Vocal Tract Contour Tracking for Real-time Speech MRI Using a Shape Model Prior

by

Sasan Asadi Abadi

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Electrical and Electronics Engineering

January 19, 2017

Vocal Tract Contour Tracking for Real-time Speech MRI Using a Shape Model Prior

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Sasan Asadi Abadi

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Assoc. Prof. Engin Erzin

_____

Assoc. Prof. Yucel Yemel

_____

Assoc. Prof. Burak Acar

Date: _____

To my parents

ABSTRACT

Human speech production is a process, which is initiating from lungs, phonating in the vocal folds and ultimately passing through the vocal tract comprising several jointly working articulators. Knowledge about the dynamic shape of the vocal tract is the basis of many speech production applications such as, articulatory analysis, modeling and synthesis. Vocal tract air-way tissue boundary segmentation in the mid-sagittal plane is necessary as a pre-step for 3D-reconstruction of the tract shape. This segmentation problem is however challenging due to poor resolution real-time speech MRI, grainy noise and the rapidly varying vocal tract shape.

In this thesis we present a robust approach to vocal tract airway-tissue boundary contour tracking by training a model for human vocal tract. We generate a dataset of manually segmented vocal tract and utilize a statistical approach to train a model for the tract. An active contour approach is employed to segment the air-way tissue boundaries of the vocal tract while restricting the curve movement to the trained shape model. Then the contours in subsequent frames are tracked using dense motion estimation methods. Experimental evaluations over the mean square error metric indicate significant improvements compared to the state-of-the-art.

ÖZETÇE

İnsan sesi üretimi, akciğerlerden başlayıp ses telleri titreşimleri ile ses yolundan geçerek ortaklaşa çalışan birkaç artikülatör tarafından oluşturulan bir süreçtir. Ses yolunun dinemik şekil bilgisi artikülasyon analizi, ses modelleme ve sentezleme gibi bir çok ses üretimi uygulamasının temelini oluşturmaktadır. Orta sagittal düzlemde (2 boyutlu görünümde) ses yolu dokusunun sınır çizgisinin bulunması, ses yolunun 3 boyutlu modellenmesi için gerekli bir ön adımdır. Ancak bu problem, gerçek zamanlı MRIın düşk çözünürlüğünden ve ses yolunun hızlı değişimlerinden dolayı zordur.

Bu tez çalışmasında, insan ses yolu için bir model eğiterek ses yolu dokusunun sınır çizgisi takibine gürbüz bir yaklaşım sunuyoruz. Manüel olarak bölünmüş ses yolunun konturlarından oluşan bir veri seti oluşturup, ses yolu modelini istatistiksel yöntemlerle eğitiyoruz. Eğrilerin hareketini eğitilen şekil modeline kısıtlayarak, ses yolu dokusu sınırlarını bulmak için aktif kontur yaklaşımı kullanıyoruz. Ardından, birbirlerini takip eden karelerde sınır çizgisini, yoğun hareket kestirimi yöntemleriyle izliyoruz. Ortalama karesel hata metriğine dayalı deneysel değerlendirmeler, sunulan yöntemin literatürde bulunan yöntemler ile karşılaştırıldığında belirgin iyileşmeler sağladığı gözlenmiştir.

ACKNOWLEDGMENTS

Many people have inspired me through my life to become who I am now. I am using this opportunity to express my thanks to everyone who has supported me, specially throughout writing this thesis.

I would first like to thank my advisor Prof. Engin Erzin for the continuous support in this two years of my masters career. I thank him for his great knowledge, patience, motivation and guidance through all steps of my research. He was willing to give his time so generously and I always found his office door open and wellcoming. I have learnt many things from him and this accomplishment would not have been possible without his priceless advice.

I would also like to acknowledge my committee members: Prof.Yucel Yemez and Prof. Burak Acar who kindly agreed to participate to my thesis defense. I thank them for their valuable remarks on my research and their recommendations for my future work.

I thank my fellow officemates who were always there to help me with the difficulties I was facing during these years. They have supported me both scientifically in resolving my technical problems, and socially so that I could easily adjust myself to the campus life as a beginner. Assistance provided by Adeel Afridi, Altynbek Isabekov and Rizwan Sadiq is greatly appreciated. They have helped me a lot with the courses I took and with my computer programming problems. I thank my other officemates Ongun Arisev, Hazan Ezgi Imer and Narjis Fatima for providing a lovely and quiet atmosphore in the office.

I thank Koç univerity for providing me with the scholorship, housing and all the other equipments to facilitate my life. Indeed Koc university has the most beautiful campus and friendly staff which all helped me to lead a peacefull like here in Istanbul.

I thank my dear friend and gym partner Dr.Vahid Akbari for all the moral supports.

He has always been there for me like a brother and assisted me in my life and in the gym from the first day I arrived in Istanbul.

My special thanks goes to my beloved girlfriend, Esin Mumcu, who has always believed in me and morally supported me the most while writing this thesis. Istanbul was an unaccustomed city for me before I knew her. She has given a new quality to my life and has provided me with the sincerest support in the hardest times.

I must express my sincerest gratitude to my parents, Soheyla Hosseinzadeh and Morteza Asadiabadi, and my siblings, Sara and Saman, for allowing me to follow my dreams in another country and tolerating all the long seperation and for all their support in all stages of my life. I am in debt to them for ever and without them, I could never have gotten to where I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

MRI	**M**agnetic **R**esonance **I**maging
US	**U**ltra **S**ound
VT	**V**ocal **T**ract
ACM	**A**ctive **C**ontour **M**odel
ASM	**A**ctive **S**hape **M**odel
PCA	**P**rincipal **C**omponent **A**nalysis
DP	**D**ynamic **P**rogramming
PDM	**P**oint **D**istribution **M**odel
MSD	**M**ean **S**um of **D**istance
RMSE	**R**oot **M**ean **S**quared **E**rror

Chapter 1

INTRODUCTION

It has been well known that the shape of the vocal tract modulates the sounds spoken. In a speech production system, the lungs and glottis act as the source, while the vocal tract produces different vowels and consonants by means of its time-varying shapes. The vocal tract consists of several articulators such as epiglottis, velum, tongue and lips which control the dynamics of the tract. The specific anatomical features of the articulators restricts their movements, thus the shape of the vocal tract.

Knowledge about time-varying changes of the vocal tract is the basis to understand the human speech production system. In the discrete concatenated tube models of the vocal tract, the resonance frequencies or, more generally, the transfer function of the vocal tract are estimated from the cross sectional area function of the tract. Knowing the time evolution of the vocal tract transfer function implies the speech being produced. Therefore the study of human vocal tract shape evolution could play a significant role in speech synthesis applications.

Magnetic Resonance Imaging (MRI) provides the full mid-sagittal view of the vocal tract tube of a subject. Segmentation of the lower and upper profiles of the vocal tract in the mid-sagittal view is the basis of 3D reconstruction of the tract shape. To understand the varying structure of the human vocal tract, the problem to be solved is the airway-tissue boundary detection of the tract in an image and tracking the boundaries in a sequences of images.

1.1 Background and Literature Review

1.1.1 Anatomy of speech

The main human body organs involved in the speech production process are *lungs*, *larynx* and the *vocal tract*, as shown in Figure 1.1. The larynx, also known as the voice box, contains the *vocal folds* and the opening of the folds is called *glottis*. The spoken sounds are categorized into two types base on the status of the glottis. The airflow through the open vocal folds produces *voiceless sounds* and the airflow through the vibrating vocal folds produces *voiced sounds*. The vocal tract shapes the source signal in the glottis into speech sounds. The vocal tract comprises several articulators such as pharynx, velum, tongue, hard palate and lips. Different sounds are produced based on the *place of artuculation* and *manner of the articulation* in the vocal tract. Place of articulation refers to the area in the vocal tract in which the airflow constriction occurs and the manner of articulation refers to how and to what degree the airflow is constricted.

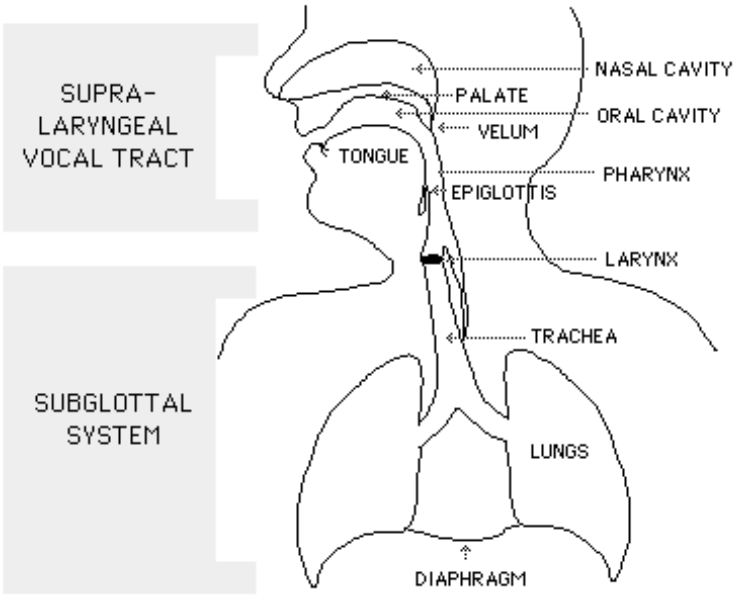


Figure 1.1: Speech production

1.1.2 Articulatory modeling

The time-varying vocal tract shape is described by the position of each articulator at any time. In a lower order description, the relative position of various articulators can provide an outline of the tract shape. Several tract variables could be defined with respect to various manners and locations of articulation. Some tract variables as defined in Saltzman and Munhall (1989) and given in Figure 1.2, include LA (Lip Aperture), TTCD and TTCL (Tongue Tip Constriction Degree and Constriction Degree) and VEL (Velum Aperture), LP (Lip Protrusion), TBCD and TBCL (Tongue Body Constriction Degree and Location), JA (Jaw Angle) and GLO (Glottal Aperture). These parameters provide a general configuration of the vocal tract. It is transparent that more variables will provide a finer description of the tract.

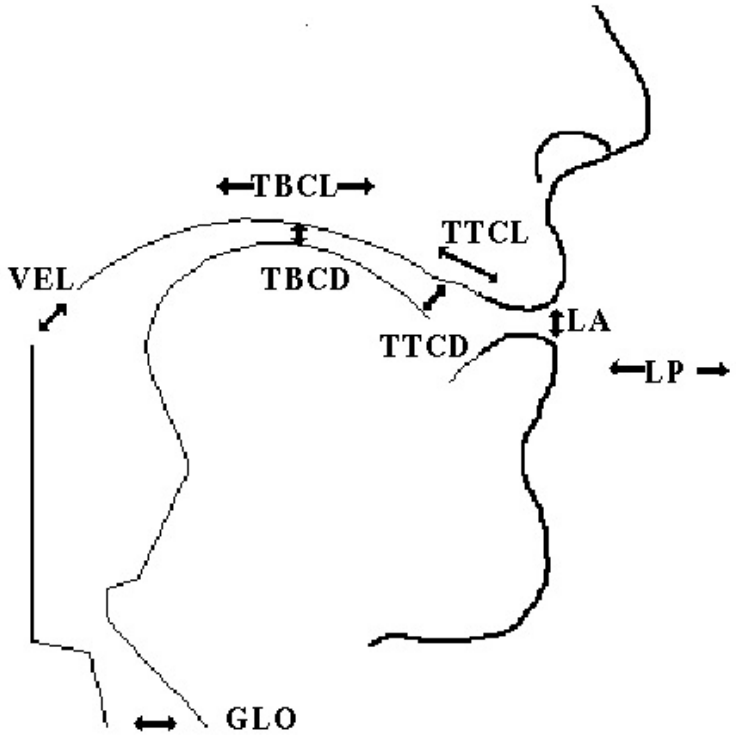


Figure 1.2: Tract variables and their range of movement. From Saltzman and Munhall (1989)

1.1.3 Statistical vocal tract modeling

Medical Imaging of the upper airway has opened new horizons for better understanding of human speech production. Extensive research has been done on vocal tract shape modeling and contour tracking using X-ray and UltraSound (US) imaging. One of the first statistical models of the vocal tract was introduced by Maeda (1990) using 1000 mid-sagittal X-ray images of the vocal tract. A semi-polar coordinate system for measuring the the lateral outlines of the vocal tract was proposed by Maeda, as shown in Figure 1.3. The hand-labeled contours extracted from the X-ray images were projected onto the semi-polar grid system and the variation of the contour points were analyzed using Principle Component Analysis. The Maeda's model for the vocal tract used seven control parameters which were allowed to vary between ± 3 of their standard deviation. The seven parameter in Maeda's model describe 98% of the shape variations in the available data.

