Adaptive Threshold Optimisation for Colour-based Lip Segmentation in Automatic Lip-Reading Systems

Ashley Daniel Gritzman

A thesis submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, September 2016

Declaration

I declare that this thesis is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

Signed this ____ day of _____ 20__

_____

Ashley Daniel Gritzman

Abstract

Having survived the ordeal of a laryngectomy, the patient must come to terms with the resulting loss of speech. With recent advances in portable computing power, automatic lip-reading (ALR) may become a viable approach to voice restoration. This thesis addresses the image processing aspect of ALR, and focuses three contributions to colour-based lip segmentation.

The **first contribution** concerns the colour transform to enhance the contrast between the lips and skin. This thesis presents the most comprehensive study to date by measuring the overlap between lip and skin histograms for 33 different colour transforms. The hue component of HSV obtains the lowest overlap of 6.15 %, and results show that selecting the correct transform can increase the segmentation accuracy by up to three times.

The **second contribution** is the development of a new lip segmentation algorithm that utilises the best colour transforms from the comparative study. The algorithm is tested on 895 images and achieves percentage overlap (OL) of 92.23 % and segmentation error (SE) of 7.39 %.

The **third contribution** focuses on the impact of the histogram threshold on the segmentation accuracy, and introduces a novel technique called *Adaptive Threshold Optimisation (ATO)* to select a better threshold value. The first stage of ATO incorporates ϵ-SVR to train the lip shape model. ATO then uses feedback of shape information to validate and optimise the threshold. After applying ATO, the SE decreases from 7.65 % to 6.50 %, corresponding to an absolute improvement of 1.15 pp or relative improvement of 15.1 %. While this thesis concerns lip segmentation in particular, ATO is a threshold selection technique that can be used in various segmentation applications.

To my parents Marcus and Vivienne,
for their passionate commitment to the value of education.

Acknowledgements

To my supervisors:

> I would like to thank Professor David Rubin ("Prof") for his guidance and support throughout the research journey. Our lively discussions on a broad range of topics will be fondly remembered.
>
> I would like to thank Professor Vered Aharonson for eagerly joining the team of supervisors to share her technical knowledge and experience. I also thank Vered for the opportunity to spend five months at the Afeka College of Engineering in Israel.
>
> I would like to thank Adam Pantanowitz for his mentorship within and beyond the scope of this research.

Funding from the following sources is gratefully acknowledged:

- National Research Foundation (NRF) of South Africa (Grant No. 97742)
- University of the Witwatersrand Postgraduate Merit Award (PMA)

I am deeply grateful to my parents, Marcus and Vivienne, for their consistent encouragement to pursue further education, and for providing every opportunity to do so.

Finally, I would like to express my heartfelt gratitude to my wife, Talia, for her love, support, and patience during the PhD journey.

Publications

1	**Automatic computation of histogram threshold for lip segmentation using feedback of shape information**
Authors	Ashley D. Gritzman, Vered Aharonson, David M. Rubin, Adam Pantanowitz
Journal	Signal, Image and Video Processing
Pages	1 – 8
Year	2015
Publisher	Springer
2	**Comparison of colour transforms used in lip segmentation algorithms**
Authors	Ashley D. Gritzman, David M. Rubin, Adam Pantanowitz
Journal	Signal, Image and Video Processing
Volume	9
Number	4
Pages	947 – 957
Year	2015
Publisher	Springer
3	**Threshold-based Lip Segmentation using Feedback of Shape Information**
Authors	Ashley D. Gritzman, Vered Aharonson, David M. Rubin, Adam Pantanowitz
Book title	Proceedings of the Conference on Facial Analysis and Animation 2015 (FAA2015)
Page	4
Date	September, 2015
Publisher	ACM

Contents

List of Figures

List of Tables

Nomenclature

ADI Accumulated Difference Image

AI Artificial Intelligence

ALR Automatic Lip-Reading

ASM Active Shape Model

ATO Adaptive Threshold Optimisation

AUC Area Under Curve

CC Chromatic Curve Map

CDF Cumulative Distribution Function

DCT Discrete Cosine Transform

DHT Discrete Hartley Transform

DWT Discrete Wavelet Transform

FLD Fisher Linear Discriminant

FN False Negative

FP False Positive

FPR False Positive Rate

FPS Frames Per Second

HDF Hue Domain Filter

HMM Hidden Markov Model

HSV Hue, Saturation and Value

ILE Inner Lip Error

LDA Linear Discriminant Analysis

LPF Low Pass Filter

LSM Lip Shape Model

LUX Logarithmic Hue Extension

MI3 Modified I3 Channel

MLLT Maximum Likelihood Linear Transform

MR Mouth Region

NN Neural Network

OL Percentage Overlap

OLE Outer Lip Error

PCA Principal Component Analysis

PDM Point Distribution Model

PHE Pseudo-Hue Transform (Eveno)

PHT Pseudo-Hue Transform (Talea)

POI Point Of Interest

PP Percentage Point

RE Red Exclusion

RGB Red, Green and Blue

ROC Receiver Operating Characteristic

ROI Region of Interest

SE Segmentation Error

SSI Silent Speech Interface

SVR Support Vector Regression

TN True Negative

TP True Positive

TPR True Positive Rate

Chapter 1

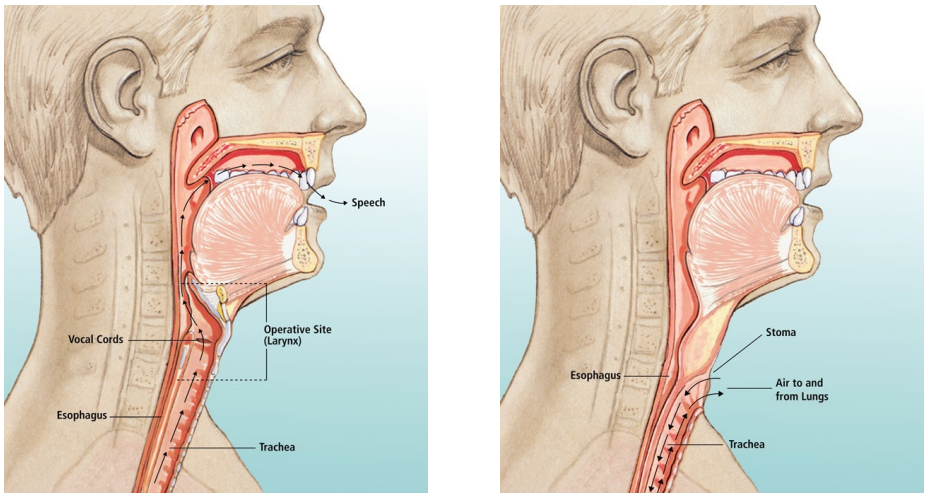
Introduction

"Just because I cannot speak, doesn't mean I have nothing to say!"

Communication is central to human life. In 350 B.C.E. Aristotle wrote: "Nature, as we often say, makes nothing in vain, and man is the only animal whom she has endowed with the gift of speech" [1]. Speech provides a common channel to express our most basic physiological needs, our most complex thoughts, our most creative ideas, and our most intimate emotions. Thus, when this faculty is taken away, the consequences can be devastating to both the victim and the family members. The victim feels cut-off from society, and often suffers from anxiety and depression [2, 3]. Loss of speech has been associated with an increased propensity towards alcoholism and suicide [4, 5]. Family members experience high levels of stress in trying to cater for the needs of the affected individual, and become frustrated when they cannot understand what the person needs or wants [3].

A *laryngectomy* is the partial or complete surgical removal of the larynx (voice box), which leaves the patient unable to speak. The upper part of the trachea is joined to a tracheostoma (opening) in the front of the neck, to provide an alternate passage of air to breathe (see Figure 1.1). A laryngectomy is usually performed to treat cancer of the larynx; however, this procedure may also be performed in the case of severe trauma (e.g. gunshot wound), or severe damage to the larynx caused by radiation treatment [6]. According to the American Cancer Society, in the USA alone 13 560 new cases of laryngeal cancer were diagnosed and 3640 people died from laryngeal cancer in 2015. Worldwide, approximately 136 000 cases of cancer of the larynx are diagnosed each year [8]. The prognosis for laryngeal cancer is better than most upper aerodigestive tract cancers, with a five year survival rate of 68 % [8]; however, many

of these patients will face the prospect of a laryngectomy, and will no longer be able to speak or produce any vocalisation.



(a) Before **(b)** After

Figure 1.1: Anatomy of the airway before and after a laryngectomy: (a) before – air flow from the lungs passes between the vocal cords producing speech; (b) after – the larynx (including the vocal cords) has been removed and air is expelled directly through a stoma (opening) in the neck. Reused with permission from InHealth Technologies [9].

Restoration of some degree of voice is crucial to the laryngectomy patient's morale, self esteem and reintegration into society.

Three conventional methods are commonly used in clinical practice to restore verbal communication following a laryngectomy procedure: the electrolarynx, tracheo-oesophageal speech, and oesophageal speech. While these techniques have achieved some success in restoring speech, various limitations and drawbacks are associated with each technique including usability, voice quality, and intelligibility.

Digital solutions to voice restoration are attracting growing interest as the processing power of portable devices increases while the size and cost decrease. Devices which facilitate speech communication in the absence of an acoustic signal are known as *Silent Speech Interfaces (SSIs)* [10]. SSIs use sensors to acquire data from the speech production process, and recognition algorithms to process and interpret the data. The digital speech representation can then be enunciated by a voice synthesiser or displayed as text. Experimental SSIs reported in the literature are based on various technologies including optical imaging, ultrasound, surface electromyography (sEMG), and electroencephalography (EEG).

Since the revolution in mobile computing, including smartphones, tablets, and laptops,

optical imaging has become integrated into many aspects of our daily lives. A digital video camera is widely accessible, anywhere, anytime; with a simple tap of an app or click of a mouse, an optical SSI complete with sensor, processing power and audio/video output, can be set up and ready to go. For hands-free use, the camera can be integrated into an audio-visual headset, as shown in Figure 1.2. The camera captures images of the lips during speech and the earpiece provides auditory feedback to the user. The audio-visual headset shown in Figure 1.2 is commercially available as part of the AudiSee system manufactured by AudiSoft Technologies Inc. [11]. The system was originally designed to aid hearing impaired students decipher the words of the teacher, but the headset can be adapted for use in a hands-free SSI device.

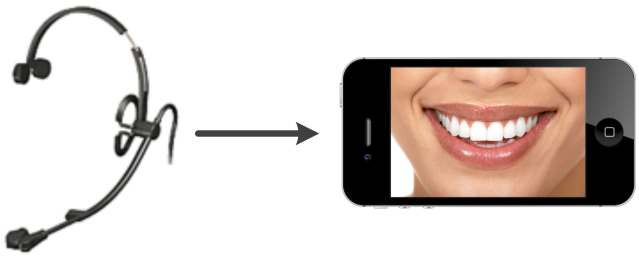


Figure 1.2: The Audio-visual headset manufactured by Audisoft can be integrated with a smartphone; the camera captures images of the mouth and the earpiece provides auditory feedback to the user (adapted with permission from: AudiSoft Technologies Inc. [11]).

In an optical SSI, the technique of retrieving speech contents from visual clues, such as the movement of the lips, tongue, and teeth, is known as *Automatic Lip-Reading (ALR)*. The premise for using ALR to restore verbal communication for laryngectomy patients is based on the lip-reading proficiency of hearing impaired people. In a study by Bernstein et al. [12], hearing impaired adults correctly identified up to 88 % of phonemes in independent sentences. When the context of the sentence is available, proficient lip-readers can achieve near perfect comprehension.

In 2004, the England and Wales Court of Appeal established the admissibility of lip-reading evidence in a landmark case, *Luttrell* [2004] EWCA Crim 1344. An expert lip-reader produced a transcript of a video conversation which aided the prosecution in obtaining a guilty verdict for Luttrell and eight others on a charge of conspiracy to commit armed robbery and dispose of stolen goods [13].

An ALR system comprises two core components: lip segmentation and recognition. Lip segmentation is the challenge of accurately discriminating between lip pixels and non-lip pixels. Recognition involves identifying meaningful speech information from

the movement of the lips. Lip segmentation is an image processing challenge, while recognition is a machine learning challenge.

This thesis addresses the image processing challenge of accurate and robust lip segmentation, and focuses on three image processing contributions to this end.

The first contribution concerns the colour transform to enhance the contrast between the lips and skin. There is much debate among researchers as to the best colour transform for this purpose, and as such, this thesis presents the most comprehensive study to date covering 33 different colour transforms. This work has been published in the journal of *Signal, Image and Video Processing* [14].

The second contribution is the development of a new colour-based lip segmentation algorithm. This algorithm is referred to as the *base algorithm*, as it forms the platform to test and analyse the third contribution.

The third contribution is the crux of this thesis, and concerns the development of a novel threshold selection technique called *Adaptive Threshold Optimisation (ATO)*. ATO uses feedback of shape information to validate and optimise the threshold value. This research has been published in the journal of *Signal, Image and Video Processing* [15], and *Proceedings of the 2015 Conference on Facial Analysis and Animation (FAA2015)* [16].

The structure of this thesis is as follows:

Chapter 2: This chapter begins with an overview of the relevant anatomy and physiology of the larynx (voice box) and the mechanisms involved in the production of speech. The main focus of this chapter is a review of voice restoration techniques for laryngectomy patients, including three conventional techniques and seven silent speech interface (SSI) technologies.

Chapter 3: This chapter provides an overview of automatic lip-reading (ALR) including the challenges, approaches, and technologies. The techniques used in existing ALR systems are summarised in four categories: face detection, mouth region detection, feature extraction, and recognition.

Chapter 4: This chapter presents the argument for using ALR to restore verbal communication for laryngectomy patients. Once this basis has been established, the scope of the research is narrowed to focus on the unresolved image processing challenge.

Chapter 5: This chapter comprises the first image processing contribution towards ALR and begins with a comprehensive survey of the colour transforms used in the lip segmentation literature. Thirty three transforms are identified, which are compared in terms of lip-skin segmentation as well as lip-oral cavity segmentation.

Chapter 6: This chapter comprises the second contribution and details the design and implementation of a new threshold-based lip segmentation algorithm. Since technologies to locate the face and mouth region are well established, the starting point for the lip segmentation algorithm is the pre-cropped mouth region. The lip segmentation algorithm is primarily a colour-based technique which exploits the best colour transforms from Chapter 5 to enhance the contrast between the lips and the skin.

Chapter 7: This chapter presents the results and analysis of the lip segmentation algorithm described in Chapter 6. The algorithm is tested on 895 mouth region images from the AR Face Database, using percentage overlap (OL) and segmentation error (SE) to quantify performance. The discussion presents examples of both successful and unsuccessful segmentation results, which leads to an understanding of the strengths and the weaknesses of the algorithm.

Chapter 8: This chapter motivates for the forthcoming work on threshold selection by describing the associated challenges from both qualitative and quantitative perspectives. The qualitative component illustrates the challenge of threshold selection by analysing two examples where Otsu's method fails to select a suitable threshold. The quantitative component computes the improvement in segmentation accuracy that can be obtained by adjusting the threshold value.

Chapter 9: This chapter comprises the third contribution, and details the design of an adaptive algorithm for selecting the histogram threshold. The algorithm reduces unnecessary overhead by first comparing the initial segmentation to a reference lip shape model to decide if optimisation is required (the *validation stage*). In cases where optimisation is required, the algorithm iteratively adjusts the threshold to reduce the segmentation error (the *optimisation stage*). This novel technique for threshold selection is referred to as *Adaptive Threshold Optimisation (ATO).*

Chapter 10: This chapter analyses the performance of the ATO algorithm by conducting three tests. In the first test, the validation stage is evaluated as a binary classifier which aims to identify poor segmentation. In the second test, the optimisation stage is evaluated by comparing the lip segmentation accuracy of the default threshold, versus the ATO threshold. In the final test, the performance of

the overall algorithm is evaluated, which includes images that are optimised as well as images that are not optimised.

Chapter 11: This chapter summarises the three image processing contributions towards ALR, and suggests several avenues for future research.

Chapter 2

Voice Restoration Techniques

2.1 Introduction

The larynx (or voice box) houses and controls the vocal cords, which are essential for phonation. Surgical removal of the larynx during a laryngectomy leaves the patient unable to speak or produce any vocalisation. The conventional methods of voice restoration – electrolarynx, tracheo-oesophageal speech, and oesophageal speech – attempt to use the remaining structures of the vocal tract to define a new speech production process. While these methods are commonly used in clinical practice, there are various limitations and drawbacks associated with each technique.

With recent advances in portable computing power, a digital solution to this problem has become feasible. Silent speech interfaces (SSIs) use digital sensors and processing to augment the existing speech production pathway, and have the potential to improve on some of the limitations of conventional techniques [10]. However, SSIs are still in the experimental phase, and face several challenges of their own. SSIs suitable for laryngectomy patients are based on seven technologies: optical imaging, ultrasound, electropalatography (EPG), electromagnetic articulography (EMA), surface electromyography (sEMG), electroencephalography (EEG), and intracranial electrode implants.

This chapter reviews the anatomy and physiology of speech production and discusses both conventional and SSI voice restoration techniques.

