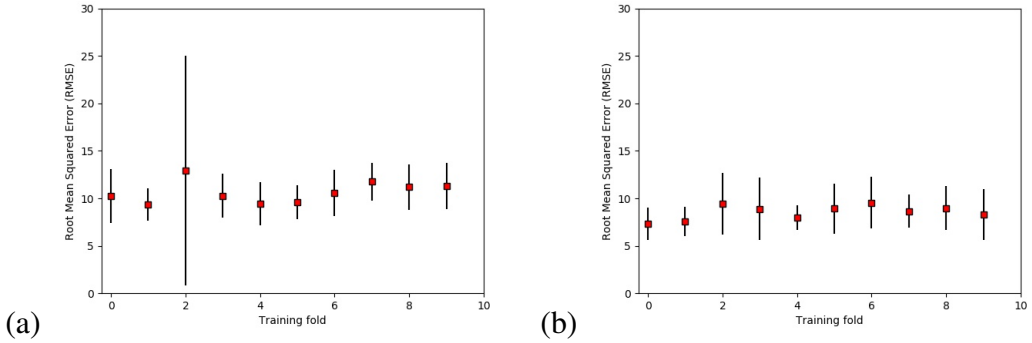


Figure 4.24: Root mean squared error (RMSE) of each fold of model representing sound [l] using: (a) ASM, (b) RFRV.

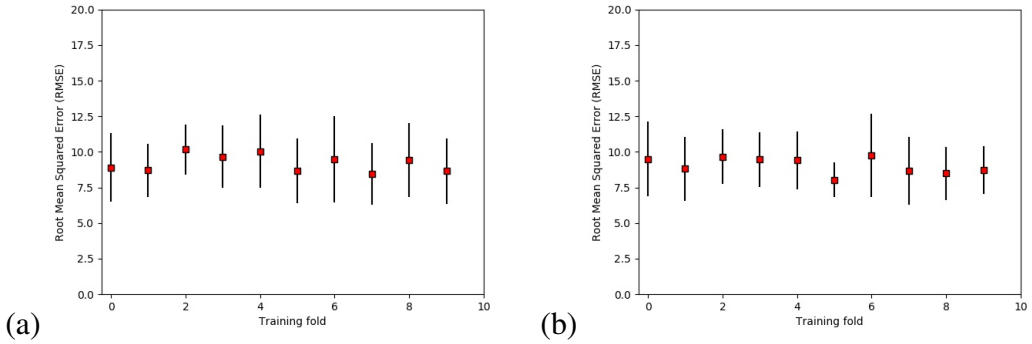


Figure 4.25: Root mean squared error (RMSE) of each fold of model representing sound [o] using: (a) ASM, (b) RFRV.