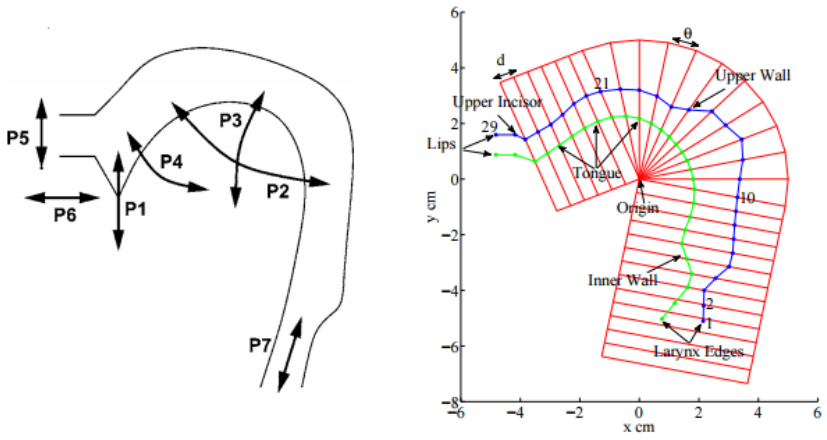


Figure 1.3: Left: Maeda's model parametes, Right: Maeda's gridline system with lower and upper profiles of vocal tract

Another parametric model of the vocal tract was proposed by Yehia et al. (1996). In his method, 519 mid-sagittal X-ray images of a female speaker were projected onto the Maeda's semi-polar grid system to obtain the cross-sectional area function along the vocal tract, from glottis to lips. An eigenvalue decomposition analysis yielded a

5 parameter model for the tract, describing over 92% of the data variation.
Several other statistical models of the vocal tract have been investigated using lateral view of VT images using X-ray, MRI and US, with the purpose of articulatory modeling with low dimensional model descriptors.

1.1.4 Vocal Tract Boundary Detection and Tracking

Intensive research has been conducted on segmentation of the vocal tract air-way tissue boundary. Li et al. (2005) developed the *EdgeTrak*, a modified active contour model-based automatic contour tracker for detecting and tracking human tongue in a sequence of US images. A multimodal acquisition of articulatory data was suggested by Aron et al. (2008). US images, electromagnetic localization sensors and sound data were utilized to implement an active contour-based algorithm for tracking the human tongue in a sequence of US images using optical flow. Roussos et al. (2009) introduced an active appearance based method for contour tracking. A shape prior model was trained for the tongue, from the manually segmented tongue contours extracted using X-ray images and an appearance model was trained for the image intensities around the tongue contours, using the US images. The contours were tracked in a sequence of US images using model fitting approaches.
Among all of the medical imaging technologies developed, attention toward the MRI has been growing since MRI is safe for the subject and can provide full mid-sagittal view images of the vocal tract as well as a 3-D outlook. Following are the most recent proposed algorithms for 2D tracking vocal tract in MRI videos.
Bresch and Narayanan (2009) present a novel approach to segmentation of upper airway real-time MRI in a frequency domain. Their algorithm uses an anatomically informed model for the vocal tract, fitting to an observed image through a gradient descent optimization procedure. Despite the visually accurate results in the paper, their algorithm suffers from a major aspect. The computational complexity of their proposed algorithm is so high that it is inaccurate for real-time research. To get well-captured contours for the vocal tract, 650 gradient descent optimization steps

are carried out and the process takes 20 minutes for each image on a modern desktop computer [Bresch and Narayanan (2009)].

A semi-automatic rapid VT air-way tissue boundary segmentation is proposed by Proctor et al. (2010). In his work the image intensity profile along Maeda's gird line system is investigated. A graph is constructed by connecting all the intensity profile local minimas on all of the grid lines. The central air-way path through the tract is then estimated by finding the optimal path to minimize a defined score through the constructed graph, using the Dijkstra algorithm. The air-way tissue boundary are estimated by locating the first point on the grid lines on either side of the center line crossing a threshold.

Kim et al. (2014) proposed an enhanced vocal tract contour tracking in a MRI videos. His approach is quite similar to Proctor et al. (2010) in the sense that the vocal tract center-line is estimated by minimizing an objective score using the Viterbi algorithm. The constructed graph in this method is not just limited to the local minimas along the grid lines, but using all of the pixels on each line, which results in a quite high run time of the algorithm. Later, at each grid-line, the VT boundaries are determined by finding the first pixels crossing a fixed intensity threshold ,from either side of the center-line. Hence, the accuracy of detecting the air-way tissue boundaries from the center-line, is directly dependent on the accuracy of center-line estimation.

1.2 Database

The USC-TIMIT database is a great source for studying the human speech production. The database comprises recorded mid-sagittal MRI videos of 10 different subjects, synchronized with the audio signal. MRI and audio data of five male and five female American English speakers reading 460 sentences from the TIMIT corpus is available in the USC-TIMIT database. The videos are recorded at a frame rate of 23.13 frames/sec and a spatial resolution of 68×68 pixels over 20×20 cm (approximately 2.9mm pixel width). All the videos utilized in this thesis are taken from the USC-TIMIT database.

1.3 Contribution

The main contribution of this thesis is presenting a robust vocal tract air-way tissue boundary tracking for real-time MRI. The traditional Active Shape Models and Active Contour Models are combined to track the air-way tissue boundary of the vocal tract in MRI videos. A new modified energy is defined to detect the air-way tissue boundary using the trained models as well as other energy terms for training imperfection compensation.

1.4 Thesis Outline

In the first chapter we review some basic research on the vocal tract contour tracking. A wide range of research has been done on the segmentation of the vocal tract for a single frame. However, the study of time-varying changes of the vocal tract is not just a contour detection problem, but also a contour tracking problem. The importance of tracking the vocal tract contours from one frame to another is in finding a correspondence between contours in subsequent frames thus being able to locate each articulator in the sequences of images for computing the time-varying tract variables. Therefore we mainly focus on previous research with a vocal tract boundary tracking nature in the literature review.

In the second chapter we describe the dynamic contour models. The Active Contour Model (ACM), introduced by Kass et al. (1988), is a powerful framework for tracking deformable objects. We discuss Active Shape Models (ASM) which are model-based methods and make use of prior shape information of what is expected in the images, making them a great tool for biomedical image segmentation.

In the third chapter we discuss our work on driving a model for the vocal tract from mid-sagittal view of MRI videos of different subjects. We build a statistical shape model for the vocal tract using Principal Component Analysis (PCA) from several manually segmented vocal tract contours, extracted from real time MRI of USC-

TIMIT database. We present our method of ASM-ACM combination to track the vocal tract boundaries in a sequence of MR-images.

Finally, in the fourth chapter we evaluate the performance of the proposed algorithm and compare the results with other contour tracking methods presented for the vocal tract. The evaluated results show that the proposed method is capable of tracking the time-varying vocal tract shapes with higher accuracy and within a reasonable time compared to other available methods, specially for the lips and tongue contours.

Chapter 2

CONTOUR TRACKING MODELS

2.1 Active Contour Models

Active Contours, also known as snakes, are energy minimizing methods evolving a curve initially placed close to a target boundary toward the region of interest. In the original formulation of the snake, introduced by Kass et al. (1988) the energy function has an internal term which applies a smoothness constraint on the curve and an external term motivating the snake to move toward the image features.

An active contour model describes a continuous curve with a set of discrete points, referred to as *snaxels*. The snake is the polygon (open or closed) formed by connecting the snaxels to each other with straight lines, thus the resolution of the curve depends on the number of points describing the curve. Each snaxel is a vertex of the formed polygon and the lines connecting them to each other are called edges. The energy at a snaxel v_j is defined as:

$$E(v_j) = E_{int}(v_j) + E_{ext}(v_j) \tag{2.1}$$

where E_{int} is the internal energy and E_{ext} is the external energy.

The internal energy E_{int} consists of two terms:

$$E_{int}(v_i) = \alpha(v_j)E_{con}(v_j) + \beta(v_j)E_{cur}(v_j) \tag{2.2}$$

where E_{con} is the continuity energy imposing a closeness constraint on the snaxels and E_{cur} is the curvature energy keeping the curve from bending. These two energies are combined with the weight factors α and β which can be set based on the nature of the problem. In the discrete representation of the snake these energies are approximated

as:

$$E_{con} \approx ||v_{j+1} - v_j||^2 \tag{2.3}$$

$$E_{cur} \approx ||v_{j+1} - 2v_j + v_{j-1}||^2 \tag{2.4}$$

Minimizing continuity energy defined in (2.4) attracts the snaxels to each other which could cause a shrinkage in the snake length. To avoid a shrinkage in the snake, a modified continuity energy is often used to keep the snaxels at a fixed distant from each other and is defined as:

$$E_{con}(v_j) = |d - ||v_{j+1} - v_j|||^2 \tag{2.5}$$

Where d is the average distance between all snake points and is computed as:

$$d = \frac{1}{N}\sum_{j=1}^{n} ||v_{j+1} - v_j|| \tag{2.6}$$

The modified continuity energy encourages the snaxels to be at a distance of d from each other i.e. favors an equidistant curve.

E_{ext} in (2.1) is the external energy forcing the snake toward image features, such as edges. In the vicinity of the edges the image gradient grows higher, therefore a negative intensity gradient is used to minimize the energy close to the target edges.

$$E_{edge}(v_j) = -|\nabla I(x, y)|^2 \tag{2.7}$$

Total energy of the snake is a weighted sum of energies of all snaxels defined as:

$$E_{total} = \sum_{j=1}^{n} (\alpha E_{con}(v_j) + \beta E_{cur}(v_j) + \gamma E_{edge}(v_j)) \tag{2.8}$$

where E_{con}, E_{cur} and E_{ext} are given in (2.4) and (2.7) respectively.

The problem of finding the target boundaries from the initial curve is now turned into an energy minimization problem. The task is achieved by finding a set of vertex position $V = (v_1, v_2, ..., v_n)$ which minimizes the energy function in (2.8).

2.2 Active Shape Models

2.2.1 Shape Model Analysis

A contour is presented by a set of n points, also called landmark points. The landmark points defining the boundary of the object are extracted from a set of N training images to construct a point distribution model (PDM) of the object. The landmark points of each training image is stored as a column vector:

$$X = [x_1, x_2, ..., x_n, y_1, y_2, ..., y_n]^T \tag{2.9}$$

N such vectors Xj are generated for various possible shapes of the target and stacked in a $2n \times N$ m matrix D as:

$$D = [X_1|X_2|...|X_N] \tag{2.10}$$

Before performing any statistical analysis to obtain a shape model, it is important to note that the shape of an object is independent from its orientation, scale and position. Therefore the generated shapes in the training set must be represented in a common co-ordinate frame. There are various approaches to align a set of shapes to a common co-ordinate frame. The popular Procrustes method [Goodall (1991)], described in details in Appendix A, aligns the shapes by minimizing sum of distance of each shape to the mean shape.

Each shape (vector) is a point in a $2n$ dimensional space, forming a distribution in the space. Having a huge training data set comprising shapes which describe a particular target, observation of some degree of correlation is expected. Calculating the inter-correlation between points can reduce the dimension of the problem. Principal Component Analysis (PCA) is a popular method to model the distribution of the points by detecting the major directions of data variation. The variation about the

mean is described by the eigenvectors of the covariance of matrix D computed as:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2.11}$$

$$S = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})^T \tag{2.12}$$

The model can generate new shapes using the equation

$$X = \bar{X} + Pb \tag{2.13}$$

where $P = [P_1|P_2|...|P_t]$ contains t eigenvectors of the covariance matrix S, corresponding to the t largest eigenvalues λ_i ($\lambda_i \geq \lambda_{i+1}$) and b is a t-dimentional vector referred as model parameters.

Any arbitrary shape X can be approximated from the mean shape setting model parameters

$$b = P^T(X - \bar{X}) \tag{2.14}$$

The variance of the i^{th} parameter, b_i, across the training set is equal to λ_i i.e. the i^{th} eigenvector of the covariance matrix S. To allow only generation of plausible shapes similar to the ones in the training set, the parameters b_i is restricted to $\pm m\sqrt{\lambda_i}$, where m is usually between 3 to 5.

2.2.2 Data Appearance Model

Given a set of parameters b, a shape model roughly describing the boundaries of the object is available. From image processing point of view, there exist several ways to find better parameter values to have a better fit of the model to the target. This can be achieved by searching in a local region around each model point to find an alternative model point that moves the model closer to the target boundary. A simple way to attain this is by searching along a perpendicular line to the current point and moving to the point with strongest edge, as visualized in Figure 2.1.

However the intensity profile along the normal is not always as shown in Figure 2.1. Several local maxima and minima along the normal profile corresponding to other

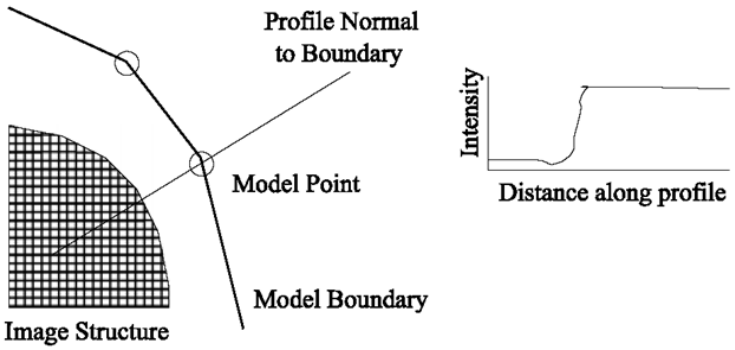


Figure 2.1: Fitting the model points to the target boundary

image structure around the target might exist. To describe the image structure around each landmark point a statistical model of the gray-level structure perpendicular to the contour is trained.

