Faculdade de Engenharia da Universidade do Porto

Computational Analysis of Magnetic Resonance Images of the Upper Airways: Algorithms and Applications

Jessica Condesso Delmoral

Dissertation submitted in fulfillment of the requirements for the conclusion of the Integrated Master in Bioengineering – Branch of Biomedical Engineering, FEUP

Supervisor:

Prof. Dr. João Manuel R. S. Tavares

Prof. Associado do Departamento de Engenharia Mecânica, FEUP

Co-Supervisor:

Prof. Dra. Sandra Rua Ventura

Professora Adjunta da Área Técnico-Científica da Radiologia, ESTSP-IPP

September 2015

Abstract

The human upper airways anatomy consists of the jaw, tongue, pharynx, larynx, palate, nasal cavities, nostrils, lips, and adjacent facial structures.

The interplay and connective movement between all the anatomical structures present in this region is complex, and basic physiological functions such as muscle activation patterns associated with chewing, swallowing, and speech production are not well understood.

Specifically, one of the least studied organs in this region is the tongue, in which the tasks of imaging and quantification of its anatomy are of great relevance for further study and analysis of the anatomic and physiological mechanisms that govern it. Furthermore, new insight can be given on other applications such as surgical planning, post-operative rehabilitation and the study of new adaptations acquired upon possible changes in function of pathological origin, for example in the presence of tongue cancer, surgical intervention or aging. Magnetic Resonance imaging (MRI) is the state of the art methodology for visualization and study of soft tissues, since it provides the best image contrast of soft tissues such as the muscular tissue of the tongue.

Under the scope of the Computational Vision field, an area that has over recent years allowed the development of new tools of analysis that can be applied to medical images, this dissertation aims to present computational algorithms for object detection and segmentation in images, suitable for application on objects such as the tongue.

The proposed methodology includes a set of algorithms developed for human tongue image processing, in order to study morphology through the building of an Active Shape Model, that captures the shape variability of the anatomy during production of various European Portuguese sounds. The developed model allowed to simulate realistically the tongues shape capturing its variability in the production of different sounds. Subsequently, this model also allowed the building of a semi-automatic detection and segmentation algorithm of this structure. The study was carried out using midsagittal plane images, since this plane is the most representative in the depicting the overall tongue shape variability, which constitutes to be especially advantageous for speech assessment purposes. The suggested model made it possible to obtain a realistic segmentation of the tongue as well as efficiently perform segmentation in new images. Furthermore, the use of such image analysis techniques allows quantitative measures

with higher precision and are particularly advantageous when speech therapists and imaging specialists need to analyze a large volume of data.

In conclusion, the identification and analysis of human structures are complex tasks, since their shapes are not constant and vary through time. However, techniques of Computer Vision and objects modeling can assist in their achievement as is demonstrated throughout this dissertation.

Resumo

A anatomia das vias aéreas superiores humanas é contituída pela mandíbula, língua, faringe, boca, fossas nasais, narinas, lábios e estruturas faciais adjacentes.

Os mecanismos de interação combinada que se associam ao movimento conexivos entre todas as estruturas anatómicas presentes nesta região são complexos, e a sua funcionalidade inclui actividades fisiológicas básicas, tais como padrões de ativação muscular associados com a mastigação, deglutição, e produção da fala.

Um dos órgãos menos estudados nesta região é a língua, e portanto as tarefas de imagiologia e quantificação da sua anatomia são de grande relevância para o estudo e análise mais aprofundado dos mecanismos anatómicos e fisiológicos que a regem. Mais ainda, este estudo poderia produzir novo conhecimento passível de ser utilizado em outras aplicações, tais como o planeamento cirúrgico, reabilitação em pós-operatório e estudo de adaptações compensatórias na produção da fala, adquiridas pela presença de cancro da língua, após intervenção cirúrgica ou envelhecimento.

No âmbito da área de Visão Computacional, uma área que tem nos últimos anos permitido o desenvolvimento de novas ferramentas de análise que podem ser aplicadas em imagens médicas, esta dissertação tem como objetivo apresentar algoritmos computacionais para deteção e segmentação de objetos em imagens, adequado para aplicações em órgãos deformáveis tais como a língua. Para o estudo de tecidos moles, o estado da arte referente às técnicas de aquisição imagiólogica, a Ressonância Magnética (RM), uma vez que proporciona o melhor contraste de imagem de tecidos moles, tais como o tecido muscular da língua.

A metodologia proposta inclui um conjunto de algoritmos desenvolvidos para processamento de imagem da língua humana, para o estudo morfológico através da construção de um Modelo de Forma Activa, que capta a variabilidade anatómica desta estrutura durante a produção de vários sons do Português Europeu. O modelo desenvolvidos permitiu simular de forma realista a língua na sua variabilidade de forma durante a produção dos diferentes sons. Seguidamente, este modelo permite ainda, uma produção de um algoritmo de deteção semi-automática desta estrutura. O estudo foi realizado utilizando imagens do plano sagital médio, constituindo o plano mais representativo da variabilidade de forma da língua global, tornando-se este estudo especialmente vantajoso para fins de avaliação dos mecanismos de produção da fala. O modelo sugerido tornou possível obter uma segmentação realista da língua, bem como

executar eficientemente a segmentação da mesma em novas imagens. Mais ainda, o uso de tais técnicas de análise de imagem pode permitir a obtenção de medições quantitativas, com uma precisão mais elevada e são particularmente vantajosos para a análise por especialistas em imagem ou em produção da fala, no sentido da análise de grandes volumes de dados.

Em conclusão, a identificação e análise de estruturas humanas são tarefas complexas, uma vez que as suas formas não são constantes e variam ao longo do tempo. No entanto, as técnicas de visão computacional e modelagem de objetos podem ajudar na sua realização como é demonstrado ao longo desta Dissertação.

Acknowledgments

Firstly, beginning by the direct participants in this work I would like to thank Professor Dr. João Tavares for all the availability, patience and counseling provided that guided me in this work, as well as, my co-supervisor, Professor Dr. Sandra Rua Ventura, for the availability and help to direct the purpose of this work and allowing me to achieve the objectives successfully.

And I would furthermore like to thank the colleagues that accompanied me during the process of making this dissertation and of the long hours in front of the computer. Also, a special thanks goes to those that tried to keep me motivated, who helped me not to forget to relax and take a break to laugh.

Secondly, I would like to thank everyone that crossed my path, and marked it somehow, during the five years that have past. To all my friends with whom I have the best memories, with whom I grew and from whom I save the best moments. And also to the more recent 'nerd' friends. You have all played your part.

I would like to thank especially my oldest friends, which are the ones that accompanied me during the process of growing up to this point in our lives. Thank you for being there, Carolina, since 1998 and Rita, since an unknown year between 1992 and 1998.

Finally I would like to thank with all my heart my family, and my parents for accompanying me and supporting me until this point, thank you for always pulling me to be better, and to grow.

"We are just an advanced breed of monkeys on a minor planet of a very average star.
But we understand the Universe and that makes us something very special"

Stephen Hawking

Contents

List of Figures

List of Tables

Acronyms

1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional
AAM	Active Appearance Model
ACR	American College of Radiology
ASM	Active Shape Model
CAD	Computer-Aided Diagnosis
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
DTI	Diffusion Tensor Imaging
EPG	Electropalatography
FID	Free Induction Decay
GG	Genioglossus muscle
GTF	Game-theoretic Framework
MI	Mutual Information
MRI	Magnetic Resonance Imaging
MRS	Magnetic Resonance Spectroscopy
NEMA	National Electrical Manufacturers Association
NMCs	Neuromuscular compartments
PACS	Picture Archiving and Communication System
PET	Positron Emission Tomography
PDM	Point Distribution Model
RF	Radiofrequency
RW	Random Walker
SPECT	Single-photon emission computed tomography
SSM	Statistical Shape Model
T	Tesla
TR	Repetition Time
TE	Echo Time
US	Ultrasound

Chapter 1

Introduction

The tongue constitutes a unique anatomical structure among all the organs integrating the human body.

It is a specialized organ located in the oral cavity, which plays an important role in mastication and swallowing (for digestion process), taste and speech production. Breathing and swallowing processes are closely interrelated in their central control at brainstem and are highly coordinated. Many muscles and structures of the aerodigestive tract have dual roles in these processes, namely the tongue.

The ability to produce of fast and precise movements during the production of vocalic and consonant sounds, and doing so that there is an extensive variety of languages, each with its characteristic sounds, makes the study of the tongue of great interest and importance. The muscle components of the tongue have the unique purpose of contracting in order to deform the body of the tongue itself, and not simply function as most skeletal muscles in the human body. These act as a force generating organ for the movement and stabilization of attached body structures.

In speech production the tongue deforms to modulate the flow and acoustic resonances of air through the vocal tract. The transport of the bolus through and around the appropriate surfaces through tongue movements, followed by its propulsion into the esophagus is the purpose of mastication and swallowing tasks, respectively.

The functions of speech production and swallowing, can affect particularly the survival and quality of life. Therefore, for this process to occur, the tongue needs to be able to execute a sequence of organized and integrated motor events, mediated by neuro-motor stimulus, which can only be feasible if the anatomical and physiological integrity of this structure is preserved. All of these functions are controlled by highly evolved neuromuscular systems under both voluntary and involuntary control.

The purpose of this dissertation is to establish a semi-automatic segmentation tool of analysis for further understanding of how the tongue changes shape in response to muscular contraction, given that researchers have remarked that our knowledge of the tongue is extremely limited.

1.1. Motivation

The combined organs and tissues of the respiratory tract and the upper part of the digestive tract, called as aerodigestive tract, in all of the associated functionalities, represent a vital mean of survival for humans. Speech production is one of the processes secured by all these organs that constitute this tract, being the vocal tract one of the most important and complex structures.

One of the most important structures of the upper airways, or also referred to, the aerodigestive tract, is the tongue, an organ controlled by complex neuromuscular mechanisms, capable of high deformations of its shape to conquer the physiological tasks in which it intervenes, in the modulation of the upper airway properties. The study of the full detailed anatomy if this organ has recently gained significant relevance, and the comprehension level towards the study of the complex system of tongue conformation during the various functions, and have proven to play a key role in its correct execution, where speech impairments, respiratory disturbances (e.g. obstructive sleep apnea), as well as other pathologic consequences need to be studied in further depth.

Magnetic Resonance Imaging (MRI) is an imaging technique first discovered in 1952 by Felix Bloch (University of Stanford) and Edward Purcell (University of Harvard), for which they received the Nobel Prize in Physics. This technique revolutionized medical imaging, having been only comparable to the invention of the X-Ray by Wilhelm Conrad Roentgens, having been first applied to medical purposes in the 1970's decade (Rinck, 2001).

Emerging researches are being carried out addressing the study of the functional, mechanical and dynamic properties, whereas it is well established that targeting specifically the tongue is a matter of high relevance.

Currently, there are no tools or exams that allow the complete characterization or evaluation of tongue motion and its modulation of the upper airways, by a non-invasive way.

The study in a Computational Vision point of view is therefore, of high importance in this field, and the objective is the creation of Computer Aided-Diagnosis (CAD) tools of modelling and quantification. Many are the advantages that derive from tongue segmentation, but its extent goes from the adequacy of imaging acquaintance through MRI to the two-dimensional and three dimensional analysis needed to understand its conformation and dynamics. Accordingly, the diagnostics and surgery planning related to the structures included in the upper airways holds a gap that can be fulfilled through the development of Computer Aided Diagnostics tool. Furthermore, the pertinence of the study of the tongue, is in practical

appliance expressed by speech therapists and imaging specialists, that in order to perform qualitative studies of the tongue, proceed to manual segmentations, done pixel by pixel, which obviously stands as highly time-consuming and subject to human error. The understanding of the mechanisms that govern the tongue, need therefore to be studied through qualitative and quantitative analysis with adequate precision, and it is accordingly advantageous to be possible to do so, in adequately large volumes of data for study validation.

1.2. Objectives

For the extraction of the tongue shape from MRI images, three key aspects must be considered:

- MRI images are usually very noisy, since this type of image is acquired through fourier transform reconstruction of the retrieved magnetic signals, that due to the presence of different tissues in each scan, are bound to present random noise;
- The tongues shape is highly deforming and cannot easily be represented by a parametric model;
- The study of 2D midsagittal tongue anatomies, would allow the performance of statistical studies of speech mechanisms;
- These studies would furthermore, allow the statistical analysis of the mechanisms acquired in pathological subjects comparatively to the normal deformation mechanisms observed in healthy subjects;
- The development of implementations with a certain automaticity, would be determinant to add value to the state of the art methodologies available.

Having the previous problem key points in mind, for the development of this work the main goals are:

- Development of the potential properties of magnetic resonance images for the analysis of the aerodigestive tract as to the 2D conformation and motion during speech production;
- Description of Landmark-based geometric morphometrics;
- Development of a semi-automatic segmentation process of the aerodigestive tract structure, specifically the tongue;
- Development of a computational analysis of the properties of the structures, namely the tongue;
- Demonstrate the viability of the segmentation results through quality measurements analysis;
- Demonstrate the viability of this analysis for the application as a Computer-Aided Diagnosis System (CAD system).

In order to establish the most adequate methodology to achieve the cited results, this dissertation included an analysis of the medical problems, focused on the tongue, that need to be tackled in the sense of defining the type of information that is pertinent to be retrieved, the methodologies reported in the literature based on Computational Vision, and the modelling techniques of identification of soft tissues in MRI images. This study also involved the selection of the appropriate platform of implementation.

1.3. Report Organization

The comprehensive analysis of human tongues anatomy and functionality will be addressed over Magnetic Resonance imaging, and the various stages of image analysis addressed, cover a wide spectrum of fields.

Chapter 2 presents an overview on the tongues full anatomy and functionality as well as the imaging technique used to acquire images of the complete aerodigestive tract. Regarding applications and previous works of computational analysis based in images the modelling and segmentation methods described in previous works are presented. An overview of the state of the art of tongue segmentation studies and techniques, from the very initial reports with poor description of the anatomy, which over the years was never very thoroughly described, and the perception that the complexity of its study has not been widely addressed. Only in recent years the developments of Computational Vision allowed that the studies address this organ with careful attention and the complexity of such anatomy as one of the most complex in the human body.

Chapter 3 introduces the developed methodology, to the modelling of the tongue, describing the standard Active Shape Model building, that is based on a Statistical Shape Model and a Profile Model, that characterize the tongues shape and boundary intensities, respectively, based on a set of training images, the tongues shape and intensity distribution based on landmark labeling.

Chapter 4 provides a thorough explanation of the developed segmentation methodology based on the building of an Active Shape Model, to segment the shape of the in new images, using the statistical models presented in the previous chapter, in new images.

Conclusions and future perspective for the dissertation work are presented in Chapter 5.

Chapter 2

Fundamentals and Related Works

2.1. Human Aerodigestive Tract Anatomy

The human aerodigestive tract is regulated by many complex mechanisms and organs that sustain important functions such as mastication and swallowing (fundamental for the digestive process), taste, respiration and speech production. The importance of tongues functionality for said abilities implies actions of (1) positioning of food in the whole vocal cavity, (2) along with the buccinator muscle maintaining food in position for the mastication tasks, (3) propelling of the food to the palate and posteriorly into the pharynx initiating deglutition, (4) change its conformation in order to alter the sounds produced during speech production. In addition, humans have taste receptors including in the upper surface of the tongue and the epiglottis. The anatomical structure of the vocal tract (Figure1) is well established, being the tongue a central organ of this system, which plays a crucial role for the correct functioning of the referred tasks. The development of the anatomical structure of the human vocal tract, continues to change after birth. Specifically, the position of the tongue changes gradually, whereas the newborn tongue is initially flat, positioned almost entirely in the oral cavity, and later, as it descends into the pharynx, acquires a posterior rounded contour, carrying the larynx down with it. Suprapharyngeal horizontal and vertical proportions undergo comparative growth that reaches maturity by the age of 6-8 years old (Lieberman et al., 2001). This is confirmed by Vorperian et al., (2009) based on a longitudinal study of 605 subjects using MRI and CT images.

However, the anatomical study of this structure has been simply forgotten, since the actual knowledge and role in the execution of the referred tasks has only been attempted to be understood in very recent turn of investigations, being also aided by the application widening of the available imaging technics towards the characterization of this organ. In the literature, reported references that confer some extent of attention to the tongues anatomy are very

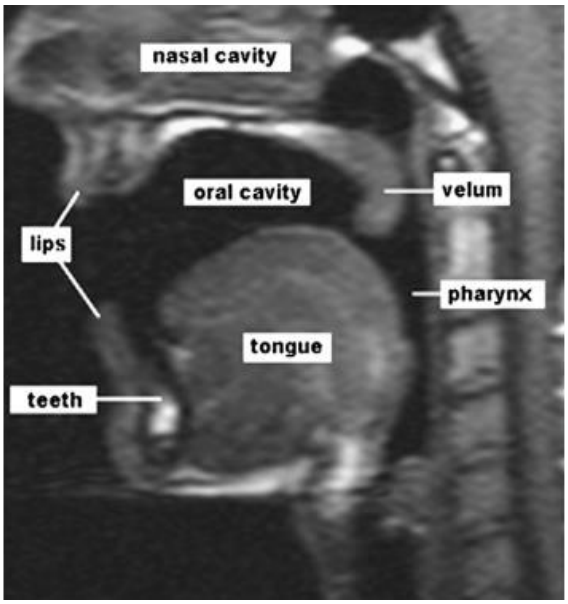

Figure 1 - MR midsagittal image (slice) indicating the vocal tracts structures (Ventura et al. (2011)).

scarce. For instance, a gross anatomy of the tongue is, in very early anatomic discoveries, to be found in full human anatomy works (Gray (1918), Salter (1852)).

2.1.1. Anatomy of the tongue

The human tongue is an organ composed primarily by skeletal muscle and located in the oral cavity, occupying a major portion of its volume. It is attached to the oral cavity through its posterior structures, namely via tendons, and other neighboring muscles as well as to its pavement through the lingual *frenulum* fold.

The tongue is attached to the support structure of bones of this region, specifically to the mandible, the hyoid bone and the styloid process of the skull. The styloid process and bone structure of the skull is shown in Figure 2, and the bone attachments of the tongue are depicted in Figure 3.

The posterior connection of the tongue is made by an attachment to the hyoid bone, which is suspended in the larynx structure, by muscles and cartilaginous tissue. Anteriorly, the tongue connects to the posterior aspect of the mandibular symphysis. The tongues base is connected by fascia to the supralaryngeal muscle that lies immediately inferior to the tongue and forms the muscular floor of the mouth, the mylohyoid.

The tongues structure is composed by a complex arrangement of muscles whereas, the muscles can be grouped in two categories: intrinsic muscles, those that are actually part of the tongue, have no bone insertions and are responsible for shape changing, flattening and up-lifting abilities, and extrinsic muscles, those that are connected to the main structure and attached to bone, responsible for protrusion and retraction, lateral movement and shape modification abilities (Seeley et al., 2008).

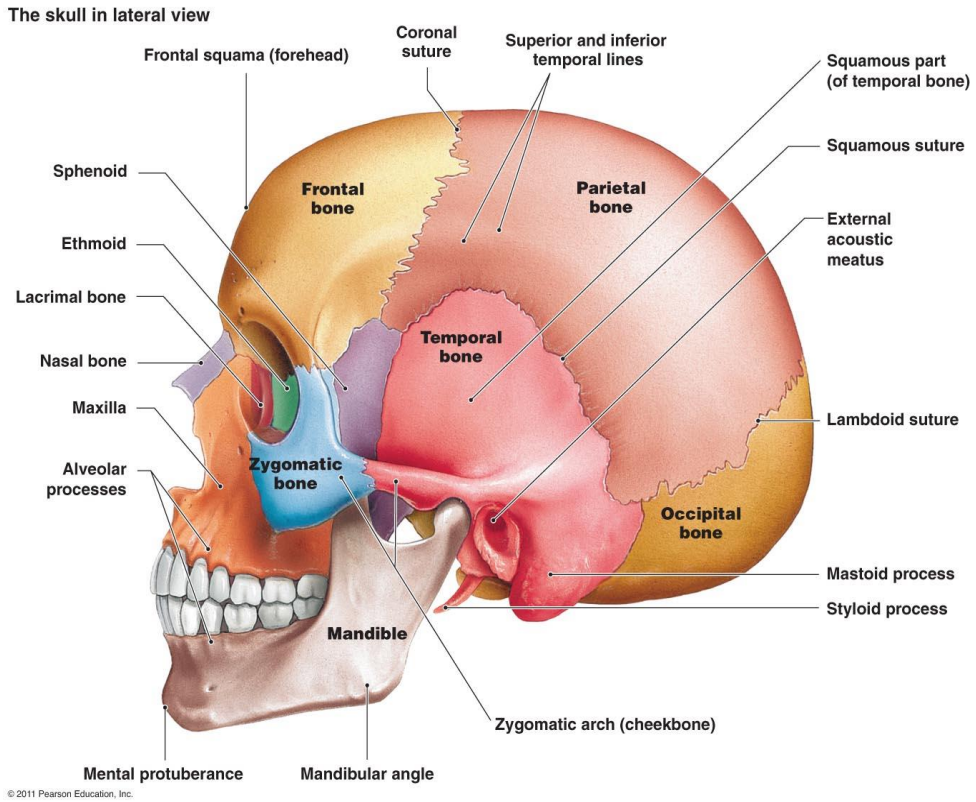

Figure 2 - Side view of the skull. The styloid process is just posterior to the mandible (Georgia Highlands College, 2013)

The extrinsic muscles are genioglossus, hyoglossus, styloglossus, and palatoglossus. The remaining muscles, transversus, verticalis, superior longitudinalis, and inferior longitudinalis, are intrinsic to the tongue.

A groove, named terminal groove, divides the tongue into two portions. The anterior portion relatively to the groove corresponds to 2/3 of the surface of the tongue being covered with taste buds, with taste receptor cells. The posterior third portion is, in contrast, deprived of taste buds, having only some taste terminal receptors on its surface, being occupied by little glands and a big agglomerate of lymphoid tissue belonging to the lingual amygdalae.

The musculature of the tongue has been described as being composed by eight paired muscles, as illustrated in Figure 4.

Genioglossus

Genioglossus constitutes the main volumetric portion of the tongue posteriorly, having a fan or wedge-shape. It is fixated through a musculo-tendinous origin from the inner surface of the symphysis menti, continuing from root to tip. Its muscular anterior fibers are arranged in a curved antero-dorsal direction that culminates in the anterior fibers of the inferior longitudinal, hyoglossus, and styloglossus muscles. Its posterior fibers run horizontally and backwards to the

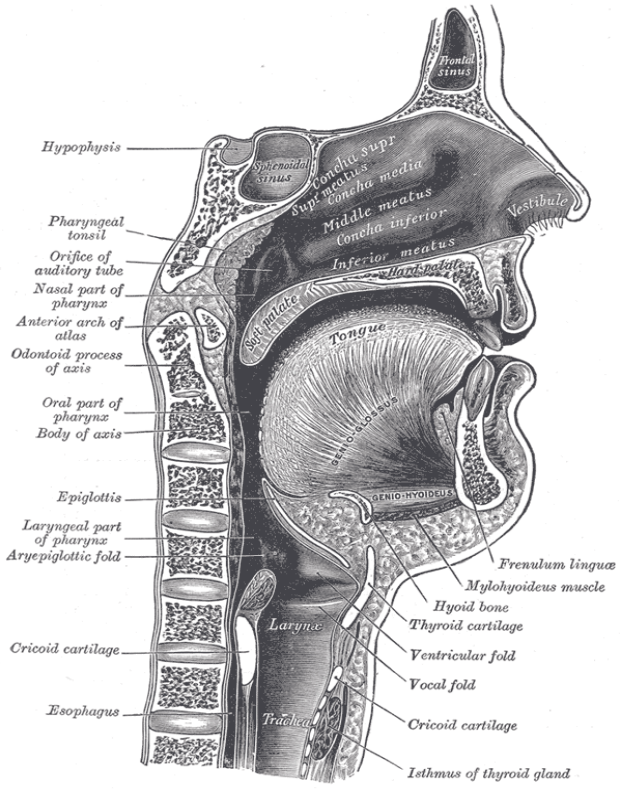

Figure 3 - Tongues attachments and neighboring structures in a sagittal anatomical view (Gray (1918)).

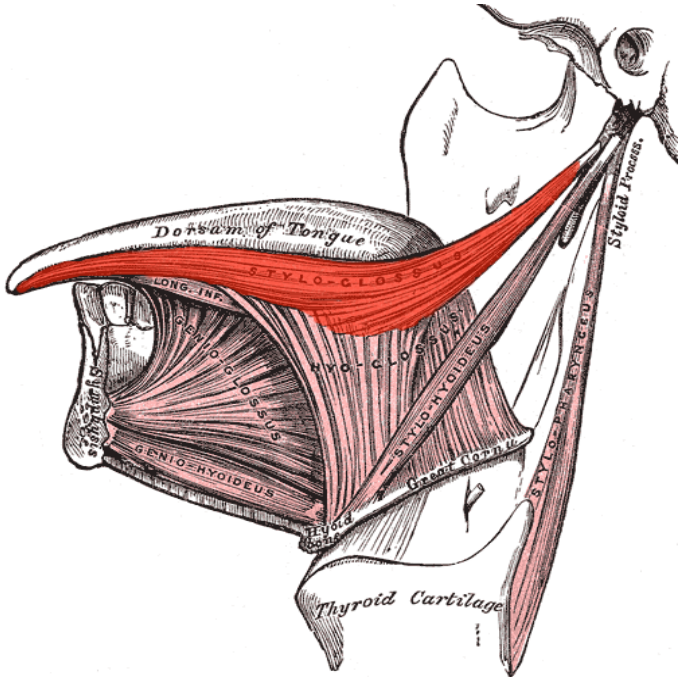

Figure 4 - Extrinsic muscle of the tongue with styloglossus visible at center top (in red) (Gray, 1918).

root of the tongue towards the anterior surface of the hyoid bone and anterior surface of the base of the epiglottis. Also, intermediate bundles of fibers diverge with different degrees of obliquity between the two mentioned portions. In parasagittal plane, it becomes possible to identify its orientation.

Hyoglossus

Hyoglossus radiates in a fan-shaped manner in its upper portion, having a quadrangular conformation in base. Anatomically, towards the other tongue muscles, it is positioned medially, between the inferior longitudinal and genioglossus muscles. It arises from the body of the hyoid bone and interdigitates at its origin with superficial and deep fibers of the geniohyoid. Fiber orientation in the posterior portion of the muscle consists in an antero-posterior radiation. The anterior fibers run and terminate in an approximately longitudinal direction towards the tip of the tongue. The posterior portion lies therefore, under cover of styloglossus, terminating in a fusion to its fibers.

Styloglossus

Styloglossus departs from an insertion in the anterior and lateral surface of the styloid process, close to its apex, continuing in a descending and forward direction into the tongue. Its deep fibers interdigitate with the body muscle of the tongue. After inserting into the tongue, the fibers divide into two bundles. An anterior bundle continues anteriorly along the inferior surface of the inferior longitudinalis, laterally to the hyoglossus, finalizing in the tip of the tongue. A posterior bundle penetrates de hyoglossus and courses medially into the lingual septum.

Transversus

Transversus is part of the bulk of the tongue, along with the Verticalis. It is located between the superior longitudinal muscle, dorsally, the genioglossus and inferior longitudinal muscles, ventrally. The more superficial muscle fibers take a dorsal direction, and the deepest ones are disposed in a ventral direction.

Verticalis

Verticalis is the other muscle that constitutes the thickness of the tongue, being in a tight joint surface with the Transversus muscle. Verticalis fibers are generally vertical, spreading at its superior and inferior portions. The Genioglossus, transversus, and verticalis partially overlap with one another.

Table 1 - Muscles of tongue movement (Seikel et al. (2009)).