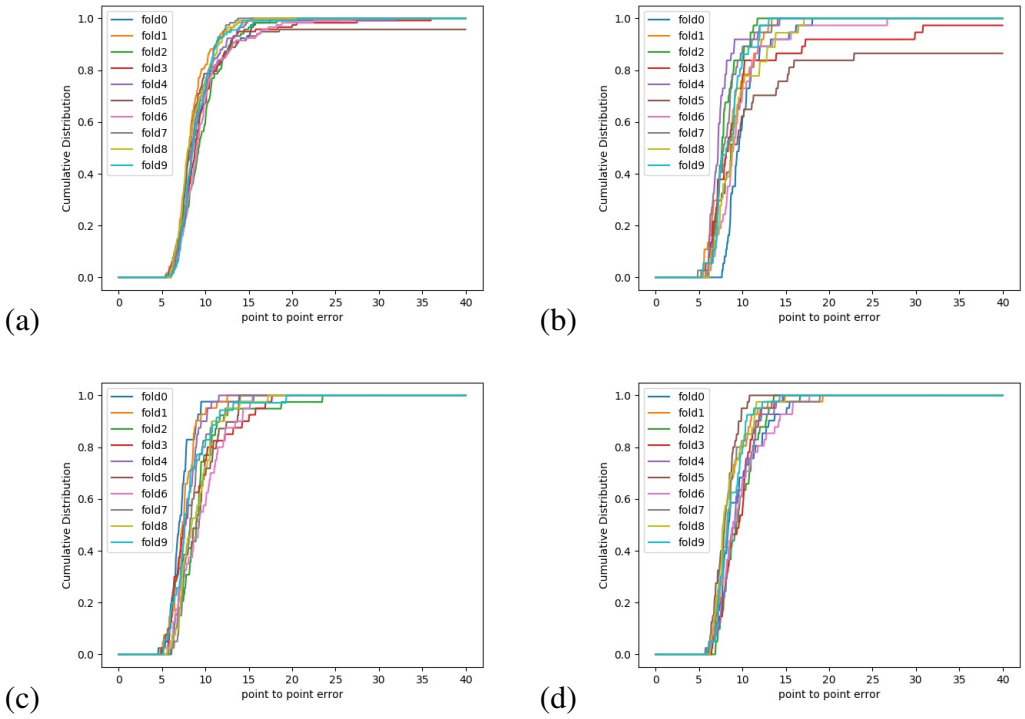


Figure 4.26: Performance graphs (Cumulative Distribution Frequency of all 10 folds) of: (a) universal model, (b) model representing sound [a], (c) model representing sound [l] and (d) model representing sound [o].

Table 4.7: 10-Fold Cross-Validation Errors.

Sound	ASM		RFRV	
	Error (pixels)	Standard Deviation (pixels)	Error (pixels)	Standard Deviation (pixels)
[a]	9.34	1.17	9.62	2.03
[l]	10.66	1.08	8.55	0.71
[o]	9.21	0.59	9.05	0.54
All	10.66	0.63	9.35	0.58

Chapter 5

Discussion

The aim of this thesis was to identify, locate and draw the contour of the tongue in ultrasound tongue images. In order to achieve this, a mean shape of the tongue was first constructed with the help of statistical shape models. Later, with the help of image search algorithms, a method was developed to find the tongue contours in UTI. This thesis intends to build a firm foundation for understanding the basic skeleton of the tongue in UTI. It also aims to provide a direction in addressing the issue of tongue lateralization. Keeping in mind the aims and objectives of this thesis, a number of research questions were drafted. In this chapter, the research questions will be addressed. It also provides an insight of the results that influence in addressing the research questions.

5.1 How can the shape of the tongue be modelled?

A basic literature review was conducted on the working of ultrasound images. This indeed provided a primitive knowledge of ultrasound images, their working and how to interpret them. Having understood the prerequisite, a tongue shape was designed (see section 3.2.1). This design addressed various parts of the tongue shape model and their significance in building tongue shape models. Through a systematic literature review, it was decided that statistical shape modelling was the right way forward in modelling the annotated tongue shape models.

The image searching methods implemented in this thesis are primarily based on the models built using statistical shape modelling. Results suggest that the choice of methods adopted for modelling tongue movements in this thesis are appropriate and adequate. Having said that, it would be interesting in further investigating the following.

- Investigating more about the ultrasound images. An in-depth knowledge of ultrasound technology and ultrasound images might help a developer in grasping the finer details. The knowledge can be replicated in either designing the shape of the tongue to be modelled or in processing the ultrasound images for better modelling.
- More attention to the design. Although the tongue shape model designed in this thesis proved to be trustworthy, it would be fascinating to experiment with the shape of the tongue. One idea is to experiment with shape models by using a 1-dimensional line to represent the tongue.
- Presuming a 1D line model, will the shape models capture the variability in the shape of the tongue? Will the model better fit the tongue in UTI? Perhaps a more detailed investigation might help answer these questions.

5.2 What amount of variation is allowed for the deformable object (modelled tongue)?

The methods adopted in this thesis to search ultrasound images for the built tongue shape model require a constraint on the flexibility of the modelled tongue. In other words, image search algorithm should know in prior, the shape it is looking for in the image and to what extent can that shape be deformed in order for it to identify the shape as the desired object. The problem is not trivial. The modelled object that needs to be searched is not rigid. Therefore, it is genuine to expect some change in its shape.

A number of attempts were made, various approaches were proposed and deliberated to tackle this problem. These attempts and approaches were reviewed in chapter 2 and a decision on how to tackle this problem was made.

The principal component analysis method captures the variation of a dataset. It computes the principal axes for reference. Varying parameters along these axes will generate all the possible deformations of a given object. However, a limit needs to be defined for this variation. In this thesis, the built tongue shape models were subjected to PCA. Experiments were conducted to understand the variation. An elaborated explanation of the results of capturing the variation in tongue shape models is in section 4.2.