2.2 Mechanisms of Natural Speech Production

The production of speech is a complex motor task which involves the coordinated use of approximately 100 muscles. The complex motor patterns produce speech sounds at a rate of 15 per second [17]. The organs involved in the production of speech can be divided into three main groups: lungs, larynx, and vocal tract. The *vocal tract* is the air passage extending from the vocal cords to the lips and consists of the pharynx, soft palate, hard palate, teeth, tongue and lips (see Figure 2.1).

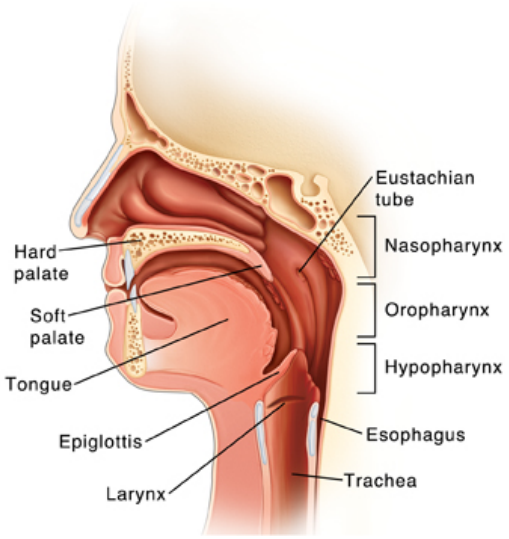


Figure 2.1: Midsagittal view of head and neck showing the organs involved in speech production (reused with permission from Krames StayWell [18]).

The production of speech is composed of five elements: respiration, phonation, resonance, articulation, and prosody.

Respiration is an aerodynamic process in which pressure differences between the thoracic cavity and the atmosphere are manipulated to alternately inflate and deflate lungs (inhalation and exhalation). The lungs provide the source of energy for speech production in the form of a steady stream of air expelled during exhalation [19]. The larynx is a continuation of the trachea with specialised muscles and cartilage structures to manipulate the vocal folds (see Figure 2.2). The vocal folds (vocal cords) stretch across the larynx and separate the pharynx from the trachea. The opening between the vocal folds is known as the glottis (see Figure 2.2b). During normal breathing the vocal folds remain wide open to allow easy passage of air through the glottis [20]. *Phonation* occurs when the glottis is closed and air is expelled from the lungs, creating a pressure drop across the larynx. The glottis expands to relieve the

pressure gradient, and contracts due to elastic and aerodynamic forces [21]. The frequency of oscillation of the vocal folds depends on the mass, length and tension of the vocal folds [21]. The muscles of the larynx control the frequency of oscillation by altering the tension of the vocal folds. Speech sounds incorporating phonation are referred to as "voiced sounds" (e.g. /z/) while those produced without phonation are referred to as "voiceless sounds" (e.g. /s/).

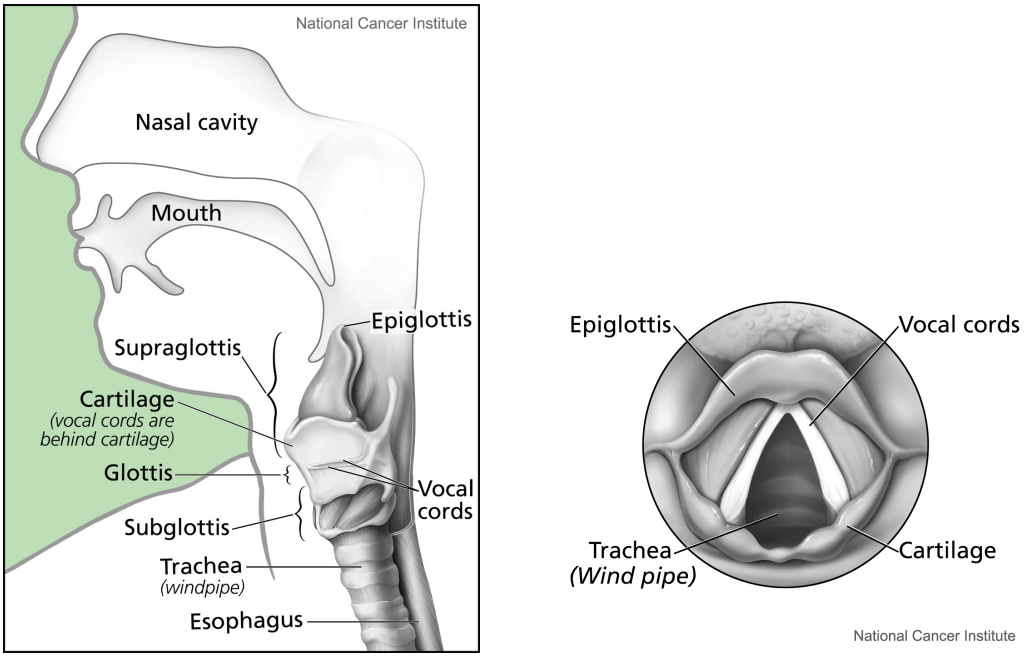


Figure 2.2: (a) Anatomy of the larynx and surrounding structures; (b) Top view of larynx with the vocal folds open. Images in public domain: Hoofring [22, 23].

The basic noise that emerges from the larynx is not speech, but a bleating sound which is modified into speech as it passes through the resonating cavities [24]. Different sounds are produced by adjusting the natural frequency of the resonating cavities such that certain harmonics are suppressed while others are accentuated [24]. The natural frequency is adjusted by changing the shape and volume of the resonating cavities: larger cavities accentuate lower frequencies (e.g. /ɑː/), while smaller cavities accentuate higher frequencies (e.g. /iː/) [24]. There are three resonating cavities: the pharynx, mouth, and nose. The dimensions of the nose are fixed, but the dimensions of the pharynx and mouth can be significantly altered by muscular action [24]. *Resonance* is therefore the process by which the nature of the sound is changed as it passes through the vocal tract [25].

During speech production, the shape of the vocal tract is varied extensively by movement of the articulators, including the velum, tongue, jaw, and lips [19]. *Articulation* refers to the movement and shaping of the vocal tract to form speech sounds, and often involves bringing two speech organs together to form an obstruction. The point of maximum obstruction is known as the *place of articulation*, and the way in which the obstruction is formed and released is known as the *manner of articulation* [26]. The sound /p/ is generated by bringing the two lips tightly together to block the air, causing a build-up of air pressure. The lips are then released suddenly, leading to a burst of sound. The place of articulation of this sound is therefore called *bilabial*, and the manner is called *stop* (also known as a *plosive*) [26].

Vowels are sounds produced with the oropharyngeal cavity system open [25]. Different vowel sounds are produced by changing the phonation and resonance factors. *Consonants* are sounds that contain noise elements generated in the mouth, and may or may not include phonation generated in the larynx [25]. The noise is created either by stopping and releasing the air stream passing through the mouth (e.g. /p/, /b/, /m/) or by forcing the airstream through a restricted space (e.g. /f/, /l/, /z/).

Prosody refers to the rhythm, stress and intonation of speech [26]. Prosody does not act at the level of individual phonemes, but rather applies to sequences of words (phrases) or sentences. Prosody may reflect various features of the speaker or the utterance: the form (statement, question, or command), the presence of irony or sarcasm, the emotional state of the speaker [26].

2.3 Conventional Voice Restoration Techniques

Following a laryngectomy, three conventional techniques are commonly used in clinical practice to restore some degree of verbal communication, namely: the electrolarynx, tracheo-oesophageal speech, and oesophageal speech. These techniques have achieved varying degrees of success in voice restoration.

2.3.1 Electrolarynx

More than half of all laryngectomy patients use an electrolarynx to restore verbal communication [27]. The electrolarynx uses an electromechanical vibrator to transmit sound waves into the oral and pharyngeal cavities [28]. The sound waves are modulated into words by the patient's articulators [29]. Two different types of

electrolarynx exist: the neck type and the intraoral type [8]. Figure 2.3 shows the neck type of electrolarynx which is placed against the side of the neck, under the chin or on the cheek [8]. The intraoral type is used when sufficient conduction cannot be achieved through the skin, so a small tube is placed directly in the posterior oral cavity. However, the electrolarynx voice is monotonic and difficult to understand [29] resulting in low intelligibility and poor listener acceptance [28].

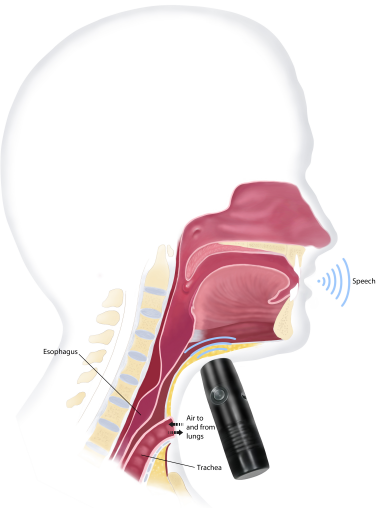


Figure 2.3: Sound waves generated by the electrolarynx are modulated into words by the patient's articulators (reused with permission from Atos Medical AB [30]).

2.3.2 Tracheo-oesophageal Speech

Tracheo-oesophageal speech is considered the gold-standard in voice restoration for laryngectomy patients [31]. Tracheo-oesophageal speech requires surgical creation of a fistula (passageway) between the trachea and the oesophagus [8, 29, 32]. A one-way silicone valve is inserted into the fistula which allows air to pass from the trachea into the oesophagus, but prevents food and liquid from entering the trachea [8]. During speech the patient must cover the tracheostoma, forcing air through the one-way valve and into the oesophagus (see Figure 2.4). The airflow through the oesophagus causes the pharyngo-oesophageal segment to vibrate producing a sound [8]. The sound is modulated into words by the patient's articulators.

Tracheo-oesophageal speakers must constantly alternate between covering the stoma to speak and uncovering the stoma to inhale, resulting in a slow speech rate [31]. The silicone valves initially perform very well; however, in many patients the valve rapidly becomes colonised by biofilm (aggregate of microorganisms on a surface) and fails after

3 – 4 months [29, 33–36]. Various valve modifications have been proposed, for example: Eerenstein et al. [37], Everaert et al. [38], and Hilgers et al. [39]; however, these approaches do not appear to provide a long-term solution [29]. Tracheo-oesophageal speech is the preferred method of voice restoration in USA, UK and certain countries in Europe; however, Jacobson et al. [40] note that non-English/French speakers experience difficulty in using this method [41]. Finally, Eadie et al. [42] observe that listeners often struggle to identify gender from tracheo-oesophageal speech.

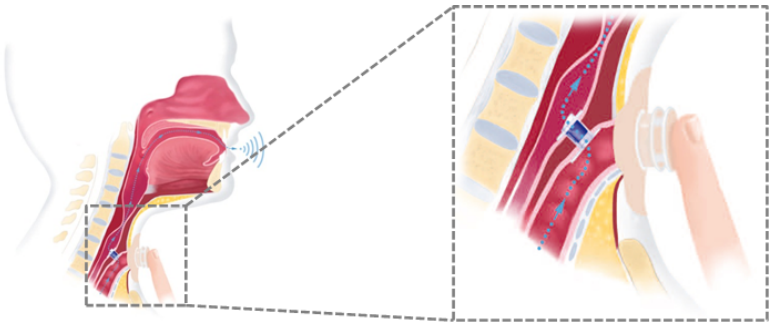


Figure 2.4: Tracheo-oesophageal speech is produced by blocking the tracheostoma which diverts air from the trachea into the oesophagus, causing the pharyngo-oesophageal segment to vibrate (adapted with permission from Atos Medical AB [30].

2.3.3 Oesophageal Speech

Oesophageal speech is similar to the process of belching (burping) and involves alternately swallowing and expelling air (see Figure 2.5). The oesophagus is used as a reservoir to store air that has been swallowed [29]. Controlled release of the air from the oesophageal reservoir (belching) results in vibration of the pharyngo-oesophageal segment, which produces a sound [41]. The sound is modulated into words by the patient's articulators. Oesophageal speech is a significant voice restoration technique in Asian countries [41]; however, it is difficult to learn and fluent speech is impossible due to the short phonation duration [29, 41]. The voice generated by oesophageal speech is perceived as harsh, gurgling, hoarse, low pitch and low volume [41].

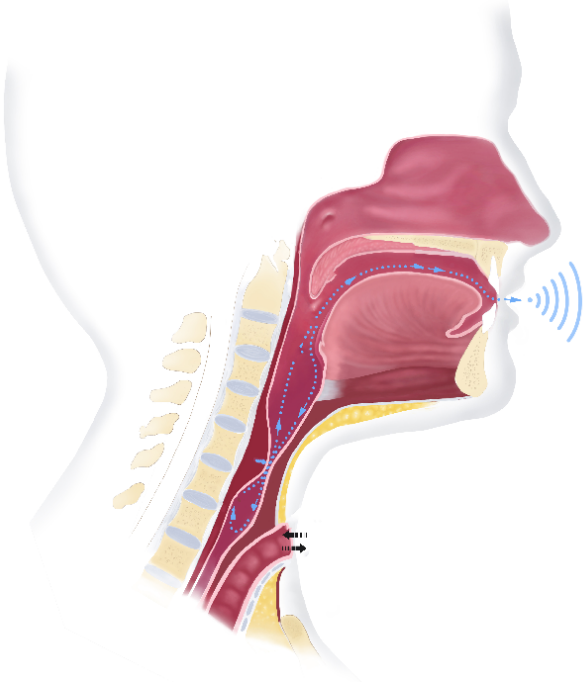


Figure 2.5: Oesophageal speech is similar to the process of belching: air is swallowed into the oesophagus and expelled in a controlled manner producing a sound (reused with permission from Atos Medical AB [30]).

2.4 Silent Speech Interfaces (SSIs)

Modern advances in electronic miniaturisation and portable computing power have paved the way for a computer-based solution to voice restoration [29]. A *Silent Speech Interface (SSI)* is a system enabling speech communication in the absence of an intelligible acoustic signal [10]. SSIs acquire sensor data from elements of the human speech production process including the larynx, articulators, neural pathways, or the brain itself [10]. Recognition algorithms use the sensor data to produce a digital representation of speech, which is then enunciated using a voice synthesiser or displayed as text. SSIs have been developed primarily to aid the speech-handicapped; however, other applications include providing privacy for telephone conversations and improving speech recognition in noisy environments [10]. SSIs are still in the experimental stage, and several challenges must be addressed before 'meaningful use' is achieved [10].

2.4.1 SSI Technologies

While SSIs acquire sensor data from various elements of the speech production process, laryngectomy patients are restricted to SSI technologies that do not depend on vibration of the focal cords. Such systems are based on seven types of technology.

Optical Imaging

In an optical SSI, a standard video camera is used to retrieve speech content from visual clues such as the movement of the lips, tongue and teeth. An optical SSI relies heavily on image processing techniques to extract useful speech information from the image stream, primarily the lip contours. Initiated by Petajan in 1984, ALR research has focused on combining visual and auditory modalities to improve speech recognition in noisy environments [44–49]. Chapter 3 presents a more detailed overview of automatic lip-reading (ALR).

Ultrasound Imaging

An ultrasound transducer emits high frequency sound waves which are reflected back at the boundaries between body structures. The returning echoes are detected by the transducer, and converted to an image. Ultrasound imaging is used to view soft tissues, muscles and internal organs. In an ultrasound-based SSI, speech information is inferred from images of the tongue obtained by placing an ultrasound probe beneath the chin [50]. In the *Ouisper Project*, an SSI was developed by combining ultrasound imaging of the tongue and optical imaging of the lips [51]. Using a combined ultrasound and optical system, Hueber et al. [52] correctly predicted 60 % of phonemes in one hour of continuous speech, without any vocabulary restrictions. While an accuracy of 60 % is insufficient to facilitate synthesis of intelligible speech, limiting the vocabulary will clearly improve the recognition accuracy which may enable 'meaningful use' of the device.

Electropalatography (EPG)

ElectroPalatoGraphy (EPG) is a technique used to monitor dynamic contact patterns between the tongue and hard palate, particularly during articulation and speech. A custom-made artificial palate containing an electrode array is moulded to fit against

the speaker's hard palate. Contact between the tongue and the electrodes causes a change in electrode conductivity, which is used to reconstruct the 2D palato-lingual contact pattern. While EPG is typically used by speech therapists to correct pronunciation issues, Russell et al. [53] propose an artificial larynx based on EPG to restore verbal communication for laryngectomy patients. Russell et al. [53] used a neural network to identify 50 common English words, and achieved an accuracy of 94.14 % with a rejection rate of 17.74 %.

Electromagnetic Articulography (EMA)

ElectroMagnetic Articulography (EMA) uses coupling between magnetic implants and sensors positioned around the head to monitor movements of points within the vocal tract. Fagan et al. [29] attach permanent magnets to the articulators (lips, teeth, tongue) and measure variations in the magnetic field using sensors mounted on a pair of spectacles. The Dynamic Time-Warping algorithm is used to classify the signals from the magnetic sensors and hence identify phonemes and words. The system achieved an accuracy of 94 % in recognising 13 phonemes, and 97 % on a set of 9 words.