The normal vector to the contour at a model point v_j is approximated as:

$$N(v_j) \approx \frac{v_{j+1} - v_j}{||v_{j+1} - v_j||} + \frac{v_{j-1} - v_j}{||v_{j-1} - v_j||} \tag{2.15}$$

and normalized to have a unit normal vector at each model point:

$$N(v_j) = \frac{N(v_j)}{||N(v_j)||} \tag{2.16}$$

For the i^{th} training image and the j^{th} model point in the image , k pixels are sampled (using linear interpolation , See [Cootes and Taylor (1993)]) each side of the point along the normal and the image intensities at the sampled pixels are stored in a $2k+1$ column vector g_{ij} as

$$g_{ij} = [g_{ij,1}, g_{ij,2}, ..., g_{ij,2k+1}]^T \tag{2.17}$$

Cootes and Taylor (1993) propose to use the normalized first derivatives of pixel intensities along the normal to reduce the effect of global intensity change. The normalized derivative intensities are stored as:

$$dg_{ij} = [dg_{ij,1}, dg_{ij,2}, ..., dg_{ij,2k+1}]^T \tag{2.18}$$

where

$$dg_{ij,1} = dg_{ij,2} - dg_{ij,1} \tag{2.19}$$

$$dg_{ij,k} = dg_{ij,k+1} - dg_{ij,k-1} \qquad k = 2, 3, \ldots, 2k \tag{2.20}$$

$$dg_{ij,2k+1} = dg_{ij,2k+1} - dg_{ij,2k} \tag{2.21}$$

and normalized as

$$dg_{ij} = \frac{dg_{ij}}{\sum_l dg_{ij,l}} \tag{2.22}$$

The model is trained for all N images available from the training set. For a model point v_j the gray-level derivative matrix dg_j is constructed from the gray-level derivative vectors computed from each training image.

$$dg_j = [dg_{1j}, dg_{2j}, \ldots, dg_{Nj}]^T \tag{2.23}$$

Mean and covariance of the (2k+1)-by-N matrix dg_j forms the statistical model describing the model point v_j:

$$\bar{dg}_i = \frac{1}{N} \sum_{i=1}^{i=N} dg_{ij} \tag{2.24}$$

$$Sg_j = \frac{1}{N-1} \sum_{i=1}^{i=N} (dg_{ij} - \bar{dg}_i)(dg_{ij} - \bar{dg}_i)^T \tag{2.25}$$

Note that dg_j is a $(2k+1) \times 1$ column vector and S_g is a $(2k+1) \times (2k+1)$ matrix. The quality of fit of a new sample, dg_s , to the model point v_j is given by

$$f(dg_s) = (dg_s - \bar{dg}_j)^T Sg^{-1} (dg_s - \bar{dg}_j) \tag{2.26}$$

which is the Mahalanobis distance from the sample to the mean model.

From a mathematics point of view, the covariance matrix in the equation above might be singular i.e. not invertible. Matrix S_g can be factored using eigenvalue decomposition to

$$S_g = Q \Lambda Q^T \tag{2.27}$$

With Q and Λ being the eigenvectors and eigenvalues of S_g, respectively.

Equation (2.26) can be rewritten as:

$$f(dg_s) = (dg_s - \bar{dg}_j)^T Q \Lambda^{-1} Q^T (dg_s - \bar{dg}_j) \tag{2.28}$$

The projection of $(dg_s - \bar{dg}_j)$ onto all eignvectors present in Q is computes as:

$$b_g = Q^T(dg_s - \bar{dg}_j) \tag{2.29}$$

Note that b_g is a $2k+1$ column vector and Λ is a $(2k+1) \times (2k+1)$ diagonal matrix with eigenvalues of S_g on the diagonal.

By inserting (2.29) into (2.28), the Mahalanobis distance can be written as:

$$f(dg_s) = b^T \Lambda^{-1} b = \sum_{i=1}^{2k+1} \frac{b_{g,i}^2}{\lambda_{g,i}} \tag{2.30}$$

The mathematical computation derived above is actually the Principal Component Analysis of the gray-level appearance model. The key idea to deal with the singularity of the S_g matrix is to use PCA to produce a compact model of the allowable variations in the gray-level profile. Using only a limited number of eigenvectors t_g, (2.30) can be approximated as [Cootes and Taylor (1993)]:

$$f(dg_s) \approx \sum_{i=1}^{t_g} \frac{b_{gi}^2}{\lambda_{gi}} + \frac{2R^2}{\lambda_{tg}} \tag{2.31}$$

where R^2 is the sum of squares of differences between the new sample and the reconstruction of this sample and can be computed as:

$$R^2 = \sum_{i=t_g+1}^{2k+1} b_{gi}^2 \tag{2.32}$$

For a new frame, given that an initial shape model in known (that can be the mean shape, or the contour from the previous frame), n_s pixels are sampled each side of the j^{th} model point along the normal. The $2n_s + 1$ points (including the current model point) are the candidate locations where the model point can move to. For each candidate point a $2k+1$ pixel derivative gray-level profile centered to that candidate point, is computed along the normal and the quality of fit of each candidate location is computed using (2.31). The point with highest fit quality (minimum Mahalanobis distance) will be the new position of the j^{th} model point.

In the traditional ASM [Cootes and Taylor (1993)], each landmark point moves independently to the best matching location. The independent displacements of adjacent

points can result in an outlier configuration. Behiels et al. (1999) uses a Dynamic Programming method for the points displacement based on minimizing the Mahalanobis distance as well as the displacement between the adjacent points. In his algorithm the the curve movement is carried out by minimizing the function

$$f(X) = K_{1,x_1} + \sum_{i=2}^{n}(K_{i,x_i} + \alpha|x_i - x_{i-1}|) \tag{2.33}$$

where K is a $n \times l$ matrix containing the Mahalanobis distance at every l pixel positions along the normal for all n landmark point and $X = (x_1, x_2, ..., x_n)$ are the displacements in pixel units along the normal to the curve.

Once a shape model and an appearance model are trained for the target object, the target boundary a segmented for a new frame using the following steps:

1. Examine normal to the contour at each landmark point to find the point best quality fit (minimum Mahalanobis distance).
2. Update the model parameters b to best fit the new found contour.
3. Put a restriction on parameters b to ensure plausible shapes.
4. Repeat until convergence.

2.3 Point Tracking Using Optical Flow

The point tracking task simply is: given a point in frame t, estimate its location in frame $t+1$. Optical flow techniques calculate the motion between two images at times t and $t+dt$. *Brightness constancy constraint* assumes that the intensity of image does not change from time t to $t + dt$. Following this constraint, the motion flow of a 2D image can be approximated as:

$$I(\mathbf{x}, t) \approx (\mathbf{x} + d\mathbf{x}, t + dt) \tag{2.34}$$

where $I(\mathbf{x}, t)$ is image intensity at pixel position $\mathbf{x} = [x, y]$ at time t. Expanding the equation using Taylor series and taking only the first order term, (2.34) is written as:

$$I_x(x, y).u + I_y(x, y).v + I_t(x, t) = 0 \tag{2.35}$$

where I_t is the temporal derivative , I_x and I_y are spatial derivatives , u and v are motion vector in x and y direction respectively. Equation (2.35) is insufficient to determine two unknown motion vectors. In the Lucas and Kanade (1981) method the motion vectors are assumed to be constant in a $K \times K$ neighborhood of pixel p. This yields a system of K^2 equations for 2D-velocity vector at pixel p and can be written as :

$$Af = b \tag{2.36}$$

Where A is a $K^2 \times 2$ matrix containing the spatial gradients of pixels in the $K \times K$ window, $f = [u, v]^T$ and b is a $K^2 \times 1$ vector containing the temporal gradient of all the neighborhood pixels. Minimum least square error solution of (2.36) is obtained by minimizing $||Ad - b||^2$ (in a Least Squares sense):

$$f = -M^{-1}d \tag{2.37}$$

where

$$M = A^T A = \begin{bmatrix} \sum I_x^2 & \sum I_x I_y \\ \sum I_x I_y & \sum I_y^2 \end{bmatrix} \tag{2.38}$$

$$d = A^T b = \begin{bmatrix} \sum I_x I_t \\ \sum I_y I_t \end{bmatrix} \tag{2.39}$$

And the summation is on all the pixels in the selected neighborhood.

The optical flow vectors obtained from (2.37) are valid for an invertible and well-conditioned matrix M. In the point tracking application of optical flow, the points with a valid motion vector are chosen for tracking. A point is called valid if its forward and backward trajectories does not differ significantly. These methods are based on the assumption that a correct tracking is independent from the direction of the flow. Simoncelli et al. (1991) shows that the uncertainty in the calculation of (2.35) decreased in proportion to the trace of M.

Various optical flow based contour tracking have been proposed. Yokoyama and Poggio (2005) used features with strong magnitude of the spatial gradient, hence more reliable motion vectors, for object detection and tracking. Fu et al. (2000)

proposed an adaptive motion snake for tracking the visible boundaries of an object. In his method an optical flow based energy term was to the active contour model to motivate the curve toward reliable motion vector estimation.

Chapter 3

VOCAL TRACT BOUNDARY TRACKING

3.1 Training a Shape Model for VT Boundary

To model the movements of the vocal tract, the tract airway-tissue boundaries are manually extracted for two male and two female speaker, from a set of training images. The original MR-images available in the USC-TIMIT database are 68×68 pixels on a 20×20 cm frame (approximately 2.9 mm pixel width) . We use a bi-cubic interpolation to increase the resolution of the frames by 6 times to obtain 408×408 pixel images (approximately 0.48 mm pixel width). The interpolated images visually look smoother and better for manual segmentation. Manual extraction is carried out by choosing 30 equidistant landmark points on the lower profile and on the upper profile of the vocal tract, as shown in Figure 3.1. Therefore each lower and upper boundary of the vocal tract is represented as a 60-dimensional vector. Totally 1400 frames were manually segmented, of them 1000 were used for training (speakers M4 and F1) and the rest were used for performance evaluation (speakers M1, M4, F1 and F3). These ground truth are labeled manually by clicking on the desired landmark points along the VT contours, using a Matlab software, Therefore the points might not be perfectly equidistant. An automatic process is applied to get a set of equidistant points from the glottis to the lips, for all of the labeled frame, Figure 3.2a. To obtain the set of equidistant contours, the curve is interpolated with a cubic spline to get a 100 point description of the contour and then 30 equidistant points points are sampled on the curve. The collected data are aligned to each other prior to training the VT shape model. All the processed MRI videos contain a mid-sagittal view of the vocal tract and no significant orientation such as head movements are observed. However the vocal tract size and position in the recorded videos might vary from one subject to another.

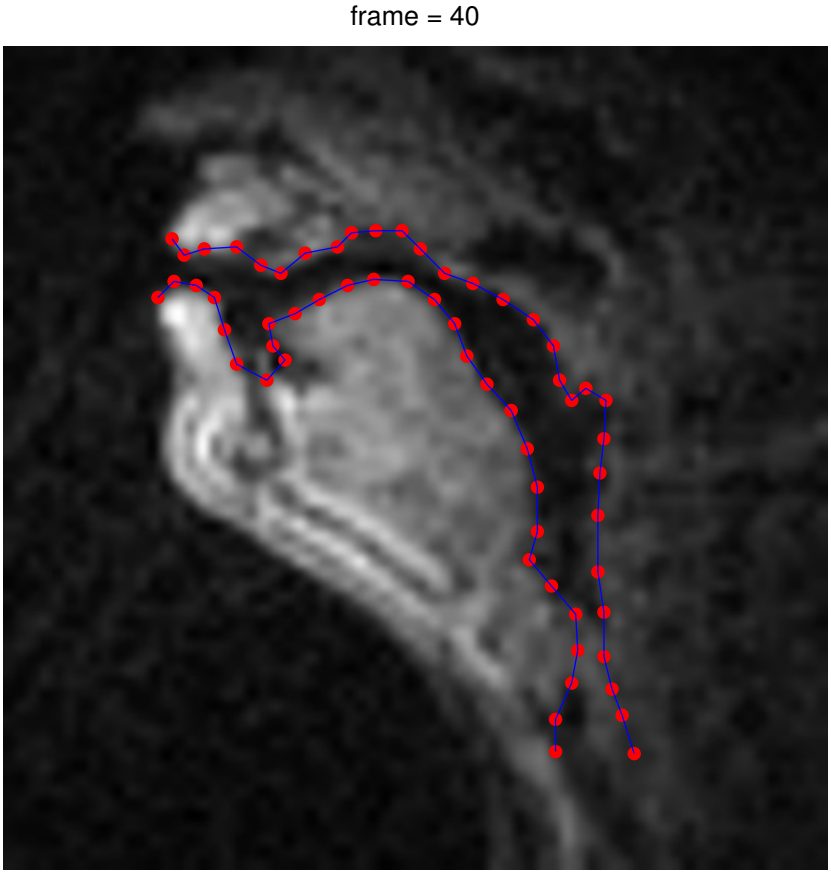
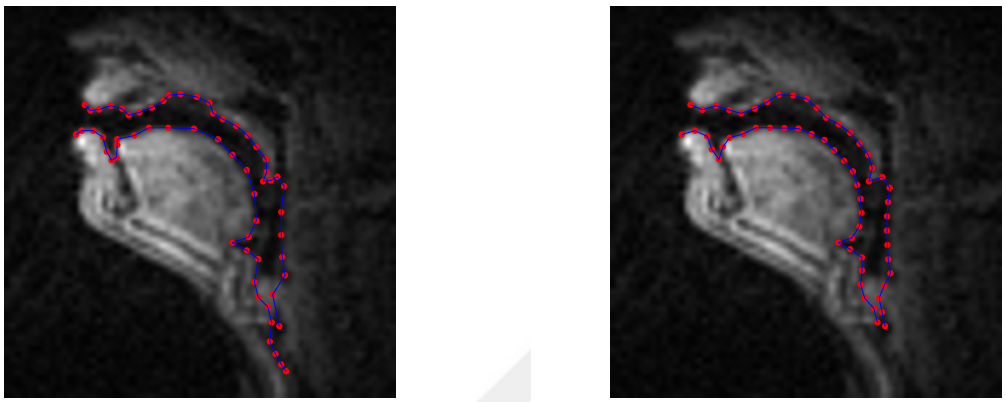