In capturing the variation of the tongue shape models, a combination of principal components that contributed to at least 95% of the variation, was selected. The image search algorithms superimpose the mean shape model over the ultrasound image. Strong edges are detected along the normal of each point and these are used as a reference to compute the model space parameters. This is a vital step in image searching due to the fact that the deformability of the shape of the identified object in the image should be restricted to a set limit.

This can be better explained using an example. If the model after detecting the strong edges, moves the points in the direction of the strong edges, there is absolutely no guarantee that the new shape will represent the desired object. Therefore, the suggested changes are used to project the shape onto the model parameter space. The model then checks if the suggested points belong to the allowable space domain. If yes, it reconstructs the shape in the image space using the suggested movements. Otherwise, it applies constraints to scale the parameters to fit the shape model, also considering the movement in the suggested direction. In order to keep the shape of the tongue in check, the allowable range of deformation is set to one standard deviation of

the computed model space parameter.

5.3 Can the tongue be identified in a given ultrasound tongue image (UTI)?

This thesis considers active shape models to be the prime method for detecting and marking tongue contours in ultrasound tongue images. In order to achieve this, that is, to identify the tongue in ultrasound images, the model builds a mean shape of the tongue. This process along with the design chosen for building the mean shape model of the tongue has already been addressed. The next step is to capture the allowable variation of the built tongue shape model, which is also addressed. The active shape model follows an iterative procedure for identifying the tongue. This procedure is explained in detail in section 3.2.4. However, there are certain issues that need to be discussed.

Finding the initial pose parameters. The first obstacle encountered in building this model to achieve the goal of identifying the tongue in ultrasound images was to detect the initial pose parameters. Each shape has three pose parameters. The scale of the modelled shape, its angle and it's translation. However, the critical question here was: how to compute these parameters? In order to compute the parameters, there needs to be a reference shape onto which it can be mapped and the pose parameters can be determined. To tackle this challenge, the centroid of the mean shape model is computed. The centroid marks the origin of the mean shape model. The initial mean shape was mapped onto the computed centroid and the pose parameters were determined.

Computing the normal. The next challenge encountered was: computing the normal. A normal cannot be computed for a single point. To overcome this challenge, the slope of the line connecting two adjacent points was computed and the slope of its normal was determined. This slope along with the coordinates of the points to which

a normal has to be computed was substituted in the equation of a line to compute a set of points along the normal. The difficulty does not end there. The slope of the normal cannot be determined for a point from which, the line joining its adjacent point is parallel to the x-axis. To overcome this problem, a condition is set. If the angle of the slope of the computed normal is in the range $(80^\circ \leq \theta \leq 100^\circ)$, the pixel values that are perpendicular to the point were considered.

Detecting strong edges. The challenge of detecting strong edges is in a way related to the previous challenge of computing normal. If the normal was computed conventionally, that is, the method specified in the previous challenge was not followed. This indicates all the points that do not fall under the previously set category. The normal to these points are inclined at an angle and therefore, a few pixels might be skipped. This implies that the detected strong edge might not be the strongest edge as there is a loss of information. To overcome this challenge, a window can be implemented. A method can be developed in which, a set of pixel values are considered when computing the normal.

Reason for poor performance. The results evidently show that the random forest regression approach performs better than the active shape model. This can be due to loss of information. The annotated points of the tongue shape are of floating point. However, when searching for points along the normal, the image values are considered. Images in python are represented as a matrix. Each pixel has an index. Therefore, the indices of any matrix are of integer type. In converting the point coordinates from floating point to integer, there is a loss associated. However, this loss is not very significant.

Addressing these challenges while following the procedure specified in section 3.2.4 simplified the process of identifying the tongue in ultrasound tongue images.

5.4 Can an alternative method apart from active shape model be used to achieve the goal?

A major part of this thesis was dedicated to reviewing shape models that were robust and adequate to achieve the aim of this project. A number of models were reviewed in chapter 2. The model that was investigated as an alternative to the active shape model was the random forest regression voting approach. The working of this method is explained in section 3.2.5. Both the models using active shape model and random forest regression voting methods were subjected to 10-fold cross validation for testing their effectiveness and performance. As shown in the results, models using the random forest regression voting method are more efficient and clearly outperforms models using active shape modelling methods (a majority of the models). To that end, random forest regression voting indisputably is a better alternative to active shape models.