Surface Electromyography (sEMG)

Muscle activity generates small electrical currents in the form of ion flows, which manifest as voltage differences at the surface of the skin. Surface ElectroMyoGraphy (sEMG) uses electrodes placed on the skin to measure these voltage differences, and thereby infer the electrical activity of muscles during contraction and relaxation. In 1985, Sugie & Tsunoda developed a speech prosthesis which used three sEMG electrodes to monitor activity of the articulator muscles during speech [54]. The system performed the recognition task in real-time and achieved an accuracy of 71 % on 5 Japanese vowels. The capacity of EMG-based systems has since increased to 101 words with an accuracy of 90 % [55]. Current research in sEMG speech recognition is focused on the following challenges: recognition of continuous sentences as opposed to isolated words [56]; recognition of phonemes to facilitate larger vocabularies [55]; and speaker independent recognition [57].

Electroencephalography (EEG)

Neurones communicate via electrical impulses which are generated by ion flows. ElectroEncephaloGraphy (EEG) is the technique of measuring brain activity using electrodes to record the voltage fluctuations along the scalp. Suppes et al. [58] were the first to show that EEG and MEG (MagnetoEncephaloGraphy) can be used to recognise isolated words. Their experiment involved presenting subjects with isolated words in the form of auditory stimuli, and attempting to detect the words from EEG and MEG recordings. Using EEG in Brain-Computer Interfaces (BCIs) typically requires the user to learn how to explicitly manipulate their brain activity [59]. Wester & Schultz [60] investigated a more intuitive system in which the user imagines speaking a word, however no sound is produced (referred to as "unspoken speech"). The system was tested on a vocabulary of up to 10 words, and obtained word error rates of 4–5 times above the chance level.

Intracranial Electrode Implants

Invasive BCI devices (Brain-Computer Interface) use electrical signals captured by intracranial electrode implants to predict intended speech information [61]. The electrodes are inserted into the cortex of the brain in a surgical procedure whereby a craniotomy is performed to expose the brain. Invasive BCI devices are based on three technologies: ElectroCorticoGrams (ECoG), Local Field Potentials (LFPs), and Single Unit Activity (SUA). While EEG and MEG capture the electric/magnetic activity produced by tens of thousands to millions of neurones, LFP represents the activity of tens of neurones and SUA represents activity of individual neurone units [61]. Brumberg et al. [62] conducted a study of attempted speech production in a patient suffering from locked-in syndrome using microelectrode recordings. The system obtained an accuracy of up to 21 % in recognising 38 phonemes (chance level: 1/38 or 2.6 %). While phoneme recognition using intracranial electrode implants have yielded promising results, the field is still very much in its infancy.

2.4.2 SSI Challenges

SSIs based on the above seven technologies are still in the experimental stage, and face several common challenges [10].

- *Sensor positioning and robustness* – sensors must be carefully positioned at the

start and a slight shift in position may necessitate recalibration of the entire system. For example, an ultrasound-based SSI is sensitive to orientation of the tongue surface relative to the probe, thus any movement of the probe will affect the entire system. Similarly, EMA, sEMG, and EEG are highly sensitive to variations in sensor positioning.

- *Speaker independence* – performance of the system is dependent on specific characteristics of the speaker, for example: optical imaging depends on the speaker's skin colour, sEMG depends on speaker anatomy, EMA depends on characteristics of articulator movement.
- *Vocabulary size* – a trade off must be made between vocabulary size and recognition accuracy: increasing the vocabulary size causes the recognition accuracy to decrease. The challenge is to construct a limited vocabulary that is rich enough to be genuinely useful to the patient, without prohibiting accurate recognition. This can be achieved by tailoring the vocabulary to specific tasks and scenarios in the patient's everyday life.
- *Cost* – complex technologies or techniques may significantly increase the cost of design and manufacture.
- *Routine-use considerations* – significant time is required to set-up and fit the SSI device, often requiring assistance from another person; the device is bulky and difficult to transport; and the device is inconvenient and uncomfortable to wear and use. As an example, consider an EEG device: the time required to set-up an EEG device includes application of conductive gel, placement of electrodes, and calibration of the device. Furthermore, apart from the discomfort caused by wearing the electrodes and wires, it is simply not feasible to carry around an EEG device.

2.5 Conclusion

The natural speech production process comprises five elements: respiration, phonation, resonance, articulation, and prosody. A laryngectomy disrupts the natural production of speech by removing the larynx, which is the source of phonation. The conventional methods of voice restoration – electrolarynx, tracheo-oesophageal speech and oesophageal speech – have achieved some success in restoring speech but suffer from various shortcomings. Researchers are currently working towards a computer-based solution to this problem whereby digital sensors and processing are used to decipher speech in the absence of an acoustic signal. These systems are

termed ‘Silent Speech Interfaces (SSIs)’ and have the potential to improve on some of the limitations of traditional techniques, while facing a different set of challenges: sensor positioning and robustness, speaker independence, vocabulary size, cost, and routine-use considerations.

Chapter 3

Automatic Lip-Reading

3.1 Introduction

Automatic Lip-Reading (ALR) is the computer-based technique of retrieving speech information from visual cues including the movement of the lips, teeth, and tongue [63]. The premise of ALR is based on the abilities of expert human lip-readers, who can achieve near perfect speech comprehension.

The concept of using a computer to lip-read has been in existence since the 1960s, and was popularised by the 1968 film "A Space Odyssey", where a "HAL 9000" computer was able to lip-read the conversations of astronauts who were plotting its destruction.

The first work on automatic lip-reading (ALR) appeared in 1984 when Petajan investigated the use of ALR to enhance speech recognition by creating a bimodal (audio-visual) speech recognition system [43]. In subsequent years, ALR research has focused on combining visual and auditory modalities to improve speech recognition in noisy environments [44–49].

The bimodal nature of speech is demonstrated by the McGurk effect: conflicting audio and visual sound components are presented to the subject, resulting in a perceived sound that may not correspond to either modality [64, 65]. For example, when a person hears the sound /ba/, but sees the articulatory movements corresponding to the sound /ga/, the person may not perceive either /ba/ or /ga/, but rather /da/.

Inclusion of visual speech information, primarily shape and movement of the lips, has been shown to significantly improve the performance of purely acoustic-based

speech recognition [43, 66–68]. This outcome can be understood by considering the sounds /m/ and /n/ which are easily confused in the audio domain, but are easily distinguished in the visual domain [65].

Applications of ALR include:

1. Speech recognition in noisy environments
2. Human-machine interfaces
3. Speaker authentication
4. Privacy for telephone conversations
5. Aid for people with a speech handicap

This chapter provides an overview of automatic lip-reading (ALR) including the challenges, approaches, and technologies. Section 3.5 presents a summary of the techniques used in existing ALR systems and details two complete ALR systems to give an indication of the typical structure and accuracy that can be expected.

3.2 Challenges of Lip-Reading

There are several challenges inherent in the task of human lip-reading, some of which are compounded when attempting to automate the lip-reading process, while others are reduced or eliminated entirely. Likewise, the task of ALR (or machine lip-reading) faces several unique challenges which are taken for granted in the context of normal human speech.

3.2.1 Human Lip-Reading

Lip-reading performed by humans is associated with six challenges:

1. *Low visibility of speech sounds* – Most of the motor (muscular) movements involved in the formation of sound occur within the mouth and cannot be detected by the eye [69]. The lip movements play a relatively minor part in the formation of sounds. It is estimated that approximately 60 % of speech sounds are either obscure or invisible [69].
2. *Homophenous sounds* – Alexander Graham Bell coined the term “homophene” to describe words that look alike, but do not sound alike (e.g. /bat/, /pat/,

/mat/). Jeffers & Barley [69] extend this definition to include sounds that look alike, (e.g. /b/, /p/, /m/). There is not a single constant sound in the English language that has a characteristic lip or jaw movement of its own and hence can be recognised on the basis of vision alone [69]. The lip-reader must guess the sound corresponding to the observed lip movement.

3. *Rapidity of normal speech* – Ordinary speech averages approximately thirteen speech sounds per second while the eye is capable of consciously seeing only eight to ten movements per second [70]. Jeffers & Barley [69] assert that the speed itself is not the primary challenge; rather, as the speed increases, the distinctive movements which differentiate the speech sounds are lost.
4. *Transition effect (co-articulation effect)* – The formation and hence appearance of sounds can be altered by sounds that precede or follow [69]. For example, the lip-reader depends on forward protrusion of the lips to identify the sounds: /sh/, /ch/, /j/ as in /shoe/, /chew/, /Jew/; however, this movement is often missing or obscured when the consonant is followed by a high front vowel as in /shape/, /chip/, /jeep/.
5. *Variation in sound formation* – Several articulation patterns produce the same speech sounds [69]. For example, the sound /t/ can be made by placing the tongue in various positions.
6. *Environmental limitations* – The lip-reader requires a clear line of sight to the lips of the speaker. Lip-reading is not possible if the speaker's back is turned, whereas hearing is possible in this situation. In addition, lip-reading requires sufficient illumination of the speaker's lips, whereas hearing is possible in complete darkness.

3.2.2 Automatic (Machine) Lip-Reading

Machine Advantages

Machine lip-reading is not subject to the limitations of the human eye which is capable of seeing only eight to ten speech movements per second [70]. The speed of an ALR system depends on the frame rate of the camera, the processing power, and the efficiency of the algorithms. An efficient ALR system can easily process 40 frames per second (fps), which would facilitate recognition of normal speech at 13 speech sounds per second [71].

Furthermore, ALR systems are not subject to the human frailties of fatigue and

limited attention span. In fact, an ALR system could theoretically operate on multiple speakers simultaneously, for an unlimited period of time.

Low visibility of speech sounds and homophenes gives rise to the fundamental challenge of human lip-reading: the interpreter is presented with incomplete speech information in the visual domain. As a result, the interpreter must make an educated guess of what has been said based on the grammar and context. For example, consider the sentence "I pat the dog" – in the visual domain the words "pat", "bat" and "mat" are indistinguishable; however, the words "bat" and "mat" do not make sense in the context of the sentence, therefore the interpreter will deduce the unknown word to be "pat".

In recent years, language prediction technology has advanced to the point that it has become integrated into our everyday activities. *Google Autocomplete* helps one find information quickly by offering search suggestions or predictions. The predictions are based on a number of factors including: popularity in search activities of all web users (frequency), content of web pages indexed by Google, geographical location of the user, and historical search activities of the user. For a user in South Africa on 20 May 2016, entering the letters "nel" into the Google search box would prompt the following predictions: "Nelson Mandela", "Nelspruit", "Nelly", and "Nelson Mandela Quotes". Another example of language prediction technology is *Next Word Prediction Tools*, which are currently used in smartphones to help users type messages faster and more accurately. The tools use natural language algorithms to predict words based on spelling, grammar and context, and incorporate machine learning tools to adapt to the specific user.

Considering the recent advances in this technology, it seems reasonable to 'predict' that given the same visual information, ALR systems will be able to perform lip-reading to the same extent, if not better than expert human lip-readers who can achieve almost perfect speech perception.

Machine Disadvantages

The fundamental challenges of ALR or "machine lip-reading" are lip segmentation and recognition. Lip segmentation is the task of accurately discriminating between lip pixels and non-lip pixels, which is an image processing challenge. Recognition involves identifying speech information from the movements of the lips, which is a machine learning challenge.

Lip segmentation

Lip segmentation is taken for granted in the human context – humans naturally identify and track the face and lips of a speaker, without much thought or effort. However, in the domain of ALR, tracking of the face and lips is not a trivial task and is subject to various limitations and challenges.

The task of lip segmentation is challenging primarily due to significant variability in speaker profile and background conditions [63]:

- Lip and skin colour (race)
- Lip shape, width, height
- Facial hair, make-up, glasses
- Amount of lip movement during speech
- Distance from the camera (scale)
- Orientation relative to the camera (pose)
- Illumination conditions: intensity, shadows, glare
- Background activity

In addition, the following limitations are specific to machine lip-reading:

- *Fixed camera position* – the camera position is usually rigid and cannot easily be changed, therefore the speaker must remain within the frame. In contrast, the human visual field can easily be adjusted by moving the eyes, head or whole body.
- *Depth perception* – the camera is typically two dimensional, whereas the human visual system includes depth perception.

Recognition

The recognition component of an ALR system is subject to two main challenges: vocabulary size and speaker independence [10].

Regardless of the recognition architecture, increasing the vocabulary size causes the recognition accuracy to decrease. To address this issue, ALR systems are built to recognise speech units as opposed to whole words. Using this approach, a relatively small group of speech units can be used to build a large vocabulary. Nevertheless, it may be necessary to limit the vocabulary size to facilitate meaningful use of the ALR system. The ensuing challenge is to construct a dictionary of words that is of limited size, but rich enough to be genuinely useful [10].

The performance of the recognition component is often dependent on specific characteristics of the speaker. For example, an ALR system is likely to be more accurate on people with more exaggerated speech movements. Speaker dependence must be carefully considered in the design of the recognition component. One approach is to train speaker-dependent recognition models, which can be continuously improved using an online learning algorithm – the more the system is used, the better it becomes. The downside is that the models become tailored to a specific speaker and lose the ability to generalise speech trends.

3.3 Speech Units

The recognition module is designed to identify a particular type of speech element, which may range from phonemes to entire sentences. Selection of an appropriate speech element depends primarily on the size of the vocabulary. For a small vocabulary, choosing isolated words as the target speech element has the advantage of simplicity in implementation and high recognition accuracy [63]. However, the isolated word approach cannot easily be extended to applications that require large vocabularies, since the increased vocabulary size prohibits adequate recognition accuracy [63]. Thus, in large vocabulary applications (such as voice restoration for laryngectomy patients) the recogniser must be designed to identify the speech units which compose the words. In the audio domain the fundamental unit of speech is the *phoneme* while in the visual domain the fundamental unit of speech is the *viseme*. The following definitions apply:

Phoneme

A phoneme is the basic unit of acoustic speech that is employed to form meaningful contrasts between utterances in the audio domain [65, 72]. If one phoneme is replaced by another, the meaning of the utterance is changed [65].

Viseme

A viseme is the basic unit of visual speech that describes the particular facial and oral movements that occur with the production of a particular phoneme [73]. A viseme may correspond to a single stationary pose or a dynamic movement; however, most visemes can be approximated by stationary poses [67]. Visemes are derived from groups of phonemes having the same appearance [68, 74].

A viseme-based recognition approach attempts to identify the individual visemes in each word. This approach has the advantage that a small number of visemes facilitates a relatively large vocabulary of words. The exact number of visemes to be recognised depends on which phoneme-to-viseme map is used. For example, the *MPEG-4* phoneme-to-viseme map defines 14 viseme classes [75]. However, since a viseme represents only a small piece of a word, it is generally more difficult to differentiate individual visemes as opposed to individual words.

3.3.1 Phoneme-to-Viseme Map

A phoneme-to-viseme map is a many-to-one map of the one or more phonemes corresponding to each viseme [74]. The map is many-to-one as phonemes cannot be individually distinguished using only visual information, but phonemes can be grouped based on visual information [74]. There are two approaches to developing a phoneme-to-viseme map: the linguistic approach and the data-driven approach.

Linguistic Approach

The linguistic approach creates visemes based on linguistic knowledge and subjective perception experiments [74, 76]. Speech reading movements form the linguistic foundations used to create visemes. A *speech reading movement* is a recognisable visual motor pattern, usually common to two or more speech sounds [69]. Speech reading movements are characterised by movements of the lips, jaws, teeth, tongue, and hyoid bone [69]. Table 3.1 shows a summary of the speech reading movements and corresponding phonemes determined by Jeffers & Barley [69]. “Visibility rank” refers to the ease with which the speech reading movement can be identified. “Frequency”

refers to the frequency of occurrence of the speech reading movement in spoken language. Table 3.1 can be used to create a phoneme-to-viseme map by simply creating a viseme class for each speech movement. Several different phoneme-to-viseme maps using the linguistic approach have been proposed, for example: Jeffers & Barley [69] and Bozkurt et al. [77].

Table 3.1: Visibility rank and frequency of speech reading movements used in conversational speech (adapted from Jeffers & Barley [69]).

Visibility Rank	Description	Phonemes	Frequency (%)
1	Lower lip to upper teeth	f, v	3.15
2	Lips puckered (narrow opening)	w, r, uː, ʊ, oʊ, ɜr	15.49
3	Lips together	p, b, m	5.88
4	Lips relaxed (moderate opening) to lips puckered (narrow opening)	aʊ	0.70
5	Tongue between teeth	θ, ð	2.90
6	Lips forward	ʃ, ʒ, tʃ, dʒ	1.20
7	Lips rounded (moderate opening)	ɔː	1.73
8	Lips back (narrow opening)	j, iː, ɪ, eɪ, ʌ, ə	20.51
9	Lips rounded (moderate) to lips back (narrow)	ɔɪ	0.08
10	Teeth together	s, z	4.36
11	Tongue up and down	t, d, n, l	21.10
12	Lips relaxed (moderate opening)	ɛ, æ, ɑː	7.79
13	Lips relaxed (moderate) to lips back (narrow)	aɪ	3.16
14	Tongue back	k, g, ŋ	4.84

Data-Driven Approach

The data-driven approach creates visemes by clustering the phonemes based on features extracted from visual speech data [74]. The data-driven approach has two main advantages over the linguistic approach: first, the performance of viseme recognition systems may be enhanced by using similar statistical techniques to create

the viseme classes and to perform recognition on the unknown visemes [74]; second, the data-driven approach is performed on continuous speech whereas the linguistic approach is focused on isolated phonemes, thus the data-driven approach can account for co-articulation effects [74]. Some examples of phoneme-to-viseme maps generated using a data-driven approach are: Hazen et al. [78], MPEG-4 [75], Mattheyses et al. [79], and Goldschen [80].