Figure 3.1: Manual segmentation: 30 points are labeled on each lower and upper tract boundaries (red points)

More specifically the VT length of an adult male is 17.5 cm on the average, while for an adult female that is less(around 16cm). Therefore alignment of the shapes into a common co-ordinate frame is necessary to eliminate the changes in tract size and position. Procrustes analysis [Appendix A] is used for the alignment purpose. Figure.3.3 shows an example of aligning two sample vocal tract shapes to each other using the Procrustes method.

We utilize a PCA method on the aligned training data to to train a model for the vocal tract shape. Analysis show that a 21 model parameter for the lower tract and



(a) Initial labeled contours (b) Equidistant point contours

Figure 3.2: Equalizing distance between landmark points

16 parameters for the upper tract are sufficient to explain 98% of the data variation in the training set. Fewer Parameters can be used to describe less variations of the shapes in the training set. The number of model parameters to explain different portions of the training set variance is given in Table 3.1. Despite the complex structure of the vocal tract, the anatomy of articulators puts a constraint on the movements and deformation. Putting a restriction on the model parameters b ensures generation of only plausible tract shapes. Figures 3.4 and 3.5 shows the deformation of the first (main) four modes of variations from the mean shape for lower and upper VT contours, varying the parameters between ± 3 and ± 2 of their standard deviation, respectively. The model for the upper boundary is restricted more to allow less shape variations, since the upper VT profile is not as dynamic as the lower one.

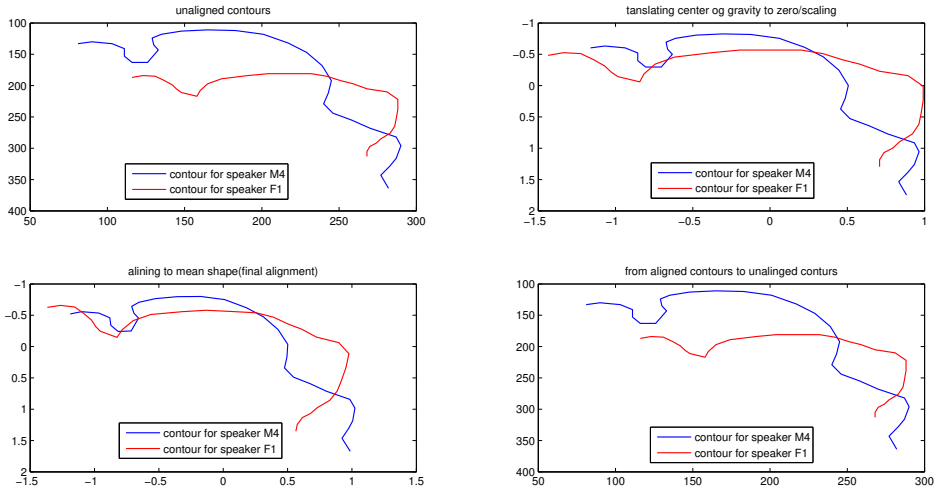


Figure 3.3: Shape alignment

Table 3.1: Number of shape model parameters required to cover different portions of the total variance.

	92%	**96%**	**98%**	**99%**
Upper	8	11	16	23
Lower	7	13	21	29

3.2 Vocal Tract Appearance Model Training

The gray-level profile around a VT contour landmark shows complex intensity variations due to the low quality MRI videos. Therefore training of an statistical model is necessary to describe the image structure around the VT air-way tissue boundary. The model is trained for each landmark point using the manually extracted contours as described in section 2.2.2. For each landmark point, $k = 8$ pixels are sampled each side of the point along normal to contour at that point. The mean ($\bar{dg}$) and

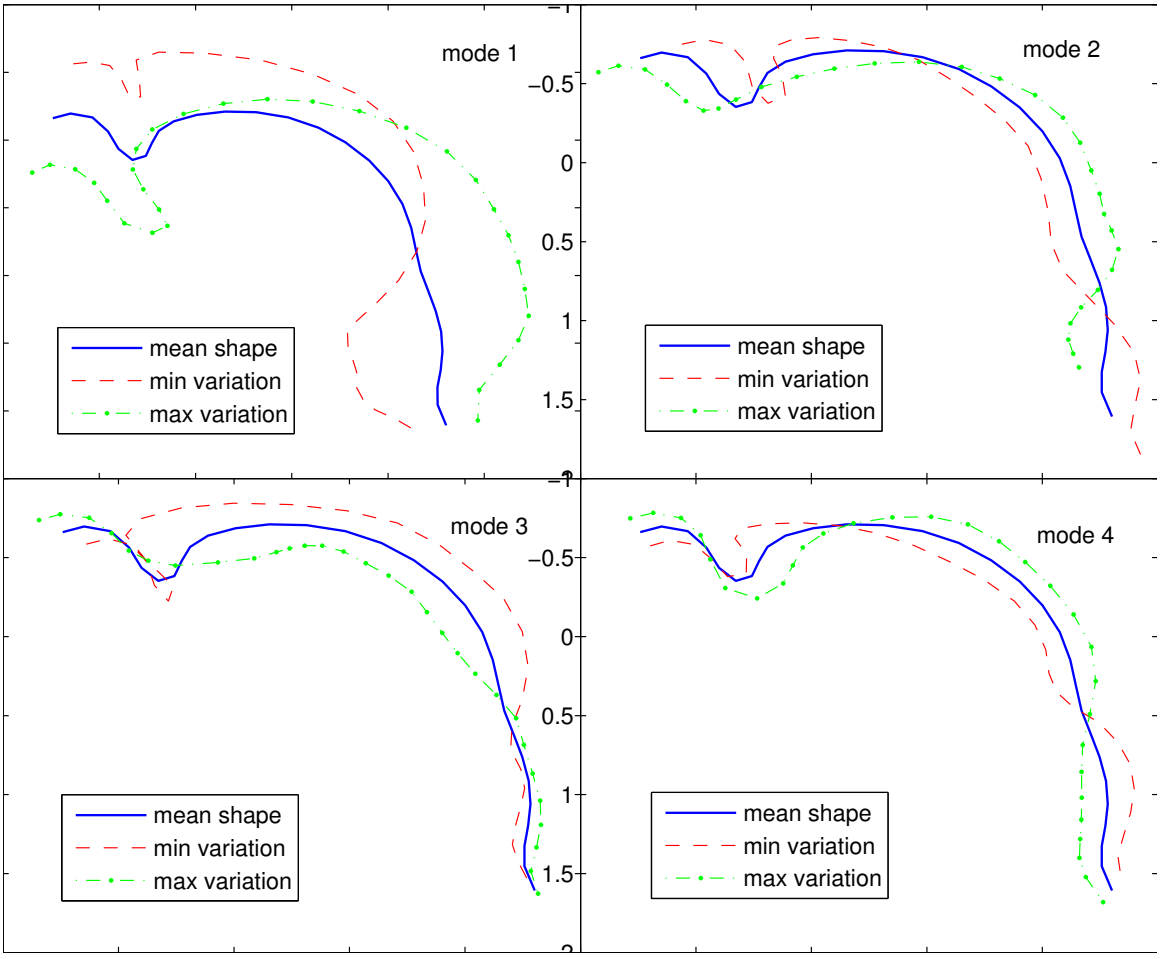


Figure 3.4: Main modes of variation for lower boundary between ±3 of the standard deviation

the Covariance (S_g) of the derivative gray-level profile form a statistical model for the appearance of the vocal tract contours. However due to the high order dimension of the covariance matrix which could possibly lead to its singularity, the variations in the extracted gray-level appearance is analyzed using PCA. The eigenvectors of the covariance matrix S_g are truncated to explain 98% of the variation in derivative profiles. The analysis reveals that from the 17 eigenvectors here, only $t_g = 7$ of them corresponding to the seven largest eigenvalues, are sufficient to describe 98% of the data. The Mahalanobis distance to measure the quality of fit of a new instance to the trained model, is computed using (2.31).

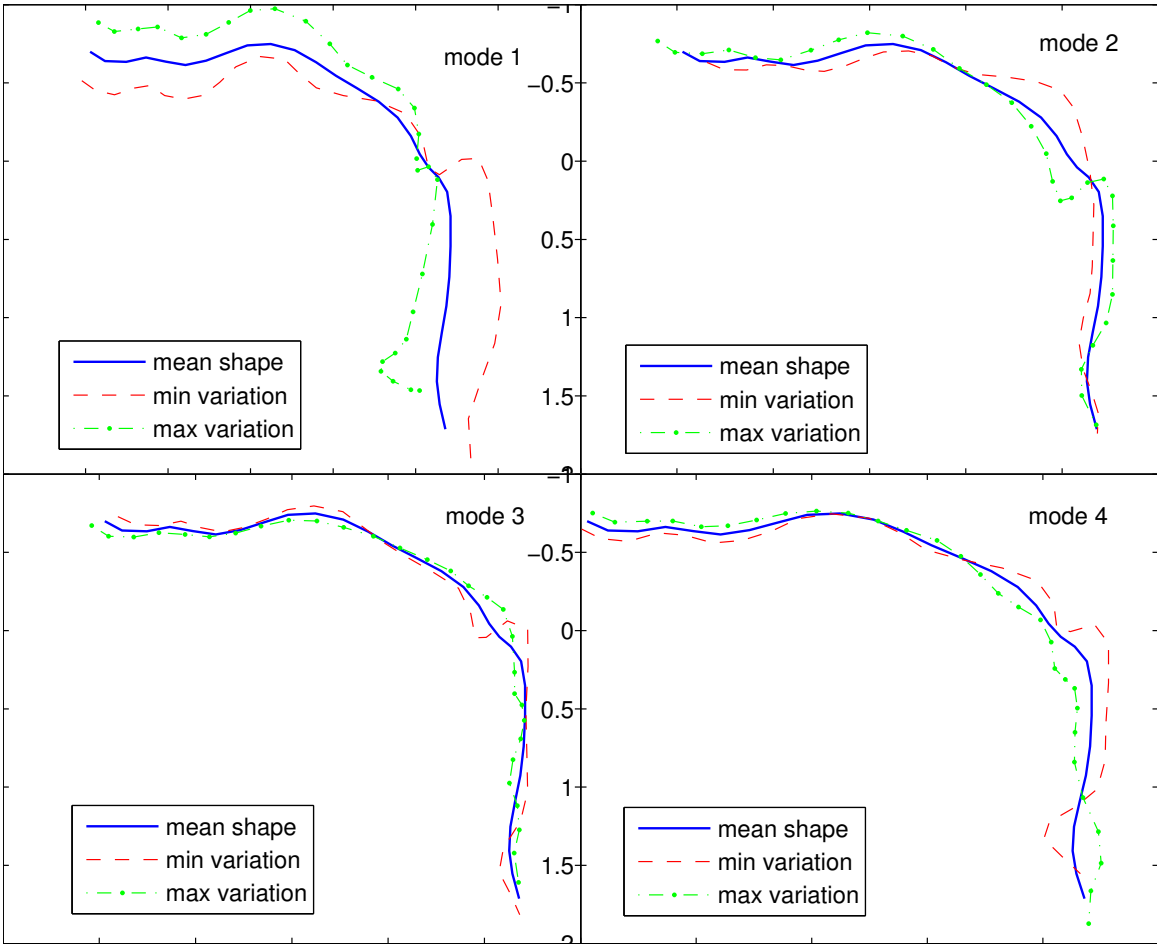


Figure 3.5: Main modes of variation for upper boundary between ±2 of the standard deviation

3.3 Proposed Boundary Estimation Algorithm

To fit the trained model to a new image, an initial estimation of the vocal tract contour is needed. For any MRI video to be segmented, we initialize the lower and upper contours at the first frame and the the upcoming frames are segmented automatically. However having a set of pose parameters for un-aligning the model mean shape to the observed frame space, the un-aligned mean shape can be set as the initial contour as well. Two example of the estimation of tract lower boundary from the mean shape is shown in Figure 3.10 and 3.11. Despite the considerable differences between the

mean shape and the observed frame, the algorithm well-detects the boundary. This figures are included here just to emphasize the robustness of the algorithm and for the rest of evaluations, We initialize the contour at the first frame.