Elevate tongue tip	Superior longitudinal muscles
Depress tongue tip	Inferior longitudinal muscles
Deviate tongue tip	Left and right superior and inferior longitudinal muscles for left and right deviation, respectively
Relax lateral margin	Posterior genioglossus for protrusion; superior longitudinal for tip elevation; transverse intrinsic for pulling sides medially
Narrow tongue	Transverse intrinsic
Deep central groove	Genioglossus for depression of the tongue body; vertical intrinsic for depression of central dorsum
Broad central groove	Moderate genioglossus for depression of the tongue body; vertical intrinsic for depression of dorsum; superior longitudinal for elevation of margins
Protrude tongue	Posterior genioglossus for advancement of body; vertical muscles to narrow tongue; superior and inferior longitudinal to balance and point the tongue
Retract tongue	Anterior genioglossus for retraction of the tongue into oral cavity; superior and inferior longitudinal for shortening of tongue; Styloglossus for retraction of tongue into pharyngeal cavity.
Elevate posterior tongue	Palatoglossus for elevation of sides; transverse intrinsic to bunch tongue.
Depress tongue body	Genioglossus for depression of medial tongue; hyoglossus and chondroglossus for depression of sides if hyoid is fixed by infrahyoid muscles.

Superior Longitudinalis

Superior longitudinalis consists of a thin stratum muscle. Its fibers are directed longitudinally along the lamina propria, although this directionality is not clearly defined, being reported with disagreement in Anatomy bibliography. The muscle has a gradual reduction in thickness as it reaches the Styloglossus, hyoglossus and inferior longitudinal muscles, laterally in the tongue.

Inferior Longitudinalis

Inferior longitudinal is a narrow muscle that extends between the paramedian septum and the medial lamella of the lateral septum. It arises medially with the genioglossus muscle, having lateral attachment from the body of the hyoid bone. It is positioned medially with the hyoglossus muscle. In the middle body of the muscle, it blends with the genioglossus hyoglossus, and Styloglossus muscles forming the tip of the tongue.

The whole description of the musculature existent in the tongue is based on the findings reported in (Abd-el-Malek, 1939). Takemoto, (2001) was able to describe and illustrate his findings on the relative positioning, especially well for the extrinsic muscles, stating the

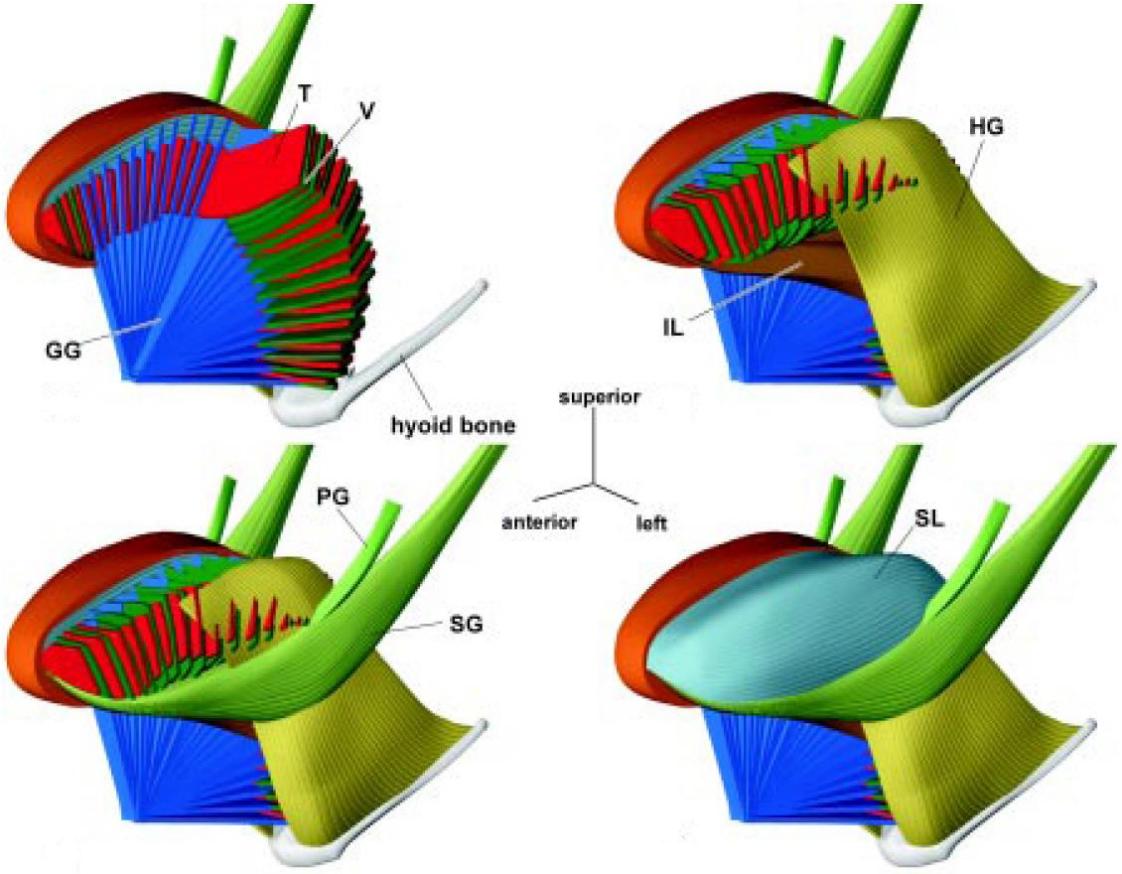

Figure 5 - Muscles of the tongue (Takemoto (2001)): GG - genioglossus, T - transversus, V - verticalis, HG - hyoglossus, IL - inferior longitudinalis, S - superior longitudinalis, PG - palatoglossus, SG - Styloglossus.

difficulties of distinction between the genioglossus, transversus and verticalis, and produced a three-dimensional tongue model based on impressions from his tongue dissections, depicted in Figure 5. Also, muscle tongue movement has been established for each constitutive muscle of the tongue, as indicated in Table 1.

Despite the unclear definition of the myoarchitecture of anatomical fiber orientation and 3D arrangement of the tongue, in the last ten years a new interest has been taken by the scientific community in the comprehensive analysis of this structure. To answer these disparities, the detailed study of the tongue, specifically of the lingual myoarchitecture has been collected with new recordings through diffusion tensor magnetic resonance imaging, or diffusion tensor imaging (DTI). This technique is very attractive for these types of studies since it enables fiber orientation imaging and analysis in vivo. Gilbert and Napadow, (2005) report imaging three human tongues statically, and Shinagawa et al., 2008 reports imaging from single sections of *in vivo* human tongues during rest and protrusion movement. Many techniques such as electropalatography (EPG), X-ray imaging, ultrasound, and cine-MRI imaging have been reported in the study of lingual function (Shinagawa et al., 2008). However, the characterization and anatomy of the tongue are not well understood, in contrast to other neighboring structures such as, for instance, the hard palate. Other attempts of imaging the surface during movement and/or oral functions, new analysis of the activation of the tongue muscle fibers for deformation of its body, and a clear understanding of these mechanisms *in vivo* has only been in recent years considered a matter of deserving attention.

2.1.2. Neurophysiological control of the tongue

Neurophysiology is an advanced field that addresses the understanding of the mechanisms that govern the motor control system, especially at the level of last-order muscular output. Since the tongue is purely a muscular structure, the understanding of its complexity may address the neural complex mechanisms of activation that rule its functionality. This analysis is of preponderant importance since the neural control on tongue movement is crucial to the function of rhythmic tasks of respiration and swallowing, whereas disruptions of these mechanisms have even been associated with the highest mortality reported among the pathological problems that may arise (Sawczuk and Mosier, 2001).

The neuromotor system is based on the activation of motor units. These consist of single motor neurons and an assortment of muscle fibers onto which it is connected. Through this connection, synapses occur, through electrical potential signals that are sent along the specific motor neurons innervating the muscle fiber bundles that need to be activated, producing a simultaneous contraction of said fibers. Motor units are organized in motor pools activated in a systematic stimulation, by the central nervous system.

Tongue muscle movement, contractile properties and generator-produced rhythmic modulation derive all from the innervation of the hypoglossal motor neuron complex. The motor neurons are clustered in the hypoglossal nucleus, part of the brainstem, from which departs

the hypoglossal nerve, the twelfth cranial nerve XII. The system of motor neurons that innervate this group of muscles is astonishing, evidencing the remarkable complexity of such an important organ in all its functions. Although the actual number of neurons that intervene in this structure is reported with high disparity, placing, for instance, the total number of myelinated fibers in 9,900 (Atsumi and Miyatake, 1987). In contrast, other muscles of higher dimensions, including biceps or rectus femoris, for instance, are innervated by an average of 441.5 and 609 motor units, respectively (Hamilton et al., 2004).

Electromyographic studies have, on the other hand, been more recently carried out in order to comprehend the complete muscle activity involved. Recent studies report that the genioglossus is the primary upper airway dilator muscle, and its internal motion activation is inhomogeneous. The neuronal control has been vastly studied in the last ten years, and punctual conclusions have been established relatively to the phases of control of the Hypoglossus. EMG findings reveal that inspiratory neuronal activity begins approximately 250ms before the inspiratory process begins, whereas, during inspiration neuronal stimulus increases, and during expiration tonus level is maintained (Cheng et al., 2008).

Although this basic neuronal source is established, the tongue is very uniquely characterized by a complex mechanism of activation that is not yet known, whereas the highest difficulty of the comprehensive process is straightly related to its anatomical complexity. In fact, the human tongue is not only of higher complexity relatively to other mammals, but its anatomical nerve activation and gross neuroanatomy are also lacking. The most extensively studied muscle among tongue muscles is the Genioglossus, responsible for protrusion and depression motion, which has been demonstrated to take part in most tongue movements carried out.

It is hypothesized, in the literature, although it has not been directly reported, that neural control of the tongue may be done, as reported in other mammals for skeletal muscles control, by means of tissue composed of neuromuscular compartments (NMCs), that are morphologically and functionally activated by distinct neuromotor pools, defined as “smallest portion of a muscle to receive exclusive innervations by a set of motoneurons” (English et al., 1993). In (Mu & Sanders, 2000) is demonstrated a compartmental organization of the canine tongue, specifically the innervation present in the genioglossus, where it is reported the presence of two compartments, with fibers horizontal and an obliquely oriented, as well as the branches subdivision departing from the main genioglossus nucleus.

This mechanism is reported to base neuromuscular control of shoulder muscles (Wickham & Brown (2012), Lucas-Osma & Collazos-Castro (2009)); however, even in said anatomically simpler muscles NMCs boundaries are not completely defined.

Unfortunately, no careful anatomical data is found in the literature describing the neuronal organization of the human tongue, compartmental or non-compartmental wise.

2.1.3. Speech Production, Respiration and Swallowing

Speech production, respiration and swallowing are the three main activities that are carried out by the aerodigestive organs, with determinant aid of tongue motion.

Among these functions, speech production is the area that has been more extensively studied by the scientific community, due to its multidisciplinary character. The human phonetic apparatus may be divided in organs responsible for sound production and organs of speech articulation. Sound production or phonation, is achieved through the vibration of the vocal folds into the airstream of the airway, a process named voicing, following their fixation into specific position that modulates the aerodynamics of airstream passage. The vocal tract acts as an acoustic filter for a source signal generated in the vocal folds within the larynx, whereas the process of speech production implies the complement of simple phonation with the execution of an extremely well-organized and integrated sequence of movements of the speech articulator organs (lips, mandible, tongue and palatal velum), shaping the resonant cavities of the vocal tract and consequently altering the resulting acoustic output (Seikel et al., 2009). Tongue deformation is directly related to vocalic sounds as well as palatal, velar and pharyngeal consonants sound production. Many are the studies that model tongue conformation, during production of specific sounds present in various languages worldwide, as shown in Figure 6 for vocalic sounds of Portuguese language.

Moreover, the cross-sectional area along the vocal tract, in its supralaryngeal section determines formant frequencies, whereas records of studies addressing the human tongue deformation during speech production, exist from over 150 years (Lieberman, 2012). The analysis of the resonance cavities involved in phonation is, as obviously understood by the scientific community that has undergone an extensive amount of research relevance to the study of speech production anatomy and mechanism, in this sense of extreme importance. In addition, more importantly for the understanding of how the mechanisms allow the diversity of

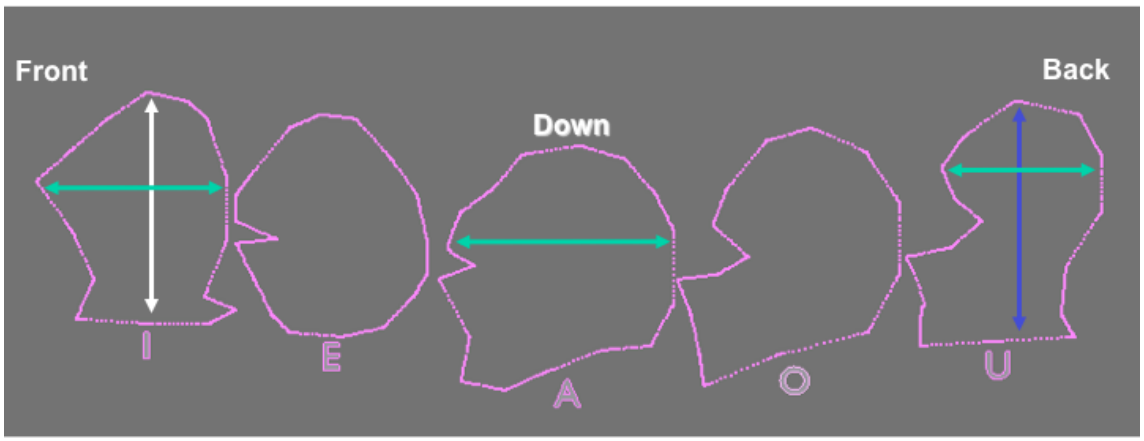

Figure 6 - Tongue contour extracted from midsagittal images, during production of vocalic sounds present in Portuguese language (Ventura et al., 2008).

phonation capacity and how disturbances of pathological origin or otherwise to the structures involved may affect their functionality.

Swallowing consists in the passage of a bolus of food through the mouth to the pharynx, and into the esophagus that will trigger a swallowing reflex as it passes into these regions. To this process be succeeds the larynx must elevate, and the epiglottis (attached to the root of the tongue) drops down to cover the aditus, avoiding choking or pulmonary aspiration can occur. Food bolus building was illustrated and explained in (Abd-El-Malek, 1955). His observation of subjects masticating nuts, gelatin and chewing gum led to the description of the following steps:

a) Preparatory stage – acquires a pouch-like form, to collect the food on its dorsum;

b) Throwing-stage - a twisting movement towards one side to deposit the bolus onto the molars;

c) Guarding stage - tongue twists even more, making contact with the upper and lower teeth, in order to keep the bolus between the molars during mastication;

d) Bolus building – after several chewing movements the cheeks move medially and the tongue moves side to side, mixing the bolus with saliva and coating it with mucus.

e) Swallowing - the tip of the tongue is raised and pressed against the posterior surface of the front teeth and the anterior part of the hard palate, as to close the mouth and pharynx;

These stages are illustrated in Figure 7.

Muscle activation during this process is automatic, and important processes regard studying swallowing to assess the stiffness of the tongues surface, or the force that the tongue is able to exert on the hard palate.

In humans, respiratory airway activity involves important tasks of patency maintenance. Substantial studies suggest that this function is provided by the tongues genioglossus muscle (GG). Airway patency is a matter of extreme importance, and delicate to control, since the human pharynx has no rigid support except at its extreme upper and lower ends where it is anchored to bone (upper extremity to hyoid bone) and cartilage (part of the larynx). Therefore, the airway depends on 20 skeletal muscles that dilate and keep the oropharynx open (Dempsey et al., 2010).

During respiration, tongue deformation has been analyzed through tagged MRI, a technique that arose later as a modality of MR imaging, allowing quantification of physiological motion. Expiratory and inspiratory tasks create pressure differences in the airway and muscle tonus of the involved structures that define its need to be able to maintain the adequate compliance. Inspiration tasks generate a negative inspiratory pressure that manifests at epiglottis level, that has been directly correlated with neuronal firing of the genioglossus (Pillar et al., 2001). In Cheng et al. (2008) is reported that the muscle movements activated throughout the respiratory cycle. Genioglossus muscle analysis indicated posterior movement during expiration as opposed to an anterior movement during inspiration, and over the geniohyioid. Geniohyioid has presented very little movement during respiration.

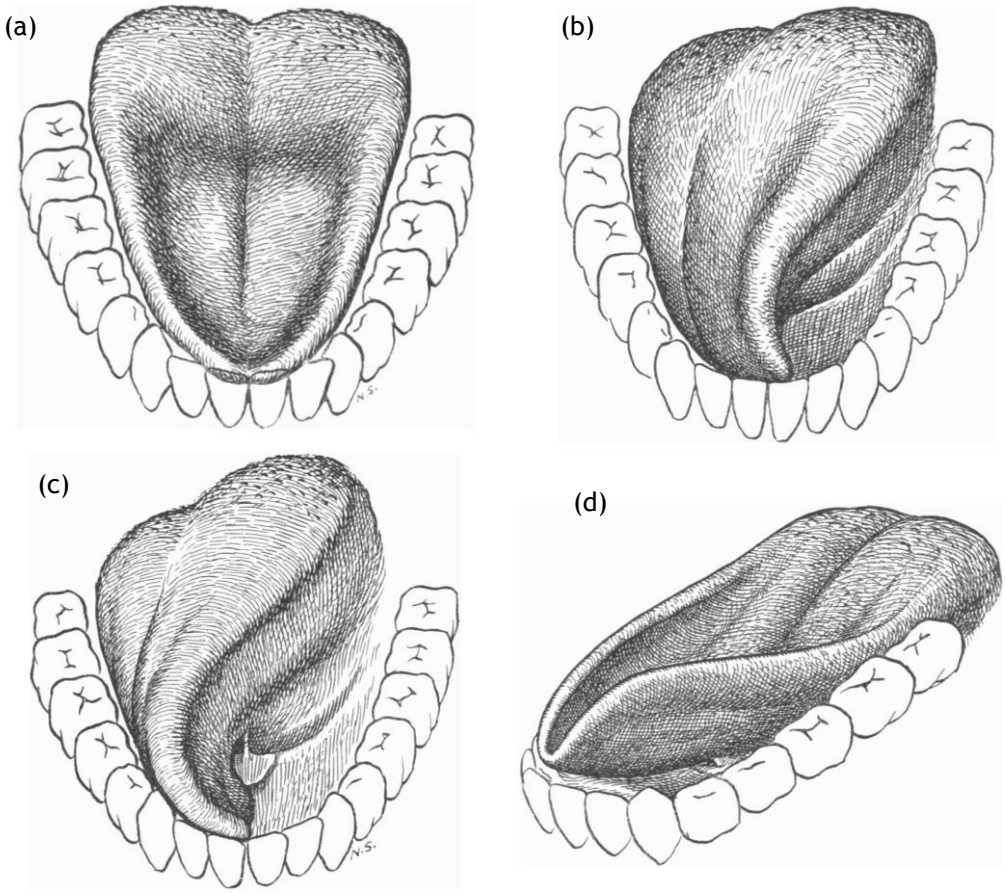

Figure 7 - Abd-El-Malek (1955) illustration of the preparatory stage of mastication (a), throwing stage of mastication (b), guarding stage of mastication (c), initial stage of deglutition (d).

2.2. Magnetic Resonance Imaging in the context of Aerodigestive Organs

Since the development of novel imaging techniques of the tissues that the *in vivo* anatomy of living organisms has been made possible.

Magnetic resonance imaging (MRI) is a diagnostic method that uses strong magnetic fields and radiofrequency (RF) waves to form images of the human body. This technique allows a non-invasive imaging method that presents a wide range of potential clinical applications.

MRI is therefore, nowadays a well-established imaging method used by physicians in the evaluation and characterization of soft tissues. The technique presents major advantages compared to conventional imaging methods: uses non-ionizing radiation, allows greater soft tissue contrast and also enables an analysis of the three-dimensional structures surrounding the upper airway. Analysis of images from MRI, relatively to other imaging techniques is characterized for being more informative in terms of output extent of data that can be

retrieved, allowing an analysis of the outputs to be oriented to the monitoring of the respiratory airway during sleep and the structures that play a determinant role in the study of normal functioning upper airway, relatively to the imaging of pathological aerodigestive tract. It allows therefore, the addition of tremendous value to screening, diagnostic, surgical planning and follow-up of patients, for a variety of pathologies developed in these organs. A particular case where this imaging technique is advantageous and necessary, precisely for the appearance of pathological scenarios during the developmental process in the aerodigestive tract, is when it is applied to children, to whom the usage of non-ionizing radiation is preferable. Despite the advantages presented, the use of MRI is not quite as common as it was idealized, being the main reason related to the high cost of the imaging technique.

In this chapter, the physical principles in which this technique is based are described, as well as the variable aspects that affect its quality and adequacy, in order to better understand the adaptability and potential in the application of imaging the human tongue.

2.2.1. Basic Principles in Magnetic Resonance Imaging

The rotational movement of protons presented in the 1H atoms nucleus – spins – implies that each of them is associated with magnetic dipolar moment (m.d.m). The most abundant atoms present in tissues are 1H atoms, with spin =1/2, being more sensitive to magnetic fields applied in Magnetic Resonance (MR). When a magnetic field is applied to the spins, these go from a state of null magnetization, to a state of magnetization where the m.d.m's tend to align themselves with the orientation of the referred field, in a given volume element, assuming a magnetization value different from zero.

This alignment is done in its majority according to a parallel direction related to the field; however, a part of these spins does not respect this behavior and its movement is named precession movement that occurs with a given frequency, called Larmor frequency (Rinck, 2001). An external pulse applied in form of oscillations of the magnetic field in the range of radiofrequencies at Larmor frequency of those spins, forces them to enter in phase precession, which originates a signal of image in RM.

The phenomenon explained in terms of physical behavior, can be examined considering, where the magnetization vector is in the Z axis, and the precession phenomenon makes the spins rotate around that axis of magnetization with a deflection angle in the vertical plane containing said axis. Therefore, into an MR equipment, a given coil is positioned in the xy plane that detects a variable electromagnetic field, producing an oscillatory signal, which corresponds to the MR image signal. This method, consequently, intends to detect the energy released by the phenomenon of Relaxation, which occurs when the radiofrequency (RF) pulse ends and the spins start to relax to the minimum energy state.

2.2.2. Relaxation Times

There are two types of relaxation of tissues, longitudinal or spin-lattice relaxation (T1 weighted time), made through the Z component of magnetization Mz after the application of magnetization in the *xy* plane, and transversal relaxation or spin-spin relaxation (T2 weighted time), that occurs by the additional effect of dephasing of magnetization induced by interactions between spins of neighbor protons, that when subjected to magnetic fields with slight differences, rotate at corresponding Larmor frequencies. This process of continuous loss of phase coherence, becomes gradually more prominent with time. The magnetization then implies that T2 relaxation time is always less than T1, and that the timeline of the process starts at a magnetization in *xy* plane that then tends to zero, followed by an increase in the longitudinal magnetization until equilibrium is achieved, in axis Z. T1 relaxation results from the interaction with the mesh of atoms in the tissue, and is characterized by a rate of magnetization Mz vector through time given by:

$$Mz\,(t) \;=\; M(0).(1 \;-\; e^{-\frac{t}{T1}}) \qquad (1)$$

This equation describes a profile, where the recovery tends to a thermodynamic equilibrium state, for which Eq. (1) given t=T1 is $[1-(1/e^{1})]$, meaning that T1 characteristic time is the time where the longitudinal magnetization recovers 63% of its equilibrium value (Rinck, 2001).

2.2.3. K-space

Spatial encoding of the image is another part of the mechanism, of acquisition that includes:

- Slice selection – implies the positioning of a gradient in the perpendicular direction to the cut to be retrieved (in the Z plane for an axial slice), the position of slice is selected by the frequency of the pulse, and the thickness by its bandwidth.

- Frequency encoding – applying a first signal according to a specific direction, the signal emitted by the different elements of volume, are characterized by different frequencies.

- Phase encoding – applying a second signal according to a determined direction, the different elements of volume according to that direction will be characterized by different phases.

Consequently, for an axial acquisition, the slice selection is done in the Z plane, the axis X and Y are responsible for the frequency and phase encoding. The two magnetic fields distributed, make for each orientation of the phase encoding gradient Gy correspondent to a line (y position), and the frequency encoding gradient dictates each columns value (x position) of that line, and in this way the (x,y) positions are stored in a matrix called K-space. Each combination is afterwards mapped in the image reconstruction to its position, and the amplitude into the corresponding intensity, by applying the Fourier transform to the 2D distribution (A. Bernstein et al., 2004).

The design of appropriate gradients, is preponderant so that k-space samples can be acquired and then inverse Fourier transformed to obtain an image of the magnetization M(x; y). K-space must be sufficiently sampled according to the Nyquist criterion to avoid object domain aliasing. The extent of k-space coverage determines the images resolution.

2.2.4. Contrast and tissue signal in RM

Contrast in MRI is due to the occurrence of specific relaxation phenomena in the different tissues, where it depends on the different times of relaxation T1 and T2, as well as different proton densities, which are characteristic and intrinsic of each type of tissue. The different tissues contain large numbers of chemical components that contribute to the measured magnetic resonance signal, and this composition characterizes each type.

Image acquisition in MRI is made through specific sequences of pulses, of RF and orientation of the phenomena of relaxation where, given the dependence on time of these phenomena, contrast can be adjusted and chosen by applying specific combinations of temporal parameters of acquisition. In the conventional MRI acquisition, these phenomena will also be influenced by the technical factors of medical acquisition, or biologically extrinsic factors. These include the magnetic field strength and homogeneity, and are crucially determined by the pulse sequence contrast influencing components TR, TE, TI and FA.

The main objective since the discovery of this technique relies in combining these parameters in order to emphasize certain contrast determining factors, or determining relaxation phenomena among others, or even a set of different factors.

2.2.4.1. TR, TE and Pulse Sequences

Pulse sequences of acquisition consist in a sequence of signals sent to the tissues, by MR machines. The pulse sequence consists in repeated RF pulses that cause a free induction decay (FID) characterized by a specific initial amplitude, mediated by the pulse sequence parameters. The two time parameters that determine this method are TR (repetition time) and TE (echo time) of the pulse sequences. TR is the time interval between two successive RF pulses, and TE is the time at which the echo signal, the signal produced by induction of the spinning protons, reaches the detector of the machine and is measured. TR can therefore determine the degree of relaxation of protons back into alignment of the magnetic field, whereas specific rates of relaxation of the tissues will imply having TR times shorter than what is needed for a full relaxation decrease the signal retrieved from the analyzed tissues.

2.2.5. Limitations and determinant considerations

The growing interest in the tongues function over all its functionalities of taste, swallowing and speech production tasks has given rise to the importance of imaging the aerodigestive tract and its structures with the best imaging technique available; whereas for the correct imaging of such complex structures, a good contrast between tissues is fundamental to allow the differentiation of the different structures at its correct boundaries.

These factors are of extreme importance for the development of the dissertation work proposed here.

Therefore, the rigorous imaging of the structures at study is determinant for the correct function of the following computational tasks of retrieval of the target structural.

In spite of the image quality conditionings referred above, MRI technique is considered as the best, a non-invasive, accurate method imaging modality available for the imaging of the muscular organ under study.

2.3. Upper airway imaging and computational analysis

Computational processing and analysis of medical images is a novel field that has gained a promising and relevant importance over the years, presenting astonishing developments in the areas of computer aided diagnostics, improving imaging technics, and imaging analysis processing of aspects that cannot be visualized and/or retrieved by plain image observation.

Volumetric imaging techniques can be used to reconstruct three-dimensional structures from serial two-dimensional images. This section provides a conceptual overview of those techniques by illustrating the reconstruction of the aerodigestive organs.