In recent years, the phoneme-to-viseme map described in the MPEG-4 standard [75] has become the most commonly used phoneme-to-viseme map [79]. The MPEG-4 mapping is shown in Table 3.2 and comprises fourteen different visemes augmented with one viseme for silence.

Table 3.2: MPEG-4 standard phoneme-to-viseme map [75].

Viseme No.	Phonemes	Examples
0	–	silent viseme
1	p, b, m	put, bed, mill
2	f, v	far, voice
3	θ, ð	think, that
4	t, d	tip, doll
5	k, g	call, gas
6	tʃ, dʒ, ʃ	chair, join, she
7	s, z	sir, zeal
8	n, l	lot, not
9	r	red
10	ɑː	car
11	e	bed
12	ɪ	tip
13	ʕ	top
14	ʊ	book

3.3.2 Trisemes

Recognition of individual visemes without taking into account neighbouring visemes is prone to errors due to the effects of co-articulation: the appearance of a sound can be altered by the sounds that precede or follow [69, 77]. The effects of co-articulation can be addressed by creating context-dependent viseme models, which depend on a sequence of visemes as opposed to isolated visemes. A common approach to

constructing context-dependent models is to use *trisemes*. A *triseme* is a triplet of visemes, with each triseme corresponding to a sequence of three visemes [69, 77]. Constructing a recogniser based on trisemes has the advantage of incorporating co-articulation effects into the model; however, the drawback of this approach is a significant increase in the number of classes that must be identified. The MPEG-4 standard defines 14 viseme classes corresponding to $14^3 = 2744$ potential triseme classes. In practice, the number of trisemes can be significantly reduced by excluding the trisemes that do not occur in the vocabulary. Goldschen [80] takes this approach one step further by grouping similar trisemes together to create "generalised trisemes".

3.4 Overview of ALR System

Figure 3.1 shows the components of an ALR system: face detection, mouth region detection, colour transformation, feature extraction, recognition, and grammar model. This section briefly discusses the approaches, technologies and considerations associated with each component.

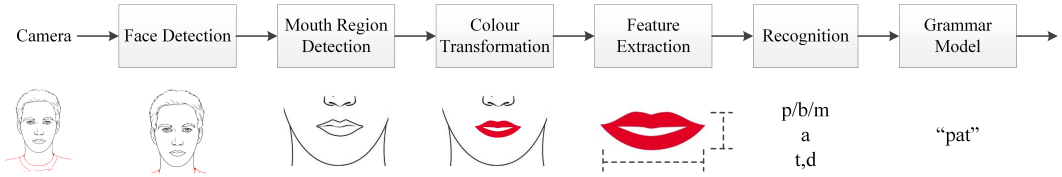


Figure 3.1: Components of ALR system.

3.4.1 Face Detection

There are three main approaches to face detection for an ALR system, which may be used in isolation or in combination. The first approach is based on skin hue, which is surprisingly independent of race and is robust to variations in lightness [81]. However, additional morphological constraints are required to reject hands and other skin-coloured objects.

The second approach is based on the characteristics of facial features, including the eyes, nose, and mouth. Facial features can be distinguished by their high edge content and low reflectance [81]. Fixed spatial relationships between facial features can be used to exclude erroneous candidates. Nostrils are generally the most robust facial feature as they are not occluded by facial hair or glasses.

The third approach to face detection is based on motion. The speaker is generally the only moving object in the frame, and the speaker's mouth is the primary source of movement. Additional morphological information is required to discriminate the face if alternate sources of movement are present in the frame.

3.4.2 Mouth Region Detection

The *mouth region* is a rectangular region containing the lips and surrounding skin. The size of the mouth region depends on the degree of mouth opening during a particular speech sound. For example, the sounds /w/ and /r/ are formed with the lips puckered, therefore the resulting mouth region is relatively small; however, for the sound /ɑː/ the mouth is wide open resulting in a large mouth region. The mouth region must be large enough to contain the entire mouth (lips, teeth and oral cavity) during the formation of any speech sound, but small enough to exclude any unnecessary information (e.g. nostrils) which will increase the difficulty of the lip segmentation stage.

WenJuan et al. [82] propose an algorithm to locate the mouth region based on the spatial relationships between the face, eyes and mouth. The algorithm uses the outline of the face and the location of the eyes to extrapolate the mouth region. 300 face images are analysed, and the following spatial relationships are observed:

1. The mouth is contained within the vertical lines defined by the eyes.
2. The mouth is contained within the bottom half of the region between the eyes and the chin.
3. The mouth is parallel to the horizontal line between the eyes and rotates accordingly.

WenJuan et al. [82] report that the algorithm to locate the mouth region is robust and unaffected by facial expressions, lip shapes or lighting conditions.

Cappelletta & Harte [83] propose a method to locate the mouth region based on nostril detection and the Accumulated Difference Image (ADI). The *Accumulated Difference Image (ADI)* is used to extract a moving object from a stationary scene based on the differences in successive frames. Cappelletta & Harte [83] use the ADI to locate the mouth, which is the primary source of motion in the frame. However, the ADI also detects movements of the head which results in lower accuracy in locating the mouth region. To address this issue, Cappelletta & Harte [83] use the

motion and orientation of the nostrils to generate a motion-compensated ADI, which compensates for movements of the head by observing movements of the nostrils. The compensated ADI is less noisy than the non-compensated ADI. Cappelletta & Harte [83] report a 100 % success rate in detecting the mouth region if the nostrils are successfully detected. The success rate for nostril detection is 74 %.

In recent years, the Viola-Jones detector has emerged as a popular technique for face detection and mouth region detection [82, 84–88]. The Viola-Jones detector was first introduced by Viola & Jones in 2001 as "a machine learning approach for visual object detection, capable of processing images extremely rapidly and achieving high detection rates". The Viola-Jones detector combines four key concepts: simple rectangular features called "Haar features", an Integral Image for rapid feature detection, the AdaBoost machine-learning method, and a cascaded classifier to combine many features efficiently [90]. The popularity of the Viola-Jones detector can be attributed to its implementation in OpenCV (an open source computer vision library), making it accessible to thousands of image processing researchers. OpenCV was designed and developed by Intel, focusing on computational efficiency and real-time application.

3.4.3 Colour Transformation

Selection of a suitable colour transform to enhance the lip-to-skin contrast is essential to the performance of the lip segmentation algorithm. An appropriate colour transform will result in a clear distinction between the lip and skin pixels, and the subsequent feature extraction will yield superior results. There is no consensus among researchers as to the most appropriate colour transform for this purpose. Chapter 5 presents a comprehensive comparison of the colour transforms used in lip segmentation algorithms and evaluates 33 different transforms: 21 channels from 7 colour space models (RGB, HSV, YCbCr, YIQ, CIEXYZ, CIELUV, CIELAB); and 12 additional colour transforms (8 of which are designed specifically for lip segmentation).

3.4.4 Feature Extraction

Feature extraction techniques are used to generate a sequence of feature vectors which characterises the movement of the lips. The techniques can be grouped into two broad categories: image-based approaches and model-based approaches [71]. The

image-based approach and the model-based approach may be used in combination to form a hybrid approach [91]. The fundamental principles behind the image-based approach and the model-based approach are discussed below.

Image-based Approach

The entire mouth region is presented to the recogniser after applying a colour transform. Effectively, every pixel in the mouth region is a feature. This approach ensures that no information is lost; however, the high dimensionality and high redundancy of the feature vector presents a significant computational processing challenge [71, 81]. Consequently, the pure image-based approach is considered unsuitable for real-time lip-reading systems [71]. To reduce the computational requirements, the dimension of feature vector can be reduced by applying a dimension reduction technique (PCA, DCT, DWT, LDA, FLD, MLLT) or by segmenting the lips and measuring various parameters (e.g. area, width and height) [92]. However, the image-based approach does not incorporate any shape or smoothness constraints, so the segmentation is often very rough and potentially unsuitable for lip-reading [93]. A further disadvantage of the image-based approach is that the system relies on the classifier to learn the non-trivial task of finding the generalisation for translation, scaling, rotation, illumination, and inter/intra speaker variability [94]. Finally, the system is highly sensitive to variations in illumination and orientation of the camera which will alter all the pixel values [81].

Model-based Approach

The model-based approach describes the lip contours by means of a small set of parameters, which are usually invariant to translation, rotation, scale, and illumination [71]. The model-based approach usually involves fitting a model to the image by minimising a cost function [91]. The important visual features can be represented in low dimensional space by constructing a feature vector comprising both low-level and high-level features [94]. Low-level features are usually parameters taken directly from the model (e.g. intensity values or landmark point coordinates); whereas, high-level features represent an abstraction derived from the model parameters (e.g. histograms or shape models based on landmark points). Challenges to the model-based approach include variations in speaker appearance, speaker sound formation, illumination, and poor colour contrast [71]. In addition, the model may inadvertently omit relevant

speech information [94]. Finally, optimising a cost function can be computationally expensive.

3.4.5 Recognition

An Artificial Intelligence (AI) recognition algorithm is used to classify unknown speech movements based on a training dataset. The input to the recognition algorithm is a temporal sequence of feature vectors obtained from the feature extraction component. Hidden Markov Models (HMMs) are the most common recognition technique in automatic speech recognition systems [95]; however, Neural Networks (NNs) have also received much attention [81].

The popularity of HMMs is primarily attributed to their inherent rate invariance and efficient algorithms for training and recognition [81]. The rate invariance property allows the system to effectively model variations in the rate of speech, which may occur with different speakers [81]. The efficient training and recognition algorithms are crucial for real time applications [96]. However, HMM-based recognisers have the following drawbacks: difficulties in modeling co-articulation effects, difficulties in determining appropriate state complexity, and inherent assumptions that may not be valid (e.g. features are uncorrelated) [81].

NNs have the advantage that they make few assumptions about the underlying data; however, training is slow and rate invariance is challenging to achieve [81, 96].

HMMs and NNs are considered "black box techniques" as they are difficult to understand and modify. Baldwin et al. [97] propose a recogniser based on fuzzy set theory which is a significantly more intuitive way of modeling speech data and is referred to as a "glass box technique". Baldwin et al. [97] achieved an accuracy of 92.71 % on a four word vocabulary (Tulips1 dataset); however, the use of fuzzy set theory for ALR has not been evaluated on larger vocabularies.

3.4.6 Grammar Model

Language is an auditory phenomenon and is severely limited as a vehicle for visual communication [69]. The fundamental challenge of lip-reading is constructing meaningful words and sentences given the incomplete speech information in the visual domain. Lip-reading is not a simple combination of elements or parts, but requires guessing and mental filling-in as well. A crucial requirement of successful lip-reading

is synthetic ability, defined by Jeffers & Barley [69] as: "the ability to make associations and to arrive at perceptual and conceptual closures when a good part of the sensory information is either missing or not perceived". *Perceptual closures* refers to the identification of words and phrases. *Conceptual closures* refers to the organisation of words and phrases into a tentative idea. Inherent in the lip-reading task is extrapolating the correct meaning from limited information, and thus a grammar model is a crucial component of an ALR system.

The recognition accuracy of an ALR system can be significantly improved by incorporating a grammar model to capture some of the rules of the spoken language [80]. The most simple grammar model is *word-pair grammar* which uses training sentences to derive a list of word-pairs that can appear in a sentence. Based on the permissible word-pairs, a candidate word is either accepted or rejected. An n-gram language model is a slightly more advanced language model, which is commonly used in speech recognition. An *n-gram* is a sequence of n symbols (e.g. words, syntactic categories, etc.) and an *n-gram language model* predicts the probability of each symbol in the sequence given its $n-1$ predecessors [98]. An n-gram language model is built by analysing the symbol frequencies in a given training text [98]. Goldschen [80] suggests using bigram grammar while Huang et al. [95] use trigram grammar.

3.5 Review of Existing ALR Systems

Stork & Hennecke [81] review 24 ALR systems and provide an overview of the image processing, feature extraction, sensory integration, and pattern recognition techniques. Table 3.3 provides a summary of these techniques [81].

Two well-known ALR systems are reviewed below to give an indication of the typical structure and accuracy that can be expected: Wang et al. [71] describe a system to recognise isolated words, while Goldschen [80] describes a system to recognise entire sentences.

Wang et al. [71] transform the RGB lip image to the CIELAB colour space and CIELUV colour space to enhance the perceptual colour difference. A fuzzy c-means clustering algorithm is employed to segment the lip image. The outer lip contour is modeled using a 14-point Active Shape Model (ASM). The ASM incorporates an alignment stage whereby the lip image is aligned to an original reference size and rotational angle. The alignment stage helps negate the effects of variation in camera setting and speaker head position. Wang et al. [71] normalise the lip image

Table 3.3: Summary of techniques used in the ALR systems reviewed by Stork & Hennecke [81].

Component	Techniques
Face detection	• Colour – similarity to a template skin colour • Motion – face located by areas of significant motion
Mouth region detection	• Colour – similarity to a template skin colour • Edge – colour or luminance edges around the lips • Value – lightness differences between the lips and skin • Lipstick – marked the lips with chroma key • Eyes and nostrils – ratios based on initial location of eyes and nostrils
Feature extraction	• Threshold – colour or luminance threshold • Dots – reflective dots applied around the face • Motion – visual motion, optical flow • Vector Quantisation (VQ) • Surface – fitting contours in high-dimensional pixel space • Principal component analysis (PCA) • Linear discriminant analysis (LDA) • Deformable templates • Active Shape Models (ASM) • Active Contour Models (ACM)
Recognition	• Linear Time Warping (LTW) • Dynamic Time Warping (DTW) • Neural Network (NN) – multilayer perceptron • Time-Delay Neural Network (TDNN) • Hidden Markov Models (HMM) • Boltzmann Zippers (BZ)

with respect to the first image in a lip image sequence, which enables dynamic monitoring of the scale (s) and rotational angle (θ). Wang et al. [71] construct a feature vector comprising the outer lip contour parameters, teeth area (T_{area}) and mouth opening (M_{open}). HMMs with a continuous probability distribution are used for classification of whole word models. A database of 10 isolated English digits (0 to 9) is used to evaluate both speaker dependant and speaker-independent recognition. The performance of the system is tested at isolated word-level. The speaker-dependent experiment yielded an accuracy of 93.3 % while the speaker-independent experiment yielded an accuracy of 84 %.

Goldschen [80] uses optical information from the oral cavity shadow of the speaker to perform continuous lip-reading. The feature vector is comprised of thirteen features, the majority of which are dynamic features. The recogniser is based on HMMs using visemes, trisemes, and generalised trisemes. The system is designed to recognise entire sentences, and does not attempt to resolve the individual words within the sentence. The system achieved a recognition rate of 25.3 % on sentences having a perplexity of 150, without the use of a grammar model (syntactic, semantic, acoustic or contextual).

3.6 Conclusion

Automatic lip-reading (ALR) is the technique of retrieving speech information from visual cues including the movement of the lips, tongue, and teeth. After almost three decades, ALR remains an active area of research with the primary goal of combining visual and auditory modalities to improve speech recognition in noisy environments. The main challenges of ALR are lip segmentation and recognition. Lip segmentation is challenging due to variability in speaker profile, speaker orientation, lighting and background activity. The recognition component is faced with the challenges of vocabulary size and speaker independence. In large vocabulary applications, the recogniser is designed to identify the visual speech units ("visemes") which make up words. A phoneme-to-viseme map is required to translate the visual information back into the auditory domain, from where words and sentences can be reconstructed.

An ALR system is comprised of six major components: face detection, mouth region detection, colour transformation, feature extraction, recognition, and grammar model. The various approaches, technologies and considerations associated with each component have been discussed in this chapter.

Chapter 4

Motivation, Scope, and Objectives

4.1 Introduction

Chapter 2 describes the voice restoration options for laryngectomy patients, covering both conventional techniques and experimental computer-based approaches. Chapter 3 presents the background to automatic lip-reading (ALR), including the challenges, techniques, and technologies. This chapter consolidates the argument for using ALR to restore verbal communication for laryngectomy patients. Once this basis has been established, the scope of the research is narrowed to focus on the unresolved image processing challenge. The dataset used to conduct the research is also described in this chapter.

4.2 Motivation – The Case for ALR

Patients suffering from laryngeal cancer may undergo a laryngectomy to remove the cancer, leaving them unable to speak or produce vocalisation. These patients generally have three options to consider for voice restoration: the electrolarynx, tracheo-oesophageal speech, and oesophageal speech. While these techniques have achieved some success in restoring speech, there are various limitations and drawbacks associated with each technique. The electrolarynx must be held by hand and is difficult to understand [29]; tracheo-oesophageal speech requires surgical intervention and the prosthetic valves fail after 3 – 4 months [29, 33, 34]; oesophageal speech is difficult to learn, and suffers from short duration and poor voice quality [29, 41].