Given an initial position of the model points at a frame, the VT-boundaries are estimated via some iterations as visualized in Figure 3.6. At each iteration, the VT contours evolve in a local neighborhood to minimize an energy function as defined in (3.1) via Dynamic Programming. For each model point we set the local search neighborhood to be $ns = 5$ pixels either side of the point along the normal to contour. To allow only plausible vocal tract shapes, the estimated contour at the end of each iteration is projected into the trained shape model space using (2.14) for a model parameters range check. The lower and upper boundary contours are restricted to ± 3 and ± 2 of their standard deviation, respectively.

If the process converges, an initial contour for the next frame is predicted from the estimated contour at the current frame, using optical flow. Convergence is checked when no significant change happens in any of the contour vertices or the iteration reaches its maximum allowed number (here set to be 10).

3.3.1 The Energy Function

Following a path as in [Behiels], we carry out the points displacements process not only with respect to the Mahalanobis distance, but also employing some extra energies. The curve evolution is executed by minimize an energy defined as

$$f(V) = \sum_{j=1}^{n} (E_{mah}(v_j) + \alpha E_{con}(v_j) + \gamma E_{edge}(v_j) + \delta E_{mot}(v_j)) \tag{3.1}$$

where $V = (v_1, v_2, ..., v_n)$ is a set of points minimizing the function f; E_{con} and E_{edge} are the traditional snake continuity and external energy, E_{mah} is the Mahalanobis Distance energy moving the curve to maximum fit quality locations and E_{mot} is the motion energy tending the points toward more reliable motion vector estimation positions.

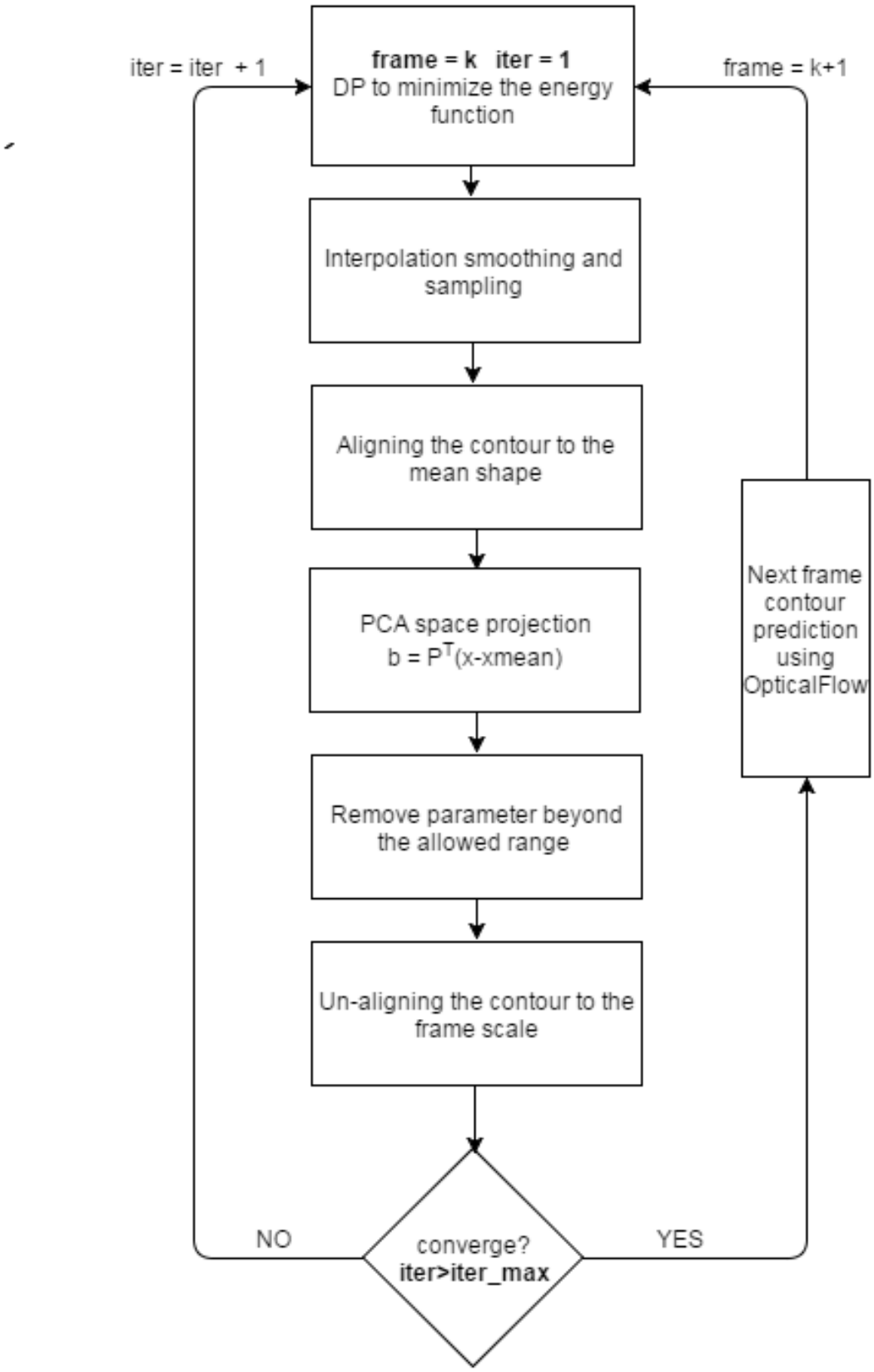


Figure 3.6: Proposed vocal tract boundary tracking algorithm

We excluded the classical curvature energy in here mainly for two reason: (1) The curvature energy prevents the curve from bending thus the snake can not follow the concave regions such as air-way region under the tongue tip and near the lower incisors. Therefor to cope with the complex structure of the tract, the weight of curvature energy β should be as small as possible. The effect of a very small curvature energy weight can be compensated with a bit larger continuity energy which also control the bending of the curve up to a point. (2) Excluding the curvature energy turns a triple potentials DP problem into a pair-wise potentials problem, since two edges are required to compute the curvature at a point. As a consequence we reduce the the complexity of the problem by turning a $O(nm^3)$ order problem to a $O(nm^2)$ order, where n is the number of contour points (here 30) and m is the number of search points (here 11).

Examining the results of tracking the VT boundary with the curvature energy excluded shows that the general shape of the VT boundary is well estimated in the tracking process, however the contours may not seem smooth enough. To have smoother curves we interpolate the snake at each iteration with a cubic spline and sample 30 equidistant points on the interpolated curve to represent the smoothed curve. Figure 3.7 shows the result of tracking for a sequence of images using the classical curvature energy vs the proposed interpolation smoothing method.

The E_{edge} and E_{mah} energies attract the curve toward the image edges; the first one only based on the intensity gradients of the current frame and the later according to the appearance model trained for the vocal tract. The accuracy with which the active shape model can locate the boundaries of the target is constrained to the trained model. The model may not fit to the boundaries if some variations of the tract shape are not included in the training process. Hence the E_{edge} energy is included to compensate a possible imperfection in the trained model.

The motion energy, E_{mot}, is applied to encourage a movement toward the points with reliable motion vectors. The motion vector at a point v_j in frame k is called reliable and will translate v_j to a new location $\hat{v_j}$ in frame $k+1$, if backward tracking of the

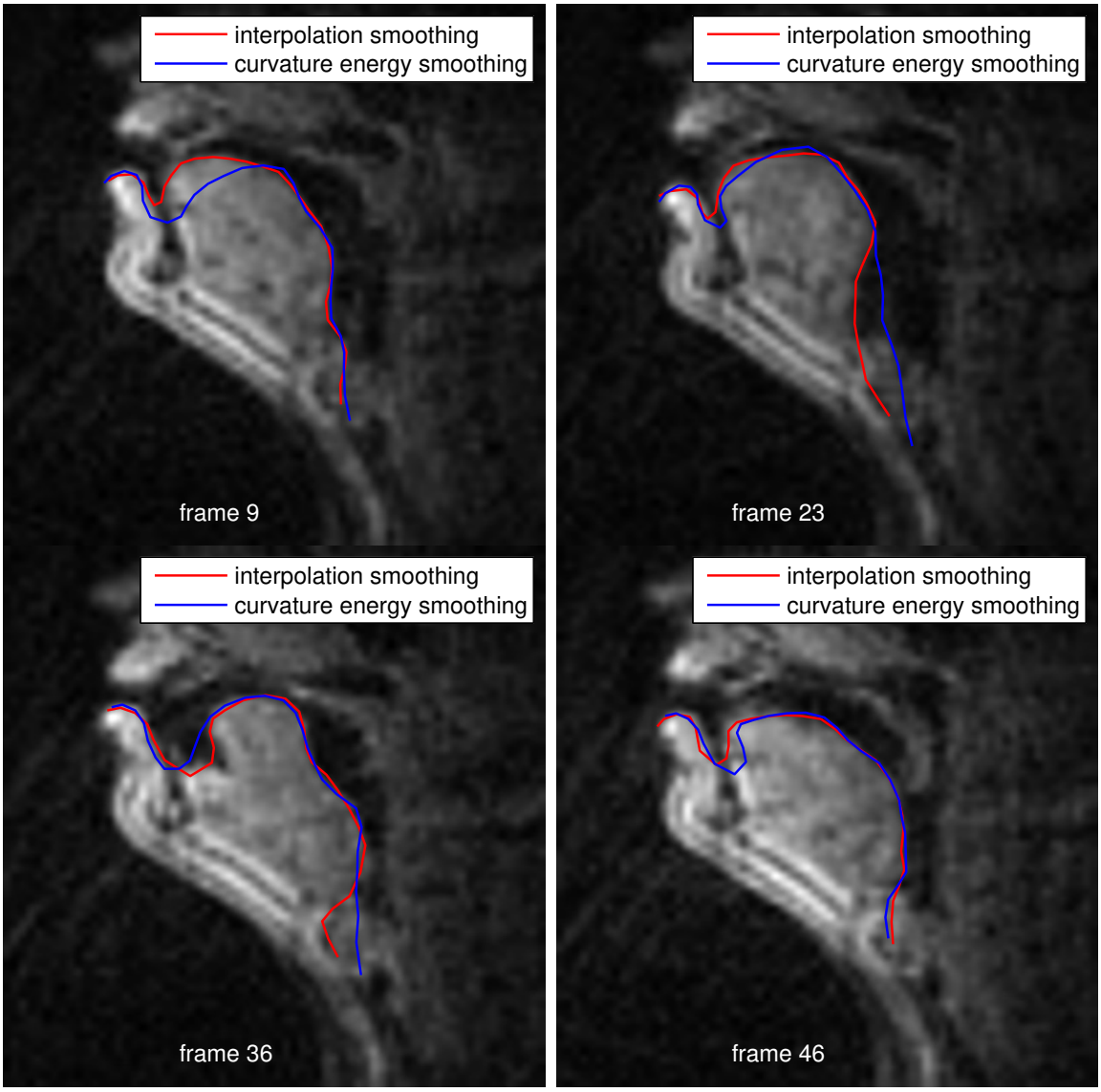


Figure 3.7: Interpolation smoothing tracking vs curvature energy tracking

point $\hat{v_j}$ yields a close point to the original point v_j. The distance between the point v_j and the computed backward tracked point $\hat{v_j}$ is called bidirectional error.

$$BE(v_j) = ||v_j - \hat{v_j}|| \tag{3.2}$$

where

$$\hat{v_j} = (v_j + f_j) + b_j \tag{3.3}$$

with f_j being the forward motion vector (from frame k to frame $k+1$) at point v_j and b_j being the backward motion vector (from frame $k+1$ to frame k) at point $v_j + f_j$.

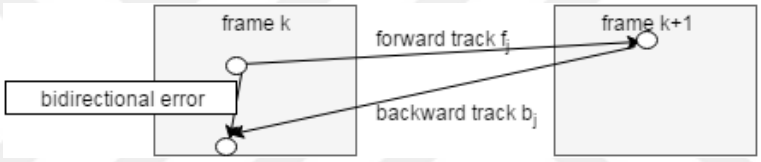


Figure 3.8: Bi-directional error

A motion energy is defined to minimize the squared bidirectional error for each snake point during snake evolution. This energy motivates the snake toward points with more reliable motion vectors and is defined as:

$$E_{mot}(v_j) = \delta||v_j - \hat{v_j}||^2 \tag{3.4}$$

We compute the optical flow vectors in a 5×5 window around each point.
The total energy function in (3.1) is then minimized using Dynamic Programming.