Segmentation of the target anatomical structures from MRI is still a challenging process. There are various reported methods of segmentation of static MR images/volumes (Balafar et al., 2010). Their applications to the particular segmentation of tongue, is reported in a scarce number of instances, highlighting the need of further studying this organ and the development of the adequate tools accordingly.

The imaging study of the tongue is a very underdeveloped field that has limited the improvement of anatomical and functional characterization of this organ. The recent development of Computer Vision and Machine Learning fields of Image Analysis in recent years have provided the availability of new tools of image computer analysis regarding 3D volume segmentation and reconstruction.

The first imaging reports of the tongue are made through ultrasound (US) imaging (Sonies, 1981), and subsequent applications towards the analysis of swallowing and articulation tasks

using snakes in (Unser and Stone, 1992), and using scale space filtering for edge detection in (Kelch and Wein, 1993). The main applied studies that address specifically this structure are extensively reported in speech studies. Therefore, US imaging presented the best imaging characteristics for a dynamic acquisition of multiple frames during speech production exercises. First tongue 3D modelling and reconstruction were reported in (Watkin and Rubin, 1989), that describes a trigonometric transformation of the 2D coordinates into a volume, and latter, more advanced segmentation methods were described by Akgul et al. (1998) and more recently for segmented 2D motion analysis applying Markov random fields in Tang et al. (2012).

Although the demonstrated applicability of US to tongue modelling, further study of its anatomy implies that a higher contrast and resolution imaging technique, such as MRI, prevails as more adequate in the intended study of the tongue.

The first reports of tongue anatomy imaging through MRI were reported in (Lufkin et al., 1983).

The analysis of tongue anatomy and physiology has been reported in studies using both static volumetric MRI, standard imaging modality for 3D imaging, Cine-MRI and even tagged-MRI imaging (another imaging modality that has been extensively used for temporal characterization of the tongues anatomy). Reported dynamic acquisition image analysis studies reinforce the necessity of a proper segmentation in 3D studies to the evaluation of the dynamic processes it is responsible for, such as swallowing and speech production (Lee et al., 2014). Other studies pretend to reinforce the study of the biomechanical modelling of this structure, and therefore, select a high resolution imaging modality such as static volumetric MRI (Harandi et al., 2014).

The emerging interest in the study of the tongues deformation and functionality has established that the requirement for an automated method of image analysis of this kind of anatomic data is expected to gain a rapid eminent relevance (Woo et al., 2012).

Reported studies on segmentation of the tongue, focus of the segmentation of static and dynamic acquisitions. Dynamic acquisition reveals to have obvious relevance in the study of tongue motion characterization. The processing needed is common since the format is usually based on 2D image segmentation. Vasconcelos et al. use statistical models to segment the tongues shape during the production of different sounds, in order to study speech production (Vasconcelos, Ventura, Tavares, & Freitas, 2009).

Stone et al. (2010) is one of the first reports that focuses on the strict tongue segmentation, and establishes the relevance of this study for motion patterns during speech production. In this 2D study, the images were to simply be registered through a landmark based transformation algorithm and aligned, following principal component analysis for the motion study.

The processes reported are usually divided into various basic phases: 1) Resolution wise pre-processing, 2) Segmentation, 3) Registration, 4) 3D Volume reconstruction.

In Lee et al. (2014) is reported an isotropic volume super-resolution reconstruction from dynamic tagged-MRI images. The images were subjected to a super-resolution volume reconstruction, in order to address inter-slice resolution. It was attempted to surpass the

limitation, extensively mentioned throughout this report, of long acquisition time, through the acquisition of three images with 6.0 mm thickness, which obviously affected the resolution in the through-plane direction. An up-sampling in the through-plane direction was developed using a fifth-order B-spline interpolation. Registration, for inter-slice alignment is reported in various studies (Lee et al. 2014, Woo et al. 2012), where the application of the Mutual information (MI) similarity measure is reported for registration of sagittal with axial and coronal volumetric image stacks. After the registration, a final intensity correction is made using a local intensity matching algorithm, following the application of the Random walker (RW) segmentation algorithm.

The Random walker algorithm, for segmentation of 3D super-resolution volumes was also cited in the literature for similar purposes, due to its attractive features in Woo et al. (2012).

Tagged-MRI is not adequate, regarding preponderant implications on volume reconstruction, to be used in these studies since the image quality is very low to when compared to static volumetric MRI.

A mesh modelling approach is reported in Harandi et al. (2014) whereas the registration technique departs from an initial source model of the tongue to whose vertices are applied external forces forcing it towards the target boundaries through a process dictated by local intensity profile registration and positions computed through normalized cross-correlation and finalized by shape matching. The advantage of this approach is that it allows user input to automatically correct the mesh nodes positioning.

The most recent study published attempted to go further in the investigation of functional behavior, and describes a novel method of segmentation of individual tongue muscles (Ibragimov et al., 2015), specifically genioglossus and inferior longitudinalis. In their work, it was implemented an adaptation to muscle segmentation of the game-theoretic framework (GTF) algorithm, based on land-mark-based segmentation.

2.3.1. MRI 3D volumes image segmentation techniques

Computer-aided modelling of the aerodigestive organs is beneficial for 3D visualization, and for the understanding of the associated physiology. Medical imaging is retrieved in a universal format, organized according to a predefined standard.

The studies that address image segmentation of the tongue are limited and therefore, an overview of this list of presented in the following points.

2.3.2. DICOM Standard Overview and Volumetric Data

The process of imaging has become extensive, including a wide variety of formats, imaging technics, and post-acquisition procedures. For this reason, in addition to the creation of a communication system and network storage used, named Picture Achieving and Communication System (PACS), a common format that allows correspondence between station and safe data transference was created.

A picture archiving and communication system (PACS) is essentially a network system for digital or digitized images from any modality to be retrieved, viewed and analyzed by an appropriate expert system, at different workstations.

This communication is safeguarded by a pattern called DICOM - Digital Imaging and communications in Medicine, a standard for the communication and management of medical imaging information and related data (ISO 12052). The DICOM format was first released in initial versions of the ACR-NEMA - version 2.0 published in 1988 - created standardized terminology, an information structure, and file encoding, whereas the version 3.0 of the standard published in 1993 finally addresses the matters of a standardized communication of digital image information, developed by the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) who in 1983 formed a working group with the objective of developing a model that would allow a fully digital workflow for image exchange. It is defined as a set of standards for treatment, storage and transfer of medical images and associated information, in an electronic format, and was created with the purpose of standardizing the formatting of diagnostic images allowing these to be exchanged among equipments, computers and hospitals (NEMA). The DICOM system has interest in a variety of medical fields, including cardiology, dentistry, endoscopy, mammography, ophthalmology, orthopedics, pathology, pediatrics, radiation therapy, radiology, surgery, etc.

From the Scientifics community point of view, this standard enabled an open architecture for imaging systems, bridging hardware and software entities and allowing interoperability for the transfer of medical images and associated information between disparate systems (Dreyer et al., 2006). Furthermore, in the field of Computational Vision, the development of image processing and analysis tools is now possible to be standardized, without any format and organizational issues.

The data structure of a DICOM file consists of a set of data elements. A header portion includes general data elements related to the image. Image data is also contained in one data element, or more data elements if there are more than one part image in this DICOM file. Each data element is stored as depicted in Figure 8.

After the header a dataset follows, which represents the content of the file. The dataset can be an image, a presentation state, a structured report or another DICOM object. For reading procedures, the format implies that a system based on a data dictionary, which stores all kinds

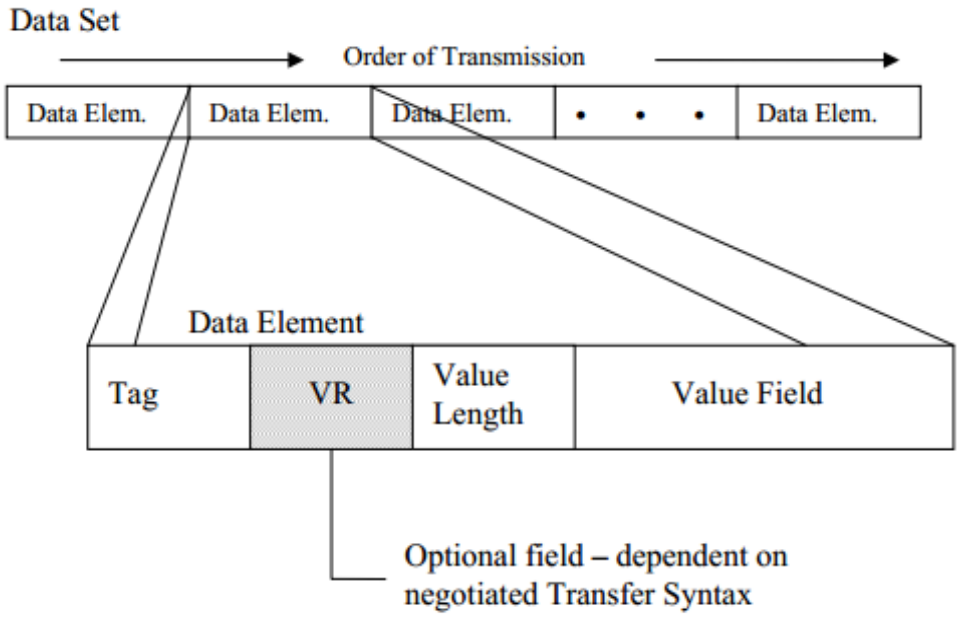

Figure 8- DICOM data set structure consists of several data elements.

of tag groups so that every data element can be read correctly. Information of each kind of image (CT - computer tomography, MR - magnetic resonance tomography) has an identifier as well as the instance of such a class. There is no definition of 3D data storage in DICOM standard.

A volume is usually presented by an ordered series of 2D DICOM files, each of which may have multiple components of the same size and representation, which are the parallel slices of the volume.

2.3.3. Image Segmentation

Computational Vision includes tasks of image segmentation whereas the objective of segmentation algorithms is to partition an image into a finite number of important regions under the image scope, such as anatomical or functional structures in medical images.

Image segmentation can be defined as the process of decomposing an image into various labeled regions that are characterized by some measure of homogeneity inside it, and heterogeneity among different regions is maximal.

When it comes to airway contour delimitation, the process is complex due to eventual non-identification of organs, anatomic parts or artificial inclusion of non-existing parts. Air-tissue boundaries of vocal tract are hard to extract due to the similarity of anatomic structures around it. High resolution MRI is known to provide good representation of muscle anatomy. However, a compromise of image quality for the acquisition of volumetric data is in many cases a balance to take into consideration upon the definition of the image acquisition protocol. This will lower boundary resolution and contract, since upon the acquisition pixel intensities are obtained, through an averaging process of signal over each TR time, over the space of the target volume.

As previously stated, one of the issues that arise from MRI acquisition, is the technical consequential issue created by the rather long times necessary for the retrieval of each 2D k-space image. Volumetric MRI data consists of a series of 2D images corresponding to a given series of slices of tissues, and a determined thickness. Each slice is acquired consecutively, in a sequential series of acquisition, whereas the process of each acquisition is therefore, very sensitive to motion of tissue, that will practically inevitably cause some degree of inter-slice misalignment.

Under the field of medical image processing and analysis this issue is currently covered by image registration, under which extensive research devotion and developments have been made over a time span of 25 years, and its relevance and attention given include applications with computed tomography (CT), magnetic resonance imaging (MRI), Positron emission tomography (PET), Single Photon Emission Computed Tomography (SPECT), and also a later increase of applications in Ultrasound (US) imaging. Registration is in many cases used to achieve the alignment or/and fusion of different types of images in order to retrieve and complement the information obtained from each one. Image registration is also used to correct for subject motion between acquisitions. Accurate registration is of great importance in this application because small perturbations in alignment can lead to visible artifacts after applying the MAP-MRF reconstruction algorithm (increases the variance of intensity values at each spatial location). Mutual information (MI) (Maes et al., 1997) is one of the most popular similarity metrics, whereas, reports show it as being successfully employed for non-rigid registration, although this metric presents also limitations.

A registration method using a mesh-to-volume technic represents a different approach to landmark generation by adapting a deformable surface model to the target volume. This registration is used in Harandi et al. (2014), based on mesh nodes position calculation through local gradient intensity profiles and normal to the mesh surface.

The problems of the segmentation of this structure may arise from the presence of poor muscle-neighboring structures interface visibility, intensity mismatches, blurring, blank regions, etc.

2.4. Related work

As stated previously in this report image quality is a determinant factor for the success of computational analysis. The technical time limitations of MRI acquisition protocols, translate into resulting limited resolution images due to its high sensitivity to motion, which increases almost inevitably the probability of movement due to swallowing motion to occur, and will automatically condition negatively the images acquired. 3D acquisitions of upper airway (head and neck imaging) takes usually at minimum 4-5 minutes. Maintaining the tongue immobilized for such time span is likely to induce involuntary motion and/or swallowing. In Woo et al. (2012), methods for correcting this problem are proposed with super resolution volumes. Super

resolution algorithms can be categorized into being based on non-uniform interpolation, frequency domain, and spatial domain analysis methods. 3D MR images of the tongue can be produced from sets of orthogonal volumetric images, acquired at a lower resolution and combined using super-resolution techniques. The production of super resolution volumes may also imply adaptations of acquisition protocols in order to obtain, for instance, volumetric acquisitions with specific/target areas of super-resolution as reported in Ibragimov et al. (2015), as an adapted kind of orthogonal acquisition from (Woo et al., 2012).

The success of this step of image processing will determine prominently the success of the following image segmentation steps.

Supervised segmentation algorithms are based on an analysis of a training data as example and produces an inferred function that allows the mapping of new data.

Supervised segmentation algorithms typically operate under one of two paradigms for guidance:

1) Specification of a portion of the boundary of the target object;

2) Specification of a small set of pixels belonging to the desired structure and (possibly) a set of pixels belonging to the image background.

Therefore, supervised algorithms only use labeled information retrieved by any of the previous methods data. Particular variants are also relevant in this study, such as semi-supervised algorithms that make use of unlabeled data for training, typically a small amount of labeled data with a large amount of unlabeled data (Xiaojin Zhu, John Lafferty, 2003).

Among these categories, image segmentation can be based on seed growing approaches, which requires an operator/user to empirically select seeds and thresholds. Pixels around the seeds are examined, and included in the region if they are within the thresholds, sometimes adding the requirement that they are sufficiently similar to the pixels already in the region. Each added pixel then becomes a new seed whose neighbors are inspected for inclusion in the region. The random walker algorithm falls under this category.

Random Walker segmentation algorithm

The Random Walker (RW) algorithm was proposed in Grady (2006), is being applied in several studies in the segmentation of the upper airway. This algorithm presents the several characteristics that confer the adequacy and suitability to this algorithm among others. It is characterized by having fast computation costs, flexibility, an easy user-interaction is required, and produces a very accurate segmentation with minimal interaction, through user-defined seeds.

The algorithm is a K-way image segmentation and semi-automatic since it requires user-defined regions correspondent to K structures. These are defined by the user, specifying a small number of pixels with user-defined labels as seeds (on the tongue and the vocal cavity). Also, the algorithm uses for graph representation, harmonic energy minimizing functions, whereas low energy corresponds to a slowly varying function over the graph has will be defined next (Zhu et al., 2003).

2.4.1. Modelization of deformable tissues

Model-based deformable models that are able to fit to new data instances have great interest in computer vision. Within the spectrum of study that this area offers, statistical modeling of shapes includes many techniques. Beginning by its simpler version, there are Active Contours or Snakes (Kass, 1987), which are energy minimizing curves that deform according to internal and external forces, followed by Active Blobs (Isidoro & John, 1998), whereas the models deformation is based on physical properties such as stiffness and elasticity modeled by Finite Element Methods (FEM), and finally reaching the more complex methods of Active Shape Models (ASM) and Active Appearance Models (AAM). Active Shape Models (ASM) also known as Smart Snakes (Cootes & Taylor, 1992) are Point Distribution Models (PDM) or Statistical Shape Models (SSM), i.e. landmark based methods, where the variability of shape is learned offline using statistical evaluation through Principal Component Analysis (PCA), and allows model fitting of new shapes by combining *a priori* knowledge about how a shape deforms and evaluated texture information along normal scanlines, driving the shape model to a fast and very accurate model fitting. Finally, Active Appearance Models (AAM) (Cootes, Edwards, & Taylor, 2001) consist of an evolved version of ASMs, respective to in addition, including a complete texture mode, to the full model of shape used to fit to new images.

The tongues anatomy has been modeled in a number of fields of study in previous works. Parametric representations of the tongue shape have been devised based on statistical methods by Badin et al. (Badin & Gérard, 1998), Engwall et al. (Engwall, 2000), and spline descriptions by Parent & King (Parent & King, 2001), Stone & Lundberg (Stone & Lundberg, 1996). A physiological representation is described by Takemoto . Dynamic models have been constructed using both mass-spring systems by Dang & Honda (Dang & Honda, 2004) and finite element methods by Gerard et al. (Gerard, Perrier, & Payan, 2006), Wilhelms-Tricarico (Wilhelms-Tricarico, 1995). A recent survey by Hiiemae & Palmer (Hiiemae & Palmer, 2003) describes these representations and applications in detail.

Approaches based on generating point distribution models that captures the shape of the object of interest and then augmenting this model with intensities near landmarks in the case of Active Shape Modeling (ASM) falls within the supervised category of modeling and analysis of anatomical variations in images. Vasconcelos et al. applies precisely this type of models statistically describing tongues shape, in an analysis of speech production of vowels (Vasconcelos, Ventura, Tavares, & Freitas, 2009). In previous works, these methods were on another approach adapted to a game-theoretic perspective as was validated by Ibragimov et al. (2012), and applied to tongues individual muscles segmentation for the first time, by the same authors (Ibragimov et al., 2015).

Game-theoretic framework for landmark based segmentation

This algorithm is based on an adaptation of an Economics theory, the game theory, which studies the decision making of player that affects the other players during a game, that was

established in Neumann and Morgenstern (1947) into the landmark position search of the ASM segmentation. In this method, candidate points are defined for each landmark, and likelihoods that each candidate point represents a specific landmark are evaluated. The landmark detection is formulated mathematically as a game, considering landmarks as players, landmark candidate points as strategies, and likelihoods that each candidate point represents a landmark as payoffs.

To the obtained combination of optimal candidate points follows the definition of the boundaries connection each pair of adjacent landmarks, formulated as an optimal path searching problem. Image intensities in the area between landmark and are filtered by a control intensity function that minimized the distance error training images to the ground truth boundary.

Landmark-based atlasing using B-spline and Demons atlasing are other possible algorithms to be used for non-rigid segmentation, based on transformations to map/align the training-defined landmarks to the landmarks identified in the new target image (Ibragimov et al., 2015).

In 2000, a 3D tongue model was developed by Engwall et al. within the Kungliga Tekniska Hogskolan (KTH) 3D vocal tract project using manually extracted tongue contours from MR images of a reference subject producing 43 sustained Swedish articulations (Engwall, 2000). The extraction of the articulatory models parameters was done by decomposing the geometrical points of the tongue in linear components, through a Linear Component Analysis, where the factors to be extracted were imposed on the model using MR images articulatory measures. Two years later, in Badin et al., a database of 3D geometrical description of tongue, lips and face was established for a speaker sustaining a set of French allophones (Badin, et al., 2002). For this, data from MRI, along with a video with and without a jaw splint were used. An important finding of this research was that, most 3D geometry of tongue, lips and face could be predicted from their midsagittal contours, at least for speech assessment purposes. Indeed, the knowledge acquired from midsagittal data and from traditional 2D models is far from obsolete.

2.4.2. Segmentation using Statistical models

When it comes to specifically study 3D organs, the more sophisticated variances of statistical models described above in the found literature of tongue modelling are necessary. However these methods are based in an image quality achieved by image interpolation methodologies, and resolution improvement.

On the other hand, for 2D studies, the dimensional reduction implies a direct simplification of the modelization. In real image segmentation, a specific SSM preserves the characteristics of an organs shape even if the image information is misleading or ambiguous. The study carried out by Vasconcelos et al. is a representative example of this (Vasconcelos, Ventura, Tavares, & Freitas, 2009). Although it should be noted that segmentation errors produced by for instance

statistical shape models, cannot be exclusively accounted to the limited generalization ability or specificity, but also to deficiencies of the model-to-shape search algorithm, that is, the ASM. In particular, the results of this methodology may be influenced by the initial placement of the SSM, as well as by the search strategy adopted, which may fail to detect certain image features.

The search algorithm therefore is key to the segmentation process:

- local image feature search computes a set of candidate positions around each landmark.
- An appearance model is used to assign a score to each candidate such that they can be ranked.

The referred features can be 1D or 2D image features, to which many processing methods are available for ideal candidate choosing, by scoring the feature candidates, or classifier ideal candidate choosing, among others. Many studies use a wide range of search algorithm strategies that is susceptible to the structure boundary features (Heimann, Wolf, & Meinzer, 2006).

Finally, M. Vasconcelos proved in a recent study that it is possible to segment the vocal tract, and capture its modulation variability upon the production of different sounds using Active Shape Models that capture the adequate variability it suffers upon the production of different sounds, including the variation of the tongue, vellum, pharynx tissue shapes. It includes the production of a variety of sounds, and the model captures the variability suffered upon the tongues dorsum and posterior wall (Vasconcelos, M., 2015).

2.5. Conclusion

Many are the applications that can profit by the study of the aerodigestive structures in images. The tongue appears to have still endless functional and physiological mysteries yet to be resolved. Its relationship with the neighboring structures is very complex and seems to be intrinsically related with other organs to the performance of the tasks that are under the vocal tracts responsibility. Various imaging methods have been reported in previous studies addressing the airway structural geometry including muscle activation, using specifically endoscopic imaging (Kuna, 2004), X-ray fluoroscopy (Wheatley et al., 1991), acoustic reflection, Computer tomography imaging (CT, in Teguh et al., 2011), optical coherence tomography (OCT, in Togeiro et al., 2010), as well as magnetic resonance imaging (MRI) (Woo, Murano, Stone, & Prince (2012), Moon et al. (2010), Arens et al. (2003)).

Another important aspect that represents a current challenge in the clinical practice of physicians, takes into consideration that large numbers of target and normal tissue structures present in the head and neck, that require manual delineation. An example includes cancer patients, where the contouring is tedious and time consuming. Also, in certain courses of treatment, such as head-and-neck intensity-modulated radiotherapy, it is required accurate delineation of those structures, implying efficiency benefits from an economical perspective, besides obvious improvement to the patients treatment.

Furthermore, the available imaging systems, cannot yet take upon the adoption of more sophisticated imaging techniques and types of acquisition. This factor is important since multi-stack 3D acquisitions are one of the latest and more sophisticated techniques for the analysis of the tongue anatomy in MRI. However, the available medical imaging procedures do not include these types of acquisitions as golden standard, implying only single plane imaging acquisitions, usually the sagittal plane, being the most representative of all the structural changes of the aerodigestive tract.

It is preponderant to address the various health issues that still need to be studied further, as well as improve the imaging. Understanding speech disorders, understanding sleep apnea, planning and practicing surgery with computer models, and understanding problems in tongue movement following surgery are some of the examples of problems that could be addressed in further studies of the tongue. No instance of an automatic tongue segmentation framework was found in the literature, whereas only semi-automatic methods were found. This is extensively referred to be a difficult task, in a number of study, due to the imaging quality disadvantage, and the image features related to the structural environment in which this organ is inserted, being in close vicinity with other structures, and having boundary segments directly connected to neighboring structures. The insufficient image contrast between the structures to be segmented, such as the tongue and adjacent soft tissues at the periphery, makes the segmentation task challenging and the boundary detection techniques in the reported segmentation methodologies present limitations, that represents a concern especially upon the analysis of the lower boundary of the tongue.

Chapter 3

Statistical Modeling of the tongue

In this chapter, the methodology developed and all the algorithms constituting it will be thoroughly explored, as well as the image dataset used in this work.

The goals of the proposed work were the development of a simple and objective system for a semi-automatic modeling and subsequent segmentation of the tongue. Always bearing in mind the medical barriers and clinical usefulness, several methods were combined to fulfil the proposed objectives and will be explained in detail in the current chapter, from the dataset construction to the set of computations to be applied.

Active Shape Models (ASMs) are widely established has algorithms presenting adequate features, for the analysis of deformable objects. These models algorithm, is based on an image segmentation technique that takes advantage of the data derived from the training set. The model is built up from shape information obtained from analyzing points along object boundaries known as landmarks. ASMs are based on the combination of a Statistical Shape model - SSM (also referred as Point Distribution Model – PDM) plus a set of local image appearance models. In this work the appearance model is based on a Profile Model, which described the intensity distribution in the landmarks neighborhood. The SSM describes the shape variability of the template and the appearance models describe the image variability around each of its points. This forms the basis in searching unknown images for target shapes, where ASMs segmentation framework, combines precisely the shape and intensity information that is described by the two models to search matches of the objects shape in new images (Figure 9). This process is only achieved if the building of each of these models is adequate and furthermore, if the information described by each of these is fully understood. The analysis of the adequacy of the models built in the context of analyzing the tongue, was addressed in this Chapter as well as analyzed the information retrieved by each of the of the models.

This chapter is divided in three main sections: describing the image datasets used and important Magnetic Resonance Imaging protocol details, an overview of the methodology of implementation of an Active Shape Model (ASM) for the study of the shape of tongue and appearance during speech production, and finally in the following two Sections each of the sub-models of the ASM are described along with the basic equations and mathematical fundaments. The description of the *classical* Active Shape Model is presented in (Cootes & Taylor, 2004) and was the methodology followed for building said models.

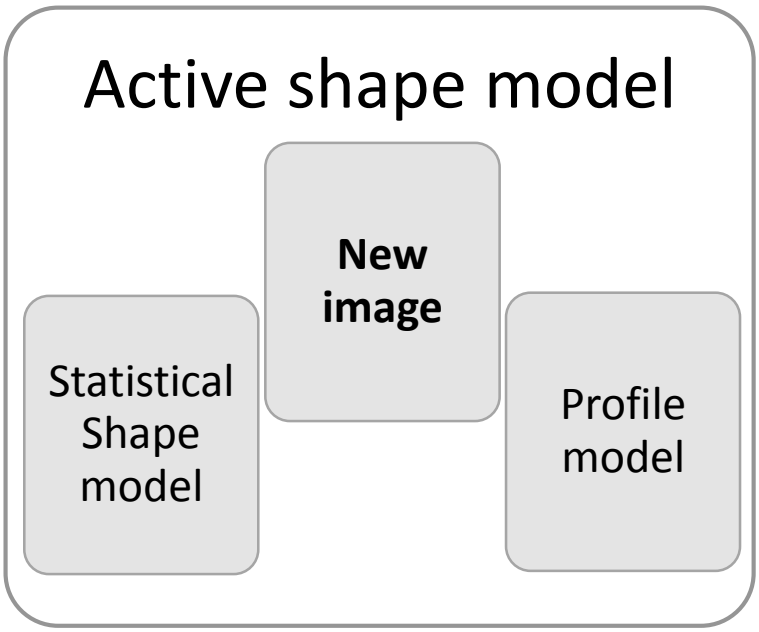

Figure 9 - Active Shape model structure scheme.

3.1. Image Dataset

In order to build the statistical model of the tongue shape it was used a training set of MRI images collected during artificially sustained articulations of Portuguese sounds.

For the analysis of the tongues configurations during sustained articulations of EP speech sounds, a dataset produced through MR 3T acquisition system was used, that comes from the original works in [Ventura 2012]. Image acquisition was performed using a Siemens Magneton Symphony 3 Tesla (3T) system and a head and neck array coil, with the subject lying in the supine position. The T2-weighted sagittal slices obtained have 3 mm thickness, by using Turbo Spin Echo Sequences, with the acquisition duration of approximately 10.6 s. Subsequently, this protocol has resulted from a compromise between the signal to noise ratio, the number of slices acquired and the time needed for subjects to sustain articulation successfully during image acquisition process.

The dataset is composed by a total of 19 images retrieved by two subjects, one male and one female. The subjects were subjected to the same pre-imaging vocal training, and were imaged under the same equipment conditions.