With recent advances in portable computing power, a technological solution to this

problem has become feasible. Silent speech interfaces (SSIs) use digital sensors and processing to decipher speech in the absence of an audio signal. Conventional methods of voice restoration use the remaining structures of the vocal tract to define a new speech production process; consequently, these techniques are not spontaneous or intuitive, and produce a foreign-sounding voice. By contrast, SSIs augment the existing speech production process with sensors and processing, and thus have the potential to be easier to learn and produce a more natural-sounding voice [10]. The voice synthesiser can be designed to mimic the patient's pre-laryngectomy voice, improving the intelligibility and patient satisfaction [53].

Laryngectomy patients are restricted to SSI systems that do not require glottal activity. Seven suitable technologies have been used to build experimental SSIs in the literature: optical imaging, ultrasound, electropalatography (EPG), electromagnetic articulography (EMA), surface electromyography (sEMG), electroencephalography (EEG), and intracranial electrode implants.

SSIs based on optical imaging offer several distinct advantages over the alternatives:

- *Hardware availability* – The hardware requirements of an optical SSI including camera, processing power, and audio/video output are readily available in any standard smartphone, tablet or laptop.
- *Cost* – The seven SSI technologies require similar processing power and output capabilities, thus the major cost differentiator is the sensor technology and supporting electronics. An optical SSI is by far the most affordable technology using a standard digital video camera.
- *Comfort* – An optical SSI is the only technology that does need to be physically attached to the speaker – the smartphone, tablet or laptop can simply be placed on a table in front of the speaker. Thus, an optical SSI is likely to be the most comfortable and unobtrusive SSI technology for regular and long-term use.
- *Set up* – An optical SSI is simple to set up and fit, and the patient does not require assistance from another person.
- *Portable* – An optical SSI is compact and can easily be carried in a pocket or handbag.
- *Health risk* – An optical SSI is non-invasive and does not pose any danger to the patient.

Importantly for optical SSIs, facial expressions do not change after a laryngectomy [84]; therefore, the system can be designed using traditional linguistic theory, and

can be tested on regular speaking individuals.

In an optical SSI, ALR is the technique of retrieving speech content from visual clues, such as the movement of the lips, teeth and tongue. The premise of ALR is based on the proficiency of human lip-reading experts who can achieve near perfect speech comprehension.

4.3 Scope

ALR research is centred around the two core components: lip segmentation and recognition. The challenge of lip segmentation resides in the domain of image processing and involves accurately discriminating between lip pixels and non-lip pixels. The challenge of recognition resides in the domain of machine learning and involves identifying meaningful speech information from the movements of the lips.

While both lip segmentation and recognition remain active areas of ALR research, this thesis focuses on the image processing challenge of accurate and robust lip segmentation. Development of the recognition algorithm falls outside the scope of this work. In a complete ALR system, the output of the lip segmentation algorithm is fed into the recognition algorithm.

Although numerous segmentation techniques have been developed over the past decades, few of these techniques have been applied successfully to the task of lip segmentation owing to the low chromatic and luminance contrast between the lips and skin [99]. Lip segmentation presents a challenging image processing problem arising from three levels of variability. First, the inherent challenge of lip segmentation is variability in the speaker profile including skin colour, lip colour, lip shape, facial hair, and make-up. Second, the contents of the region of interest (ROI) is not static and the visibility of the teeth, tongue, and oral cavity changes as the lips move to form facial expressions and speech sounds. Finally, non-ideal environmental conditions including lighting, speaker orientation, and background create a third layer of complexity.

4.4 Objectives and Contributions

The first stage of lip segmentation involves applying a colour transform to enhance the contrast between the lips and the surrounding skin; however, there is much debate

among researchers as to the most suitable colour transform. The first objective of this thesis is to determine the best colour transform for lip segmentation by evaluating and comparing 33 different transforms.

The second objective is to develop a new lip segmentation algorithm that utilises the best colour transforms from the comparative study. The resulting algorithm is referred to as the *base algorithm*, as it forms the platform to analyse the impact of the threshold value on the segmentation accuracy.

The final objective is to improve the segmentation accuracy by selecting a better threshold value. While this thesis concerns lip segmentation in particular, the objective is to develop a threshold selection technique that can be used in various segmentation applications.

Thus, this thesis comprises three contributions to the field of image processing:

1. Comparison of 33 colour transforms used in lip segmentation algorithms
2. Development of a new colour-based lip segmentation algorithm
3. Development of a novel threshold selection technique called *Adaptive Threshold Optimisation (ATO)*

4.5 Dataset

The dataset used in lip segmentation research should satisfy three criteria:

1. The dataset should cover the range of shapes formed by the lips during regular speech, which includes: lips relaxed, lips pressed together, lips puckered, lips rounded, lips forward, and lips back.
2. The dataset should contain sufficient inter-speaker variability.
3. Corresponding ground truth should be available.

The AR Face Database [100], constructed and curated by Prof. Aleix Martinez from The Ohio State University, is the de facto standard in lip segmentation research since the database satisfies all three necessary criteria. The AR Face Database is publicly available and is free for research purposes (`http://www2.ece.ohio-state.edu/~aleix/ARdatabase.html`).

The database comprises full-face images from 112 subjects (58 male, 54 female), and includes four different expressions: neutral, smile, anger, and scream. The four facial expressions cover the maximum and minimum dimensions of the lips that can be expected during regular speech:

- Scream – the mouth is wide open corresponding to maximum height and area
- Smile – the lips are stretched in the horizontal direction corresponding to maximum width
- Anger – the lips are pressed tightly together corresponding to minimum height, width and area

The AR Face Database includes significant variability in speaker profile, featuring male and female speakers from diverse range of racial and ethnic groups, as well as varying degrees of facial hair and make-up. Each subject participated in two separate sessions fourteen days apart, which increased the variety of clothing, hairstyles and make-up. The participants were asked to repeat the four facial expressions in each session.

The dataset used in this research comprises 895 images from the AR Face Database. The corresponding manual markings were produced by Ding & Martinez [101], who obtained labels of the facial features from three independent human judges. Figure 4.1 shows the facial key points, including the outer lip contour which is labelled with 20 points. The within-judge variability was 3.8 pixels or 1.2 %.

Since the technologies to locate the face and mouth region are well established [75, 102], many researchers in the field of lip segmentation begin with pre-cropped images of the mouth region [96]. In line with this approach, the manual markings from Ding & Martinez [101] are used to crop the full face images to a rectangular mouth region including only the lips, oral cavity, and surrounding skin as shown in Figure 4.2.

The 895 cropped mouth regions, along with corresponding manual markings of the lip contours, form the dataset used in all subsequent testing and analysis.

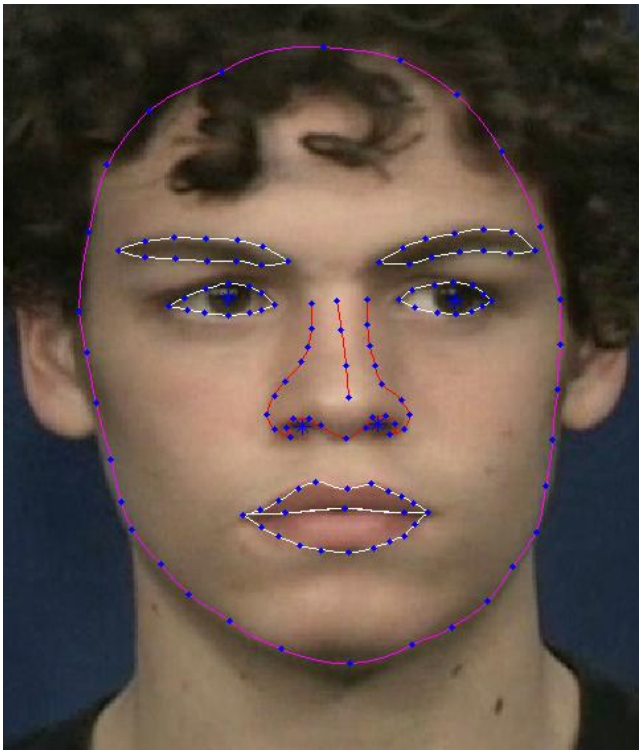
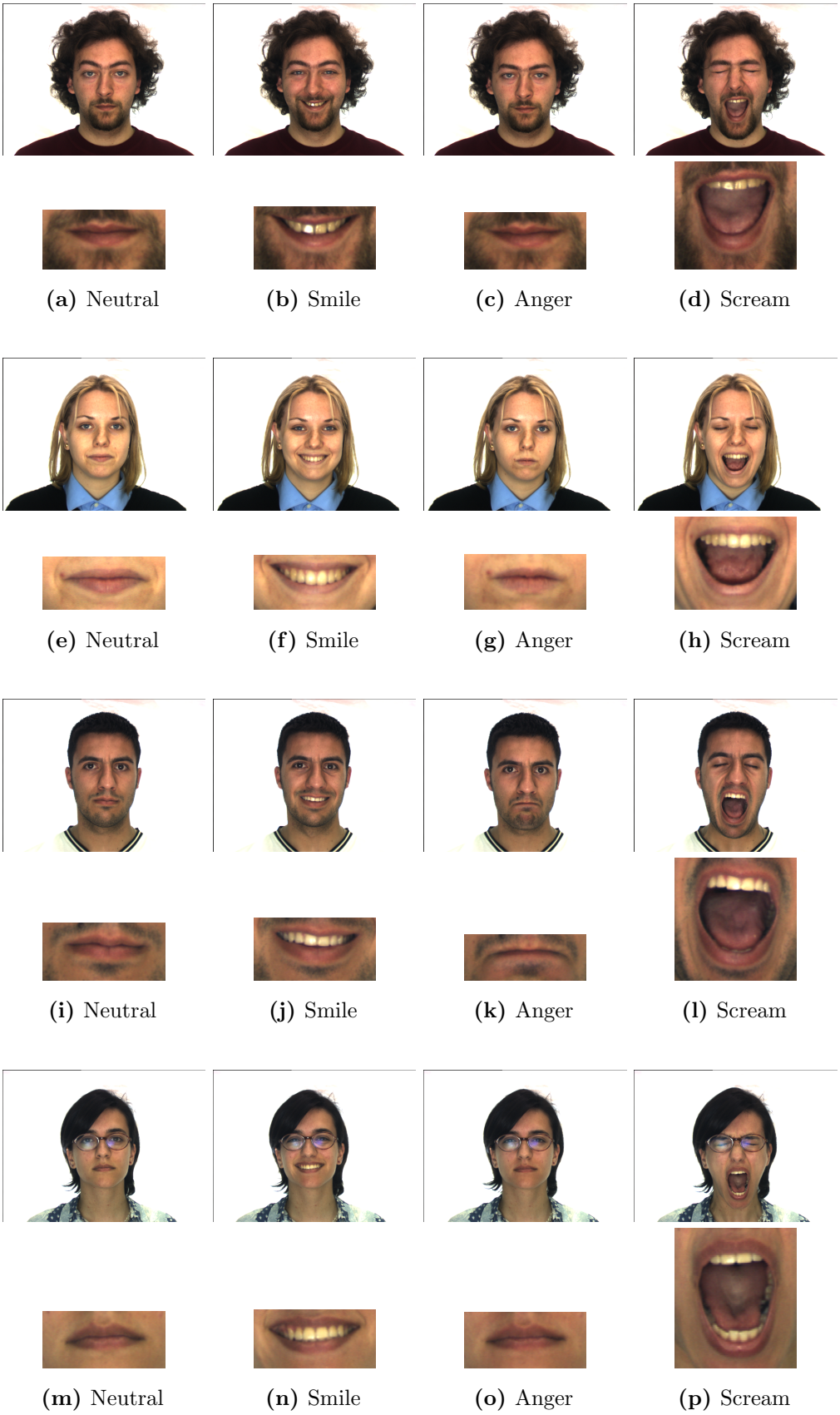


Figure 4.1: Marking of keypoints on the AR Face Database by Ding & Martinez [101].



(a) Neutral **(b)** Smile **(c)** Anger **(d)** Scream

(e) Neutral **(f)** Smile **(g)** Anger **(h)** Scream

(i) Neutral **(j)** Smile **(k)** Anger **(l)** Scream

(m) Neutral **(n)** Smile **(o)** Anger **(p)** Scream

Figure 4.2: The manual markings by Ding & Martinez [101] are used to crop the mouth region from the full-face images of AR Face Database [100].

4.6 Conclusion

This chapter presents the motivation and justification behind research into ALR as a method of voice restoration for laryngectomy patients. This thesis is focused on the image processing challenges of ALR, resulting in three contributions to the field: first, the debate surrounding the best colour transform for lip segmentation is resolved by evaluating and comparing 33 colour transforms used in lip segmentation algorithms; second, a new colour-based lip segmentation algorithm is developed; third, a novel threshold selection technique called *adaptive threshold optimisation (ATO)* is presented. ATO can be used in various segmentation applications in which prior shape information is available. The dataset used to test and validate the contributions of this research comprises 895 pre-cropped mouth regions from the AR Face Database.

Chapter 5

Comparison of Colour Transforms

5.1 Introduction

Lip segmentation is the task of accurately discriminating between lip pixels and non-lip pixels. Lip segmentation is a fundamental system component in a wide range of applications including: automatic lip-reading, virtual face animation, biometric speaker identification, and emotion recognition [103]. The task of lip segmentation is challenging due to variability in speaker profile (skin colour, lip colour, lip shape, facial hair, make-up); speaker orientation (scale, pose); illumination (intensity, shadows, glare); and speaker background [63, 104].

The preliminary phase of lip segmentation involves location of the face, and subsequent location of the mouth region – a rectangular region containing the lips and surrounding skin (see Figure 5.1). Once the mouth region has been located, the lip pixels are segmented from the surrounding skin pixels. The technologies to locate the face and mouth region are well established [75, 102]; however, there is no consensus as to the most suitable technique to extract the lips from the mouth region [14]. Consequently, many researchers in the field of lip segmentation begin with images containing the pre-cropped mouth region [99].

There are two main approaches to segmentation of the lips from the mouth region: the region-based approach and the model-based approach. The region-based approach typically involves a colour transform to enhance the contrast between the lips and surrounding skin, followed by thresholding and subsequent morphological operations (e.g. dilation and erosion). The model-based approach again uses a colour transform to obtain an intensity image, followed by minimising a cost function to fit a predefined

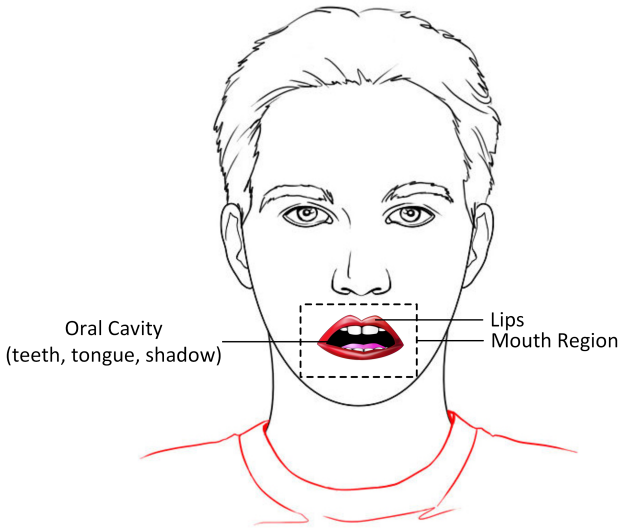


Figure 5.1: Mouth region, lips and oral cavity (teeth, tongue, shadow).

model to the lip contours. The main techniques used to construct an appropriate lip model are: deformable templates, active shape models, and active contour models (snakes).

Figure 5.2 is a histogram showing the RGB components of the lips and skin respectively. The histogram clearly illustrates the inefficiency of the RGB colour space in lip segmentation – significant overlap between the lips and skin exists in all three channels. Selection of a suitable colour transform to enhance the lip-to-skin contrast is essential to the overall performance of a lip segmentation system. An appropriate colour transform will result in a clear distinction between the lip and skin pixels, and the subsequent region-based or model-based techniques will yield superior results. However, an unsuitable colour transform will result in the loss of significant contrast which will limit the accuracy of the subsequent segmentation techniques.

There is no consensus among researchers as to the most appropriate colour transform to enhance lip-to-skin contrast. Some authors propose using a single chrominance channel from various colour space models, for example: hue channel from HSV [105, 106], or the a^* channel from CIELAB [71, 82, 85, 107]; while others present specialised colour transforms for lip segmentation based on various observations or assumptions, for example: red exclusion [96] or pseudo-hue [108].