3.4 Dynamic Programming

DP is used to find a set of *globally* best positions for a minimal energy contour [Amini et al. (1990)]. Since the internal energy is a local property of the neighboring vertices, the total energy function can be witten as:

$$E_{total}(v_1, v_2, ..., v_n) = \sum_{j=1}^{n-1} E_j(v_j, v_{j+1}) \tag{3.5}$$

Where

$$E_j(v_j, v_{j+1}) = E_{con}(v_j, v_{j+1}) + E_{edge}(v_j) + E_{mah}(v_j) + E_{mot}(v_j) \tag{3.6}$$

Hence the energy of the snake is decomposed into the energy of $(n-1)$ segments forming the snake as shown in Figure 3.9. Energy of each segment,E_i, comprises the edge, Mahalanobis and motion energy at the start point of the segment and the continuity energy between the start point and end point of the energy.

The Dynamic Programming computes the optimal position of the predecessor, at each vertex and for all of the candidate positions of that vertex. The problem is solved by defining intermediate variables S_i as (See Amini et al. (1990))

$$S_2(v_2) = \min_{v_1} E_1(v_1, v_2) \tag{3.7}$$

$$S_3(v_3) = \min_{v_2}(S_2(v_2) + E_2(v_2, v_3)) \tag{3.8}$$

$$S_4(v_4) = \min_{v_3}(S_3(v_3) + E_3(v_3, v_4)) \tag{3.9}$$

$$\vdots$$

$$S_n(v_n) = \min_{v_{n-1}}(S_{n-1}(v_{n-1}) + E_{n-1}(v_{n-1}, v_n)) \tag{3.10}$$

In the above set of equations, each $S_k(v_k)$ contains the lowest total energy for the first $k-1$ vertices of the snakes for any given v_k. Therefore the minimum energy of the snake is equal to $E = \min_{v_n} S_n(v_n)$. Once the minimum energy in the last vertex is found, the optimal location of each vertex is obtained with backward tracking.

3.5 Contour Tracking

Estimated boundaries at the frame k is a rough estimation of the boundaries for the upcoming frame $k+1$. The displacement in vocal tract contours might be remarkable from one frame to another due to the rapid changes of vocal tract, specially quick tongue deformations. To have a better initial guess for boundary segmentation in the upcoming frames, an optical flow motion estimation is applied on the estimated contour in frame k , to predict an initial contour for subsequent frame $k+1$.

Once the upper and lower air-way tissue boundaries are estimated for the current

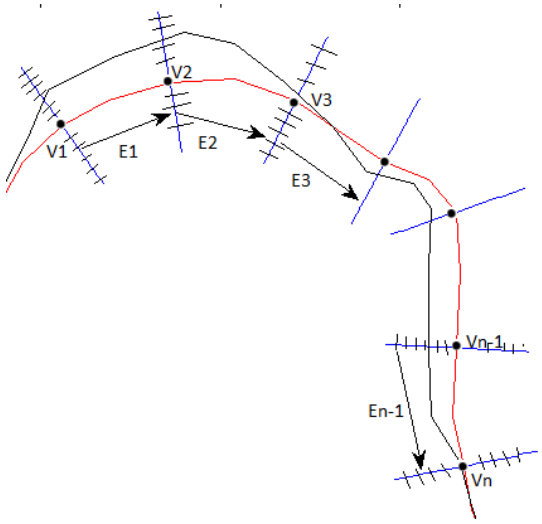


Figure 3.9: Dynamic Programming for energy minimization

frame, the contour points of this frame are updated with respect to the estimated motion vectors to predict the initial contour for the upcoming frame. Here we simply use a point tracker and predict the next frame point using:

$$\tilde{v_j} = v_j + f_j \tag{3.11}$$

where f_j is the forward motion vector at point v_j.

Results from the proposed tracking system are shown in Figure 3.12 and Figure 3.13. 25 sample frames of a 70 frame video, tracked with the proposed algorithm are given in these figures. Speaker M1 (Figure 3.12)

Figure 3.14 shows a brief comparison between the tracking results of the proposed algorithm with the method presented in Kim et al. (2014). A careful examination of the sequence reveals that the epiglottis contour and the pharyngeal wall are always intersecting in their results.