From these images, it is possible to observe different vocal tract configurations for EP vowels production, as well as for some oral sounds. Oral sounds were chosen since the greater movement produced by the tongue for speech production is made upon the production of precisely this type of sounds. The dataset includes vowel production by simple production of the sounds and in other cases by the production of the vowel sound preceded by the occlusive bilabial consonant 'p'. This second strategy was used to mark the initialization of the vowel sustentation, since it does not influence the production of vocalic sounds, and allowing the production of more natural sounds.

Comparing the several morphological configurations of the subjects during the articulation of the EP sounds, individual differences of various organs involved in the upper airway morphology upon the production of said sounds are revealed, whereas a wide range of configuration variability of the tongue would be captured.

So, the tongue moves from front-high positions to a central-low position on the oral cavity for the vowels [i, and a] and from this position to back-high positions for the vowels [u], respectively. Examples of the MR images from the datasets acquired are depicted Figure 10.

Figure 10 - Examples of images from the 3.0T image dataset used, of imaging of the oral sounds a (A), i (B) and u (C).

3.2. Methodology

In the present section it is describes the implementation of the models built for the study of the shape and appearance of the tongue.

The algorithms to create the statistical deformable model were developed in MATLAB (http://www.mathworks.com), namely a Statistical Shape Model and a Profile Model, and latter the ASM presented in the next Chapter that combines the two. Therefore, the methodologies presented in the present Chapter, were conceived always bearing in mind the final purpose of building the Active Shape Model presented in Chapter 4.

Using a number of images described in the previous section, each one contributes with two components to the models:

1. A set of landmark points that will be aligned to the mean shape.
2. A set gray-level profiles of the contour normals.

The terms mean shape, indicating the mean of all landmark points in an entire set of annotated images, and mean gray level indicating the mean of all vectors obtained by sampling the gray levels along all contour normal, will be used throughout this document.

The sets of landmark points was used to build a Statistical Shape Model, and the gray-level profiles to build a statistical gray-level model or Profile Model. The procedures for building the shape and gray-level models are very similar, whereas the key method is principal component analysis (PCA).

The Statistical Shape Model is first trained on a set of manually landmarked images. By manually landmarked it is meant that a medical image professional had to mark all the images to allow a correct validation of the basis of the model to be produced.

A Profile Model for each landmark, which describes the characteristics of the image around the landmark. The model specifies what the image is expected to "look like" around the landmark. During training, we sample the area around each landmark across all training images to build a profile model for the landmark. During search, we sample the area in the vicinity of each tentative landmark, and move the landmark to the position that best matches that landmarks model profile. This generates tentative new positions for the landmarks, called the suggested shape.

After training we can use the Active Shape Model, the combined Statistical Shape and Profile Models, to an automatic search of this structure in a test image. The general idea is (1) try to locate each landmark independently, then (2) correct the locations if necessary by looking at how the landmarks are located with respect to each other. This resulting model will be presented in the next Chapter.

3.3. Shape Model

A shape model defines an allowable set of shapes. In this document, shape models have a fixed number of points and a matrix formula which specifies the relationship between the points. We will use Point Distribution Models, which learn allowable constellations of shape points from training examples and use principal components to build the model. A concrete example will be presented shortly.

The application of statistical models, such as deformable and active models, to characterize and reconstruct the tongue during speech production was taken into consideration in the present work. The development of active models to represent the vocal structures from a global perspective is here presented.

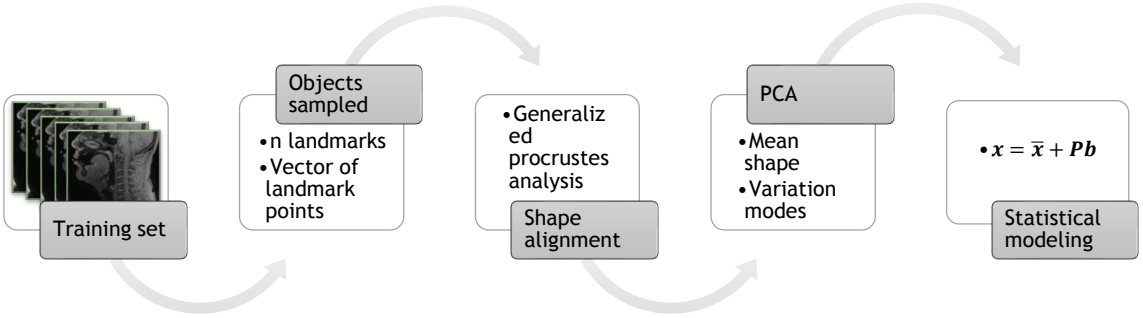

Figure 11 -Statistical shape model building scheme.

The objective of a Point Distribution Model or Statistical Shape Model is to describe statistically the shape variations allowed by a shape described by a set of points of a non-rigid object. The overview of the steps for building a Statistical shape model is a four-step process Figure 11. The first step is to capture information from the training set, accomplished by labeling the images with landmark points. This is followed by alignment of these landmark points, that describe training shapes, into a common referential, and finally Statistical analysis is then performed on the aligned shapes, by Principal Component Analysis (PCA) that will allow the description of the set of landmark-based shapes by a non-linear combination of a mean shape and the statistical parameters produced. These steps are further discussed in the following subsections.

The building process of such model begins with the selection and acquisition of the shape information from a set of images representing the object to be modelled, the training set of images. From each of said training images, the shape of the target object should be represented by a set of labeled landmark points, whose positions are required to be similarly and orderly defined so that the variations of the boundaries defining said object is possible to analyze in the following steps. In 2D images ($k = 2$), a shape defined by n landmarks $\{(x_i, x_i) : i = 1,\ldots,n\}$, define the $2n$ vector of coordinates:

$$c = (x_1, x_2 \ldots., x_{n-1}, x_n, y_1, y_2, \ldots. y_{n-1}, y_n) \quad (2)$$

The model implies statistical comparison of the training shapes and with that in mind, the first step to build the model consists of aligning the shapes, centering them into a common grid of coordinates. The formulation of the problem is then made as a shape model consisting of an average shape and allowed distortions of the average form:

$$\hat{x} = \bar{x} + Pb \quad (3)$$

$\hat{x}$ is the generated shape vector, containing all the shapes the object can acquire, and that are described by the model.

$\bar{x}$ is the mean shape, produced by the averaging of each landmarks position.

P is the matrix of the eigenvectors of the covariance matrix S of the training shape points obtained by:

$$\frac{1}{n_{shapes}-1}\sum_{i=1}^{n_{shapes}}(x_i-\bar{x})(x_i-\bar{x})^T \quad (4)$$

B is the vector of eigenvalues correspondent to each eigenvector column in P.

We can use Equation (3) to generate different shapes by varying the vector parameter b. By keeping the elements of b within adequate limits we ensure that generated shapes are within plausible instances.

The relative size of the eigenvalues tells us the proportion of variation captured by the corresponding eigenvectors. We can capture as much variation of the input shapes as we want by retaining the appropriate number of eigenvectors.

3.3.1. Landmarks

The ASM is first trained on a set of manually landmarked images. Each shape from the training set was represented by a set of labeled landmark points, which usually represent important zones of the boundary or significant internal locations of the object. The manual process of labeling an object is normally the simplest one, however, this considers the premise that the user has a technical knowledge about the object involved in order to choose the best locations for the landmarks and consequently, be able to mark them correctly in each image of the training set. Images were annotated by a medical imaging specialist and further cross-checked by the author, to detect possible inconsistencies or missed landmarks. In the labeling process, sixteen points were defined to characterize the tongue shape:

- Two points in the lingual frenulum (anterior and posterior);
- One point in the tongues tip;
- One point in the tongues root;
- Six points along tongues body;
- Six points along the inferior surface of the tongue.

A straightforward way to improve the fit is to increase the number of landmarks in the model. In this work the main landmarks retrieved were interpolated, spline interpolation by a factorization method, in order to improve this aspect of segmentation using MathWorks, 2015a. An algorithm was developed in order to allow the factorized increase of the number of landmarks, that vary by different factors to allow a study of the variation of the number of landmarks, since it is a priori knowledge that this factor influences the quality of the fitting result.

3.3.2. Shape alignment

In order to get statistical validity, it is crucial that all shapes are represented on the same referential. A shape can be aligned to another shape by applying a transform which yields the

minimum distance between said pair of shapes. For this purpose, it is suitable to remove the location, scale and rotation effects inherent to each of them in their image referential.

Cootes and Taylor's Appendix B gives methods to align two shapes by using a least-squares procedure (Cootes & Taylor, 1992). Accurate alignment may be deemed more important for certain points than for others, and points can be weighted accordingly during alignment. The classical solution of align two shapes is the Procrustes Analysis method. It align shapes with the same number of landmarks with one-to-one point correspondences, which is sufficient for the ASM standard formulation.

3.3.2.1. Procrustes Analysis

In this course of application, to align two shapes, f_1 to f_2, the process consists on finding the parameters of the transformation T, i.e. scale, s, rotation, θ and translation, (tx, ty) that, when applied to f_1 best aligns it with f_2, minimizing the Procrustes distance metric:

$$D_{procrustes}(f_1, f_2) = \sqrt{\sum_{i=1}^{n}(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (5)$$

with respect to s, θ and (tx, ty), for n corresponding landmarks of each shape.

Step 1 consists on redefining the referential, by calculating both shapes centroids, and centering them into the origin:

$$x_{1c} = x_1 - \overline{x_1} \quad (6)$$

with $\overline{x_1}$ being the centroid coordinates.

Following by their normalization by isomorphic scaling:

$$\hat{x} = \frac{x_{1c}}{\|x_{1c}\|} \quad (7)$$

This produces a matrix $S = [\hat{x}|\hat{y}]$ of size $nx2$ with each pair of origin-centered landmark coordinates. From this point the statistical comparisons between the shapes can be performed correctly.

Step 2, consists on calculating the rotation matrix to be applied, which is formulated considering a shape vectors S a translation vectors as column vectors to which it is applied, as therefore is represented as:

$$R = \begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix} \quad (8)$$

And is calculated by:

$$E = RS_2 - S_1 \quad (9)$$

where S_1 is the unaligned shape and S_2 is the reference shape, both represented in the form given by Equation (6).

The optimal rotation matrix that aligns both shapes, minimizing the distances given by Equation (5), is given by using a Singular Value Decomposition (SVD) on matrix S_2S_1, where:

$$SVD(S_2 S_1) = USV^T \tag{10}$$

And more specifically:

$$R = UV^T \tag{11}$$

The aligned new shapes can be obtained by the calculated Tranformation through the following formula:

$$T\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x_{tranlate} \\ y_{tranlate} \end{pmatrix} + \begin{pmatrix} s\,cos\theta & s\,sin\theta \\ -s\,sin\theta & s\,cos\theta \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \tag{12}$$

Matlab offers a direct function of this approach through the *Procrustes* function (Mathworks, 2015b) present in the Statistics and Machine Learning Toolbox.

3.3.2.2. Generalized Procrustes Analysis

The process of aligning various shapes present in the image dataset has been described to be successful through the method described in the previous section as an iterative optimization process, named Generalized Procrustes Analysis (GPA). The method consists in sequentially align the instances of the dataset shapes in pairs through Procrustes alignment, using a reference shape, the mean shape to which others are aligned. After the alignment a new estimate for the mean is recomputed and again the shapes are aligned to this mean.

This procedure is performed repeatedly until there are no significant alterations to the recalculated mean shape. Algorithm described in Figure 12.

Input: set of unaligned shapes

1. Choose a reference shape (usually the first shape)
2. Translate each shape so that it is centered on the origin
3. Scale the reference shape to unit size. Call this shape $\bar{x}_0$, the initial mean shape.

repeat

4. Align all shapes to the mean shape
5. Recalculate the mean shape from the aligned shapes
6. Constrain the current mean shape (align to $\bar{x}_0$, scale to unit size)

until convergence (i.e. Euclidean distance between previous and present mean shape is minimal)

Output: set of aligned shapes, and mean shape

Figure 12- Generalized Procrustes Analysis algorithm outline.

3.3.3. Principal Component Analysis

The Principal Components Analysis (PCA) is a statistical technique that allows data dimension reduction. This procedure searches for directions in the data that has largest variance and subsequently project the data onto it. Mathematically is defined as an orthogonal

linear transformation that projects data into a new coordinate system defined by the data variance axis. The dimension reduction is done by holding data that contribute more for the variance ignoring remaining, less important characteristics.

Considering a dataset with N vectors: $x_i : i = 1,\ldots,N$, where each $x_i$is a n dimensional vector. It is required that the number of samples is greater that the number of dimensions ($N > n$).

A Principal Component Analysis is performed by:

- Computing the N vectors average:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{13}$$

- The maximum likelihood estimation of the covariance matrix is given by:

$$C = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T \tag{14}$$

The basic purpose of performing PCA is to find an orthogonal coordinate system for our data cloud, in such a way that the greatest variance lies on the longest axis (first principal component), the second greatest variance lies on the second longest axis, and so on. A way of calculating principal components is by computing the eigen-decomposition of the covariance matrix, resulting in eigenvectors and eigenvalues of the covariance matrix and sorting them by their corresponding eigenvalues magnitude. This was done using Mathworks, 2015c.

3.3.4. Eigenvectors and eigenvalues

The eigenvectors, Pi, and the associated eigenvalues λ of the covariance matrix are computed and ordered in such a way that $\lambda i \geq \lambda i + 1$.

The principle to have in mind in this type of analysis is that the eigenvector which correspond to the highest eigenvalue, represents the direction of largest variation. The second eigenvalue corresponds to the largest variation in a direction orthogonal to the first.

Some eigenvalues are very small, so they do not contribute much to the total variance and can be ignored. Therefore, the data can be approximated as a linear combination of the most relevant eigenvalues and corresponding eigenvectors that result in data compression (The remaining eigenvalues represent noise in the form of numerical errors). The analysis of the resulting eigenvalues must be pondered, since it is variable according to the nature of the data, and therefore, it influences the number of resulting eigenvalues, and also determine the presence or absence of an abrupt cutoff point in their magnitude. The relative size of the eigenvalues tells us the proportion of variation captured by the corresponding eigenvectors. We can capture as much variation of the input shapes as wanted, by retaining the appropriate number of eigenvectors.

Calculating the weights of each eigenvalue relatively to all others, it was possible to organize in descendent order of the relative retained variance, and therefore the model can be defined through the t most important eigenvalues (t is the number of modes of variation). Any instance in the training set can be closer to the original data through Equation (3) discussed in the section 3.1.

The number of modes of variation to hold, t, usually is chosen in such a way that the model represents a user defined variance of the total data. Each eigenvalue, λi, gives the variance of data in the direction of the correspondent eigenvector, Pi, and the total variance of the training data is given by the sum of all eigenvalues, $V_N = \sum_{i=1}^{N} \lambda i$. The t higher eigenvalues are chosen in order that:

$$\sum_{i=1}^{N} \lambda i \geq pV_T \tag{15}$$

where p is a portion of the total variation registered in all shapes and $\sum_{i=1}^{N} \lambda i$ the vector of cumulative sums of the vector of eigenvalues. This vector here notated by λ corresponds to the vector $\boldsymbol{b}$ in the notation used along this work.

3.3.5. Representing a given shape by the shape model

In the reverse direction, given a suggested shape x on the image, we can calculate the parameter $\boldsymbol{b}$ that allows Equation (3) to best approximate x with a model shape $\hat{x}$. This is achieved by seeking the $\boldsymbol{b}$ and T that minimize:

$$distance(x, T[\hat{x} + Pb]) \tag{16}$$

T is a similarity transform, that includes scaling, rotation and translation, which maps the model space into the image space. The transform is needed because the shape x could be anywhere in the image plane, but the model works off scaled upright shapes positioned on the origin. Cootes and Taylor section 4.8 describes an iterative algorithm for finding b and T (Cootes & Taylor, 2004). After calculating $\boldsymbol{b}$, we reduce any out-of-range elements b_i to ensure that the generated shape conforms to the model, and yet remains close to the suggested shape.

3.4. Profile Model

A profile model for each landmark, which describes the characteristics of the image around the landmark was built. The model specifies what the image is expected to “look like” around the landmark. During training, we sample the area around each landmark across all training images to build a profile model for the landmark. This is needed so that during search, we sample the area in the vicinity of each tentative landmark, and move the landmark to the

position that best matches that landmarks model profile. This generates tentative new positions for the landmarks, called the suggested shape.

This section is divided into the sub-sections regarding the pre-processing of the image for noise removal, and the profile model building methodology adopted.

The objective of the profile model is to be used in the final Active Shape Model to take an approximate tongue shape and produce a better suggested shape by template matching at the landmarks. As suggested by Cootes & Taylor, the search model is started with the mean tongue shape from the shape model (Cootes & Taylor, 2004). This shape was aligned and positioned manually in this step of the study in order to allow the correct formulation of the model and model search method. In this step of the process the images quality and quality of the tissues represented has presented itself as a disadvantage of the localization of appropriate shapes, since the basic principle of the process lies in the correct detection of exactly the right boundaries of the object to be modelled. This detection is only based on the image intensities. As stated in the Literature Review section of this work, this is one of the major problems presented in the modelling of various structures presented in MRI imaging, specifically of structures represented in MRI of head and neck, whereas it is very difficult to discern the different tissues apart, since the corresponding intensities are very similar in the various structures, which is once more due to the compositions of said structures by the same or similar types of tissues. For the matter at stake, this translates into the principle disadvantage that the boundaries of said structures are difficult to detect. By simple analysis of any un-processed MRI image presented, it is possible to visualize this effect. This is obviously added to the fact that there is always inherent noise present in MRI images essentially due to the methodology and equipment used in the acquisition process. For these reasons it was necessary to apply image processing methods necessary to eliminate the noise inherent and improve the boundaries of the structures and image intensities of the tissue-tissue and tissue-air. It is common practice to assume the noise in magnitude MRI images is described by a Gaussian distribution with zero mean. The power of the noise is then often estimated from the standard deviation of the pixel signal intensity in an image region with no NMR signal.

An appropriate pre-processing of the images is crucial in order to obtain appropriate results of the search model. This preprocessing focuses essentially in the image quality improving, since the original images contain a relevant amount of noise that for segmentation purposes, disturbs the definition of the boundaries of the structures present in the images. Therefore the objective was to obtain a clean boundary, to improve their gradient force to be well defined and therefore to be correctly identified. A number of filtering techniques have been studied in the literature including anisotropic diffusion, wavelet filtering, Non-local Mean (NLM), and many others (Buades, Coll, & Morel, 2005a). Even though they may be very different in their approach formulation it must be emphasized that a wide class share the same basic principle, in which denoising is achieved by averaging at a local level. This implies a loss of finer details in the image.

3.4.1. Nonlocal Means Denoising

Nonlocal means algorithm was first described in (Buades, Coll, & Morel, 2005a). The NLM algorithm was chosen due to its excellent performance. It presents the main characteristics that any denoising algorithm should include and it does not compromise fine details therefore, does not alter the original image objects and conformations. It has been reported to present good results for Gaussian additive and multiplicative noise (Tristan-Vega, Garcia Perez, Aja-Fenandez, & Westin, 2012).

NLM is a nonlinear filter, based on a weighted average of the pixels inside a search window, whereas in the original formulation this includes the whole image, and hence explains the usage of the term *nonlocal*. It has however, the big downfall of implying heavy computational work, which is also the biggest and basically the only disadvantage comparing to other algorithms (Buades, Coll, & Morel, 2005b).

In this works was used a more robust and computationally lighter algorithm described by Tristan-Vega et al., that uses a Weighted Average (WA) of pixel inside a search window, whose weights are defined by a quantification of the similarity, through Mean Squared Differences (MSD), of minor patches that surround the pixel being compared with the pixel of interest.

Buades et al. (2005) introduced the work on full window areas instead of single pixels. The NLM algorithm compares the local area with patches all over the image to find reference values for the denoising.

The main objective of the work developed in (Tristan-Vega, Garcia Perez, Aja-Fenandez, & Westin, 2012) was to accelerate the computational cost taken by the MSD distances between patches.

The NL-means algorithm is defined by the simple formula:

$$NL\,\hat{u}(xi) = \sum_{x_j \in \Omega_i} w(x_i, x_j).u(x_j) \tag{17}$$

where Ω_i is the search window centered at pixel x_i, and $w(x_i, x_j)$ is the weight assigned to pixel x_j with respect to the pixel of interest x_i. In an initial formulation of this method the search window was the entire image.

The weighting filter is forwardly calculated through the similarity measure between the two patches centered respectively in the pixel of interest and the pixel compared, N_i and N_j, respectively:

$$w(x_i, x_j) = \frac{1}{Z_i} e^{-\frac{\frac{1}{N}\left\|u(N_i)-u(N_j)\right\|^2_{2,a}}{h^2}} \tag{18}$$

where Z_i is a normalizing constant so that $\sum w(x_i, x_j) = 1$ over all the pixel x_j of the patch N_i. and u(N_i) denotes a vector with all the intensities $u(x_j)$ of size N. The distance between two neighborhoods is computed by a Gaussian weighted Euclidean distance. Parameter h represents a statistical value, that regulates the decay of the weights according to the distance calculated.

Therefore, it needs to be proportional to the expected value distance between the patches, and therefore is related to the noise power present in the image σ^2 and a parameter $\beta \in [0.8, 1.2]$, by the following equation:

$$h^2 = \beta^2 . \sigma^2 \tag{19}$$

This parameter needs to be suitably estimated since it can over-smooth or under-denoise the image, if overestimated or underestimated. Therefore speed up of this method was ached through two approaches detailed bellow.

Voxel preselection

In this implementation the image is divided into overlapping blocks where all intensity vectors $u(N_i)$ are arranjed into one matrix, to which is applied Singular Value Decomposition (SVD) using (Mathworks, 2015d), based in the knowledge that the optimal representations of the patches of interest are those corresponding to the resulting largest Singular Values (SV) discarding dissimilar pixels. The acceleration comes from the elimination of the pixel from the WA weighting process. However this only provides some extent of computational unloading, which highly depends on the SNR present over the image, and therefore translating into unpredictable speedup.

Weighted Average

The problem still remains in the high computational cost related to the computations of the distance $d(x_i, x_j) = \frac{1}{N} \|u(N_i) - u(N_j)\|_2^2$.

The main idea is, instead of accurately calculating the distances, to estimate them, having the knowledge that said distance can be estimated from a small number of features that describe the local structure of the patches, whose computational costs is significantly lower.

The feature retrieval implies the conversion of the image inside the central patch of the pixel of interest N_i, as a local Taylor series expansion of type:

$$u(s_j, t_j) = c_0 + c_s s_j + c_t t_j + \frac{1}{2} c_{ss} s_j^2 + \frac{1}{2} c_{tt} t_j^2 + \frac{1}{2} c_{st} s_j t_j + \cdots \tag{20}$$

where c_s and c_t are the local gradient variances, c_{ss}, c_{tt} and c_{st} third order moments. s_j and t_j are the offsets between there feature of each pixel x_j with respect to x_i, that ultimately form a feature space. The problem can be formulated as a Least Squares (LS) problem:

$$\begin{bmatrix} 1 & s_1 & t_1 & \frac{1}{2}s_1^2 & \frac{1}{2}t_1^2 & s_1 t_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & s_N & \cdots & \frac{1}{2}t_N & s_N t_N \end{bmatrix} \begin{bmatrix} c_0 \\ c_s \\ c_t \\ c_{ss} \\ c_{tt} \\ c_{st} \end{bmatrix} = \begin{bmatrix} u(s_1, t_1) \\ \vdots \\ u(s_N, t_N) \end{bmatrix} \Leftrightarrow X.c \cong u \tag{21}$$

The feature vector c can be formulated as:

$$c = (X^T X)^{-1} X^T u \tag{22}$$

And therefore each interpolated patch surrounding each x_i pixel is obtained by:

$$\widetilde{u_i} = X.c_i \tag{23}$$

Having this mathematical formulation established, the problem was simplified by computing this LS fitting instead of the actual pixels, whose MSD distance can be classically formulated as:

$$\tilde{d}(x_i, x_j) = \frac{1}{N}(\widetilde{u_i} - \widetilde{u_j})^T(\widetilde{u_i} - \widetilde{u_j}) = \frac{1}{N}(c_i - c_j)^T X^T X(c_i - c_j) \tag{24}$$

which is simplified when the polynomials order is 1, and X^TX is a diagonal matrix of

Input. Noisy image v, filtering parameter h.

Output. Denoised image

1. Set parameter t × t: dimension of the feature patches: **rc**=[3,3];
2. Set parameter w × w: dimension of search zone: **rs**=[5,5] ;
3. Set parameter ps of the preselection thresholding: ps=2;
4. Compute the local features of the whole image, in every comparison patch of size txt:
    a. Gaussian filtering with size txt;
    b. Gradient patches computation with size txt;

for similar patches, and **for** each pixel i

5. Get the reference pixel i patch of features around it, of size w×w.
6. Get the comparison pixel j patch of features;
7. Calculate the distance:

$$d = (mi - mj)^2 + (gxi - gxj)^2 * f1 + (gyi - gyj)^2 * f2;\ \ \mathrm{d} = \frac{\mathrm{d}}{\mathrm{h}^2}$$

where $f1$ and $f2$ scale the feature offsets to the gradient.

8. Preselect with threshold **ps** value, as maximal distance, ignoring above distances with zero weighting;
9. Recover pixel intensity by:

$$u_{final}(i,j) = d(i,j) * u(i,j)$$

end

Figure 13- Nonlocal Means Algorithm outline.

type:

$$\tilde{d}(x_i, x_j) = (c_{0i} - c_{0j})^2 + (c_{si} - c_{sj})^2 \overline{s^2} + (c_{ti} - c_{tj})^2 \overline{t^2} \quad (25)$$

The algorithm outline is described Figure 13.

3.4.2. Forming a profile

To form the profile vector $\boldsymbol{f}$ at a landmark, we sample image intensities along a one-dimensional *whisker*. In this work it is denominated a whisker has a vector centered at a landmark point, which is orthogonal to a shape edge, therefore consisting of the intensities displayed along the normal direction of the shape contour it integrates (Figure 14).

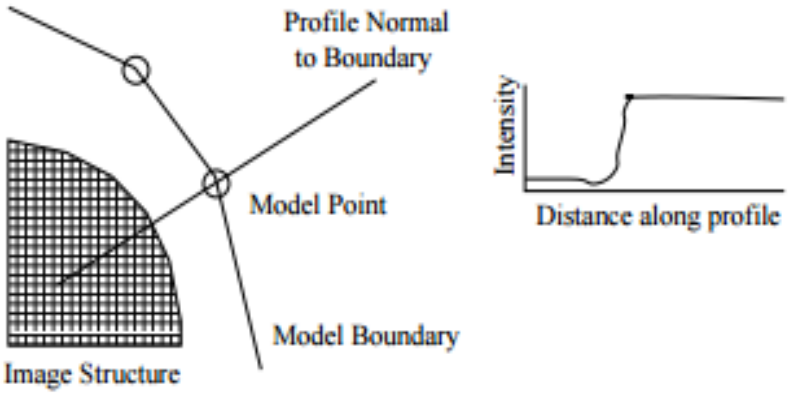

Figure 14 - The profile model takes upon the normals to the boundaries of the shape, at a given landmark (Cootes & Taylor., 2004).

To describe the appearance of an object this approach uses statistical modeling analysis, usually learned for each landmark separately. Training features are extracted by sampling image features from the training images at the landmark positions. In the problem presented, it is assumed that a training example is an intensity vector $\boldsymbol{f}i \in |R|^2$ and describes a one-dimensional profile. In the classical approach described by Cootes and Taylor the Mahalanobis distance to the learned mean appearance $\bar{\boldsymbol{f}}$ is used to quantify the fitness of features. Note that this distribution can be learned from intensity profiles $\boldsymbol{f}$ or gradient profiles $\boldsymbol{g}$. In this work a different approach was also taken up into consideration, that analyses the profiles variability by a PCA based method similar to the one used to analyze the shape variability.