5.5 How do the results vary and what factors influence the results?

The results obtained and the factors due to which they vary are discussed in chapter 4. A number of factors that influence the results are discussed here. First, the choice of data. Ultrasound images, in general, contain a lot of noise or unwanted speckles. This can lead to loss or improper recording of information during the data acquisition phase. Second, the annotator. This is perhaps the most crucial influential factor. How the results vary, depend on how the data is annotated. Any misplacement of the landmark points while annotating will add a great deal of variation to the mean shape model. The annotated data is considered to be the gold standard for evaluating the results. Therefore, this plays a crucial role in indicating the goodness-of-fit of a model. Third, the mean shape of the tongue: Once the points are marked and aligned, the mean tongue model is generated. There is a possibility that the model fails to converge to generate the mean shape of the tongue. Fourth, the design of the shape of the object to

be modelled. As discussed earlier, the design can also have an impact on the results. Fifth, captured variation. A careful consideration of all the possible outcomes with an increase in variation and their impact on the results must be analysed. Sixth, choice of search methods. Different methods yield different results. It is difficult to point out a single parameter that will play a major role in influencing results. The result yielded is the combined effect of the chosen set of parameters.

Chapter 6

Conclusion

6.1 Conclusion

Modelling tongue movements a computer vision domain related project aims at identifying and locating tongue contours in ultrasound tongue images (UTI). This is achieved in order to further investigate and address the issue of tongue lateralization. An extensive literature was reviewed. The literature was narrowed down to the research addressing the role of actuators in phonetics. The role of the tongue in phonetics more in particular. However, no literature was found on tongue lateralization. Therefore, examination of shape models was required to address a set of fundamental questions and build a primitive model that can be extended in addressing the issue of tongue lateralization.

This thesis studies statistical shape models for building tongue shape models. The data was annotated and was used for building tongue shape models. During this process, an essential research question of how to design and model tongue was addressed. An in-depth discussion leaves an open-ended question about the design of the tongue shape models. Nevertheless, the tongue shape model built using statistical shape modelling turned out to be adequate.

The mean shape models obtained after training and aligning the dataset were investigated for deciding on the constraints of deformability. A number of observations were made during this process. Factors affecting variations were realized. Methods were derived to restrict the variation to a degree so as to preserve the meaning of the modelled shape. In this case, to preserve the meaning of the modelled tongue. In this process, another crucial research question on capturing the variation of the modelled shape was discussed. This thesis succeeds in justifying the decision of restricting the variation of the modelled points in the model parameter space to just one standard deviation. It also keeps the variation in check by computing Mahalanobis distance in the model parameter space and restricting it to a set maximum. A combination of these two important formulae manages to preserve the shape of the tongue during the process of image search.

This thesis proceeds further in investigating the effectiveness of two image search approaches. Namely, active shape models (ASM) and random forest regression voting (RFRV) methods. The literature of these two methods was studied and implemented. Several diverse challenges were encountered. These challenges are acknowledged and addressed in this thesis. A test and evaluation plan was drafted to evaluate the effectiveness of the two methods. The models using both the methods were trained and tested using 10-fold cross validation.

The root mean squared error (RMSE) of the two methods are compared. The error rate of models using active shape modelling approach was large. The error rate of the universal model representing the shape models of all the sounds had an error rate of 9.35 ($\pm$0.58) pixels using random forest regression voting as opposed to 10.66 ($\pm$0.63) pixels using active shape modelling. The model representing the sound [a] had an error rate of 9.34 ($\pm$1.17) pixels using active shape modelling as compared to 9.62 ($\pm$2.03) pixels using random forest regression voting. A commendable error rate of 8.55 ($\pm$0.71) pixels using RFRV against 10.66 ($\pm$1.08) pixels using ASM for the

model representing the sound [l]. Finally, the model representing the sound [o] had an error rate of 9.21 ($\pm$0.59) pixels using active shape model method which is in close range to the error rate of 9.05 ($\pm$0.54) pixels using the random forest regression voting method. With sufficient proof, it can be concluded that the random forest regression voting method is a more efficient, adequate, optimal and robust method compared to the active shape modelling method.