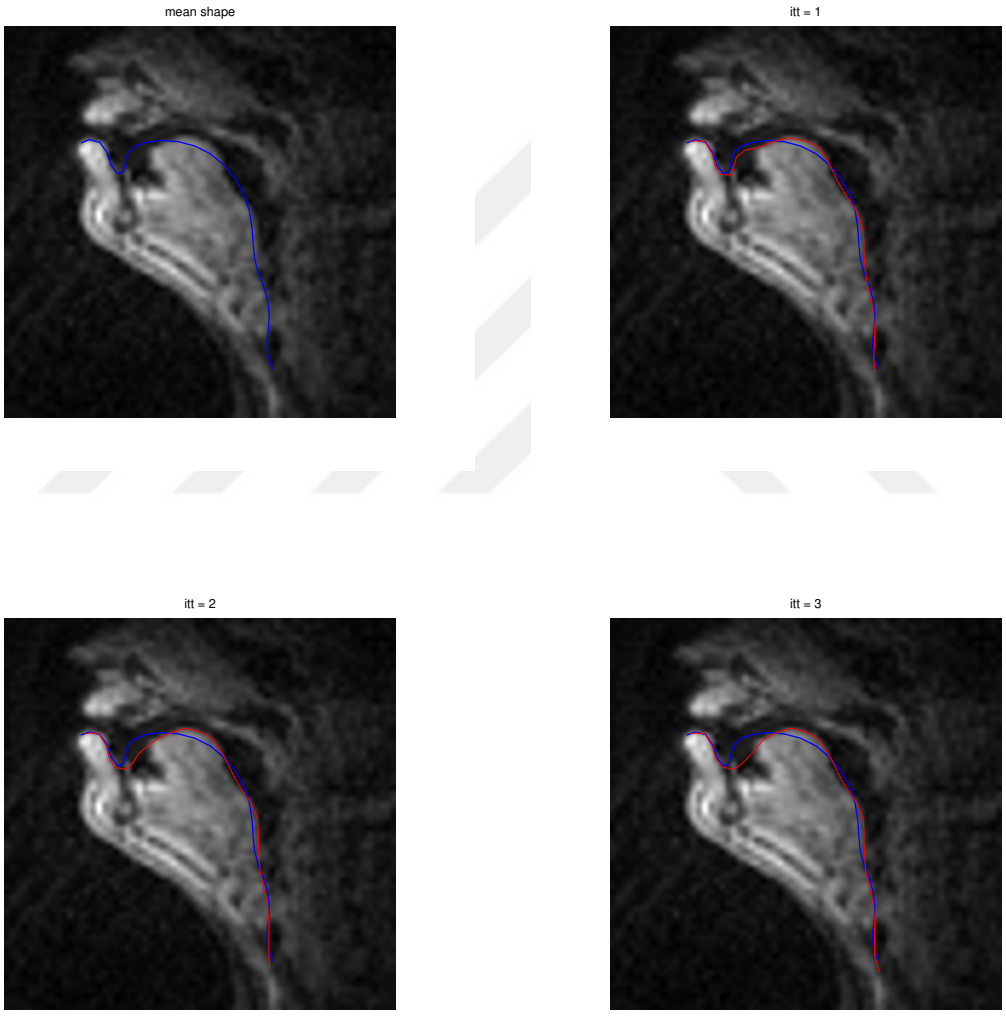


Figure 3.10: Estimating boundary from the mean shape

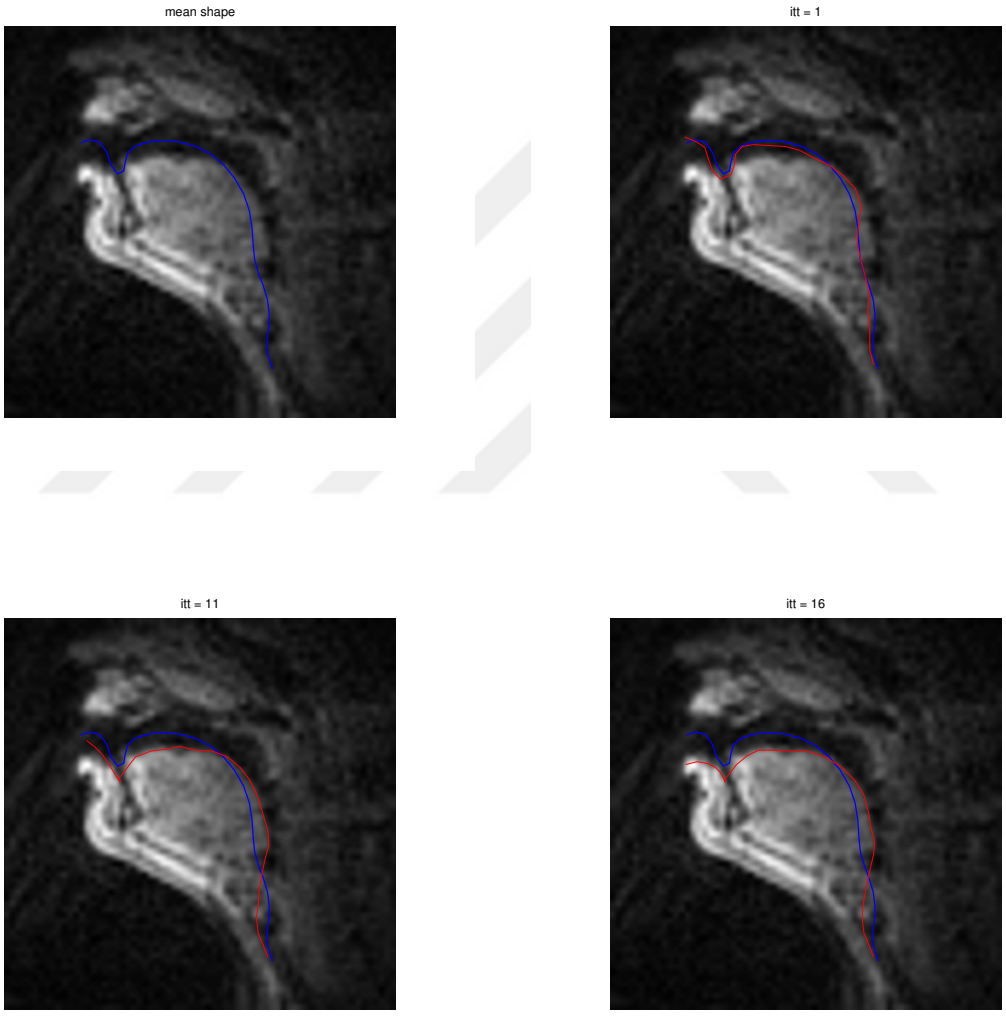


Figure 3.11: Estimating boundary from the mean shape,another example

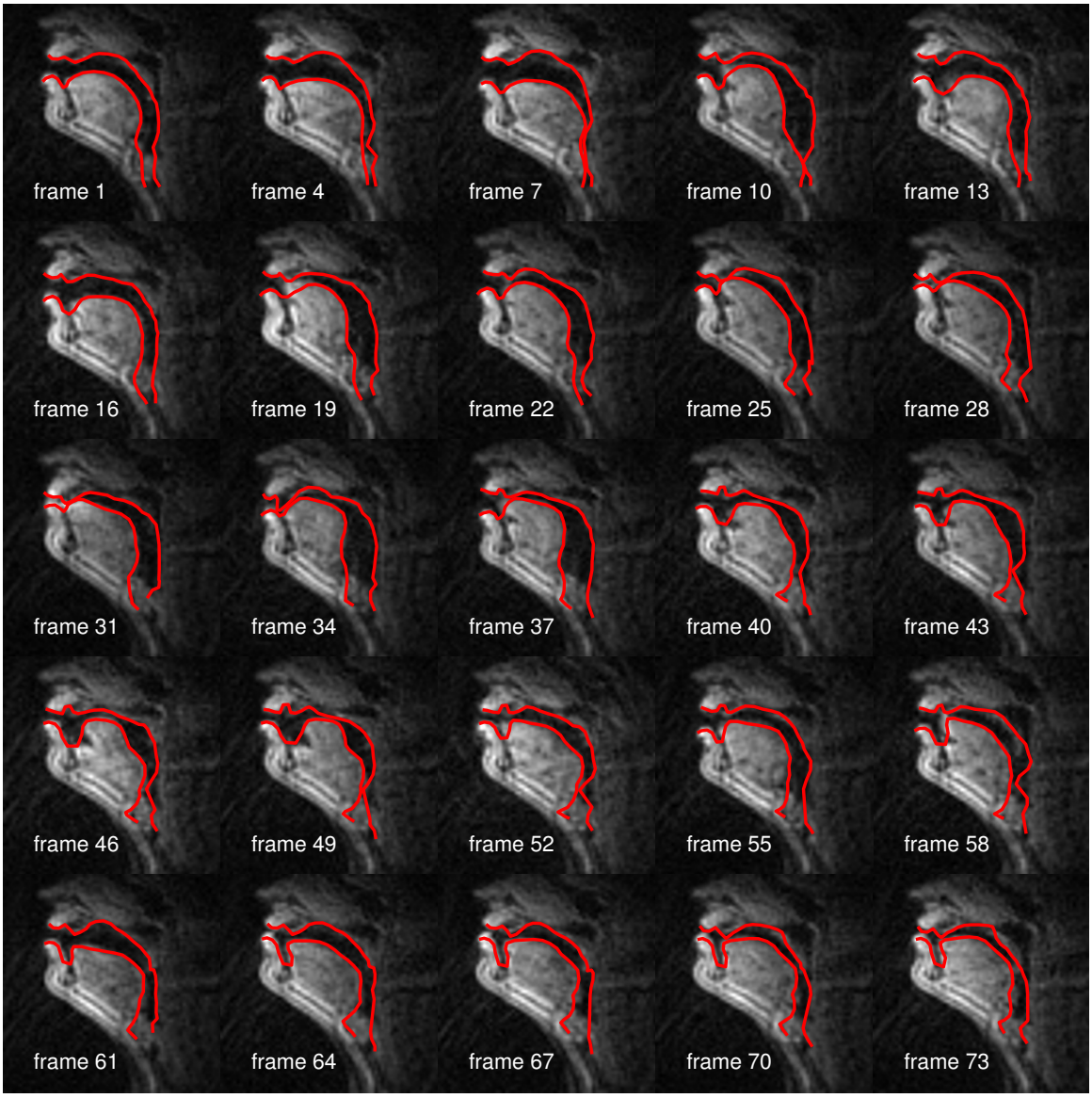


Figure 3.12: Speaker M4: Tracking results for sequence "I know I didn't meet her early enough" shown for 25 frames.

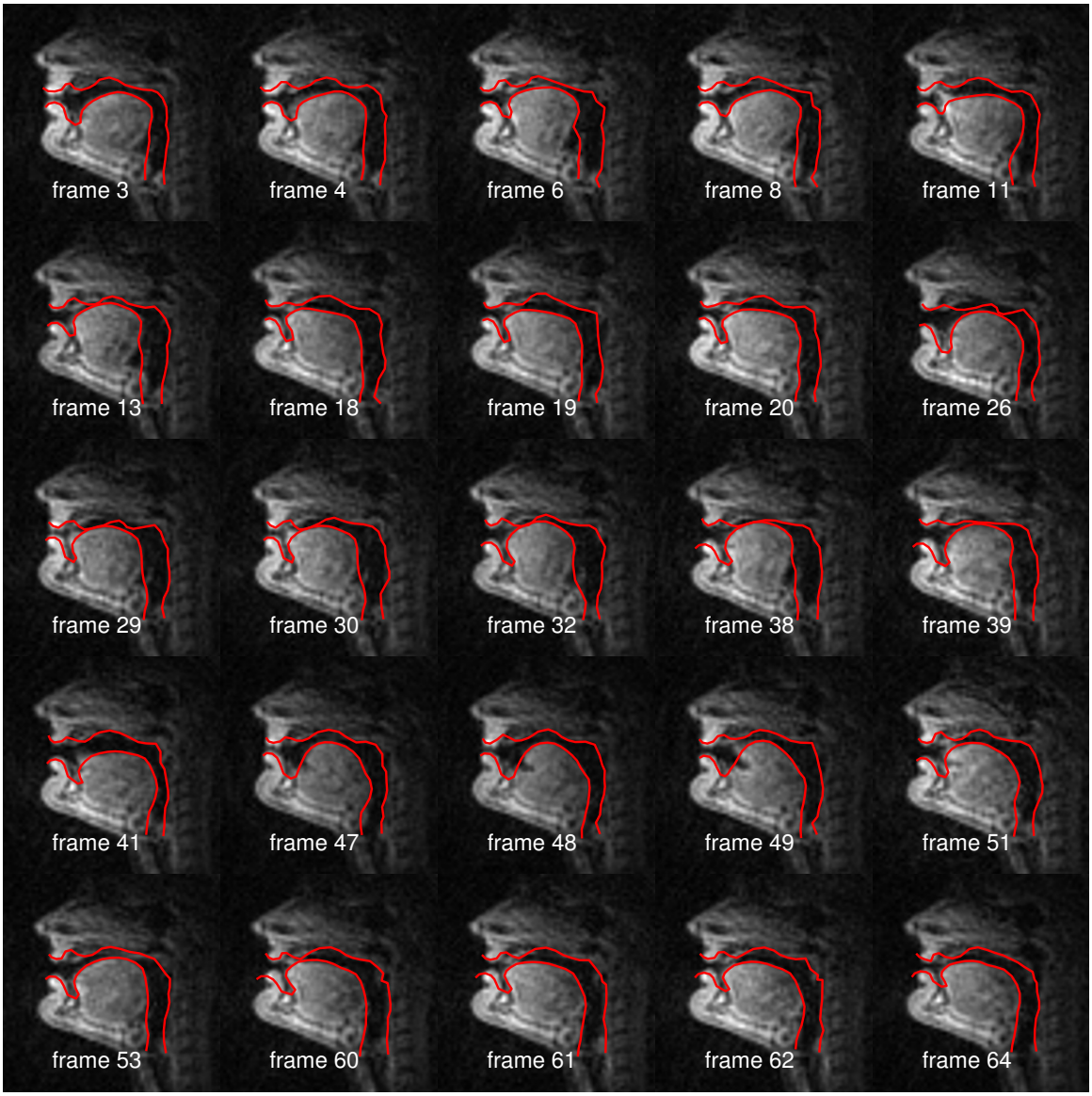


Figure 3.13: Speaker M1: Tracking results for sequence "Will you please describe me idiotic predicaments" shown for 25 random frames.

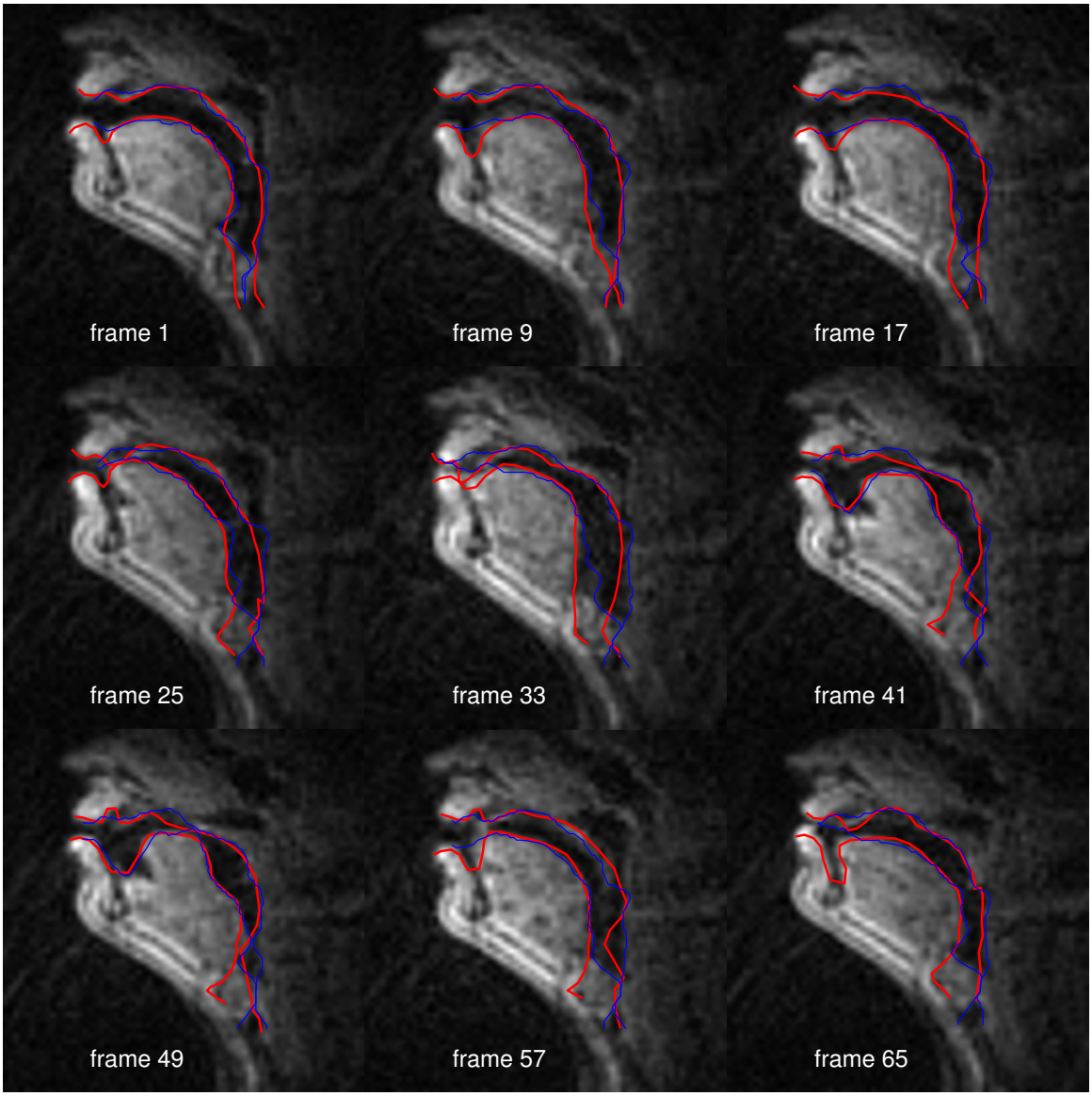


Figure 3.14: Tracking result for the same sequence as in Figure 3.12. Red: proposed method. Blue: Kim et al. (2014)

Chapter 4

EXPERIMENTAL EVALUATIONS

4.1 Error Estimation

To evaluate the performance of the proposed method, a root mean squared error analysis is done for each landmark point in a sequence of images. The error is set to be the distance between the estimated contours and the manually labeled ground truth, at each point from glottis to the lips. Following Kim et al. (2014) the Ohman grid lines system is constructed over the vocal tract and the error is evaluated for each grid line as shown in Figure 4.1.

The distance from a manually detected contour to the automatically estimated contour at any grid is computed by finding the intersection points of the grid line with both contours. The manual and automatic contours are interpolated with a cubic spline to find the intersection points with a higher accuracy.

The error analysis is performed on total 400 frames. Two videos containing 50 frames were chosen for four speakers from the USC-TIMIT database, speakers M1,M4,F1,F3. The chosen videos contain sentence with high variations in the phones involving various constriction degree and locations of the articulators also including resting frames. The vocal tract is divided into three sub-regions and the mean squared error is evaluated in each sub-region. Following Kim et al. (2014) the sub-regions are defined in a similar way as (1) grid lines 1-19 for pharyngeal region, (2) grid lines 20-72 for velar and tongue region and (3) grid lines 73-92 for labial constriction region.

We should note that the error analysis was evaluated on the original 68×68 pixel scale i.e. we down-size the estimated contours form the proposed algorithm by 6 times, to overlay the contours on the original MRI frame scale. We choose the weight factors in (3.1) as: $\alpha = 0.6, \gamma = 0.8, \delta = 0.2$.

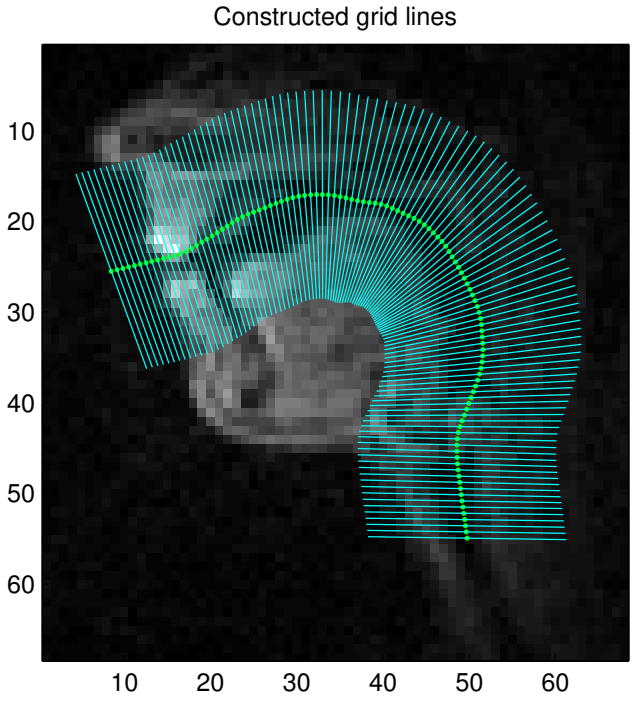


Figure 4.1: Ohman semi-polar grid lines constructed on the vocal tract. from Kim et al. (2014)

Figure 4.2 shows an error comparison between the proposed method with the method introduces in Kim et al. (2014). The results show that except the epiglottis (region 1-lower), the proposed algorithm performs with higher accuracy, specially in the tongue region which is the most dynamic articulator in the vocal tract. The computed error in the pharyngeal region is almost equal for both algorithms. The complex anatomy of the epiglottis and the poor quality of MRI videos in the pharyngeal region, are believed to be the reason of poor accuracy in the epiglottis sub-region. We should note that in the method proposed in Kim et al. (2014), the epiglottis contour frequently intersects with the pharyngeal wall. However in the ASM-ACM presented algorithm the end points of the contour are attracted to tongue root. Overall the proposed algorithm reduces the error in terms of RMSE as reported in Table 4.2.