During training, a model for each landmark is built by creating a mean profile $\bar{\boldsymbol{f}}$ and a covariance matrix S_f of all training profiles, retrieved from each image, at each landmark i. The assumption is that the profiles are approximately distributed as a multivariate Gaussian, and thus can be described by their mean and covariance matrix. This process is in its simpler theorization, the process of looking along profiles normal to the model boundary through each model point. Expecting the model boundary to correspond to an edge, a tissue-air boundary, represented by a grey-black intensity variance profile, respectively, upon which we can simply

locate the strongest edge (including orientation if known) along the profile. The position of this gives the new suggested location for the model point. However, model points are not always placed on the strongest edge in the locality, and furthermore for the purposes of the structure in study in this work, they may represent a weaker secondary edge or some other image structure. The best approach is to learn from the training set what to look for in the target image. Suppose for a given point we sample along a profile k pixels either side of the model point in the ith training image.

A normal of a surface, in this case defined by a 2D vector, is an object, in this case a vector, that is perpendicular to said surface. Considering the model landmark coordinates are known,

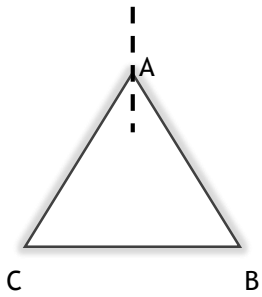

Figure 15- Example shape defined by vertexes A, B and C.

each boundary normal is calculated taking into account the forward and backward boundary vectors.The shape can be interpreted as a set of vector and the normal to each vertex are influenced by the two vector that form that vertex of the shape. The calculations below lead to the simple calculation of each normal.

In the following example (Figure 15), is considered a shape defined by three vertexes A,B and C, each defined by a pair of (x,y) coordinated.

The process of calculating the normal direction of said example shape, in for instance the Normal direction of vertex A (this direction is depicted in Figure 15), considering each of the shape lines as vectors, is through the following calculations:

Vector 1 = Vertex A - Vertex B
Vector 2 = Vertex B - Vertex C
Vector 3 = Vertex C - Vertex A

$$N.x_A = \frac{Vector1.\Delta y + Vector3.\Delta y}{Normalization\ factor},$$

$$N.y_A = -\frac{Vector1.\Delta x + Vector3.\Delta x}{Normalization\ factor}$$

$$where\ Normalization\ factor = \sqrt{(Vector1.\Delta y + Vector3.\Delta y)^2 + (Vector1.\Delta x + Vector3.\Delta x)^2}$$

The resulting normal is composed by its $N.x$ and $N.y$ components in the 2D space of the shape. This process was done for every shape model landmark of the training shapes. Having set the nomal unit vector that defined the normal direction, the study of the image in that

direction is allowed. This is done by defining the number of intensity profile pixels, that was set to k=8 pixels inward and outward from the landmark (in the normal direction) and generating a profile length of $2 * k + 1$ pixels. We have $2k + 1$ samples which can be put in a vector fi.

To reduce the effects of global intensity changes, varying image lighting and contrast, in some formulations of the profile model is appropriate to sample the derivative along the profile, rather than the absolute grey-level values. Heimann and Meinzer have evaluated three different Gaussian appearance models for liver segmentation in CT scans, studying in particular, the vectors $\boldsymbol{f}i$ containing either intensities, gradients or normalized gradients, whereas the results led to concluding that the best results are obtained with normalized gradient profiles in this application (Heimann, Wolf, & Meinzer, 2006).

The following algorithm describes the building of these profiles:

- Interpolate the positions of each profile point in the image, along the directions of $[-k, +k] * Ni + Pi$, where Pi is the landmark i coordinates.
- Replace each profile element by the intensity gradient. This is done by replacing the profile element at each position i with the difference between itself and the element at position $k - 1$.
- Divide each element of the resulting vector by the sum of the absolute values of all vector elements - Normalization.

Having the profiles defined, we will notate each gradient profile vector as $\boldsymbol{g}i \in |R|^2$ relative the ith landmark, the training phase of this models building consists in collecting each set of profiles belonging to each landmark. This information is used to build correlation matrices for each landmark, similarly as the building method of the shape model, but describing the variation of the gradient profiles. By creating a mean gradient profile $\overline{\boldsymbol{g}}$ and a covariance matrix $\boldsymbol{Sg}$ of all training profiles at that landmark. The assumption is that the profiles are approximately distributed as a multivariate Gaussian, and thus can be described by their mean and covariance matrix. If there are 64 landmarks then there will be 64 separate mean gradient profile $\overline{\boldsymbol{g_i}}$ and 64 covariance matrixes $\boldsymbol{Sg_i}$ that describe the profile variances, retaining 98% of cumulative variability.

3.5. Results and Discussion

In this chapter, all the experimental results obtained by the two models formulated by the approaches described in the methodology, are thoroughly explained, and their respective discussion is presented. The information retrieved by the statistical shape model results will be the first to be analyzed followed by the Profile model results obtained with and without the pre-processing stage highlighting the need of its implementation.

3.5.1. Landmark assignment

The generated landmark constellation is represented in the example subject depicted in Figure 16. The user-generated landmarks mapped to the image, comprising a total of 16 landmarks, positioned at key important anatomical limits that represent the tongue shape in this plane of view. In this constellation the landmarks were labelled into a numbering order whose correspondence to the anatomical structure is as follows:

- Two points in the lingual frenulum: anterior and posterior, corresponding to landmarks 1 and 2;
- One point in the tongues tip: corresponding to landmark 3;
- Six points along tongues body: corresponding to landmarks 4 to 9;
- One point in the tongues root: corresponding to landmark 10;
- Six points along the tongues inferior surface: corresponding to landmarks 11 to 16.

This outline correspondence of the initial landmark constellation will allow the result observation in its correlation to the anatomy and the image features that describe it and will be mentioned thoughout the following discussion.

This initial landmark outline was followed by a spline interpolation step, of factor 4, resulting in a total of 64 final landmarks. This step was necessary for the followed model contruction and fitting, not affecting the results of the ASM model obtained. The resulting flandmark connectivity outline is depicted in Figure 17.

Following this step the dataset was centered to the origin aligned through procrustes distance minimization methods, in order to validate the following statisticall analyis.

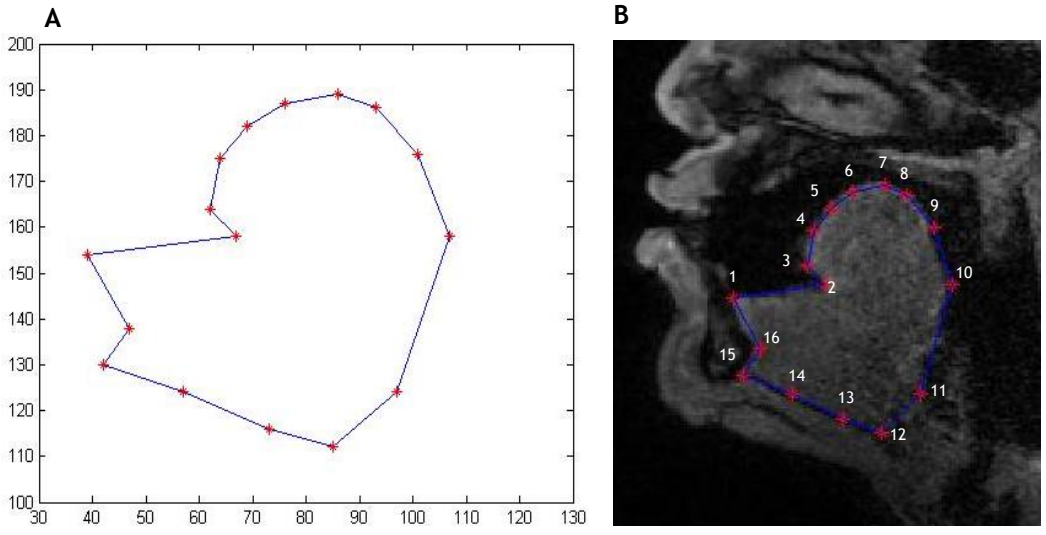

Figure 16 - Initial landmark map defined by hand representing the landmark connectivity (A), and in the image referential (B).

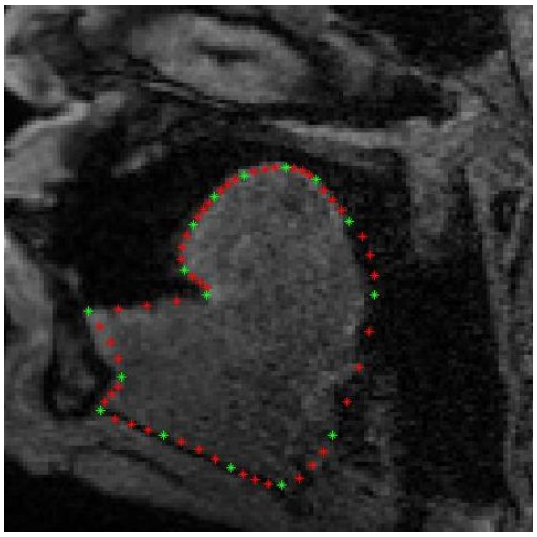

Figure 17 - Interpolated landmark map, with interpolation factor 4, depicting the original landmark constellation points in green and the interpolated points in red.

The initial shape dataset and the corresponding mean shape is depicted Figure 18-A whereas at the end of 3 iterations uppon which the minimal mean shapes euclidean difference threshold criteria is reached, and the shapes were aligned as depicted in Figure 18-B. The observation of the results leads to the comprehension of this step since only when this is achieved, the differences between the shapes can be evaluated.

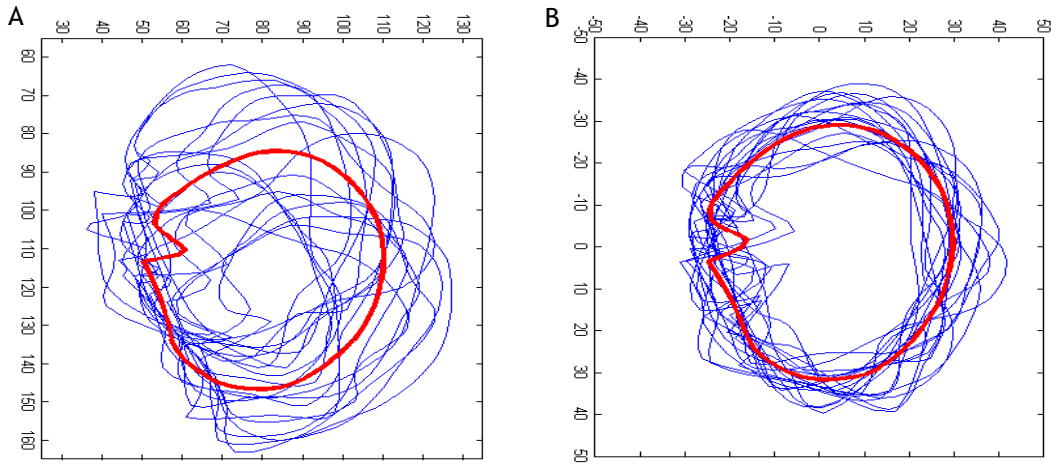

Figure 18- Raw shapes (A) from dataset (blue) and initial mean shape (red), and aligned, origin centered shapes through Generalized Procrustes Analysis (B), with final shape dataset (blue) and final mean shape (red). Images plotted in image referential.

3.5.2. Statistical Shape model

Eight modes of variation were identified with mode 1 accounting for half the total variance and mode 1 and 2 accounting for approximately 71% of total variance. This study highlights the

potential of active shape modeling to advance understanding of factors underlying morphologic and pitch-related functional variations affecting vocal structures.

The shape parameter vector $\boldsymbol{b}$ is (as usual with active shape models) allowed to vary in the range of ± 3 standard deviations, therefore in the range of $-3\sqrt{\lambda} \leq b \leq 3\sqrt{\lambda}$.

We can observe that the first four modes of the shape model built could explain 92% of all shape variance of the tongue. The first six modes explain 95% of all shape variance and with eleven modes of variation it is possible to explain 99% of all shape variance of the tongue.

Holding approximately 99% of the total variance of shape data, the final model produced, presents a total of eleven variation modes. From a first overall analysis of the resulting modes of variations of the model it is possible to conclude that, as expected, the first eigenvalues retain a higher extent information associated to the data, represented by the cumulative retained variation, which therefore is expected to cause a bigger movement between landmarks

Table 2 - First seventeen modes of variation of the model obtained and their retained percentages, describing 99.9% of shape variation.

Mode	Retained (%)	Cumulative Retained (%)
$\boldsymbol{\lambda_1}$	75.711	75.711
$\boldsymbol{\lambda_2}$	8.789	84.500
$\boldsymbol{\lambda_3}$	5.028	89.529
$\boldsymbol{\lambda_4}$	2.631	92.160
$\boldsymbol{\lambda_5}$	1.956	94.117
$\boldsymbol{\lambda_6}$	1.738	95.856
$\boldsymbol{\lambda_7}$	1.117	96.973
$\boldsymbol{\lambda_8}$	0.904	97.878
$\boldsymbol{\lambda_9}$	0.735	98.613
$\boldsymbol{\lambda_{10}}$	0.379	98.992
$\boldsymbol{\lambda_{11}}$	0.350	99.343
$\boldsymbol{\lambda_{12}}$	0.223	99.566
$\boldsymbol{\lambda_{13}}$	0.175	99.742
$\boldsymbol{\lambda_{14}}$	0.106	99.848
$\boldsymbol{\lambda_{15}}$	0.057	99.905
$\boldsymbol{\lambda_{16}}$	0.054	99.960
$\boldsymbol{\lambda_{17}}$	0.024	99.984

positions. Accordingly, the lower significative modes of variation cause a more local variation of the data. This is also observable, by the analysis of the decay of the total variance as a function of the associated eigenvalue, depicted in Figure 19. The first three eigenvalues describe approximately 89% of the total variance whereas the remaining 11% were distributed

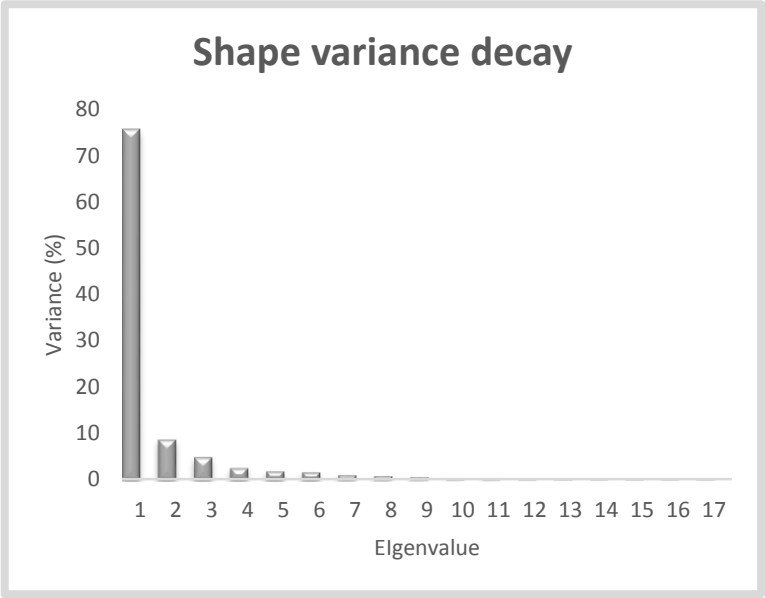

Figure 19 - Shape variance decay as the number of eigenvalues increases.

by the following 15 eigenvalues. From these 15 eigenvalues, the cut-off to the formulated model was set to 99%, i.e., in the eight eigenvalue, corresponding to the eleventh eigenvalue of the total model eigen-formulation, since the remaining eigenvalues represented residual variation modes. This is the advantage represented by statistical models, where a complex shape and almost all of the variations allowed is captured, being mathematically formulated by the combination of finite number of independent variables. The data would otherwise, only be represented by more complex methods and not hold the variable reduction achieved by PCA. This variable reduction refers to the reduction of the model into a finite number of eigenvalues and corresponding eigenvector that represent the shape variables, capturing a quantifiable total variance retained.

As established before, any instance in the training set can be closer to the original data by the correct definition of Equation (3).

The evaluation of the model in the sense of the variation captured can be assessed by the resulting model shape variability upon eigenvalue variation and was established by varying each eigenvalue individually along a suitable interval, that was defined, as reported in the literature, as $[-3\sqrt{\lambda i}, +3\sqrt{\lambda i}]$. The first seven variation mode are depicted in Figure 20 (A-F).

By the observation of the eigenvalues variation individually, it was established that the variance to be kept to allow an adequate description of the shapes were the first eight eigenvalues and corresponding eigenvectors that describe the eight modes of variation of the model. The remaining eigenvalues, corresponded to residual shape variations.

The first mode is associated to movements of the whole tongues body, associated with the rotation of its shape, and namely of the tongues frenulum and tip upwards and downwards present in the different training shapes, withholding the greatest shape variations in these directions that influences the whole constellation of landmark.

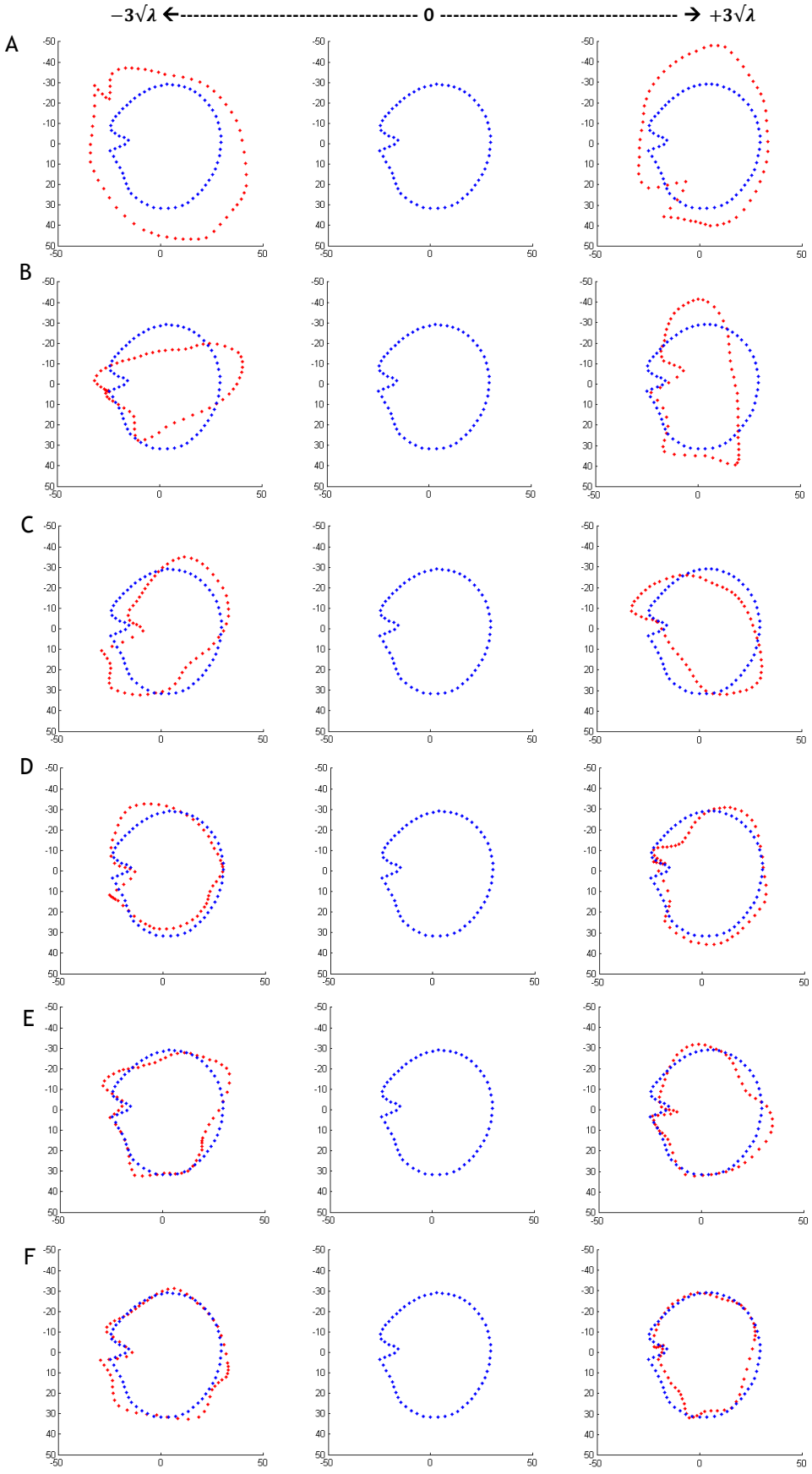

Figure 20 - Representation of the first six modes of variation plotted the model mean shape (blue) and the mean shape deformed by the model eigenvalues (red).

The second mode of variation, is associated with movements of the whole tongues body along the vertical and horizontal axis combined, specifically in the spreading and narrowing combined with upward and downward movements of the tongues upper boundary, present in the training set.

In the third mode of variation, it is possible to observe the alterations to the movements of the upper section of the tongue, that is accompanied by a compensative opposite movement of the lower section of the surfaces of the tongue. This shape conformation comes for instance, from the production of the close back rounded vowel ***[u]*** from the Portuguese word /tu/ (you), that implies the upper posterior movement of the tongues dorsum and lower anterior movement of the tongues base.

In the fourth mode of variation, are captured the movements of the upper walls curvature, namely the rise and lowering of the tongues tip.

The fifth mode of variation describes changes in the horizontal width of specifically the upper portion of the tongues that is associated with the horizontal spreading in the production of the open front unrounded vowel [*a*] in Portuguese words like *casa* (home), and on the other hand moves presents narrowing of the upper posterior wall for the pronunciation of the close front unrounded vowel [*i*] in Portuguese words such as *riso* (laghter).

Finally, the sixth mode of variation, captured in the horizontal width of specifically the lower portion of the tongue, that complements the movements described in the third mode of variation, in minor pose details.

The only similar study found in the literature was made by Vasconcelos et al. (Vasconcelos, Ventura, Tavares, & Freitas, 2009), where a statistical shape model similar was developed using an image dataset representative of oral vowels, consisting of a set of nine images from one subject, in which they were able to thoroughly characterize the speech production of said sounds, with a 7 modes of variation point distribution model, with 99% variability retention.

Direct comparison with the model developed in this work is therefore possible, whereas a more variability was sought to be analyzed in the present work, with the usage of a dataset including a set of 19 images and the presence of female and male subjects, that confer anatomic variability, inherent to gender anatomy differences, that most relevantly are related to dimension of head and neck and proportional differences of the structures present, fat distribution in head and chin. Therefore, reducing the model into an eleven variation modes model is relatively positive in comparison to the study referred.

3.5.3. Image processing

In the present section, are presented the pre-processing results from the Non-Local means denoising algorithm, detailed in Section 3.3.1. The evaluation of the advantages this algorithm presents will be outlined.

The choice of this algorithm, was based on the assumption it presents the ability to eliminating Gaussian noise while preserving adequately the edges of the structures present,

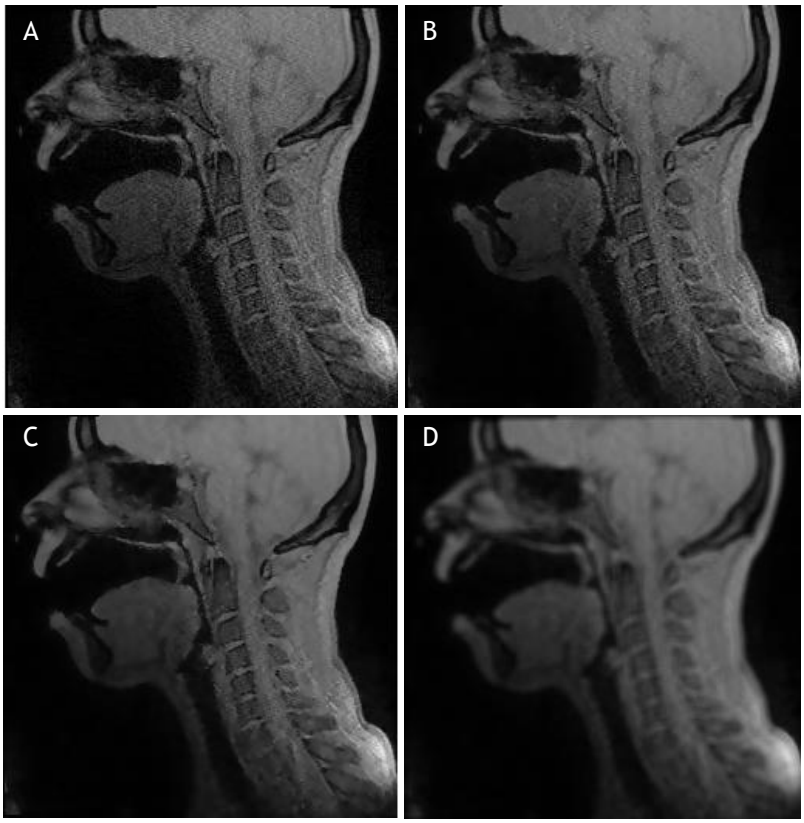

Figure 21 - Non-local means denoising results, respective the original image (A), and denoised images with h parameter set to 0.05 (B), 0.1 (C) and 0.5 (D).

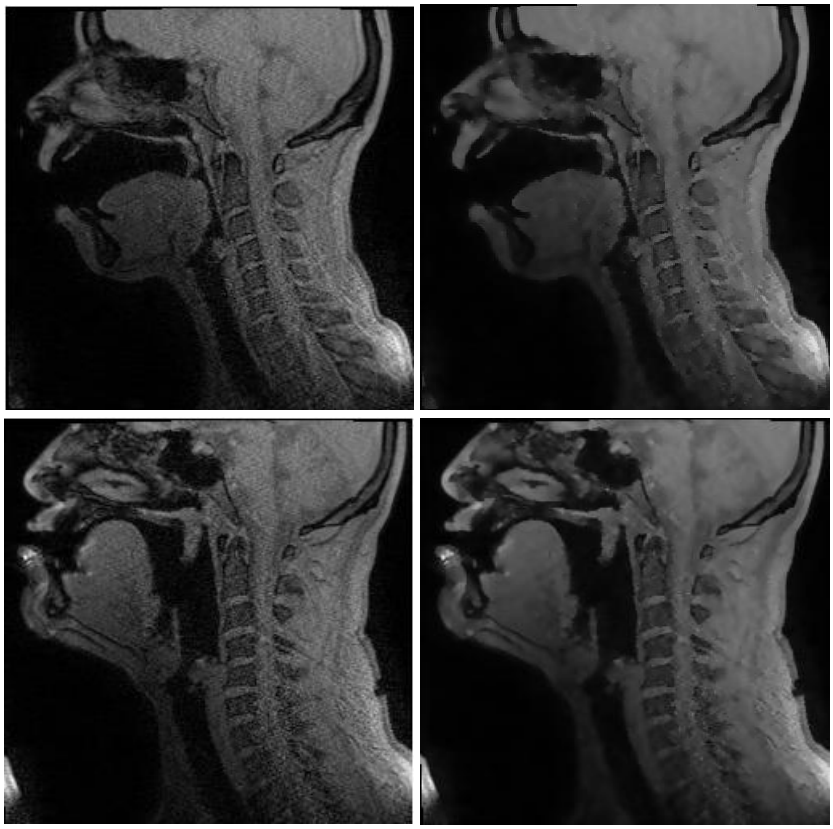

Figure 22- Example image of female (top row) and male (down row) subject before and after non-local means denoising, with h parameter set to 0.1.

which constitutes to be for this work, and to not eliminate image information and intensity homogeneity already present in the images.