Zhang & Mersereau [109] attempt to address this issue by investigating three different colour space models for lip segmentation (RGB, HSV and YCbCr). They evaluate the histograms of full-face test images, and do not pre-crop the mouth region before comparing the colour transforms. Consequently, the test images include the eyes,

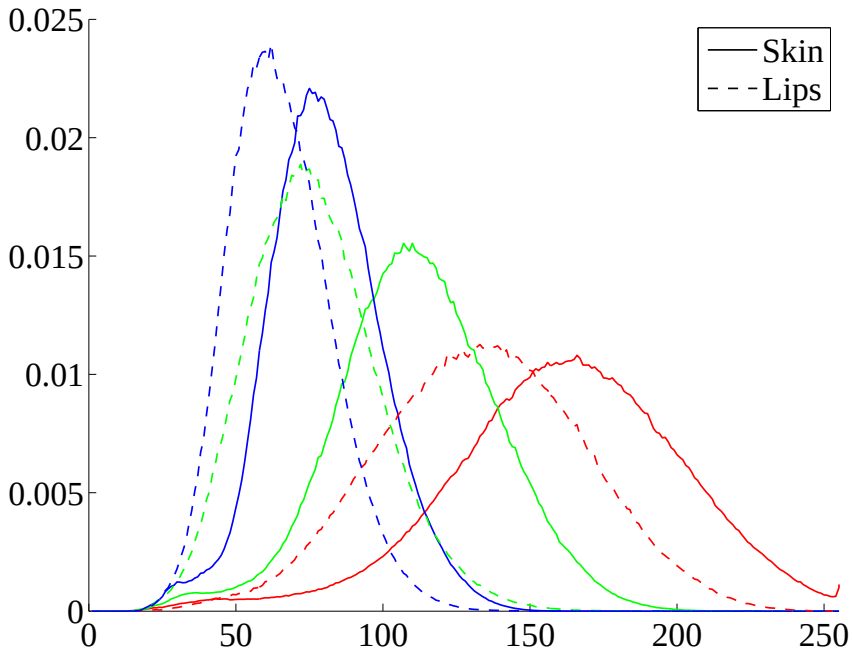


Figure 5.2: Histogram showing the R, G and B components of lips and skin.

nose, mouth, and neck, as well as clothing and background pixels. Again, algorithms to crop the mouth region from full-face images are well established, hence the approach adopted by Zhang & Mersereau [109] does not necessarily yield the most appropriate colour transform for lip segmentation. Furthermore, Zhang & Mersereau [109] draw their conclusions through qualitative observations on the histograms – no objective quantitative metric or analytic technique is used. Zhang & Mersereau [109] conclude that the hue component of HSV exhibits the least overlap between lip and non-lip pixels, while remaining relatively constant across test subjects and lighting conditions.

Caplier et al. [103] introduce an objective metric to determine the most appropriate colour transform for lip segmentation. Caplier et al. [103] use intra-class variance and inter-class variance to compare three colour space models (RGB, HSV, YCbCr) and two transforms designed for lip segmentation (pseudo-hue [110] and LUX [104]). Caplier et al. [103] build a database of skin and lip pixel samples to facilitate comparison of the colour transforms; however, the database is not publicly available and thus, it is not possible to verify the diversity of the database: skin colour, lip colour, facial hair, glasses, expression, etc. Caplier et al. [103] note that the three colour space models (RGB, HSV, YCbCr) are not particularly efficient, while the specific colour transforms (pseudo-hue and LUX) achieve better results.

The comparison of colour transforms presented in this chapter is the most comprehensive study to date and evaluates 33 different transforms: 21 channels from 7 colour space models (RGB, HSV, YCbCr, YIQ, CIEXYZ, CIELUV, CIELAB); and 12 additional transforms (8 of which are designed specifically for lip segmentation). The contrast between the lips and the skin is used to obtain the outer-lip contour; while the contrast between the lips and the oral cavity (teeth, tongue, shadow as in Figure 5.1) is used to obtain the inner-lip contour. Therefore, it is necessary to find a transform that adequately performs both tasks, alternatively two separate colour transforms are needed, one for lip-skin segmentation and the other for lip-oral cavity segmentation. It is worth noting that several automatic lip-reading systems exclude the lips entirely, and are developed around segmentation of the oral cavity exclusively [111].

The objective of this chapter is to determine the best transform for the following scenarios:

1. Lip to skin segmentation
2. Lip to oral cavity segmentation
3. Combined lip-skin and lip-oral cavity segmentation

This work has been published in *Signal, Image and Video Processing* [14].

5.2 Survey of Colour Transforms

This section presents a comprehensive survey of the colour space models and specialised colour transforms used for lip segmentation in the literature. The survey includes seven colour space models (RGB, HSV, YCbCr, YIQ, CIEXYZ, CIELUV, CIELAB), as well as eight additional transforms designed specifically for lip segmentation. Table 5.1 shows a summary of the 33 transforms to be evaluated.

5.2.1 Colour Space Models

RGB

The Red, Green and Blue (RGB) colour space model is an additive colour system based on the trichromatic theory [112]. The tri-chromatic theory is derived from the three types of colour-sensitive cones in the human visual system, and states that

Table 5.1: Summary of 33 colour transforms to be evaluated.

No.	Transform	Description
1	RGB-R	R channel from RGB colour space
2	RGB-G	G channel from RGB colour space
3	RGB-B	B channel from RGB colour space
4	HSV-H	H channel from HSV colour space
5	HSV-S	S channel from HSV colour space
6	HSV-V	V channel from HSV colour space
7	YCbCr-Y	Y channel from YCbCr colour space
8	YCbCr-Cb	Cb channel from YCbCr colour space
9	YCbCr-Cr	Cr channel from YCbCr colour space
10	YIQ-Y	Y channel from YIQ colour space
11	YIQ-I	I channel from YIQ colour space
12	YIQ-Q	Q channel from YIQ colour space
13	XYZ-X	X channel from XYZ colour space
14	XYZ-Y	Y channel from XYZ colour space
15	XYZ-Z	Z channel from XYZ colour space
16	LUV-L	L channel from LUV colour space
17	LUV-U	U channel from LUV colour space
18	LUV-V	V channel from LUV colour space
19	LAB-L	L channel from CIELAB colour space
20	LAB-A	a* channel from CIELAB colour space
21	LAB-B	b* channel from CIELAB colour space
22	DHT-C1	C1 channel from discrete Hartley transform
23	DHT-C2	C2 channel from discrete Hartley transform
24	DHT-C3	C3 channel from discrete Hartley transform
25	GL	Grey level
26	RE	red exclusion (Lewis et al., 2002)
27	PHE	pseudo-hue transform (Eveno et al., 2001)
28	PHT	pseudo-hue transform (Talea et al., 2011)
29	MI3	modified I3 Channel (Canzler et al., 2002)
30	HDF	hue domain filter (Gong et al., 1995)
31	Cb-Cr	Cb-Cr difference (Hsu et al., 2002)
32	CC	chromatic curve map (Eveno et al., 2001)
33	LUX-U	logarithmic hue extension (Lievin et al., 2004)

any visible colour can be formed by combining three independent colour channels [112, 113]. RGB is easy to implement and is widely used throughout computer graphics for capture, storage and display. However, the RGB colour space has two major drawbacks: first, RGB does not separate luminance and chrominance information; second, RGB is non-linear with visual perception – a 10 % change in stimulus does not produce a 10 % change in perception. As shown in Figure 5.2, using any single channel independently (R, G or B) does not adequately enhance the lip-to-skin contrast; nevertheless, since image processing systems use RGB to capture, store and display images, the RGB colour space forms the starting point of subsequent colour transformations.

HSV, HSI, HSL

The Hue, Saturation and Value/Intensity/Lightness colour space is a non-linear transformation from a Cartesian-coordinate representation (RGB) to a cylindrical-coordinate representation. HSV is designed to represent colour in an intuitive manner, thereby simplifying the task of quantifying a perceived colour using H, S and V values. This is achieved by separating luminance (S, V) and chrominance (H) components; however, HSV is not perceptually linear [113]. Stork & Hennecke [81] note that skin hue remains "surprisingly constant" across race, while lightness varies significantly. Eveno et al. [114] state that skin hue and lip hue remain relatively constant and well separated across a range of speakers, and thus, can be used effectively to discriminate between lips and skin. Consequently, the hue (H) channel is used extensively throughout the lip segmentation literature [105, 106, 109, 115].

YCbCr

The YCbCr colour space is a linear transformation of the RGB colour space, and is primarily used as a way of encoding RGB information. YCbCr decouples luminance (Y) and chrominance (Cb, Cr) components and is designed to improve storage and transmission efficiency by exploiting perceptually meaningful information [112]. For example, humans are more sensitive to luminance than chrominance; therefore, the efficiency of a system can be improved by storing the luminance with high resolution while compressing the chrominance components [85]. Hsu et al. [116] note that lip pixels often exhibit an increased chrominance component Cr, which can be exploited for lip segmentation.

YIQ

The YIQ colour space is a linear transformation of the RGB colour space and is used in the NTSC colour TV system [113]. YIQ is designed to exploit the colour response characteristics of the human eye to maximise the use of a fixed transmission bandwidth. The human eye is highly sensitive to changes in luminance (Y); therefore, the Y component is allocated a greater share of the transmission bandwidth. Furthermore, the human eye is more sensitive to changes in the orange-blue (I) range than the purple-green (Q) range; therefore, the I component is allocated greater bandwidth than the Q component. Thejaswi & Sengupta [117] observe that the lips are generally brighter in the Q channel and the face is generally brighter in the I channel; therefore, the I and Q components can be used to discriminate between lip and non-lip pixels. Thejaswi & Sengupta [117] also note that YIQ decouples the luminance and chrominance, providing an illumination-invariant approach.

XYZ

The XYZ (or CIEXYZ) colour space model was created in 1931 by the International Commission on Illumination (CIE) and forms the foundation of all colorimetry [113, 118]. The CIEXYZ colour space was derived from a series of experiments in which participants were required to match the colour of a test light by adjusting the contributions of three primary beams (red, blue, green) [118]. The commission used these results to create a mathematical representation of the way humans perceive colour. The CIEXYZ colour space defines all visible colours ("human gamut") using only positive values; consequently, the primaries X, Y, and Z ("tristimulus values") are not themselves visible. CIEXYZ is non-linear with visual perception and is highly unintuitive (difficult to quantify a perceived colour using X, Y, and Z values). CIEXYZ is not used directly in lip segmentation algorithms; however, several colour space models derived from CIEXYZ are widely used in lip segmentation.

L*u*v*

The L*u*v* (or CIELUV) colour space was recommended in 1976 by the CIE as an attempt to achieve perceptual linearity [113]. CIELUV is a non-linear colour space designed to mimic the logarithmic response of the eye. Wang et al. [71] use CIELUV to extract colour features for an Active Shape Model (ASM) of the lip contours, and use the u^* channel to detect the visibility of teeth.

L*a*b*

The CIE simultaneously adopted the L*a*b* (or CIELAB) colour space along with the CIELUV colour space in 1976. CIELAB represents a second attempt to achieve a perceptually linear colour space by mimicking the nonlinear response of the eye [113]. The lightness component (L^*) closely matches the human perception of lightness; a^* and b^* are colour-opponent dimensions (cannot be perceived together due to antagonistic physiological responses). CIELAB has received increased attention in recent years from lip segmentation researchers: Liang & Du [107] and WenJuan et al. [82] use the a^* component of CIELAB for lip segmentation, and report that the a^* component is highly robust to variability in skin and lip colour.

5.2.2 Specialised Colour Transforms

Several authors have proposed specialised colour transforms designed specifically to enhance the lip-to-skin contrast. This section introduces the fundamental observations and assumptions used in formulating the following specialised colour transforms: red exclusion [96]; pseudo-hue [63, 114]; modified I3 [119]; hue domain filter [105, 120]; Cb-Cr difference [116]; chromatic curve map [114]; logarithmic hue extension [104]; and discrete Hartley transform [121].

Red Exclusion

The *red exclusion (RE)* technique is based on the assumption that since both the lips and the surrounding skin are predominantly red, any contrast that may develop requires the exclusion of the red colour component [63]. Equation (5.1) shows the implementation of RE, incorporating only G and B colour components; β is a lip-skin threshold value [96, 122]. RE has been shown to be a simple and effective lip segmentation method [63, 96, 122]; however, WenJuan et al. [82] assert that RE only performs well on the white population.

$$\log\left(\frac{G}{B}\right) \leq \beta \tag{5.1}$$

Pseudo-Hue

Eveno et al. [114] develop a *pseudo-hue (PH)* transform based on several observations on the RGB colour composition of the lips and skin (see Figure 5.2) [108]:

- R is prevalent in both lips and skin
- For skin pixels: $G > B$
- For lip pixels: $G \approx B$
- The difference between R and G is greater for the lips than for the skin: $(R_{Lips} - G_{Lips}) > (R_{Skin} - G_{Skin})$. Therefore, the skin appears more yellow than the lips.

Based on these observations, Eveno et al. [108] design the pseudo-hue transform shown in (5.2) to emphasise areas where large differences between R and G occur. The pseudo-hue transform is bijective (exact pairing of elements between the two sets, i.e. inverse function exists), while standard hue is not [108]. Talea & Yaghmaie [63] propose a similar pseudo-hue transform shown in (5.3).

$$h = \frac{R}{G + R} \tag{5.2}$$

$$h = \frac{R^3}{R^3 + G^3 + 1} \tag{5.3}$$

Modified I3

Canzler & Dziurzyk [119] design a new colour transform for lip segmentation by modifying the I3 feature proposed by Ohta et al. [123]. They obtain the I3 feature by weighting the R, G and B colour components in an attempt to find an effective colour feature for region segmentation in a diverse group of images (e.g. building, seaside, home). Canzler & Dziurzyk [119] select the I3 feature for the following reasons: no singularities; good separation between brightness and colour information; simple and fast conversion from RGB. Canzler & Dziurzyk [119] modify the I3 feature specifically for lip segmentation by attenuating the B component, which is less prevalent in both the lips and skin (5.4).

$$I = \frac{1}{4}(2G - R - 0.5B) \tag{5.4}$$

Perhaps somewhat inadvertently, Canzler & Dziurzyk [119] weight the R, G, and B components in accordance with the respective lip-skin discriminating power. In Figure 5.2, the average of the skin R component is 161.44, while the lip R component is 132.74. The difference between average skin R and the average lip R is 28.70. Similarly, the difference in the G component is 34.94, and the difference in the B component is 15.17. Canzler & Dziurzyk [119] construct their transform such that the G component (which exhibits the greatest distinguishing power) is amplified, while the B component (which exhibits the least distinguishing power) is attenuated.

Hue Domain Filter

Gong & Sakauchi [120] propose combining a second order polynomial with convolution to effectively filter out a given colour feature in the hue domain. Gong & Sakauchi [120] report that this approach diminishes noise and improves the integrity of the segmented region (decreases holes scattered throughout the object and fragments scattered throughout the background). Coianiz et al. [105] use this approach to construct a hue domain filter (HDF) for lip segmentation, shown in (5.5).

$$f(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & if |h - h_0| \leq w \\ 0 & otherwise \end{cases} \tag{5.5}$$

h is the pixel hue; h_0 is a predefined value representative of lip hue; w controls the width of the filter (see Figure 5.3). Ideally, the values of h_0 and w should be tuned experimentally to achieve the best results; however, to facilitate fair comparison across the various transforms, parameter tuning is precluded and the HDF is implemented with the values suggested by Coianiz et al. [105].

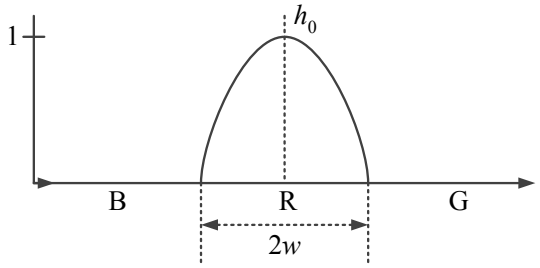


Figure 5.3: Hue domain filter with $h_0 = R$.

Cb-Cr Difference

Hsu et al. [116] analyse the colour composition of lip and skin pixels in the YCbCr colour space. Hsu et al. [116] note that the lip pixels are characterised by chrominance component Cr (red) greater than chrominance component Cb (blue); furthermore, the lip pixels have lower response in the Cr/Cb feature, but higher response in the Cr^2 feature. Based on these observations, Hsu et al. [116] propose the colour transform shown in (5.6).

$$I = \frac{Cr}{Cb} - Cr^2 \tag{5.6}$$

Chromatic Curve Map

Eveno et al. [114] use the Michaelis-Menten law to reduce the luminance dependence, followed by enhancing the lip pixels using a chromatic curve map. The luminance correction transform is derived from the Michaelis-Menten law which models the adaptation of human vision. The correction law is shown in (5.7) and enhances chromatic information such that colours are brought out of shadowy areas. a and b are parameters which control the weight of the luminance correction law. The chromatic curve map is created by defining a function which maps each pixel to a specific parabola. The parameters of the mapping function are determined by minimising a cost function to ensure that lip parabolas have a high curve and non-lip parabolas have a low curve.

$$X_{Cor} = \frac{X}{X + (b - a)L + a} \tag{5.7}$$

Logarithmic Hue Extension

Lievin & Luthon [104] propose a new colour space model called the logarithmic hue extension (LUX). LUX is a nonlinear colour space model inspired by physiological considerations (cone distribution in the fovea and nonlinear signal transduction pathways) and the Logarithmic Image Processing (LIP) model. Lievin & Luthon [104] use a simplified implementation of the U channel to enhance the lip-skin contrast, as shown in (5.8).

$$\hat{U} = \begin{cases} 256 \times \frac{G}{R} & \text{if R} > \text{G} \\ 255 & \text{otherwise} \end{cases} \tag{5.8}$$

Discrete Hartley Transform

The discrete Hartley transform (DHT) is closely related to the Discrete Fourier Transform; however, the DHT transforms real inputs into real outputs [124]. The three dimensional DHT is used in the field of image processing to perform frequency domain manipulations. Guan [121] observes that the C_3 component of the three channel DHT (shown in equation 5.9) is increased in the lip region, and exploits this observation to perform lip segmentation.