6.2 Future Work

This thesis discusses the challenges of using ultrasound tongue images in modelling tongue movements. An alternative to this can be explored and investigated. The ultrasound tongue images can be refined so as to reduce the number of unwanted noise/speckles in the images. The results of this project are very sensitive. In that, they depend on the set gold standards. This can be considered a major vulnerability of this project due to the fact that the data is annotated by an individual. The gold standard has the human factor. Therefore, the results are sensitive to the way the data is annotated. This crucial issue needs to be addressed and more reliable methods are to be developed to set a standard.

This thesis implements and studies the effectiveness of two different methods. Namely, active shape modelling and random forest regression voting methods. Although the results obtained provide evidence that the opted model is efficient, optimal and robust, few alternatives can be explored. One approach to this problem can be implementing segmentation. The ultrasound tongue images can be subjected to segmentation. There is a large literature available on segmentation. This can be investigated and an optimal solution can be yielded to address the problem of locating the tongue in ultrasound tongue images.

A second approach would be to implement active appearance models (AAMs). The

random forest regression voting approach determines an optimal position of each landmark point in a defined radius. A voting of these candidate positions yields an optimal position of each point. However, it does not consider the texture of the surroundings of each point. Active appearance models (AAMs) on the other hand, capture not only the variation of the shape model but also capture the variation in the texture of the shape model. This is a promising approach and might yield better results in locating the tongue in ultrasound tongue images.

Keeping the implementation of alternative methods aside, the existing random forest regression voting method can be improved to yield better results. The method used in this thesis uses a single stage model with a maximum of 16 modes that captures 95% proportion of variance and a frame width of 15 pixels around each model point. A two-stage coarse-to-fine model can be investigated with a frame width of 15 pixels in the first stage and a frame width of 30 pixels in the second stage. This setting might contribute a great deal in achieving optimal results for the position of the landmark points. The investigation can be carried further by experimenting with the set of parameters to train the constrained local models to yield better results.

Bibliography

[1] J. Stuart-Smith J. M. Scobbie S. Nakai Lawson, E. Seeing Speech: an articulatory web resource for the study of phonetics.university of glasgow, url: http://seeingspeech.ac.uk, 2015.

[2] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[3] Inc. Infoplease © 2000-2017 Sandbox Networks. "how many spoken languages are there in the world?". url: https://www.infoplease.com/askeds/how-many-spoken-languages/.

[4] B. Kitchenham and S Charters. "guidelines for performing systematic literature reviews in software engineering", 2007.

[5] Maureen Stone. A guide to analysing tongue motion from ultrasound images. *Clinical Linguistics & Phonetics*, 19(6-7):455–501, January 2005.

[6] Craig C. Freudenrich. "how ultrasound works",
url: https://www.physics.utoronto.ca/~jharlow/teaching/phy138_0708/lec04/ultrasoundx.htm.

[7] Glossary of ultrasound terminology, url: https://link.springer.com/content/pdf/bbm%3a978-1-4612-5805-6%2f1.pdf.

[8] I. A. Hein and W. D. O'Brien. Current time-domain methods for assessing tissue motion by analysis from reflected ultrasound echoes-a review. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 40(2):84–102, March 1993.

[9] Bruce W. Drinkwater and Paul D. Wilcox. Ultrasonic arrays for non-destructive evaluation: A review. *NDT & E International*, 39(7):525 – 541, 2006.

[10] G. Chollet, R. Landais, T. Hueber, H. Bredin, C. Mokbel, P. Perrot, and L. Zouari. Some experiments in audio-visual speech processing. In Mohamed Chetouani, Amir Hussain, Bruno Gas, Maurice Milgram, and Jean-Luc Zarader, editors, *Advances in Nonlinear Speech Processing*, pages 28–56, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[11] Yuanyao Lu and Qingqing Liu. Lip segmentation using automatic selected initial contours based on localized active contour model. *EURASIP Journal on Image and Video Processing*, 2018(1):7, 2018.

[12] Michel T-T Jackson and Richard S McGowan. Predicting midsagittal pharyngeal dimensions from measures of anterior tongue position in swedish vowels: Statistical considerations. *The Journal of the Acoustical Society of America*, 123(1):336–346, 2008.