The Non-local Means (NLM) method averages neighboring parts of the central pixel but the averaging weights depend on the similarities between a small patch around the pixel and the neighboring patches within a search window. This search window was defined to be a 4x4 pixel window. The search window is a parameter that will influence the production of more homogeneous regions with similar intensities, whereas the larger searching window possibly allows more similar patches to be found and thus yield a filter that better preserves the features. The larger window will however produce loss of textures, and so an over-blurring was not the objective of this step of the work developed. The main idea was to confer homogeneity to background regions, specifically the vocal tract, and the tongues intensities. The search window was defined by balancing over-blurring and wanted homogeneity of image regions.

The noise estimation approach was idealized for a Gaussian type of noise present MRI images, and significant image information was lost.

Therefore, the results presented allow the evaluation of the resulting image as well as their Peak Signal to Noise Ratio (PSNR). PSNR is a calculative measure of the image quality of the images produced, and therefore allows comparison of the images produced relatively to the original ones, in terms of noise suppression.

As previously stated, the success of the denoising result of this algorithm highly depends on the estimation of parameter h. For an initial evaluation of the algorithms power of noise elimination, the analysis of the influence of the parameter h in the images was studied and is presented in Figure 21, depicting a range of denoising runs with different h parameter values. It is possible to observe the blurring effects on the resulting image (Figure 21-D) when theres an overestimation of this parameter, and an underestimation leads to an insuffiecient denoising effect (Figure 21-B).

The images were analyzed with h parameter fixed to 0.1, for the entire dataset. These results are presented in Figure 22 for one male and one female example image. Furthermore, by means of image subtraction of the original and filtered image, whose result is depicted in Figure 23, it was intended to highlight that no edge details or important data from the images is lost by the application of this algorithm, and depicts the speckled noise removed, that implied intensity alterations that ranged from 0 to 0.06 in a scale of [0, 1]. Moreover, by image subtraction, it was observed that only in high overestimations of parameter h this effect is seen. The computational load of the algorithm, was not sensed in this analysis since the image number of the dataset did not make this factor relevantly disadvantagous.

The total computational cost of the pre-processing for the model contruction was of 28.34 seconds with a mean processing of each image of 1.08 seconds. Signal to noise ratio values (**Table 3**) do not allow a very explicit comparison parameter between each denoised image, since the images obviously lose quality with the comparison to the original one with increase of the parameter value. Mean squared differences however, do translate the mean improvement in the images that indicates the resulting visual effect on the images. However,

this does not reflect the visual effect the objective of this method which was to produce cleaner images, homogeneous regions, which however lowers the peak intensities, and therefore PSNR values, with the increase of the parameter value. Even though the loss of texture is verified in these results, the results the method was chosen to be adequate for the purposes at stake, whereas the edges now present a direct dependence of the intensity differences between the tongues intensities and the background intensities, not being dependent of noise interferences. This was idealized for the correct functionalization of the search of local boundaries methods that will be detailed in the next Chapter.

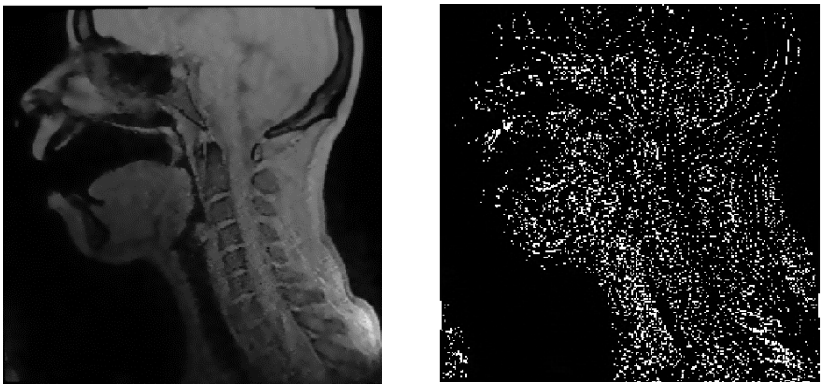

Figure 23 - NLM denoised example image and binarized image subtraction image result.

Therefore the established h parameter value for all images was 0.1, which revealed as suitable in terms of efficient denoising of the entire dataset. The results presented show that the NLM denoising method is a good approach for removing the noise in the image dataset and confer homogeneity to the features. Furthermore, the resulting effects of NLM denoising in the Profile Model, respectively the boundaries intensity distributions, will be presented in the following section.

Table 3 – Mean peak signal to noise ratio (PSNR) and mean square error (MSE) of denoised images with NLM algorithm using different h parameters.

h	PSNR (dB)	MSE
0.05	86.438	6.774e-04
0.1	80.817	5.305e-04
0.5	75.446	0.003

3.5.4. Profile Model

The model profiling was done for each of the 64 landmarks, producing a total of 64 covariance matrixes and 64 mean profiles, characterizing the intensities variability present in each landmark. The normal to each landmark was profiled obtaining a 17 pixels long profile or whiskers for each of said landmarks, centralized in the landmark pixel. The profiling example

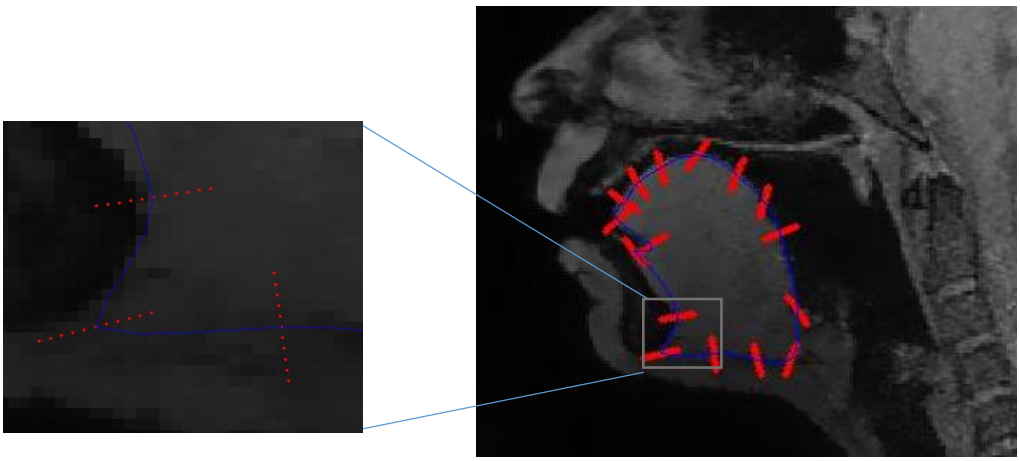

Figure 24- A one-dimensional profile of each of the 16 initial hand-labeled landmarks of an example image of the training dataset. The blue line is the shape boundary. The red line are the whiskers, orthogonal to the boundary.

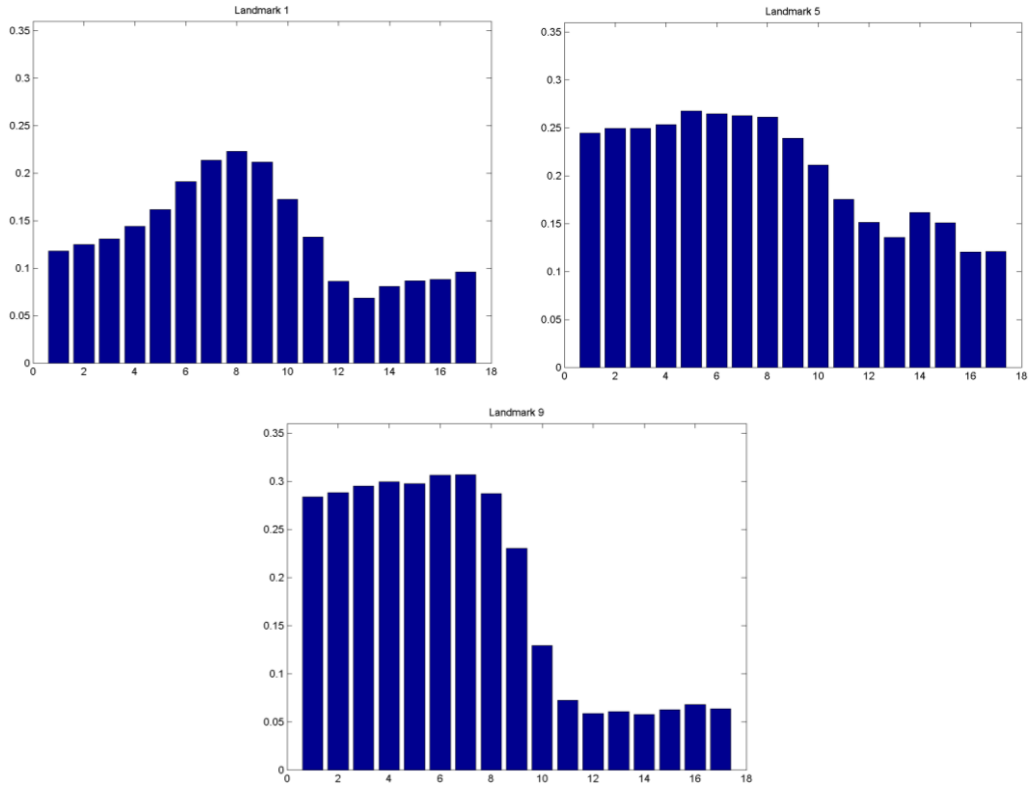

Figure 25 - Mean intensity profiles of landmarks 1 through 3 in the initial landmark constellation, corresponding to each labeled number in interpolation constellation, correspond to the tongues frenulum (anterior-posterior ends) and the tongues tip.

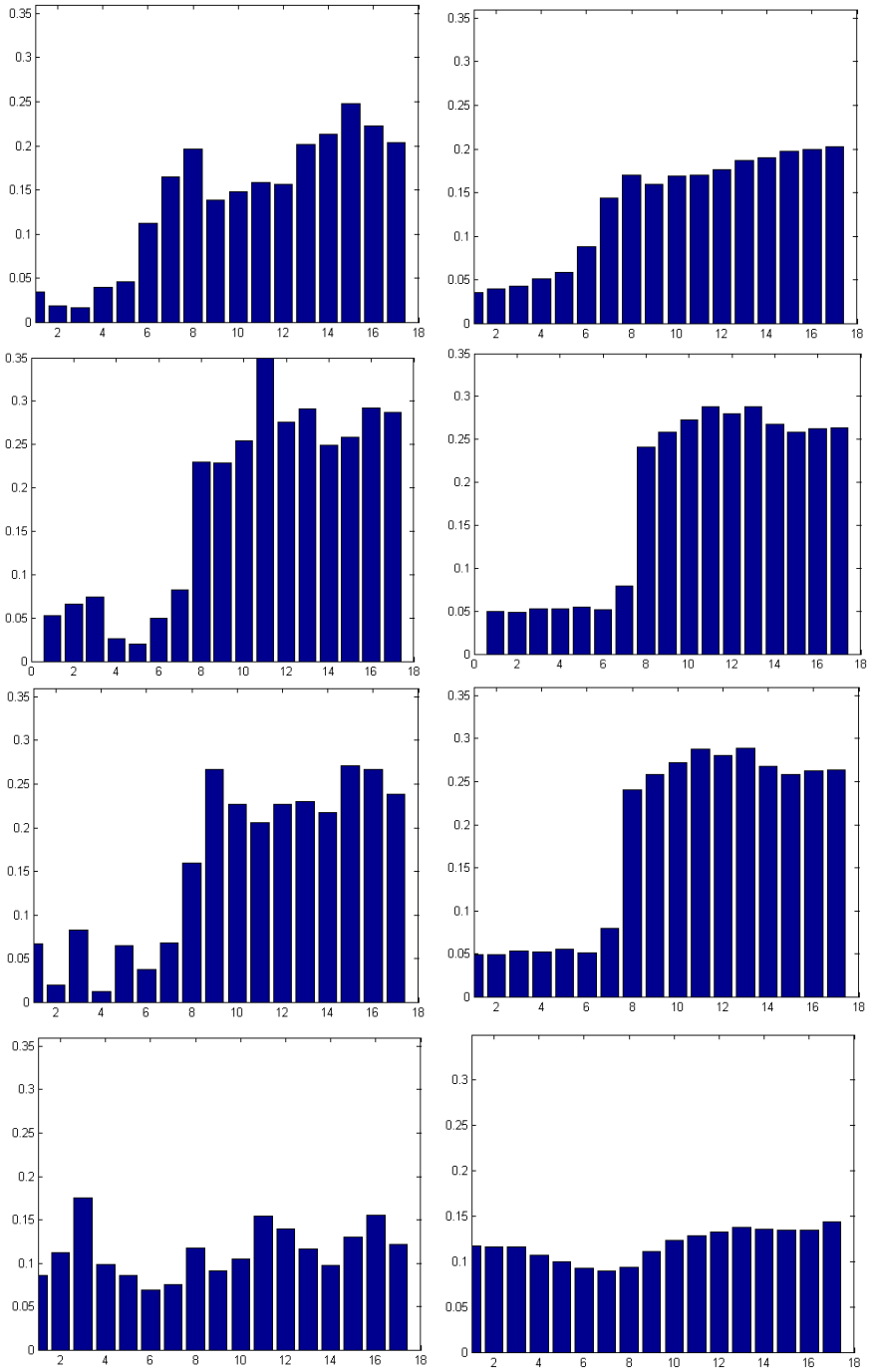

Figure 26- Unprocessed (left column) and processed (right column) image profiles examples, of each of the anterior, top, posterior and lower bounds (from top to bottom rows, respectively).

on an example image is depicted Figure 24. The landmark is at the intersection of the shape boundary and the whisker. The inset shows the pixel positions along the whisker. In practice, the ASM uses the normalized gradient of the image intensity.

The 64 mean profiles generated can be critically analysed, and allow the characterization of the intensities along the surface that delineates the tongue and local neighbouring regions.

Firstly, in order to analyse the results to the boundary intensities, in the generated profiles, an analysis of the original boudary profiles was also carried out to be compared against the

ones retrieved after non-local means denoising. Four profile examples of the each os the walls of the tongue retrieved from the exact same location in noisy and denoised images are presented in Figure 26. The main objective of the pre-processing step of the images was to achieve homogeneity of the different regions that constitute the image, namely the tongue and background regions. This is done so that when profiling is performed, the boundary intensities or intensity gradients are sufficiently discriminant relatively to the pixel position to which it corresponds. What was achieved was a clear segmentation of the intensities of the profiles centered in the boundary pixel, the eighth pixel in the profiles shown. Relatively to the lower boundary it is possible to observe that the intensities in the section of the object have a lower range of intensity variability, and are influenced by the tongue and the immediate tissue that is bellow it. However, the homogeneity achieved in each if the sides of the profile relatively to the central pixel, confer a better discriminant intensity of the boundary location.

Regarding the properties of the profiles generated regarding their correlation with the tongues regions, the analysis is made based on the results presented in Figure 25 representing the tongues frenulum (anterior and posterior ends) and the tongues tip, in Figure 27 the following tongues dorsum of the upper posterior boundaries, in Figure 28 related to the tongues root, and finally in Figure 29 related to the final 6 landmarks corresponding to the lower anterior boundaries.

For the purpose of maintaining clarity in the results presented, specifically in the following intensity profiles analysis, said profiles were analyzed in the main hand-labeled 16 landmarks, as representatives of each profiling region of the shape contour, whose correspondent landmark number, in the interpolation generated landmark constellation, is presented in their labeling title. The profiles are used here as descriptors of the intensity profiles along the contours, and intuitively since these contours are supposed to be boudaries of the segmented structure their gradient profiles theoretically represent the highest magnitude in the central pixel or in their immediate vicinity pixels, with decreasing magnitude along each side of the profile. This is also basis for the simple boundary detection based on gradient magnitude, which when describing isolated objects in a distinct background works perfectly to just study the pixel intensity variability in the normal direction detecting their position. The analysis of the produced profiles shows that some do not represent a very discriminant direction of the boudary properties in that region, since their normal direction is in some cases extending to the insertion and/or intersection of the tongue with the neighbouring structures. This occurred almost sistematically in the profiling of, for instance landmark 15 in the initial constellation (landmark 57 in Figure 29). This profiling information reveals the nature of the intensity environment of the lower boudary of the tongue where it clearly has a lower discriminant intensity profiles, which intuitively is expected to generate lower gradient magnitudes. This is very clearly visible when analysing these profiles in Figure 29, in contrast with the ones in Figure 27, correspondent to the upper boundary of the tongue and the vocal tract, whereas the range of the first set of profiles has maximum magnitude of 0.2 and the second maximum magnitude of 0.3. The vocal tract is an air-filled cavity and therefore is represented by lower intensities that the tongue,

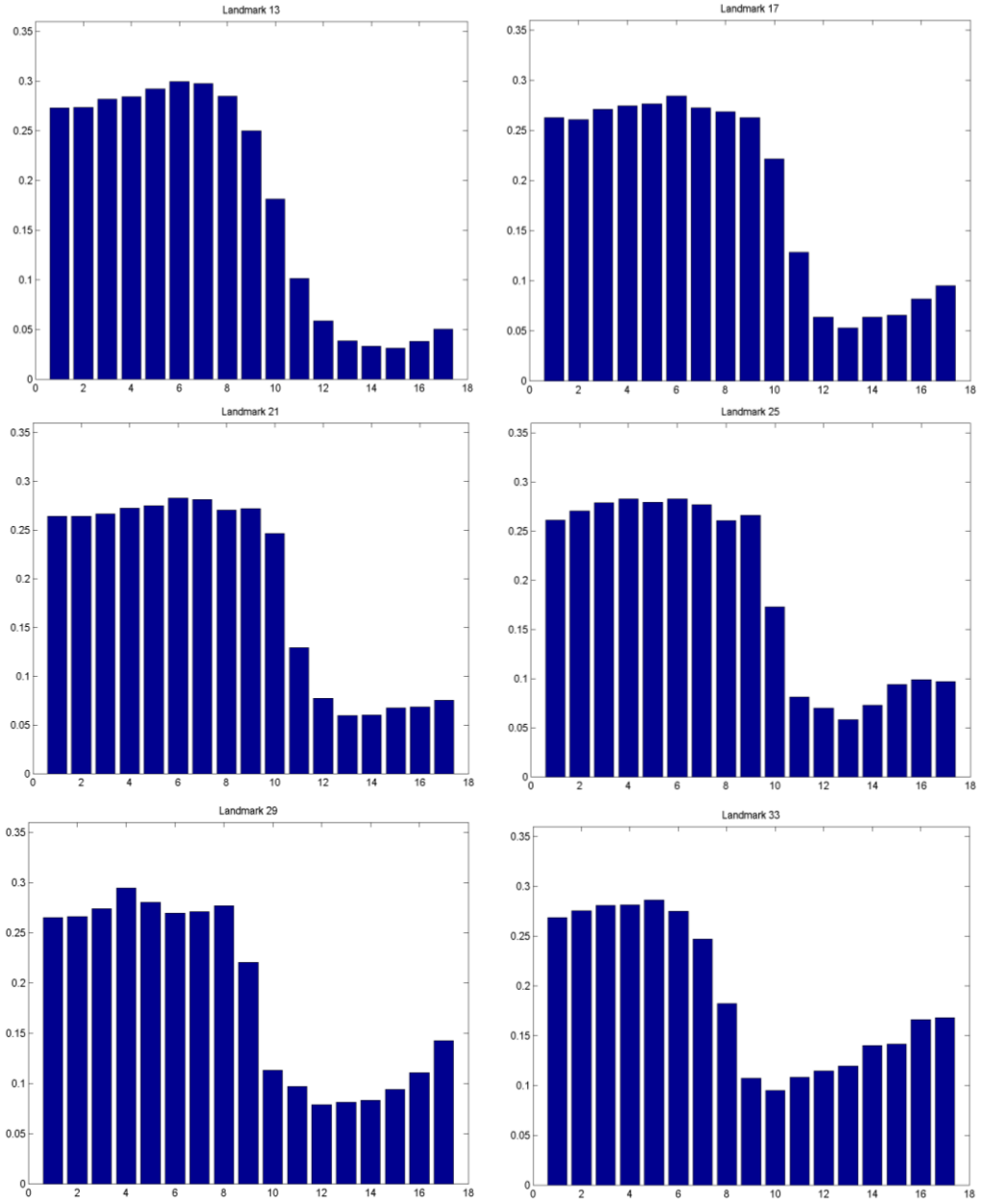

Figure 27 - Mean intensity profiles of landmarks 4 through 9 in the initial landmark constellation, corresponding to each labeled number in interpolation constellation, correspond to the tongues dorsum, or upper-posterior boundary.

and therefore the gradient along this boundary profile is represented by higher magnitudes. It should also be noted that the homogeneity presented in the profiles in Figure 29, of the following profiling values into low magnitudes, on profile pixels from 8 (central pixel) to 17, is related to the vocal tract low intensities, related to the cavity, which is observably more homogeneous that the tissue of the tongue, represented in the profiling pixels 1 to 7, where theres oscilations of intensities. This is also visible in the posterior end of the *frenulum* and tongues tip profiles. The gradient profiles obtained are only shown in Figure 30 with representative examples of each boudary walls of the tongue. Theoretically, the idel gradient profiling is a representative of the ideal magnitude descriptor of the boudary, consisting of

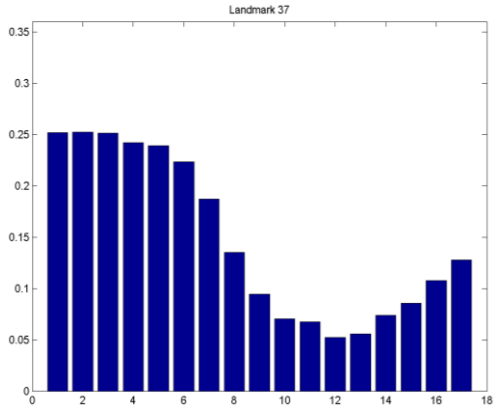

Figure 28 - Mean intensity profile of landmark 10, correspondent to the tongues root.

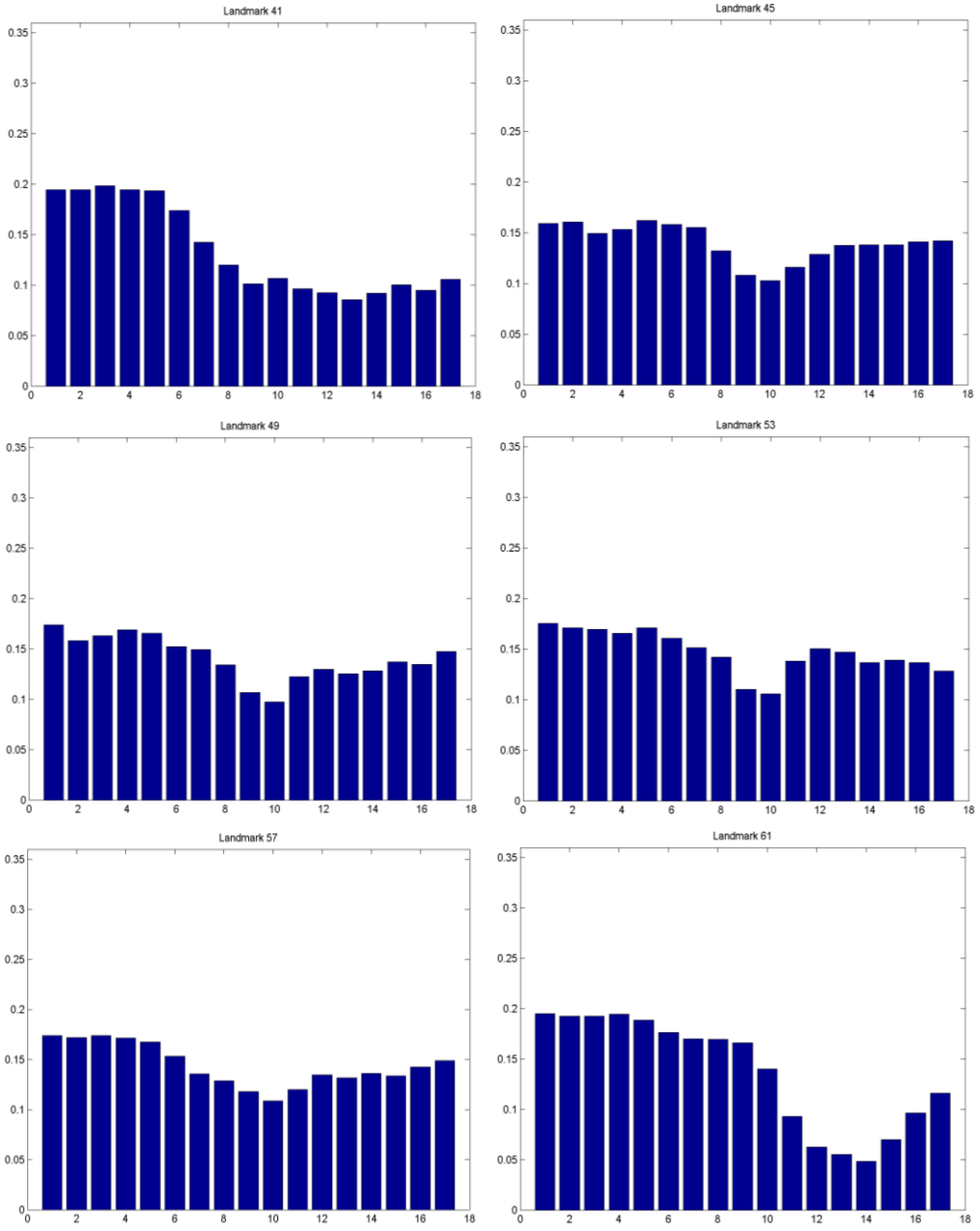

Figure 29 - Mean profiles of landmarks 11 through 16 in the initial landmark constellation, corresponding to each labeled number in interpolation constellation, correspond to the tongues dorsum, or lower and sub-frenulum-anterior boundary.

central pixel maximum magnitude with gradual decrease along the outward directions along the normal. However, this is not the case, whereas a lot of minor intensity inconsitensies produce inconsistent graduent whereas the maximum scoring is not in the central pixel or its vicinity, or is not discriminant considering the other gradient values in the profile. Therefore, analyzing the gradient profiles and intensity profiles presented, it is possible to observe the higher discriminant power the original boudary profiles have representing a better method for the search in the building of the active shape model discussed in the next Chapter.

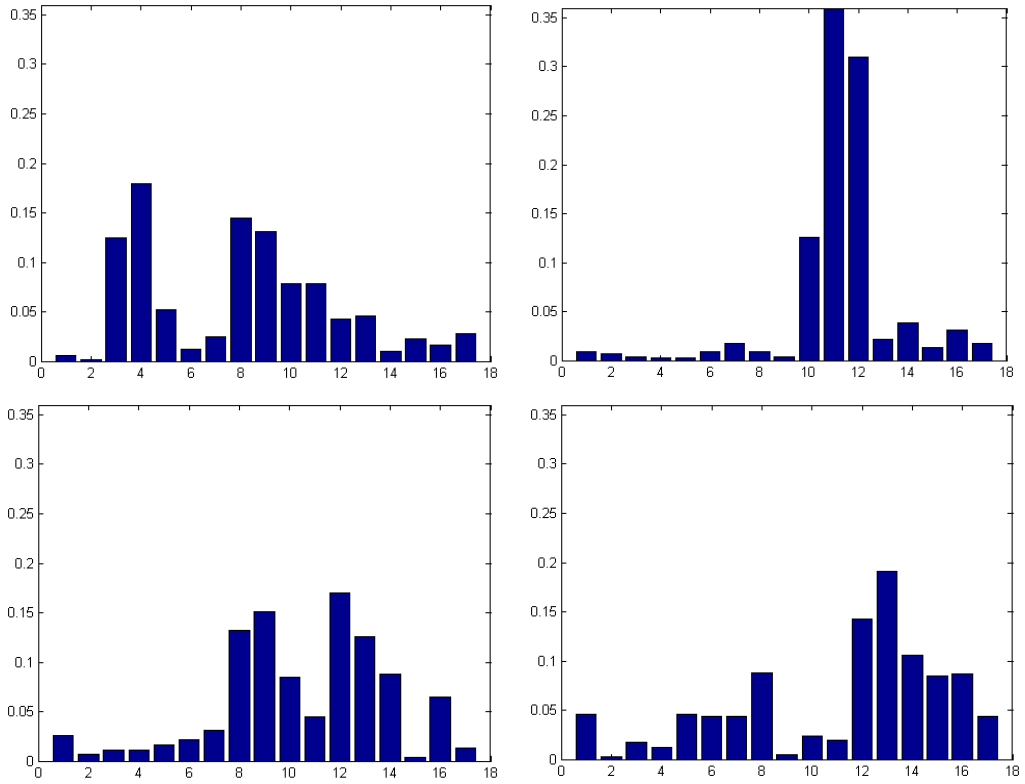

Figure 30 - Gradient examples of each of the anterior, upper, posterior and lower walls of the tongue.