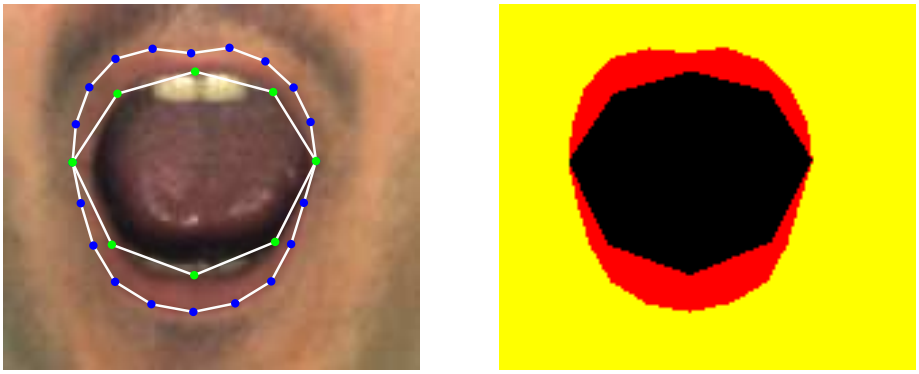
$$\begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} = \begin{bmatrix} 0.5773 & 0.5773 & 0.5773 \\ 0.5773 & 0.2113 & -0.7886 \\ 0.5773 & -0.7886 & 0.2113 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{5.9}$$

5.3 Set-up

This section describes the experimental set-up to compare the colour transforms, including the dataset, metrics, and methods employed.

5.3.1 Dataset

The dataset comprises 895 pre-cropped mouth regions from the AR Face Database [100]. The outer-lip contour is interpolated from 20 points while the inner-lip contour is interpolated from only 8 points; consequently, the outer-lip contour is more accurate than the inner-lip contour. The inner and outer-lip contours are used to generate the ground truth comprising three regions: lips, skin and oral cavity (see Figure 5.4).



(a) Manual markings **(b)** Ground truth

Figure 5.4: Manual markings from Ding & Martinez [101] used to interpolate the lip contours and generate the ground truth.

5.3.2 Metrics

Caplier et al. [103] use *intra-class variance* and *inter-class* variance to evaluate various colour transforms for lip segmentation. Otsu [125] first introduced the concepts of intra-class variance (5.10) and inter-class variance (5.11) to evaluate the "goodness" of a *single* threshold in segmenting a bimodal image (an image with two distinct peaks in the histogram). From a segmentation perspective, the ideal bimodal histogram contains two compact (low intra-class variance) and distinct (high inter-class variance) groups of pixels; therefore, Otsu [125] designed the discriminant criterion shown in (5.13) as a measure of class separability: η tends to 0 for highly similar classes, or 1 for highly differentiated classes. Otsu's discriminant operates most effectively on bimodal histograms, requiring a *single* threshold. Otsu used the discriminant criterion to determine an optimum threshold to segment images into nearly homogenous regions. Otsu's discriminant function is frequently referred to in the literature and has been widely used in a variety of applications [126] – for example, the GREYTHRESHOLD function in MATLAB^{TM} is an implementation of Otsu's algorithm [127].

$$\sigma^2_{Intra} = w_1\sigma^2_1 + w_2\sigma^2_2 \tag{5.10}$$

$$\sigma^2_{Inter} = w_1 w_2(\mu_2 - \mu_1)^2 \tag{5.11}$$

$$\sigma^2_T = \sigma^2_{Intra} + \sigma^2_{Inter} \tag{5.12}$$

$$\eta = \frac{\sigma^2_{Inter}}{\sigma^2_T} \quad 0 \leq \eta \leq 1 \tag{5.13}$$

where

w_i = probability of a pixel belonging to class i
σ^2_i = variance of class i
σ^2_T = variance of the total image
μ_i = mean of class i

Histogram intersection is another useful metric to quantify the distinguishing power of a colour transform as the intersection directly quantifies the number of pixels which overlap between the object and the background. Equation (5.14) is used to calculate the number of intersecting pixels between two histograms (H_1, H_2), where N is the number of bins. In the case of segmentation where H_1 represents the object and H_2 represents the background, $S_{Intersection}$ can be normalised by the total number of pixels (object + background) to give the overlap probability between the object and the background. The normalised intersection ranges from 0 for histograms that can be entirely differentiated (0 % overlap) to 1 for identical histograms (100 % overlap).

The intersection value indicates the upper limit on the segmentation accuracy (5.15) that can be achieved by implementing a simple thresholding scheme (bimodal, trimodal or multimodal). For example, consider the 5-bin histogram shown in Figure 5.5 which is divided into object and background. The intersection between the object and the background is 15 pixels $(2 + 5 + 3 + 4 + 1)$ or 30 %, meaning that the thresholding scheme cannot achieve a minimum overlap below 30 % or maximum accuracy above 70 %. However, the intersection measure does not reflect the required complexity of the subsequent thresholding scheme (bimodal, trimodal or multimodal); rather, intersection simply measures the maximum segmentation accuracy attainable with the ideal thresholding scheme. To address this drawback, the histogram intersection metric is complemented with Otsu's discriminant as the secondary metric, used to measure the separability attainable using a *single* threshold.

$$S_{Intersection} = \sum_{i=1}^{N} min(H_1(i), H_2(i)) \tag{5.14}$$

$$SA_{Max} = 1 - S_{Intersection} \tag{5.15}$$

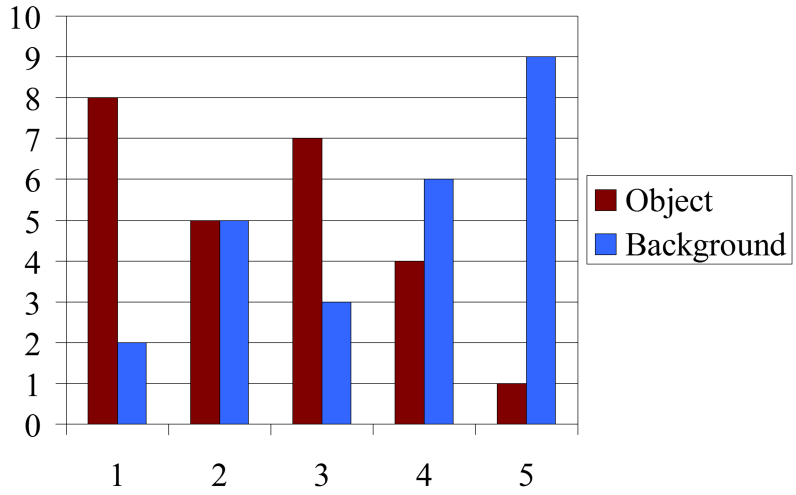


Figure 5.5: 5-bin histogram separated into object and background.

The choice of whether to focus on histogram intersection or Otsu's discriminant depends on the subsequent segmentation method. If the segmentation method involves thresholding the histogram using one threshold, then Otsu's discriminant will be of primary interest. If the segmentation method involves multiple thresholds, or a model-based approach, then histogram intersection will be of primary interest. The segmentation algorithm in Chapter 6 is designed around thresholding the histogram using one threshold, so Otsu's discriminant is considered in selecting the colour transform.

5.3.3 Method

Figure 5.6 shows an overview of the experimental method to compare the colour transforms, which consists of the following steps:

1. Use the manual markings to crop the mouth region from the full face images.
2. Apply a Gaussian low pass filter (LPF) to remove high frequency noise.
3. Interpolate the lip contours from the manual markings and segment the mouth region into lips, skin and oral cavity.
4. Apply the 33 colour transforms.
5. Scale values to range 0 to 1.
6. Divide values into 256 discrete bins.
7. Calculate metrics: Histogram Intersection and Otsu's Discriminant

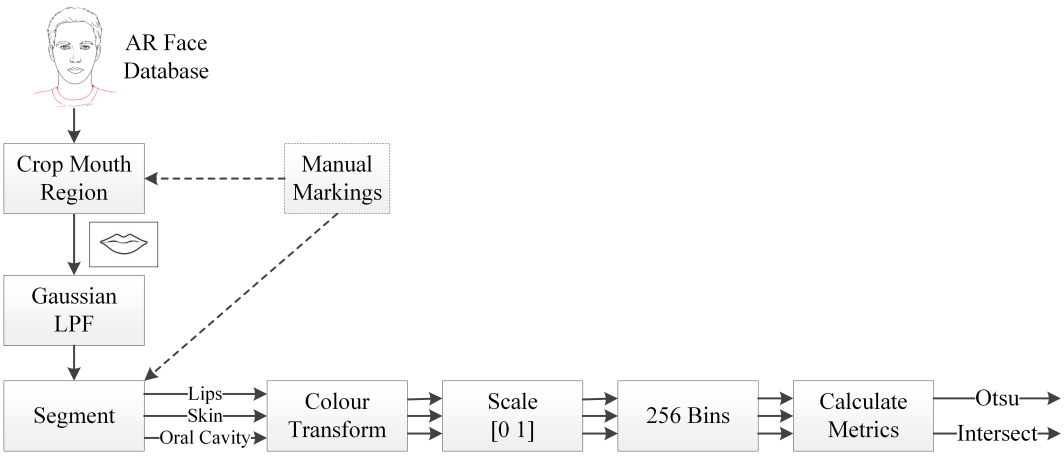


Figure 5.6: Block diagram of method to evaluate colour transforms.

5.4 Results and Analysis

The contrast between the lips and the skin is used to obtain the outer-lip contour; while the contrast between the lips and the oral cavity (teeth, tongue, shadow as in Figure 5.1) is used to obtain the inner-lip contour. In Section 5.4.1 histogram intersection and Otsu's discriminant are used to find the best transform to perform lip-skin segmentation; in Section 5.4.2 the metrics are used to determine the best transform for lip-oral cavity segmentation.

5.4.1 Lip-Skin Segmentation

Table 5.2 shows the results of histogram Intersection and Otsu's discriminant for lip-skin segmentation. The key to the transform names can be found in Table 5.1. The highlighted transforms (grey) indicate the transforms that have been directly used for lip segmentation in the literature.

The validity of the experimental procedure is immediately confirmed by noting the performance of the highlighted transforms: 11 of the 12 best performing transforms have been used for lip segmentation in the literature. Only the Cb-Cr difference proposed by Hsu et al. [116] falls outside of the top 12 transforms. The histogram intersection measures the overlap percentage between the lips and the skin – the top ranked transforms obtain an overlap as low as 6.15 %, while the bottom ranked transforms obtain an overlap as high as 18.61 %. This indicates that selection of an appropriate colour transform for lip-skin segmentation can immediately improve the accuracy of the system by up to three times. There is a high correlation between the histogram intersection and Otsu's discriminant in that transforms either perform well according to both metrics, or perform poorly according to both metrics.

According to the histogram intersection similarity metric, the hue channel from the HSV colour space (HSV-H) exhibits the greatest discriminating power, with an overlap of only 6.15 % (segmentation accuracy of 93.85 %). The impressive performance of HSV-H in lip-skin discrimination ratifies the extensive use of the hue channel for lip segmentation in the literature [105, 106, 109, 115]. Furthermore, HSV-H also performs well according to Otsu's discriminant (ranked 3rd), which indicates that not only does HSV-H have significant discriminating power, but also that the segmentation can be achieved with a single threshold. Figure 5.7 shows a histogram of HSV-H for the lips and skin respectively. It is clear that the lips and skin are well differentiated into two compact and distinct groups of pixels. The lips appear more red with a mean of 134.85, while skin is more yellow with a mean of 144.31. The variance of the lips and skin are 0.071 and 0.067 respectively. Figure 5.8 shows the HSV-H transform applied to the mouth region to enhance the contrast between the lips and the skin. HSV-H significantly enhances the lip-skin contrast; however, HSV-H is ineffective in discriminating between the lips and the oral cavity.

The hue domain filter (HDF) proposed by Coianiz et al. [105] is ranked 2nd with an overlap of 6.15 %; while, the pseudo-hue transforms proposed by Talea & Yaghmaie [63] (PHT) and Eveno et al. [114] (PHE) are ranked 4th and 5th respectively. It is interesting to note that 4 of the top 5 transforms are associated with hue. This

Table 5.2: Evaluation of colour transforms for lip-skin segmentation – highlighted transforms (grey) have been directly used for lip segmentation in the literature.

	Intersection		Otsu	
Transform	Rank	Value (%)	Rank	Value
HSV-H	1	6.15	3	0.4391
HDF	2	6.15	6	0.4029
MI3	3	7.37	2	0.4646
PHT	4	8.37	8	0.3700
PHE	5	8.37	4	0.4258
LUX-U	6	8.45	7	0.3896
YIQ-Q	7	8.84	1	0.4666
RE	8	9.88	5	0.4118
CC	9	12.15	9	0.3539
RGB-G	10	15.63	10	0.2085
LAB-A	11	16.50	19	0.1243
DHT-C3	12	16.78	20	0.1218
YCbCr-Cb	13	17.24	11	0.1959
GL	14	17.48	15	0.1465
YIQ-Y	15	17.48	16	0.1465
YCbCr-Y	16	17.48	17	0.1465
LAB-L	17	17.54	14	0.1468
XYZ-Y	18	17.54	18	0.1269
LUV-V	19	17.80	24	0.0939
XYZ-Z	20	17.80	25	0.0939
DHT-C1	21	17.97	21	0.1210
RGB-B	22	18.16	26	0.0896
XYZ-X	23	18.39	23	0.0987
LAB-B	24	18.40	12	0.1774
LUV-U	25	18.51	30	0.0659
LUV-L	26	18.55	13	0.1524
YCbCr-Cr	27	18.60	32	0.0050
HSV-S	28	18.61	33	0.0045
Cb-Cr	29	18.61	22	0.1037
DHT-C2	30	18.61	27	0.0873
HSV-V	31	18.61	28	0.0757
RGB-R	32	18.61	29	0.0757
YIQ-I	33	18.61	31	0.0079

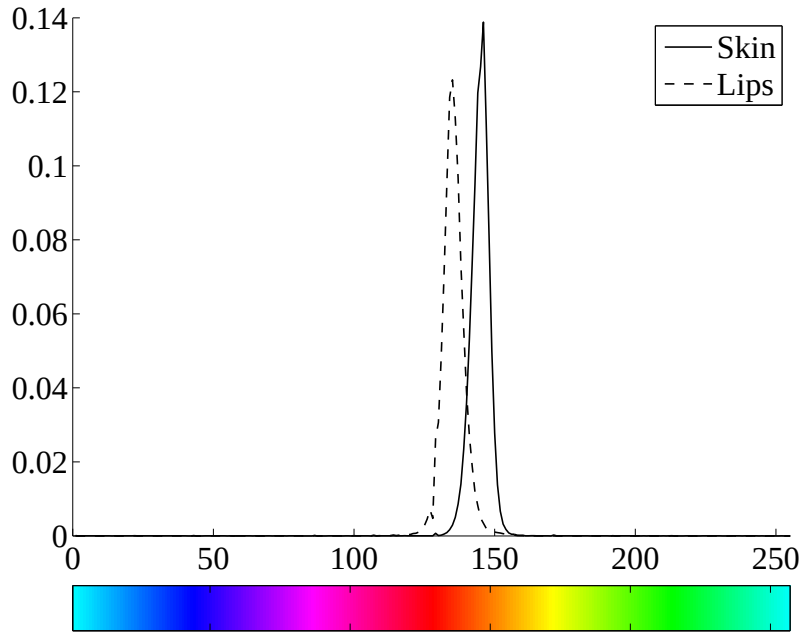


Figure 5.7: Histogram showing HSV-H component of lips and skin.

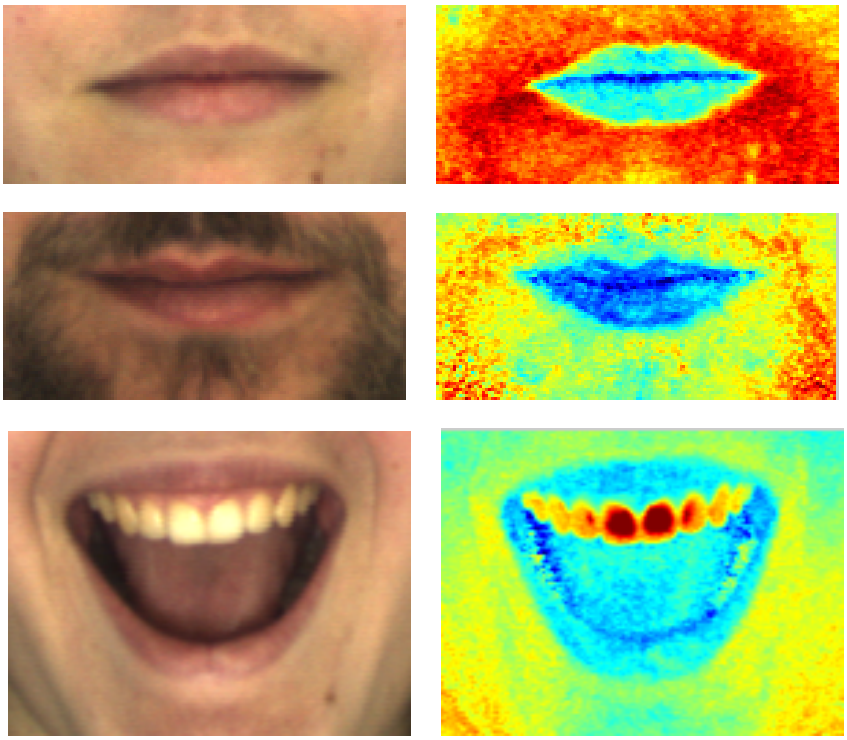


Figure 5.8: HSV-H transform used to enhance lip-skin contrast.

result concurs with Eveno et al. [114] and Stork & Hennecke [81] in their assertion that skin and lip hue remain relatively constant and well separated across a range of speakers, and thus can be used effectively to discriminate between lips and skin.