[13] Maureen Stone, Julie M Langguth, Jonghye Woo, Hegang Chen, and Jerry L Prince. Tongue motion patterns in post-glossectomy and typical speakers: A principal components analysis. *Journal of Speech, Language, and Hearing Research*, 57(3):707–717, 2014.

[14] T. Hueber, G. Aversano, G. Cholle, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone. Eigentongue feature extraction for an ultrasound-based silent speech interface. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 1, pages I–1245–I–1248, April 2007.

[15] J. Berry and I. Fasel. Dynamics of tongue gestures extracted automatically from ultrasound. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 557–560, May 2011.

[16] L. Tang and G. Hamarneh. Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 154–161, June 2010.

[17] Yin Yang and Xiaohu Guo. Tongue visualization for specified speech task. In *ACM SIGGRAPH 2012 Posters*, SIGGRAPH '12, pages 128:1–128:1, New York, NY, USA, 2012. ACM.

[18] George Nagy and Naomi Nagy. Tongue in cheek. In Vittorio Murino and Enrico Puppo, editors, *Image Analysis and Processing — ICIAP 2015*, pages 332–342, Cham, 2015. Springer International Publishing.

[19] Lisa Tang, Ghassan Hamarneh, and Tim Bressmann. A machine learning approach to tongue motion analysis in 2d ultrasound image sequences. In Kenji Suzuki, Fei Wang, Dinggang Shen, and Pingkun Yan, editors, *Machine Learning in Medical Imaging*, pages 151–158, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[20] Samuel Silva and António Teixeira. Automatic annotation of an ultrasound corpus for studying tongue movement. In Aurélio Campilho and Mohamed Kamel, editors, *Image Analysis and Recognition*, pages 469–476, Cham, 2014. Springer International Publishing.

[21] Hua Lin, John H. Esling, Scott R Moisik, and John H Esling. A study of laryngeal gestures in mandarin citation tones using simultaneous laryngoscopy and laryngeal ultrasound (sllus). *Journal of the International Phonetic Association.*, 44(1):21–58, 2014.

[22] Matthieu Loosvelt, Pierre-Frédéric Villard, and Marie-Odile Berger. Using a biomechanical model for tongue tracking in ultrasound images. In Fernando Bello and Stéphane Cotin, editors, *Biomedical Simulation*, pages 67–75, Cham, 2014. Springer International Publishing.

[23] A. Roussos, A. Katsamanis, and P. Maragos. Tongue tracking in ultrasound images with active appearance models. In *2009 16th IEEE International Conference on Image Processing (ICIP)*, pages 1733–1736, Nov 2009.

[24] Roland T Chin and Charles R Dyer. Model-based recognition in robot vision. *ACM Computing Surveys (CSUR)*, 18(1):67–108, 1986.

[25] W. Eric L. Grimson. *Object Recognition by Computer: The Role of Geometric Constraints*. MIT Press, Cambridge, MA, USA, 1990.

[26] Alan L. Yuille, Peter W. Hallinan, and David S. Cohen. Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111, Aug 1992.

[27] P. Lipson, A. L. Yuille, D. O'Keeffe, J. Cavanaugh, J. Taaffe, and D. Rosenthal. Deformable templates for feature extraction from medical images. In O. Faugeras, editor, *Computer Vision — ECCV 90*, pages 413–417, Berlin, Heidelberg, 1990. Springer Berlin Heidelberg.

[28] Andrew Hill and Christopher J Taylor. Model-based image interpretation using genetic algorithms. *Image and Vision Computing*, 10(5):295–300, 1992.

[29] A. Beinglass and H. J. Wolfson. Articulated object recognition, or: how to generalize the generalized hough transform. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 461–466, Jun 1991.

[30] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[31] Geoffrey E Hinton, Christopher KI Williams, and Michael D Revow. Adaptive elastic models for hand-printed character recognition. In *Advances in neural information processing systems*, pages 512–519, 1992.

[32] G. L. Scott. The alternative snake - and other animals. In *The 1987 Stockholm Workshop on Computational Vision, Stockholm. Dept. of Numerical Analysis and Computing Science, Royal Institute of Technology, TRITA-NA-P8714 CVAP 47*, pages 341–347, 1987.