3.6. Conclusions

In this chapter were described the statistical shape model and the profile model developed with an image dataset comprising a wide variability of tongues shape caused by different speech sounds production. The potential of this study is therefore widened in terms of range of future applications that go from speech imaging analysis to pathological anatomy analysis. The primary goal consisted on the analysis of the tongues shape during the articulation of European Portuguese sounds, and consists also as a preliminary study for the further application in the building of a Statistical Shape Model for semi-automatic segmentation of this organ. From the experimental results obtained, it is possible to conclude that the statistical deformable model built is capable of efficiently characterizing the behavior of the tongues shape modeled from

the MR images studied. Also, the modes of variation of the model built provide a partitioned explanation of the actual movements involved in the EP speech sounds considered. The model was also built in order to capture the optimal conditions to be applied in a following model application. The landmarks were augmented by spline interpolation methods in order to allow the appliance in the following search model step.

The analysis of the resulting alignment results, allowed the study of the training shapes in a common referential, eliminating pose, or rotation variations that can be inherent to the original images, and therefore allowing the partition of the shapes into individual sets of movements described by each eigenvalue.

The analysis of the model variables, allowed the assessment of an appropriate evaluation of the results, where the model was reduced to a finite number of variables that describe such a complex shape. Moreover, the resulting shape variations inherent to each eigenvalue are in accordance to the premise that eigenvectors are mutually orthogonal, a characteristic derived by the PCA properties. This premise translates into having subsequent eigenvalues describing shape variations in orthogonal directions. This is visible in the data retrieved, specifically in the variations described by the second and third eigenvalues, corresponding to vertical and horizontal specific movements, respectively, or by the fourth and fifth eigenvalues with minor vertical and horizontal movements of the tongues tip and superior section boundaries, respectively.

The model was trained with a number of different shapes that captured a sufficient extent of the variability this organ can acquire, and therefore offers all the conditions necessary so that it can easily be used to reconstruct the shape of the tongue in the articulation of speech sound, in new subjects. Furthermore, it is possible to conclude that the model built in this work allows a more clear understanding of the dynamic nature of the tongue in speech production events involved during sustained articulations. It also allowed the understanding of the boundary properties that delimit the organ at study and the regional diversity of their properties, relatively to the image features that describe them.

Chapter 4

Active shape modeling and segmentation of the tongue

This chapter describes the Active Shape Model developed that result by the actual combination of the shape model and profiled intensities model information retrieved in the sub-models described in the previous Chapter. The produced information described by the shape model allows the production of plausible shapes of the structure to be segmented within a certain range of variability imposed by its shape parameters constraints. This characteristic reveals itself as the most important advantage in the segmentation process, where the model is used to constrain the set of feasible shapes to those which are statistically plausible with respect to the patterns extracted from the training shapes. Similarly, the intensity profiles present in each landmark that characterize the training images in each of their locations is statistically analyzed to select the positions within new image profiles that better fit the image characteristics locally, at each landmark. The basic principle of the model consists in the building of an iterative search method, given a method for predicting the parameter correction needed to be made possible to achieve a better fit. The steps regarding the search of new shapes includes the initialization of the model, the search model variables, image feature search based on the appearance data retrieved from training and finally model fitting to the new landmark constellation produced based on the shape model produced from training. An algorithm scheme outline is presented in Figure 31. With the models already built up, it is now possible to find the desired shape in unknown images. This phase is generally known as the search algorithm. The process starts with positioning the model to an initial location in the test image. In the image the Profile model allows the estimation of the best movement locations for each model point, and from the mew shape points produced the model moves, rotates,

resizes and deforms until it finds the shape it was designed for, that is produced by the Statistical shape model. This is based in an iterative adaptation to the image that unites various factors to generate plausible segmentation results, and includes adequate methodologies that algorithm from collapsing, into unacceptable results. In this chapter the methods and specific algorithm steps taken to the segmentation of the tongue using the global model framework will be presented and the segmentation results will be assessed by comparison with true boundary points hand-labeled. Model initialization appears to be a key aspect in this study, and the main objective of the methods developed in this section were to generate the best possible methodology, being this step the key to the development of a semi-automatic segmentation algorithm. The performance the model developed will be accessed in the last section and therefore the adequacy of this methodology for purposes of analysis of tongue shape in MRI images, for speech studies.

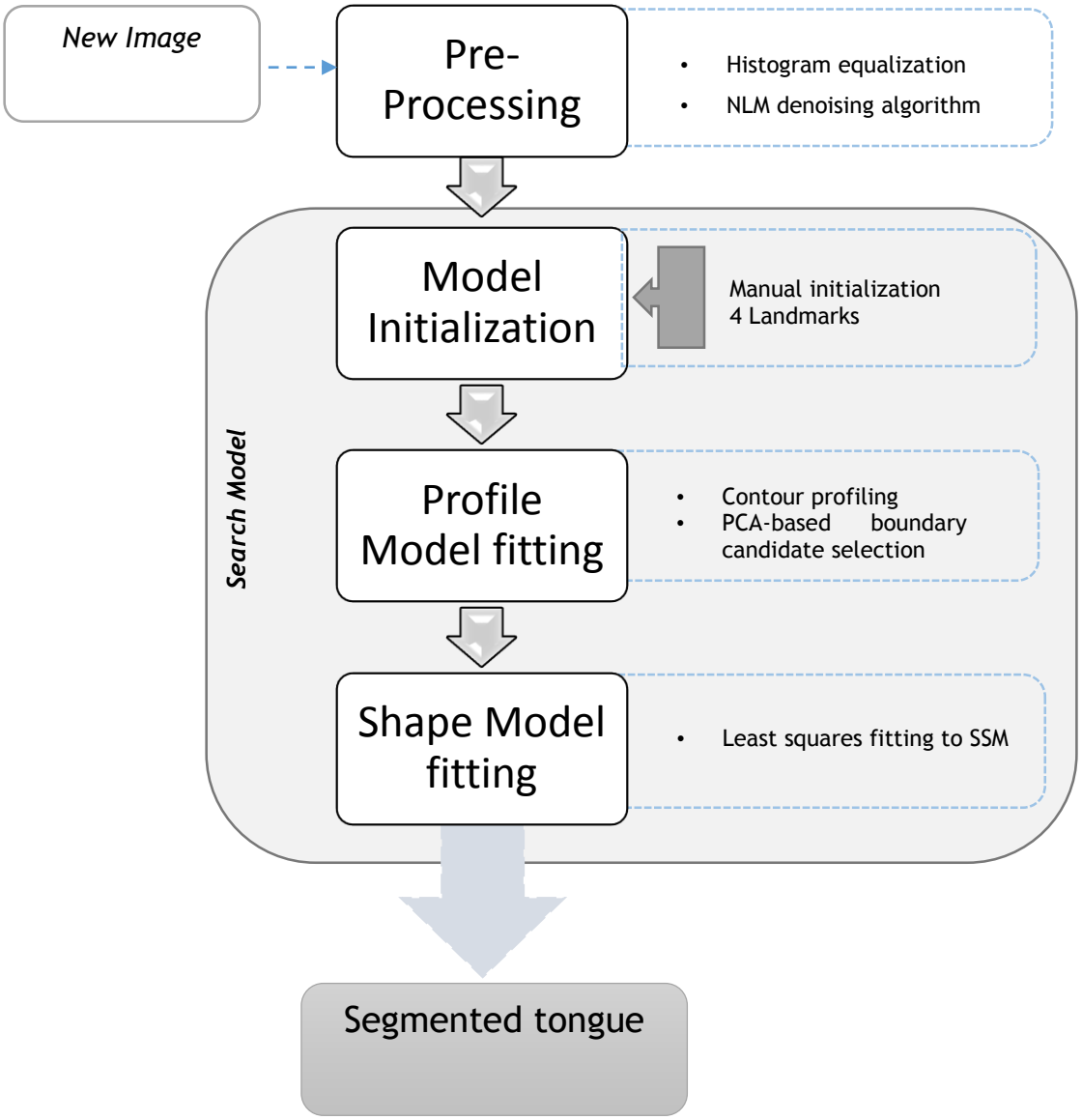

Figure 31 - Schematic model representing the final workflow of the Active Shape Model implemented, comprising the methods detailed in Chapter 3 and the present Chapter.

4.1. Search algorithm

The ASM is an iterative algorithm for image segmentation, in which the two different kinds of models were employed: A shape model, which encodes prior knowledge about the objects shape, and a profile model, which describes the appearance of the objects boundary in images. The latter model is necessary to adapt the model points to the image. The algorithm outline is described in Figure 32, which formulates the Active Shape Model in a generic way. For each step, there exist various alternative approaches, which are discussed in the sections below. The algorithm starts with placing an estimate of the targets shape — usually the model mean shape — into the image (detailed in Section 4.2). In each iteration of the main loop, the appearance model is used to perform a local search in the vicinity of each landmark in order to search for optimal image features. The shape is then deformed by displacing the landmarks to the detected image feature. Since this deformed shape does not necessarily correspond to a plausible object instance, the deformed shape is constrained with the Statistical Shape Model.

Input Place a valid instance of the shape model into the image (mean shape)

1. Generate the start location of the shape

repeat

2. Suggest a new shape by profile matching around each shape point

3. Adjust the suggested shape to conform to the Shape Model

until convergence (i.e. until no further improvements in fit are possible)

Output shape giving the (x,y) coordinates of the tongues landmarks

Figure 32 - Search model algorithm outline.

4.1.1. Multi-resolution search

Regardless of which local search strategy is used, the capture range of the shape model is restricted by the size of the local neighborhood. A popular way for increasing the capture range while retaining robustness is to use a hierarchical adaption scheme with a multiresolution image pyramid (Cootes & Taylor, 1994). This strategy has been widely used in many ASM implementations (Heimann & Meinzer, 2009). Before the search begins, an image pyramid is built, and repeat the ASM search at each level, from coarse to fine resolution. Each image in the image pyramid is a down-scaled version of the image above it. The start shape for the first search (made in the coarsest image) is the shape initially generated the user interaction initialization steps algorithm. The start shape at subsequent levels is the best shape found by the search at the level below. Usage of multi-resolution approach is more efficient, more

robust, and converges correctly to the correct shape from further away than searching at a single resolution. Two pyramid levels were used in this step.

4.1.2. Number of landmarks

A straightforward way to improve the fit is to increase the number of landmarks in the model. This is based on the assumption that the error decreases as the number of landmarks increases according to previous studies. Although the computational time increases, especially in search time because search time is proportional to the number of landmarks, it is a penalty suffered by this methodology, however is a necessary feature to consider. Goodness of fit is intuitively improved if the profiling is made over more points of the surface, whereas a less number of landmarks implies a rougher estimation of the boundaries in the inter-landmark spaces.

4.2. Model Initialization

The ASM model is a local search algorithm with limited capture range. For robust and accurate segmentation, it is necessary to determine an initial position of the model in the image, that is, the model must be placed roughly onto the target structure. Model initialization can be achieved by two methods:

1. By requiring user interaction, for example by letting the user define a set of relevant points. For instance, Kelemen et al. for the model-based segmentation of structures present in 3-D Neuroradiological image data, initializes the segmentation with user definition of anterior and posterior commissure (Kelemen, Székely, & Gerig, 1998).
2. By computationally analyzing the image for domain knowledge or additional application specific information, such has locating relevant features, structures or points automatically. For instance, Toth et al. initialize their prostate segmentation algorithm for T2-weighted MRI scans by exploiting additional information from a second modality, Magnetic Resonance Spectroscopy (MRS) (Toth, et al., 2011).

Having these two courses of analysis established, it is clear that the first leads to a semi-automatic segmentation methods, which is the method used in this work, whereas the initialization is explained in the following section. This is of course in all cases that the accuracy of the actual segmentation is not affected and/or diminished.

4.2.1. Manual initialization

Firstly, the model was built and placed in action by manually placing the mean contour shape of the model into the image. Intuitively, this approach only requires the user to select a

central single position as possible into which the mean shape model central points is determined and from which the shape positions are defined relatively to the initial ones. This approach places the shape contours very closely distanced to the real contours and in most cases inside the tongues area, producing a fitting that produced dislocations of the points to outside the normal direction of each of the shape vertexes. This is followed by the selection of 4 points, into which the first iteration of the model fitting is made in their direction. The user selects the lowest point of the anterior wall of the tongue, the point of the tongues tip, the highest point of the tongues dorsum and finally the tongues root point (Figure 33). The search in these directions is enforced in all landmarks by portioning the influence of each, in the adequate of landmarks.

This method is theoretically, expected to produce the fastest convergence to a fitting result, since it implies a user correct placement of a close boundary-to-shape distance.

Figure 33 - Points selected in the manual initialization of the model.

4.3. Image feature search

The local image feature search computes a set of candidate positions around each landmark. The Profile model is used to assign a score to each candidate such that they can be ranked. Finally, one of the candidates is chosen as new position for the landmark. The obvious strategy which is employed in most ASM implementations is to select the candidate with the highest score. This approach was used since the distance to the true boundary is preserved in the profiling if initialization is made correctly, thus being the model at a relevant distance from the true boundaries. This is also compensated in each iteration, where even though at some extent the model cannot find the true boundary, it is enforced by neighboring landmarks reassignments their correspondent true boundaries and subsequent plausible shapes generation to dislocate into new possible locations that *a priori* will be closer to the true boundary onto which these last landmarks were able to locate and adapt towards. The features to be searched represent the normalized gradient profiles, similar to the technique used in the *classical* ASM

formulation made by Cootes et al., and the original intensity profiles, through the methods detailed in Section 3.3.1. However, two methods to defines the goodness of fit along the the profile pixels were used: profiling of the derivatives used a scoring based on the Mahalanobis distance and profiling of the original image intensities with PCA analysis was also carried out. Comparison of the results was and consequent segmentation shapes generated was analysed. This method was therefore, theoretically, designed to be minimally susceptible to assign the best score to image features that do not correspondent to the boundary. Such outliers of the profile model may degrade the robustness of the active shape model, and implies that the neighboring landmarks move inconsistently. This feature was evaluated in the results.

4.4. Imposing shape constraints

By imposing shape constraints with a statistical shape model, one can ensure that the shape of the segmentation results is similar to the training shapes. Thus, leaking of the object into neighboring structures can by this step, at least partially, be avoided. In the following, the methodology for imposing shape constraint is presented.

Firstly, having present that the statistical shape model has been learned in Procrustes space, based on a affine transform calculation, including rotation, scale and translation components, it is important to know the deformed shape is in the coordinate system of the image and therefore, in order to impose shape constraints, pose parameters must be computed which define an affine transformation between these two coordinate systems. Usually, pose parameters are handled as external parameters. After this, the method adopted is once more described the one described by Cootes et al. (Cootes & Taylor, 2004):

1. Pose parameters are estimated, and the deformed shape $\hat{x}$ is mapped from the image to the model coordinate system.
2. The deformed shape is projected into the principal subspace using the equation:

$$\boldsymbol{b} = P.T\,(\hat{x} - \bar{x}) \tag{26}$$

In another notation, we can calculate the parameter b that allows to best approximate $\hat{x}$ with a model shape generated by Equation (3). We seek the $\boldsymbol{b}$ and $\boldsymbol{T}$ that minimizes:

$$distance(\hat{x}, T\,(\bar{x} + Pb))$$

3. Constraints are imposed on the bounds of the principal components bi computed, usually by enforcing that:

$$-3\sqrt{\lambda i} \leq bi \leq 3\sqrt{\lambda i} \text{ for all } i = 1, \ldots, t.$$

4. From the constrained shape parameters in $\boldsymbol{b}$, a plausible shape is generated by simply computing the result of Equation (3).
5. The constrained shape x is mapped back to the image coordinate system using the estimated pose parameters calculated in point 1.

4.5. Segmentation validation

To validate the active shape models segmentation quality, the values of mean and standard deviation of the Euclidean distances between the landmark points of the final shape of the models and the desired segmentation shapes were calculated.

4.6. Results and discussion

To obtain a quantitative evaluation of the performance of the algorithm we trained a model on 19 hand labelled head and neck MRI images, and tested it on a different set of 6 labelled images.

On each test image, it was defined by user interaction the central point of the tongue, in a rough manner, to which the model was dislocated, centralizing it in said position. Secondly, the labeling of the four landmarks of initialization, inferior point of the anterior wall, tongues tip, highest point if tongues dorsum and tongues root was carried out to which the landmarks were mapped onto, and an initial plausible shape with the horizontal and vertical dimensions adjusted by these points, was produced with the statistical shape model. At this point the initial shape is ready to be ran in the multi-resolution search.

The search proceeded through the image pyramid, from a low resolution image to the original image resolution along two image scalings of 0.5 and 1 (Figure 34), that is applied to the image and initial shape, and is correlated to the profile produced in this pyramid range in the training step of the profile model building.

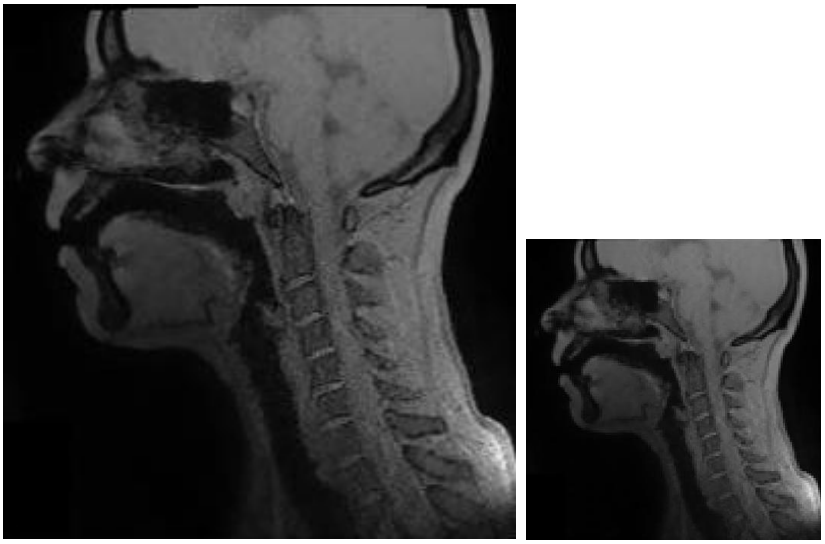

Figure 34 - An image pyramid. The first level [128x128]px is a half of the resolution of the image above it [256x256]px..

4.6.1. ASM Results with User-initialization

This section presents the results of manual initialization and segmentation results of the workflow that includes it in the model Initialization step. The suitability of active models to segment the shape of the tongue in new images is analyzed in the present section. On each test image, it was defined by user interaction the central point of the tongue, in a rough manner, towards which the model was dislocated, centralizing it in said position.

While this technique is very fast and easy to implement, it is also one reason for the limited flexibility and delineation accuracy of these types of models. This is because it partitions the shape space into two classes: allowed shapes and disallowed shapes. Allowed shapes are a subset of the set of all linear combinations of the training shapes that are inside the allowed constraints of the vector parameter b. They are necessarily elements of the affine PCA subspace, and must additionally be close to the mean. Conversely, any shape that cannot be expressed as a linear combination of the training shapes is disallowed. However, one cannot expect that all possible shape variations are present in the training shapes.

Results from successful model fittings, as well as failures, will be showed. The model fitting procedure is run after the generation of the initial shape.

Three key variables were tested in the result analysis of the active shape models produced: retained variance percentage, type of search and number of search profile pixels. Thus, two active shape models were built with 95% and 99% of retained variance and with search profiles of 7, and 17 pixels. It is important to state before any analysis that the two types of search, by gradient profiles combined with mahalanobis distance and the intensity profiles combined with PCA analysis, produced the similar segmentation results, whereas for robustness and consistency the final method used was the search fitting by PCA analysis of the original intensities.

Afterwards, 6 MR images of 3 distinct EP speech sounds, and from two different subjects (one male and one female), which were not considered in the set of training images used, were segmented by the active shape models built.

As stopping criterion of the segmentation process, a maximum of 10 iterations on the first resolution level and 15 iterations on the second. This was considered for all segmentation runs, whereas two resolution levels were used, summing to a total of 25 iterations performed in each test image. This maximum number of iterations was chosen due to its quality results obtained within the six test images. The initialization shapes will be demonstrated in the following result presentation.

The final active shape model developed under this study adopted a gray level profile of 7 pixels long, that is considering 3 pixels from each side of the landmark points. The profiling furthermore considers a search profile of 13 pixels long. This search outline, means that the search was measured moving the profile whisker of 7 pixels within 3 positions offsets in each direction. This profiling presented to be the best option for fitting, producing the best results of the study.

In Figure 35 is depicted an example of the segmentations obtained for one test image. In this figure it is possible to observe the initial shape localization indicated by user interaction and is followed by some of the iterations of the segmentation process by the active shape model built with 99% of retained variability: it starts with a raw estimation on the shape outline by forcing the first iteration of the profiling to adjust towards the correspondent landmarks selected by user interaction in the image (1st iteration), downwards each multiresolution level until converges into the final the final shape at the end of 25 iterations, where from that point on the shape does not present relevant statistical differences relatively to the one presented in the previous iteration. This figure demonstrates the difficulty of the model to converge into the tongues root and lower bounds, that is afterwards impeded by the shape constraint of the shape model. Nevertheless, this is an example of the model segmenting one of the most complex shapes in the test dataset, since it implies relevant dependence from a wide number of the modes of variation retained by this model. The performance of segmentation of this image with the model retaining 95% of variability significantly lower.

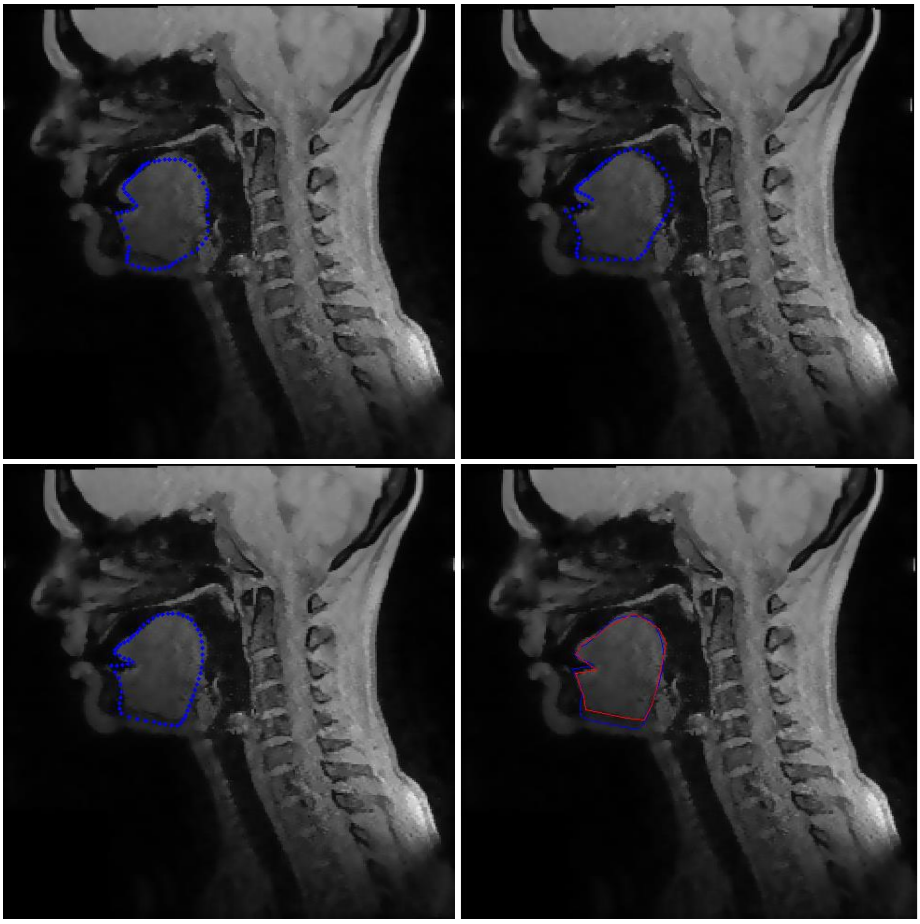

Figure 35- Segmentation process (ASM with 99% of variability) with the initial position of the shape model built overlapped (A) and the results after the 22nd (B) and 25th (C) iterations of the of the segmentation process by the active shape model. In D the produced shape (red) is overlapped with the hand-labeled shape (blue).

Table 4 - Mean and standard deviation (mean ± std) errors of the shapes segmented by the deformable models built in each test image with 7 pixels long profiles.

Image	ASM with 95% of variance	ASM with 99% of variance
male_pa	4.238 ± 3.369	3.554 ± 1.024
male_pu	5.520 ± 4.852	1.541 ± 2.786
female_pu	4.870 ± 5.401	2.326 ± 3.450
male_pi	1.987 ± 1.342	1.556 ± 2.786
female_pa	3.652 ± 4.204	2.941 ± 3.244
female_pi	7.987 ± 1.342	1.541 ± 2.005

Table 5 - Mean and standard deviation (mean ± std) errors of the shapes segmented by the deformable model built in each test image with 17 pixels long profiles.

Image	ASM with 95% of variance	ASM with 99% of variance
male_pa	8.511 ± 6.352	4.369 ± 4.342
male_pu	7.520 ± 4.778	8.671 ± 1.577
female_pu	5.870 ± 4.401	5.684 ± 3.584
male_pi	9.785 ± 3.657	6.684 ± 1.535
female_pa	6.576 ± 4.293	7.565 ± 7.64
female_pi	9.547 ± 4.576	9.684 ± 1.342

The other active shape model built consisted in a gray level profile of dimensions equal to 17 pixels were also built. However, these active shape models were not able to segment successfully the modeled organ in the testing images. These results are rather intuitive in the sense that since the images used during the segmentation process, at each landmark point is considered a segment of 23 pixels long in the active search and therefore, it can easily contain intensities of neighboring structures with a better profile fit, which added to the contrast homogeneity that is caused by the homogeneity of the tissues of the various structures present in the head and neck, does not favor of clear discriminant analysis of the true edges, and consequently, the model built can easily diverge.

Furthermore, the size of the images considered, namely 256x256 pixels, is relatively small and contributes that long profiles will fall under the influence of neighboring structures which will difficult even further the correct detection of the true boundaries and as stated, diverge. An example of the segmentation obtained using profiles with 17 pixels long is depicted in Figure 36. This example depicts the process of wandering of the model shapes produced by the adaptation into image intensities belonging to the palate. This type of model, is very sensitive to such results, when the profile length is not adequate, and furthermore, because the tissues present in the head and neck in these images, specifically the neighboring structures around the tongue present very similar intensities.