4.1.1 Mean Sum of Distances

Another comparison between the proposed method and the algorithm presented in Kim et al. (2014) was evaluated based on a the general similarity between two contours. Mean Sum of Distances (MSD) measures the difference between two contours by computing the distances between closest vertices of each contour. The MSD between two contours $U = [u_1, u_2, ..., u_n]$ and $V = [v_1, v_2, ..., v_n]$ is defined as:

$$MSD(U,V) = \frac{1}{2n}(\sum_{i=1}^{n} \min_j |v_i - u_j| + \sum_{i=1}^{n} \min_j |u_i - v_j|) \tag{4.1}$$

The mean MSD between the contours tracked from the proposed method and the manually labeled contours is compared to the MSD of Kim et al. (2014) for a 60 frame MRI video. The MSD is measured on the original 68×68 pixel scale as well and the results are reported in table 4.3

Table 4.1: RMSE for lower and upper boundary in different sub-regions [pixel units].

Region	**Proposed method**	**Kim et al. (2014)**
1-lower	2.26	2.29
1-upper	0.94	1.28
2-lower	0.61	1.57
2-upper	0.72	0.96
3-lower	1.11	1.73
3-upper	0.85	1.27
total-lower	1.06	1.58
total-upper	0.80	1.10

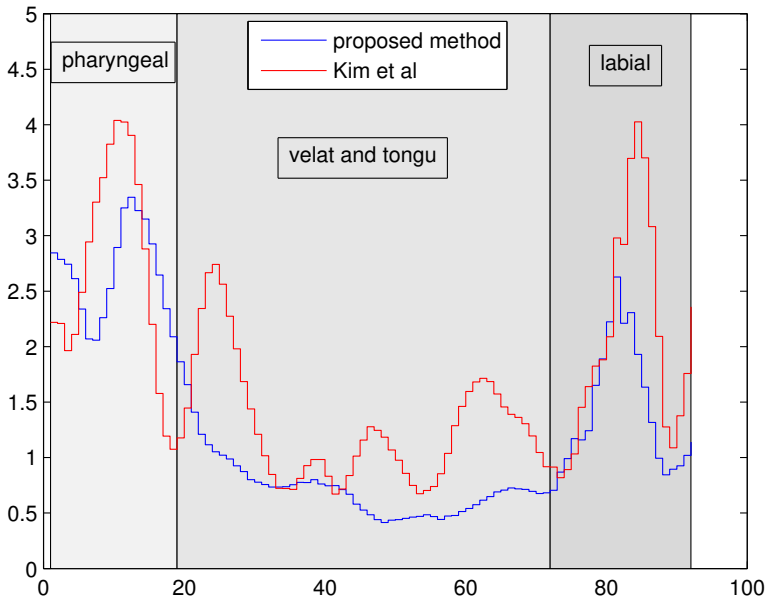
Table 4.2: MSD for lower and upper boundary [pixel units].

	Proposed method	**Kim et al. (2014)**
MSD-lower	1.16	6.21
MSD-upper	1.00	7.26

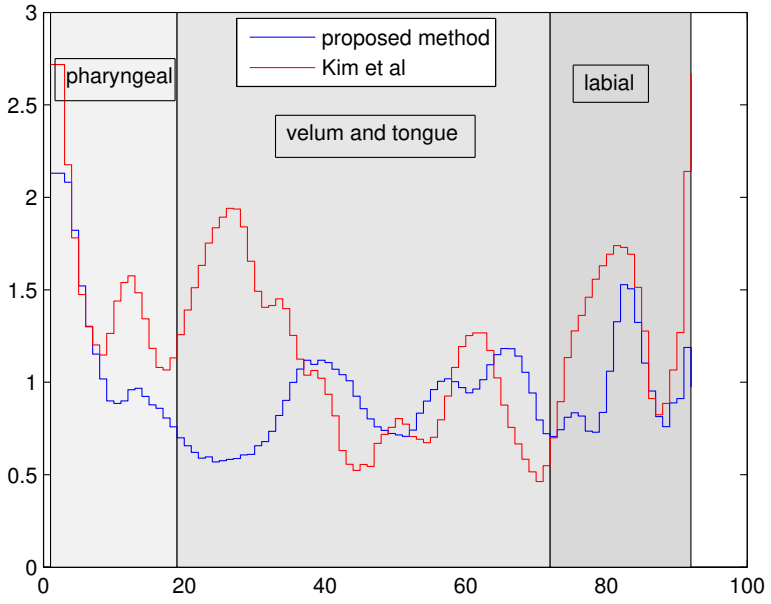
We have briefly compared our proposed algorithm to the method presented in Bresch and Narayanan (2009) over a 57 frame video. Although there is no remarkable improvement in the mean error sense, the computational complexity in the proposed algorithm is much less.

Table 4.3: RMSE and MSD lower boundary [pixel units].

	Proposed method	**Bresch and Narayanan (2009)**
RMSE	0.71	0.95
MSD	0.88	1.13



(a) Mean squared error in pixel units for each grid line, lower boundary



(b) Mean squared error in pixel units for each grid line, upper boundary

Figure 4.2: RMSE comparison between proposed method and Kim et al. (2014)

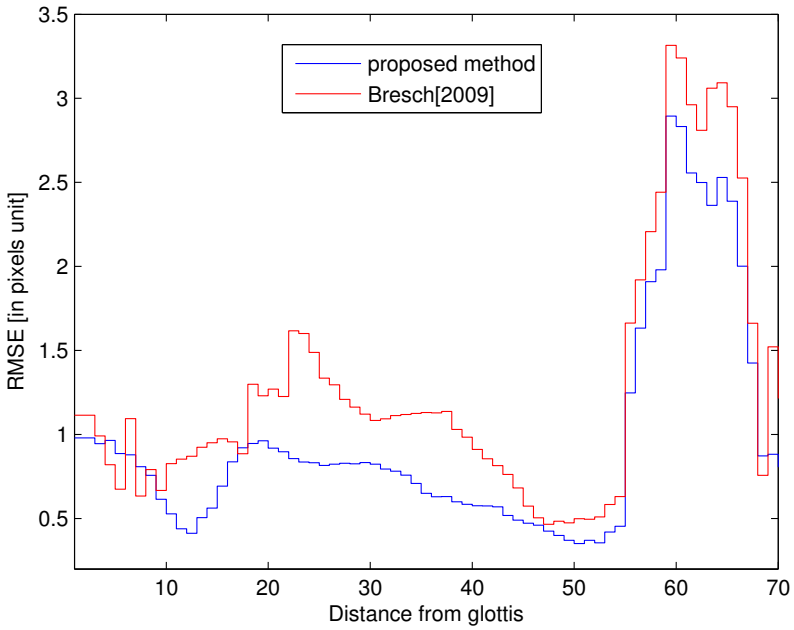


Figure 4.3: RMSE comparison for lower boundary between proposed method and Bresch and Narayanan (2009)

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this thesis we presented a robust contour tracking method applied to human vocal tract air-way tissue boundary detection and tracking in a sequence of MRI images. A parametric shape model was built for the vocal tract using statistical modeling. The vocal tract is described with 21 or 16 control parameters in the shape model for lower and upper profiles respectively, instead of the original 60-dimensional representation. The traditional active shape models were used to extract the image intensity variation around the vocal tract contours. The vocal tract contours were then detected and tracked in a sequence of images using a new-defined snake energy. The experimental analysis shows the higher accuracy of the proposed algorithm over other presented methods in the toot mean square (RMSE) and mean sum of distance (MSD) error metrics.

Among different sub-regions in the vocal tract, the tongue region was detected and tracked with the highest accuracy compared to the other available methods. The major detection errors occurred around the epiglottis region due to the poor quality of recorded videos. However we believe the error can be reduced using pre-processing methods such as image enhancement and a more precise model shape training for the epiglottis region.

5.2 Future Work

As a next step we aim to carry out the vocal tract contour tracking in a multi-modal approach. Acoustic features of the speech can provide valuable information about the vocal tract configuration. Predicting the vocal tract shape from the speech content and considering the predicted shape in the curve evolution, might increase the performance accuracy. The predicted contour could be taken into account as a new shape energy term. Also a closed contour shape modeling of the vocal tract might result in a better performance, since the closed contour snakes are expected to be more noise robust.

Appendices

Appendix 1

SHAPE ALIGNMENT USING PROCRUSTES ANALYSIS

A.1 Aligning Training Set

All the shapes are aligned to the mean shape shape in the follwing iterations:

1. Translate each shape so that the center of gravity is at the origin.

2. re-scale each shape to have equal size.

3. Choose an initial estimate of the mean shape (e.g. the first shape in the set).

4. align all the shape to the mean shape.

5. Re-estimate the mean from aligned shapes. 6. If not converged (the estimated mean has changed) return to step 2.

A.2 Aligning two shapes

Given two shapes $\mathbf{x}_1$ and $\mathbf{x}_2$, the goal is to scale and rotate $\mathbf{x}_1$ by pose parameters(s, θ) to minimize $|s\mathbf{A}\mathbf{x}_1 - \mathbf{x}_2|$, where $\mathbf{A}$ performs a rotation of a shape by θ. The pose parameters are computed as:

$$s^2 = a^2 + b^2 \tag{A.1}$$

$$\theta = tan^{-1}(b/a) \tag{A.2}$$

where

$$a = (\mathbf{x}_1.\mathbf{x}_2)/|\mathbf{x}_1|^2 \tag{A.3}$$

$$b = (\sum_{i=1}^{n}(x_{1i}y_{2i} - y_{1i}x_{2i}))/|\mathbf{x}_1|^2 \tag{A.4}$$

BIBLIOGRAPHY

Amini, A., Weymouth, T., and Jain, R. (1990). Using dynamic programming for solving variational problems in vision. *IEEE transactions on pattern analysis and machine intelligence*, 9(12).

Aron, M., Roussos, A., and Berger, M. (2008). Multimodality acquisition of articulatory data and processing. *16th European Signal Processing Conference - EUSIPCO*.

Behiels, G., Vandermeulen, D., Maes, F., and Dewaele, P. (1999). Active shape model-based segmentation of digital x-ray images. *in lecture notes in computer science.*, pages 128–137.

Bresch, E. and Narayanan, S. (2009). Region segmentation in the frequency domain applied to upper airway real-time magnetic resonance imaging. *IEEE Transaction on Medical Imaging*, 28(3).

Cootes, T. and Taylor, C. (1993). Active shape model search using local grey-level methods: a quantitative approach. *Proc. British Machine Vision Conferenc*, pages 639–648.

Fu, Y., Erdem, A. T., and Tekalp, A. M. (2000). Tracking visible boundary of objects using oclussion adaptive motion snake. *IEEE transactions on image processing*, 9(12).

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society B*, 53(2).

Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *international journal of computer vision*, pages 321–331.

Kim, J., Kumar, N., Lee, S., and Narayanan, S. (2014). Enhanced air-way tissue boundary segmentation for real-time magnetic resonance imaging data. *In proceedings in Interspeech*, 1(4).

Li, M., Kambhamettu, C., and Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics and Phonetics*, 6(19):545554.

Lucas, B. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. *in Proc. Intl Joint Conf. Artificial Intelligence.*

Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. *Speech Production and Modellling.*

Proctor, M., Bone, D., Katsamanis, N., and Narayanan, S. (2010). Rapid semi-automatic segmentation of real-time magnetic resonance image for parametric vocal tract analysis. *In proceedings in Interspeech*, 1(4).

Roussos, A., Katsamanis, A., and Maragos, P. (2009). Tongue tracking in ultrasound images with active appearance models. *16th IEEE International Conference on Image Processing (ICIP).*

Saltzman, E. and Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychol*, 1(4).

Simoncelli, E., Adelson, E. H., and Heeger, D. J. (1991). Probability distributions of optical flow. *IEEE conference on computer vision and pattern recognition*, pages 310–315.

Yehia, H., Takeda, K., and Itakura, F. (1996). An acoustically oriented vocal-tract mode. *IEICE TRANSACTIONS on Information and Systems*, E79-D(8).

Yokoyama, M. and Poggio, T. (2005). A contour-based moving object detection and tracking. *ICCCN '05 Proceedings of the 14th International Conference on Computer Communications and Networks*, pages 271–276.