[33] Lawrence H Staib and James S Duncan. Parametrically deformable contour models. In *Computer Vision and Pattern Recognition, 1989. Proceedings CVPR'89., IEEE Computer Society Conference on*, pages 98–103. IEEE, 1989.

[34] H. Isil Bozma and James S. Duncan. Model-based recognition of multiple deformable objects using a game-theoretic framework. In Alan C. F. Colchester and David J. Hawkes, editors, *Information Processing in Medical Imaging*, pages 358–372, Berlin, Heidelberg, 1991. Springer Berlin Heidelberg.

[35] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 285–339, 1991.

[36] U. Grenander, Y. Chow, and D. M. Keenan. *Hands: A Pattern Theoretic Study of Biological Shapes*. Springer-Verlag New York, Inc., New York, NY, USA, 1991.

[37] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[38] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054 – 3067, 2008.

[39] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *International journal of computer vision*, 61(1):55–79, 2005.

[40] Jason Saragih and Roland Goecke. A nonlinear discriminative approach to aam fitting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[41] Philip A Tresadern, Patrick Sauer, and Timothy F Cootes. Additive update predictors in active appearance models. In *BMVC*, volume 2, page 4. Citeseer, 2010.

[42] Patrick Sauer, Timothy F Cootes, and Christopher J Taylor. Accurate regression procedures for active appearance models. In *BMVC*, pages 1–11, 2011.

[43] Shaohua Kevin Zhou and Dorin Comaniciu. Shape regression machine. In Nico Karssemeijer and Boudewijn Lelieveldt, editors, *Information Processing in Medical Imaging*, pages 13–25, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[44] Dana H Ballard et al. Generalizing the hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981.

[45] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on statistical learning in computer vision, ECCV*, volume 2, page 7, 2004.

[46] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1022–1029, June 2009.

[47] Michel Valstar, Brais Martinez, Xavier Binefa, and Maja Pantic. Facial point detection using boosted regression and graph models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2729–2736. IEEE, 2010.

[48] Tim F Cootes, Mircea C Ionita, Claudia Lindner, and Patrick Sauer. Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision*, pages 278–291. Springer, 2012.

[49] Fred L Bookstein. *Morphometric tools for landmark data: geometry and biology*. Cambridge University Press, 1997.

[50] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[51] Douglas M Hawkins. Multivariate statistics: A practical approach, 1990.

[52] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2001.

[53] Tim Cootes. "a tool to create points and curves". url:http://uomqvxl.sourceforge.net/qmsm/qmsm_index.html.

[54] Claudia Lindner, Shankhar Thiagarajah, J Wilkinson, G Wallis, T Cootes, et al. Fully automatic segmentation of the proximal femur using random forest regression voting. *IEEE transactions on medical imaging*, 32(8):1462–1472, 2013.

Appendix A

Computing Changes In Shape

Mathematics involved in computing the change in shape parameters due to a change in pose shape parameters for mapping the mean shape onto the ultrasound tongue image.

The initial position of the point is represented by (A.1). To calculate the set of adjustments $d\mathbf{x}$ in the local model coordinate frame to meet the conditions specified in equation (A.2), an equation is obtained.

$$\mathbf{X} = M(s,\theta)[\mathbf{x}] + \mathbf{X}_c \tag{A.1}$$

$$M(s(1+ds),(\theta+d\theta))[\mathbf{x}+d\mathbf{x}] + (\mathbf{X}_c + d\mathbf{X}_c) = (\mathbf{X}+d\mathbf{X}) \tag{A.2}$$

This can be simplified to,

$$M(s(1+ds),(\theta+d\theta))[\mathbf{x}+d\mathbf{x}] = (\mathbf{X}+d\mathbf{X}) - (\mathbf{X}_c + d\mathbf{X}_c) \tag{A.3}$$

Since,

$$M^{-1}(s,\theta)[] = M(s^{-1},-\theta)[] \tag{A.4}$$

Finally, $d\mathbf{x}$ can be computed using,

$$d\mathbf{x} = M((s(1+ds))^{-1},-(\theta+d\theta))[\mathbf{y}] - \mathbf{x} \tag{A.5}$$

where,

$$\mathbf{y} = M(s,\theta)[\mathbf{x}] + d\mathbf{X} - d\mathbf{X}_c \tag{A.6}$$