In Figure 37 it is possible to observe the limitations the Active Shape Model represents when segmenting a new instance structure shape. The modelling into the positions of the two frenulum end and tongues tip is visible but the shape constraint is not able to model the shape into fine variations presented in the local angle made between the 11 landmarks that form it. This example segmentation is however, a representative of a shape that even though was segmented with the 95% of retained variance model, the sound 'pi', is inherent to precisely shape variations captured by the first 6 modes of variation. Therefore this model presented particularly good results due to no variabilities were eliminated by the retention of lesser number of modes of variation. Nonetheless, is should be noted, also in this image that the vertical variability associated with the second mode of variation of the model, achieved its maximum bound in the attempt of segmentation of the shape and therefore that inaccuracy of the segmentation in the upper boundary of the tongue is due to the shape constraint. This fact means that the upward movement of the tongue in sustention of this sound was not so extensive in any instance of the training shapes, and therefore, constitutes an example of how the absence of a given shape in the training set of shapes will not allow the segmentation of said shape, because it will be constrained, i.e. considered implausible by the model. Therefore instances where the two ends of the frenulum are coincident, the model fails to adapt to this shape. Therefore, it is concluded that the model did not capture this variance, although one instance shape in the training dataset had similar properties. This is also observable by the results presented in the previous Chapter, where the shape variance captured by each mode of

Figure 36 - Segmentation process with the initial position of the shape model built overlapped (A) and the results after the 2nd (B) and 12th (C) iterations. ASM using 17 pixels long profiles.

variation was evaluated separately, and where none presented to capture this shape. In Table 4, the values of the mean and standard deviation in which it is possible to analyze the quality of the segmentation obtained in each test MR image by the active shape models built. In this table it is possible to compare the differences between the segmentation results of the Active Models using a search profile of 13 pixels long and profiles of 7 pixels. Regarding the differences of the segmentation results related to the number of profile pixels, the analysis can be made by the computation of the mean squared differences and standard deviations of the models using 17 pixels long profiles that are presented in Table 5. Establishing comparisons of the mean errors presented in two tables mentioned, confirm the results obtained in the latter case to not be acceptable, whereas the wandering of the model for other regions, implies bigger error of segmentation relating the segmented and true boundary positions. Inside each of the profile size cases presented the errors obtained with the 95% of retained variance models were in every case worse, which was expected, and verified by the resulting segmentations observations where in this case the segmentation quality is highly damaged by the shape constraints and variability captured in the model which contains a lesser number of modes of variation

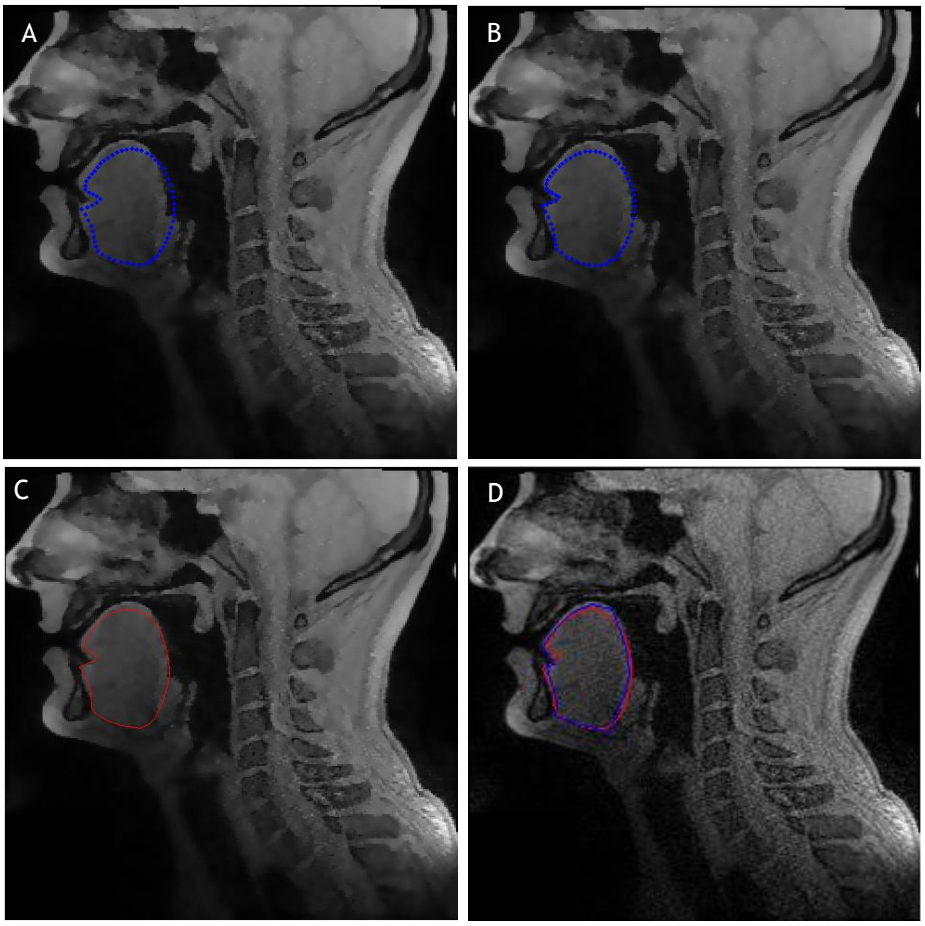

Figure 37 - Segmentation process with the initial position of the shape model built overlapped in image (A) and the results after the 2nd (B) and 12th (C) iterations of the of the segmentation process by the active shape model. In D the produced shape (red) is overlapped with the hand-labeled shape in the original image (blue).

4.7. Conclusion

In summary, the ability of the Active Shape Model developed to segment properly the tongue represented in MR images was assessed in this Chapter.

The primary goal consisted on the analysis of tongue that included a significant component of shape variability by the usage of MR acquisitions during articulation of European Portuguese sounds, followed by the evaluation of the results concerning the semi-automatic segmentation of the modeled tongue shape in new images. However for real-clinical purposes this study would have to widened to an image dataset that includes instances of resting tongue shape. The dataset is only representative and accuracy is only improved with the application of this methodology to a wider dataset.

After the tests performed it was possible to conclude that the method developed is very sensitive to wandering off of the wanted segmentation needed. The profile whisker directions in the classical ASM, is determined by the order in which the landmarks are numbered. This is because the whisker direction is determined by the position of previous and next landmarks relative to the current landmark. This is rather arbitrary and is one of the reasons by which the model is very susceptible to lose the objective boundary.

Nevertheless, the initialization process, when done correctly allows the model to adapt into the right boundaries, at the end of the chosen number of iterations. This is also achieved by the number of landmarks used in the production of the shapes, that allows the adaptation to the boundaries that are within the real boundaries enforcing their adjustment over the fewer number of landmarks that are positioned in location where there is no discriminant power in the whisker direction to set the adequate boundary pixels.

This study can be useful for speech rehabilitation purposes, namely, to recognize the compensatory movements of the articulators during speech production.

Chapter 5

Conclusion and Future Work

This Dissertation aimed to present computational algorithms for object segmentation and analysis in images suitable for application in objects such as the human tongue in images.

This work, thus needed two independent datasets: a training set for building and a test set for validation. The training is obviously, as it should, the largest, because the training set defines nearly all characteristics of the model. The validation and test sets should be big enough for the variance of the measured results to be acceptable. The unknown ideal sizes for these sets depend on the unknown complexity of the underlying function and amount of data noise. Moreover, the flexibility of this Models used in this work is highly dependent in the variability present in the training dataset, and these are automatically included in the produced results, meaning that the model can be adapted to various types of studies of the structure analyzed, the tongue.

The models developed revealed that they could easily be used to reconstruct the shape of the tongue.

5.1. Conclusion

From the experimental results obtained, one may state that the point distribution model built can adequately extract the main characteristics of the movements of the tongue from magnetic resonance images, although with discrepant accuracy on different boundaries that constitute it. While active shape models consider the information around each landmark point of the modeled object, active appearance models use also the gray level information of the object. Consequently, the former type of models is more informative and possibly more efficient than the latter. However, the segmentation process taken into consideration focused

in the accuracy of boundary detection, and therefore was concentrated in the model boundaries candidate analysis methodology of Active Shape Models. Nevertheless, both active shape models obtained remarkable results, either in terms of translating the movements and configurations involved in the different shape conformations depicted, as well as in the segmentation of the tongue in new images.

Based on the results obtained the performance could be assessed by the statistical Euclidean distance evaluation from true contours, defined for the test images, by professional anatomist. The studies results are promising and allow its inclusion in future works for an expansion of the study into a 3D analysis when combined with better image quality that needs the inclusion of a more sophisticated imaging acquisition model. Therefore, the model built can be accurate and efficient tools to be used towards the automatic study of the tongue from magnetic resonance images during speech production.

It can also be used for several studies in articulatory phonetics and in the tongue modeling for speech synthesis, with applications to speech pathology, linguistics and artificial speech. This study was also meant to improve medical analysis of images, in the sense of being used to evaluate changes in the morphologic structure at study in healthy and diseased patients.

5.2. Future work

One of the premises for acquiring an efficient deformable model, and consequently obtaining good results concerning the segmentation of the modeled object, is extremely related to the quality of the images to be studied. In this study were used images taken from 3D acquisitions with high in plane resolution, however the through-plane fails to be sufficient for the building of 3D models, where the axial contour information is extremely lost and therefore accuracy would never be possible. Previous studies present solutions for this problem that include multi-plane stacks acquisitions with following interpolation. This study was restricted to a 2D approach due to the image quality needed for a 3D expansion.

Regarding the 2D Active Shape Model described in this work, the following imminent course of development, would focus on the full automatization of the method, more specifically, the creation of a fully automatic initialization step, by the detection of auxiliary points of reference. The automatic initialization would comprise: frontal facial boundary detection, followed by mouth opening detection, airway upper and lower boundary detection using active contours, finalizing with initialization of the model shape using the lower boundary of the upper airway. This would produce an initial estimate of the tongues dorsum or upper and posterior boundaries, to which the model initial shape would be mapped into. The full automatization of the global framework of this study would allow the development of an image analysis software that would enable the study of the tongue and subsequently improve the professional practice of speech specialists among other medical professionals.

Another major improvement needed in this study concerns to the amount of data studied. The variability was sought by the inclusion of different anatomies and sound articulations, however it is merely representative. The further validity of this study would be achieved by the subjection to a higher dataset for analysis.

As previously stated, while active shape models consider the information around each landmark point of the modeled object, active appearance models use also the gray level information of the object. Therefore the expansion of this model into the following Active Model would promisingly allow improvements in the segmentation and performance of the model.

It can also be used for several studies in articulatory phonetics and in the vocal tract modeling for speech synthesis, with applications to speech pathology, linguistics and artificial speech.

This study was also meant to improve medical analysis of images, in the sense of being used to evaluate changes in the morphologic structure at study in healthy and diseased patients.

Furthermore, the demonstration of connections to other models such as jaw and airway, as well as the application of this model into the segmentation of other structure that play key roles in the tasks taken upon the aerodigestive tract, such as tonsils and the *velum,* within the framework could enforce the knowledge that models developed separately may be connected to build more complex biomechanical models towards a complete aerodigestive tract.

References

A. Bernstein, M., F. King, K., & Xiaohong, Zhou, J. (2004). *Handbook of MRI Pulse Sequences*. *Handbook of MRI Pulse Sequences* (pp. 443–490). Elsevier. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780120928613500182

Abd-el-Malek, S. (1939). Observations on the morphology of the human tongue. *Journal of Anatomy*, *73*, 201–210.3.

Abd-El-Malek, S. (1955). The part played by the tongue in mastication and deglutition. *Journal of Anatomy*, *89*(1), 250–254.1.

Akgul, Y. S., Kambhamettu, C., & Stone, M. (1998). Extraction and tracking of the tongue surface from ultrasound image sequences. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)* (pp. 298–303).

Atsumi, T., & Miyatake, T. (1987). Morphometry of the degenerative process in the hypoglossal nerves in amyotrophic lateral sclerosis. *Acta Neuropathol*, *73*, 25–31.

Badin, P., & Gérard, B. (1998). A Three-Dimensional Linear Articulatory Model Based on MRI Data. *Proceeding of the International Conference of Spoken Lnaguage*, 14-20.

Badin, P., Bailly, G., Revéret, L., Baciu, M., Segebarth, C., & Savariaux, C. (2002). Three-Dimensional Linear Articulatory Modeling of Tongue, Lips and Face, Based on MRI and Video Images. *Journal of Phonetics*, (3):533-53.

Buades, A., Coll, B., & Morel, J. (2005). A non-local algorithm for image denoising. *Computer Vision and Pattern Recognition*, (2):60 - 65.

Buades, A., Coll, B., & Morel, J. (2005b). A Review of Image Denoisinging Algorithms, with a New One. *Multiscale Modeling & Simulation*, 4(2):490–530.

Bai, W., Shi, W., O'Regan, D. P., Tong, T., Wang, H., Jamil-Copley, S., … Rueckert, D. (2013). A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images. *IEEE Transactions on Medical Imaging*, *32*(7), 1302–1315.

Balafar, M. a., Ramli, a. R., Saripan, M. I., & Mashohor, S. (2010). Review of brain MRI image segmentation methods. *Artificial Intelligence Review*, *33*, 261–274.

Can, A., Stewart, C. V., Roysam, B., & Tanenbaum, H. L. (2002). A feature-based, robust, hierarchical algorithm for registering pairs of images of the curved human retina. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3), 347–364.

Cheng, S., Butler, J. E., Gandevia, S. C., & Bilston, L. E. (2008). Movement of the tongue during normal breathing in awake healthy humans. *The Journal of Physiology*, *586*, 4283–4294.

Cootes, T., Edwards, G., Taylor, C. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681-685

Cootes, T., & Taylor, C. (1992). Active shape models - smart snakes. *British Machine Vision Conference*.

Cootes, T., & Taylor, C. (1998). Constrained active appearance models. *International Conference on Computer Vision*.

Cootes, T. F., & Taylor., C. J. (2004). *Statistical Models of Appearance for Computer Vision*. Obtido de Technical Report: The University of Manchester School of Medicine: http://www.face-rec.org/algorithms/AAM/app_models.pdf

Cootes, T., & Taylor, C. (1994). Using grey-level models to improve active shape model search. *IEEE International Conference on Pattern Recognition*, (1):63–67.

Dang, J., & Honda, K. (2004). Construction and control of a physiological articulatory model. *Journal of the Acoustical Society of America*, (2): 853-870.

Dempsey, J., Veasey, S., Morgan, B., & O'Donnell, C. (2010). Pathophysiology of Sleep Apnea. *Physiological Reviews*, *90*, 47–112.

Dreyer, K. J., Hirschorn, D. S., Thrall, J. H., & Mehta, A. (2006). *PACS: A Guide to the Digital Revolution* (2nd Editio.). Springer.

English, a W., Wolf, S. L., & Segal, R. L. (1993). Compartmentalization of muscles and their motor nuclei: the partitioning hypothesis. *Physical Therapy*, *73*(12), 857–867.

Engwall, O. (2000). A 3D tongue model based on MRI data. *Proceedings of the International Conference of Spoken Language (ICSLP)*.

Fei, B., Duerk, J. L., Boll, D. T., Lewin, J. S., & Wilson, D. L. (2003). Slice-to-volume registration and its potential application to interventional MRI-guided radio-frequency thermal ablation of prostate cancer. *IEEE Transactions on Medical Imaging*, *22*(4), 515–525.

Georgia Highlands College. (2013). Biology 2121 Human Anatomy and Physiology.

Gerard, J., Perrier, P., & Payan, Y. (2006). 3D biomechanical tongue modelling to study speech production. *n J. Harrington & M. Tabain (eds). Speech Production:Models, Phonetic Processes, and Techniques*.

Gilbert, R. J., & Napadow, V. J. (2005). Three-dimensional muscular architecture of the human tongue determined in vivo with diffusion tensor magnetic resonance imaging. *Dysphagia*, *20*, 1–7.

Grady, L. (2006). Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(11), 1768–83.

Gray, H. (1918). *Anatomy of the Human Body*. (Philadelphia: Lea & Febiger., (Twentieth Ed.).

Hamilton, A. F. D. C., Jones, K. E., & Wolpert, D. M. (2004). The scaling of motor noise with muscle strength and motor unit number in humans. *Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale*, *157*, 417–430.

Harandi, N. M., Abugharbieh, R., & Fels, S. (2014). 3D segmentation of the tongue in MRI : a minimally interactive model-based approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*.

Heimann, T., & Meinzer, H.-P. (2009). Statistical shape models for 3D medical image segmentation: A review. *Medical Image Analysis 13*, (4)543–563.

Heimann, T., Wolf, I., & Meinzer, H.-P. (2006). Active shape models for a fully automated 3D segmentation of the liver – an evaluation on clinical data. *Mediacal Image Computing and Computer-Assisted Intervention*, Larsen R., Nielsen M., Sporring J.,(Eds.), vol. 4191 of LNCS, Springer, pp. 41–48.

Hendargo, H. C., Estrada, R., Chiu, S. J., Tomasi, C., Farsiu, S., & Izatt, J. a. (2013). Automated non-rigid registration and mosaicing for robust imaging of distinct retinal capillary beds using speckle variance optical coherence tomography. *Biomedical Optics Express*, *4*(6), 803–21.

Hiiemae, K. M., & Palmer, J. B. (2003). Tongue Movements in Feeding and Speech. *Critical Reviews in Oral Biology and Medicine*, (6):413-429.

Hopp, T., Dietzel, M., Baltzer, P. a., Kreisel, P., Kaiser, W. a., Gemmeke, H., & Ruiter, N. V. (2013). Automatic multimodal 2D/3D breast image registration using biomechanical FEM models and intensity-based optimization. *Medical Image Analysis*, *17*(2), 209–218.

Ibragimov, B., Likar, B., Pernuš, F., & Vrtovec, T. (2012). A game-theoretic framework for landmark-based image segmentation. *IEEE Transactions on Medical Imaging*, *31*(9), 1761–1776.

Ibragimov, B., Prince, J. L., Murano, E. Z., Woo, J., Stone, M., Likar, B., … Vrtovec, T. (2015). Segmentation of tongue muscles from super-resolution magnetic resonance images. *Medical Image Analysis*, *20*, 198–207.

Isidoro, S., & John, S. (1998). Active blobs. *In International Conference on Computer Vision*.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. M. (2002). Improved optimisation for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*, 825–841.

Jiang, S., Xue, H., Glover, A., Rutherford, M., Rueckert, D., & Hajnal, J. V. (2007). MRI of moving subjects using multislice Snapshot images with Volume Reconstruction (SVR): Application to fetal, neonatal, and adult brain studies. *IEEE Transactions on Medical Imaging*, *26*(7), 967–980.

Joshi, A. a., Hu, H. H., Leahy, R. M., Goran, M. I., & Nayak, K. S. (2013). Automatic intra-subject registration-based segmentation of abdominal fat from water-fat MRI. *Journal of Magnetic Resonance Imaging*, *37*, 423–430.

Kass, M. (1987). Snakes: Active contour models. *International Journal of Computer Vision*.

Kelch, J., & Wein, B. (1993). Segmentation of the tongue surface in ultrasonic images using modified scale space filtering. *Proceedings of IEEE Ultrasonics Symposium*, 947–950.

Kelemen, A., Székely, G., & Gerig, G. (1998). Elastic model-based segmentation of 3-D neuroradiological data sets. *IEEE Transactions on Medical Imaging 18*, (10):829-839.

Kuna, S. T. (2004). Regional effects of selective pharyngeal muscle activation on airway shape. *American Journal of Respiratory and Critical Care Medicine*, *169*, 1063–1069.

Lee, J., Woo, J., Xing, F., Murano, E. Z., Stone, M., & Prince, J. L. (2014). Semi-automatic segmentation for 3D motion analysis of the tongue with dynamic MRI. *Computerized Medical Imaging and Graphics : The Official Journal of the Computerized Medical Imaging Society*, *38*(8), 714–24.

Lieberman, D. E., McCarthy, R. C., Hiiemae, K. M., & Palmer, J. B. (2001). Ontogeny of postnatal hyoid and larynx descent in humans. *Archives of Oral Biology*, *46*, 117–128.

Lieberman, P. (2012). Vocal tract anatomy and the neural bases of talking. *Journal of Phonetics*, *40*(4), 608–622.

Lucas-Osma, A. M., & Collazos-Castro, J. E. (2009). Compartmentalization in the triceps brachii motoneuron nucleus and its relation to muscle architecture. *Journal of Comparative Neurology*, *516*(October 2008), 226–239.

Lufkin, R. B., Larsson, S. G., & Hanafee, W. N. (1983). Work in progress: NMR anatomy of the larynx and tongue base. *Radiology*, *148*(1), 173–5.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, *16*(2), 187–198.

MathWorks, R. a. (Accessed in 15 June of 2015a). "*interp1*". Retrieved from http://www.mathworks.com/help/matlab/ref/interp1.html

Mathworks, R. a. (Accessed in 15 June de 2015b). "*Procrustes*". Retrieved from http://www.mathworks.com/help/stats/procrustes.html

Mathworks, R. a. (Accessed in 15 June of 2015c). "*Eig*". Retrieved from http://www.mathworks.com/help/matlab/ref/eig.html

Mathworks, R. a. (Accessed in 03 May of 2015d). "*Svd*". Retrieved from http://www.mathworks.com/help/matlab/ref/svd.html

Miyawaki, K. (1974). A study on the musculature of the human tongue. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics, Univerisity of Tokyo*, *8*, 23–50.

Monti, L., Renifilo, E., Profili, M., & Balzarini, L. (2008). Journal of Cardiovascular Magnetic Resonance Cardiovascular magnetic resonance features of caseous calcification of the mitral annulus. *Journal of Cardiovascular Magnetic Resonance*, *5*, 1–5.

Moon, I. J., Han, D. H., Kim, J.-W., Rhee, C.-S., Sung, M.-W., Park, J.-W., … Lee, C. H. (2010). Sleep magnetic resonance imaging as a new diagnostic method in obstructive sleep apnea syndrome. *The Laryngoscope*, *120*(12), 2546–54.

Mu, L., & Sanders, I. (2000). Neuromuscular specializations of the pharyngeal dilator muscles: II. Compartmentalization of the canine genioglossus muscle. *Anatomical Record*, *260*(July), 308–325.

Oliveira, F. P. M., & Tavares, J. M. R. S. (2012). Medical image registration: a review. *Computer Methods in Biomechanics and Biomedical Engineering*, (January 2015), 1–21.

Parent, R., & King, S. (2001). A 3D Parametric Tongue Model for Animated Speech. *JVCA 12*, (3):107-115.

Pillar, G., Fogel, R. B., Malhotra, A., Beauregard, J., Edwards, J. K., Shea, S. a., & White, D. P. (2001). Genioglossal inspiratory activation: Central respiratory vs mechanoreceptive influences. *Respiration Physiology*, *127*, 23–38.

Rinck, P. A. (2001). *Magnetic Resonance in Medicine Basic Textbook of the European Magnetic Resonance Forum* (4th ed.). Oxford: Blackwell Scientific Publications.

Salter, H. (1852). *The Cyclopaedia of Anatomy and Physiology* (Tongue. In., pp. 1120 – 1163). London: Longman, Brown, Green, Longmans, & Roberts.

Sawczuk, a, & Mosier, K. M. (2001). Neural control of tongue movement with respect to respiration and swallowing. *Critical Reviews in Oral Biology and Medicine : An Official Publication of the American Association of Oral Biologists*, *12*(l), 18–37.

Seeley, R., Stephens, T., & Tate, P. (2008). *Anatomia e Fisiologia* (8ª Edição.). McGraw-Hill.

Seikel, J., King, D., & Drumright, D. (2009). *Anatomy and physiology for speech, language, and hearing*. Delmar, Cengade learning.

Shinagawa, H., Murano, E. Z., Zhuo, J., Landman, B., Gullapalli, R. P., Prince, J. L., & Stone, M. (2008). Tongue muscle fiber tracking during rest and tongue protrusion with oral appliances: A preliminary study with diffusion tensor imaging. *Acoustical Science and Technology*, *29*(4), 291–294.

Sonies, B. C. (1981). Ultrasonic visualization of tongue motion during speech. *The Journal of the Acoustical Society of America*, *70*(3), 683. Retrieved from http://scitation.aip.org/content/asa/journal/jasa/70/3/10.1121/1.386930

Stone, M., & Lundberg, A. (1996). Three-dimensional tongue surfaces from ultrasound images. *SPIE Proceedings*, (2709):168-179.

Stone, M., Liu, X., Chen, H., & Prince, J. L. (2010). A preliminary application of principal components and cluster analysis to internal tongue deformation patterns. *Computer Methods in Biomechanics and Biomedical Engineering*, *13*(February 2015), 493–503.

Takemoto, H. (2001). Morphological analyses of the human tongue musculature for three-dimensional modeling. *Journal of Speech, Language, and Hearing Research : JSLHR*, *44*(1), 95–107.

Tang, L., Bressmann, T., & Hamarneh, G. (2012). Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves. *Medical Image Analysis*, *16*(8), 1503–20.

Teguh, D. N., Levendag, P. C., Voet, P. W. J., Al-Mamgani, A., Han, X., Wolf, T. K., … Hoogeman, M. S. (2011). Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. *International Journal of Radiation Oncology Biology Physics*, *81*(4), 950–957.

The DICOM Standard. (2015). PS3.5 - Data Structures and Encoding.

Togeiro, S. M. G. P., Chaves, C. M., Palombini, L., Tufik, S., Hora, F., & Nery, L. E. (2010). Evaluation of the upper airway in obstructive sleep apnoea. *The Indian Journal of Medical Research*, *131*(February), 230–235.

Toth, R., Tiwari, P., Rosen, M., Reed, G., Kurhanewicz, J., Kalyanpur, A., MadabhushI, A. (2011). A magnetic resonance spectroscopy driven initialization scheme for active shape model based prostate segmentation. *Medical Image Analysis 15*, (2): 214–225.

Tristan-Vega, A., Garcia Perez, V., Aja-Fenandez, S., & Westin, C. (2012). Efficient and Robust Nonlocal Means Denoising of MR Data Based on Salient Features Matching. *Computer Methods and Programs in Biomedicine, 105(2)*, (131-44).

Unser, M., & Stone, M. (1992). Automated detection of the tongue surface in sequences of ultrasound images. *The Journal of the Acoustical Society of America*, *91*, 3001–3007.

Ventura, S. R., Diamantino, R. F., & Tavares, J. M. (2008). Three-Dimensional modeling of tongue during speech using MRI data. *CMBBE 2008—8th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering*, (49), 49–58.

Vasconcelos, M., Ventura, S., Tavares, J., & Freitas, D. R. (2009). Analysis of Tongue Shape and Motion in Speech Production using Statistical Modeling. *SEECCM 2009 - 2nd South-East European Conference on Computational Mechanics*, 96-103.

Vasconcelos, M., 2015, *Computational Algorithms for image analysis: Applications on human vocal tract and silhouette*. Ph.D thesis, Faculdade de Engenharia da Universidade do Porto

Von Neuman, J., & Morgenstern, O. (1994). *Theory of Games and Economic Behavior* (p. 776). Princeton University Press.

Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., … Gentry, L. R. (2009). Anatomic development of the oral and pharyngeal portions of the vocal tract: an imaging study. *The Journal of the Acoustical Society of America*, *125*, 1666–1678.

Wang, J., & Gu, X. (2013). High-quality four-dimensional cone-beam CT by deforming prior images. *Physics in Medicine and Biology*, *58*, 231–46.

Watkin, K. L., & Rubin, J. M. (1989). Pseudo-three-dimensional reconstruction of ultrasonic images of the tongue. *The Journal of the Acoustical Society of America*, *85*(1), 496–9.

Wheatley, J. R., Kelly, W. T., Tully, a, & Engel, L. a. (1991). Pressure-diameter relationships of the upper airway in awake supine subjects. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *70*, 2242–2251.

Wickham, J. B., & Brown, J. M. M. (2012). The function of neuromuscular compartments in human shoulder muscles. *Journal of Neurophysiology*, *107*(October 2011), 336–345.

Wilhelms-Tricarico, R. (1995). Physiological modeling of speech production: methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, (5):3085-98.

Woo, J., Murano, E. Z., Stone, M., & Prince, J. L. (2012). Reconstruction of high-resolution tongue volumes from MRI. *IEEE Transactions on Bio-Medical Engineering*, *59*(12), 3511–24.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Combining Active Learning and Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. (*20*): 912–919.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. *Machine Learning-International Workshop Then Conference-*, *20*, 912.