The modified I3 channel (MI3) proposed by Canzler & Dziurzyk [119] performs impressively according to both histogram intersection (3rd) and Otsu's discriminant (2nd); while the YIQ-Q transform was placed 1st according to Otsu's discriminant, but 7th according to histogram intersection.

The R, G, and B components of the RGB colour space are ranked 32, 10 and 22 respectively, with G providing the greatest discriminating power and R providing the least discriminating power. These results concur with the premise of the red exclusion (RE) technique which states that any contrast between the lips and the skin requires the exclusion of the R channel. The RE transform is ranked 8th, above the R, G and B channels. RGB-G performs impressively for a transform that has not been used for lip segmentation in the literature, and attained a ranking of 10th according to both histogram intersection and Otsu's discriminant. RGB-G performs better than LAB-A and DHT-C3, transforms which have been used for lip segmentation in the literature. Despite this impressive ranking, the overlap of RGB-G (15.63 %) is more than twice that of HSV-H (6.15 %).

The Cb-Cr difference proposed by Hsu et al. [116] is the only transform used for segmentation in the literature that performs particularly badly, obtaining a ranking of 29th with an overlap of 18.61 %. This result suggests that the observations and assumptions upon which Hsu et al. [116] build their transform should be reconsidered.

5.4.2 Lip-Oral Cavity Segmentation

Table 5.3 shows the results of histogram Intersection and Otsu's Discriminant for lip-oral cavity segmentation. The key to the transform names can be found in Table 5.1. The highlighted transforms (grey) indicate the transforms that have been directly used for lip segmentation in the literature; however, there is no literature on transforms for differentiating specifically between the lips and the oral cavity.

The intersection for lip-oral cavity segmentation ranges from 20.76 % to 32.44 %, which is considerably greater than the intersection for lip-skin segmentation, 6.15 % to 18.61 %. There are two main factors which contribute to this discrepancy. First, the outer-lip contour is labelled with 20 points, while the inner-lip contour is labelled with only 8 points. As a result, interpolation of the inner-lip contour (used for lip-oral

Table 5.3: Evaluation of colour transforms for lip-oral cavity segmentation – highlighted transforms (grey) have been directly used for lip segmentation in the literature.

	Intersection		Otsu	
Transform	Rank	Value (%)	Rank	Value
LAB-A	1	20.76	1	0.3711
HSV-S	2	20.95	3	0.3327
DHT-C3	3	21.33	2	0.3548
YCbCr-Cr	4	23.79	4	0.3119
LUX-U	5	23.86	5	0.2912
LUV-L	6	23.93	22	0.0177
PHT	7	24.06	6	0.2881
PHE	8	24.08	7	0.2876
LAB-B	9	25.25	14	0.0451
YCbCr-Cb	10	25.55	21	0.0186
YIQ-Q	11	26.05	8	0.2493
MI3	12	26.91	10	0.2188
YIQ-I	13	27.10	9	0.2412
Cb-Cr	14	27.22	13	0.0687
HSV-H	15	27.66	18	0.0295
DHT-C2	16	27.89	12	0.1046
CC	17	29.05	15	0.0356
LAB-L	18	29.65	29	0.0012
XYZ-Y	19	29.65	23	0.0152
XYZ-X	20	29.73	27	0.0025
YCbCr-Y	21	29.88	31	0.0001
GL	22	29.88	32	0.0001
YIQ-Y	23	29.88	33	0.0001
HSV-V	24	29.90	16	0.0306
RGB-R	25	29.90	17	0.0306
HDF	26	29.98	11	0.1202
RGB-G	27	30.17	25	0.0077
DHT-C1	28	30.24	30	0.0009
RE	29	31.03	28	0.0022
LUV-V	30	31.76	19	0.0232
XYZ-Z	31	31.76	20	0.0232
LUV-U	32	31.84	24	0.0081
RGB-B	33	32.44	26	0.0038

cavity segmentation) is less accurate than the interpolation of the outer-lip contour (used for lip-skin segmentation). Figure 5.9 illustrates the result of labelling the inner-lip contour with only 8 points – the true inner-lip contour is shown in green and the linear interpolation is shown in white. The discrepancies between the true contour and the approximate contour result in incorrect labelling of the pixels shown in red. Second, when the mouth is open, the tongue may be visible depending on the illumination. Visibility of the tongue decreases the accuracy of lip-oral cavity segmentation, as the tongue is generally of similar colour to the lips (see Figure 5.11). Nevertheless, the relative performance of the transforms can be used to determine the best transform for lip-oral cavity segmentation.

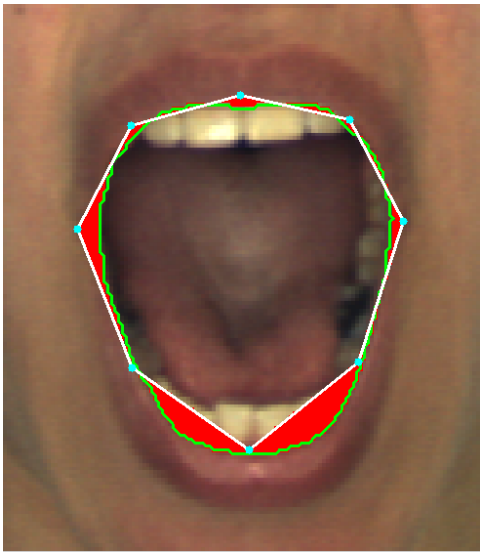


Figure 5.9: Interpolation of the inner-lip contour from 8 points. Inaccurate labelling of lip and oral cavity pixels is shown in red.

It is interesting to note that the colour transforms used directly for lip segmentation in the literature (highlighted in grey) do not perform particularly well in discriminating between the lips and the oral cavity. Of the top 10 transforms for lip-oral cavity segmentation, only 50 % have been used for lip segmentation in the literature, as opposed to 90 % for lip-skin segmentation. 7 of the 12 transforms used for lip segmentation in the literature fall outside of the top 10. Furthermore, not one of the top 5 transforms for lip-skin segmentation appear in the top 5 transforms for lip-oral cavity segmentation. These results indicate that it is incorrect to use the same colour transform for lip-skin segmentation and lip-oral cavity segmentation.

The Saturation channel from the HSV colour space (HSV-S) achieved a rank of 2nd according to the histogram intersection measure. This seems to agree with intuition – the oral cavity contains the teeth (white) and the oral cavity shadow (black) which

are characterised by a low saturation; while, the lips are primarily red which is characterised by a high saturation.

The analysis reveals a somewhat surprising result in that the transform exhibiting the greatest power in discriminating between the lips and the oral cavity is in fact the $a*$ channel from the CIELAB colour space (LAB-A). Low values of $a*$ are green, while high values of $a*$ are magenta. Figure 5.10 shows a histogram of LAB-A for the lips and oral cavity respectively. The lips appear more magenta with a mean of 153.77, while the oral cavity is more green with a mean of 111.79. The variance of the lips and oral cavity are 2.48 and 3.33 respectively. The discrete and non-linear nature of the CIELAB transform results in a significant number of pixels at a value of 128, corresponding to lip and oral cavity pixels which are black, white or varying shades of grey (depending on the value of L). Figure 5.11 shows the LAB-A transform applied to the mouth region to enhance the contrast between the lips and the oral cavity. LAB-A significantly improves the lip-oral cavity contrast; however, LAB-A does not effectively discriminate between the lips and the skin. In fact, LAB-A performs relatively poorly in lip-skin segmentation, obtaining a rank of 11th. This again emphasises the notion that different transforms should be used for lip-skin segmentation and lip-oral cavity segmentation.

If a single transform is to be chosen for both lip-skin segmentation and lip-oral cavity segmentation, then two possible candidates exist, which perform reasonably well in both cases: PHT which is ranked 4th in lip-skin and 7th in lip-oral cavity; and LUX-U which is ranked 5th in lip-skin and 6th in lip-oral cavity.

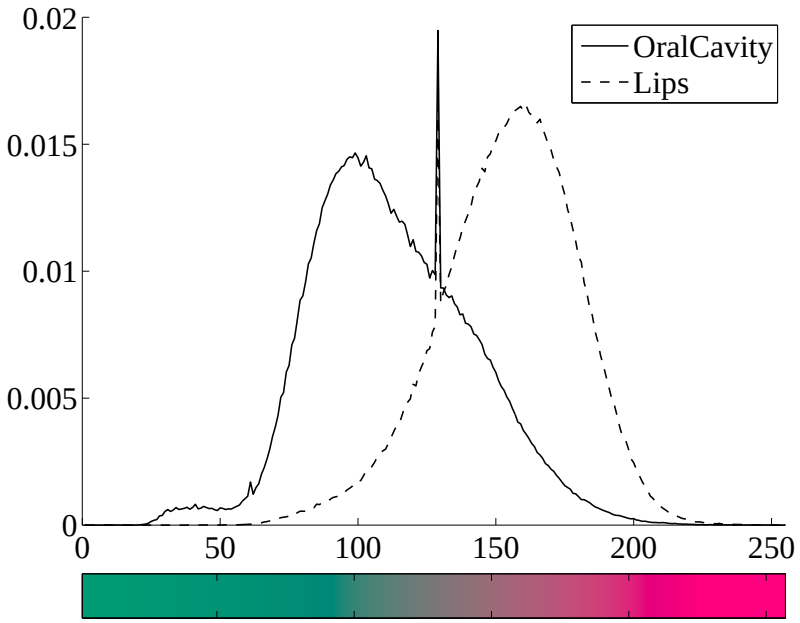


Figure 5.10: Histogram showing LAB-A component of lips and oral cavity.

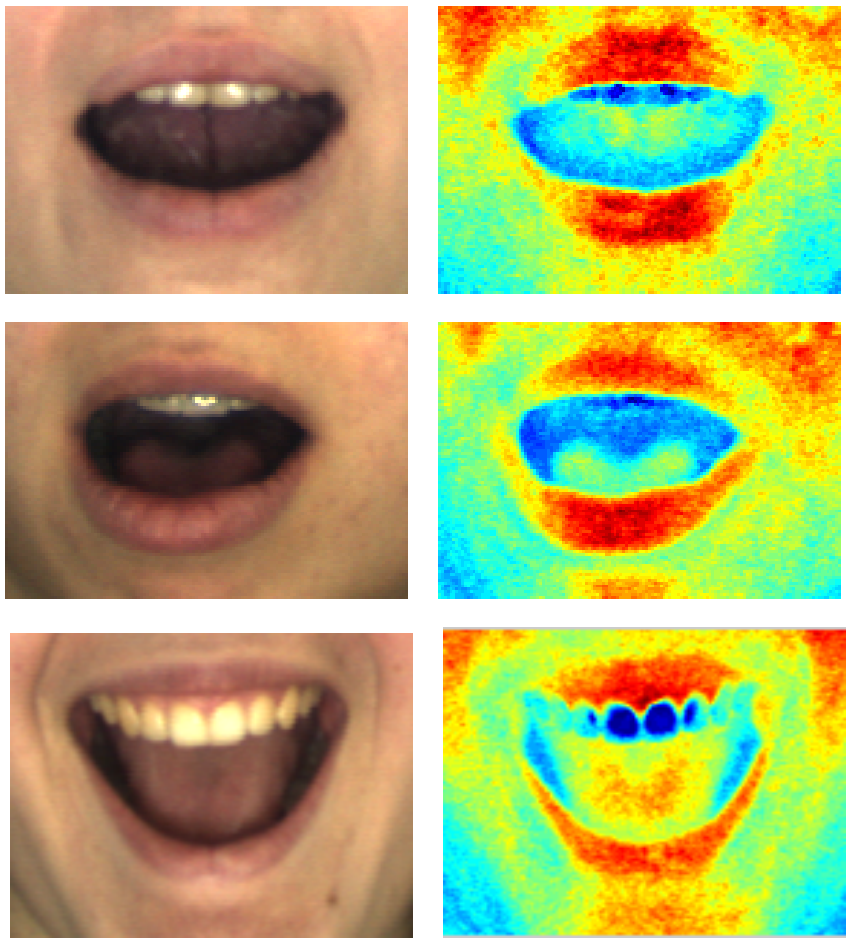


Figure 5.11: LAB-A transform used to enhance lip-oral cavity contrast.

5.5 Conclusion

The first stage in lip segmentation often involves applying a suitable colour transform to enhance the contrast between the lips and the surrounding skin. Selecting an appropriate transform can improve the segmentation accuracy by up to three times. However, no consensus exists among researchers as to the best colour transform for lip segmentation.

The comparison of colour transforms presented in this chapter is the most comprehensive study to date and evaluates 33 different transforms: 21 channels from 7 colour space models (RGB, HSV, YCbCr, YIQ, CIEXYZ, CIELUV, CIELAB); and 12 additional transforms (8 of which are designed specifically for lip segmentation). The contrast between the lips and the skin is used to obtain the outer-lip contour; while the contrast between the lips and the oral cavity is used to obtain the inner-lip contour. As such, this chapter identifies the transforms appropriate for lip-skin segmentation and for lip-oral cavity segmentation.

Histogram intersection quantifies the maximum segmentation accuracy attainable prior to morphological processing; thus, histogram intersection is well suited to compare and evaluate the colour transforms. However, histogram intersection does not reflect the required complexity of the subsequent thresholding scheme. To address this drawback, the histogram intersection metric is complimented with Otsu's discriminant as the secondary metric, used to measure the separability attainable using a *single* threshold.

Results for lip-skin segmentation validate the experimental approach, as 11 of the top 12 transforms are directly used for lip segmentation in the literature. The hue-based transforms (including pseudo-hue and hue domain filtering) occupy 4 of the top 5 positions. The hue channel from HSV emerges as the best transform to enhance lip-skin contrast with a segmentation accuracy of 93.85 %. The lip-oral cavity results reveal that the top 5 lip-oral cavity transforms are entirely different from the top 5 lip-skin transforms – this indicates conclusively that the same transform should not be used for both lip-skin segmentation and lip-oral cavity segmentation. The a^* component of CIELAB, and Saturation component of HSV are ranked 1st and 2nd respectively for lip-oral cavity segmentation. Finally, if a single transform must be chosen for both lip-skin segmentation *and* lip-oral cavity segmentation, then pseudo-hue and the LUX transform are the best candidates.

Chapter 6

Threshold-based Lip Segmentation Algorithm

6.1 Introduction

This chapter details the design and implementation of a new lip segmentation algorithm. Since technologies to locate the face and mouth region are well established, the starting point for the lip segmentation algorithm is the pre-cropped mouth region – a rectangular region containing the lips and surrounding skin. In recent years, the Viola-Jones detector has emerged as a popular technique for face detection and mouth region detection, which is capable of processing images extremely rapidly while achieving high detection rates.

The lip segmentation algorithm is primarily a colour-based technique which utilises the best colour transforms from Chapter 5 to enhance the contrast between the lips and skin. The major components of the lip segmentation algorithm include: preprocessing, colour transformation, thresholding, morphological processing, and contour smoothing.

This algorithm is referred to as the *base algorithm* in future chapters, as it forms the platform to develop and test adaptive threshold optimisation (ATO).

6.2 Existing Techniques

Since the 1970s, numerous image segmentation techniques based on different theories and methodologies have been proposed (see surveys [128–131]). However, few of these techniques have been applied successfully to the task of lip segmentation owing to the low chromatic and luminance contrast between the lips and skin [99]. As a result, a range of segmentation methods specific to the task of lip segmentation have been developed.

Lip segmentation techniques can be classified into two broad categories: colour-based techniques or model-based techniques. Elements from both categories can be combined to form a third category called 'hybrid techniques' [91]. This section discusses these three approaches to lip segmentation.

6.2.1 Colour-based Approach

Colour-based techniques operate at pixel or neighbourhood level, and attempt to differentiate between lip and skin pixels based on colour features. Lip segmentation algorithms usually start by transforming the RGB image to an intensity (or 'gray-scale') image by applying a suitable colour transform. Chapter 5 evaluates 33 colour transforms for lip segmentation.

Wark et al. [132] and Chiou & Hwang [133] segment the lips by applying upper and lower limits to threshold the R/G channel. In a similar manner, Coianiz et al. [105] and Zhang & Mersereau [109] apply fixed thresholds to the H channel. While these techniques are simple and efficient, the major limitation is the automatic computation of robust thresholds [103]. Fixed thresholds cannot be generalised due to variability in speaker appearance and lighting conditions, hence the threshold parameters must be calibrated for the specific speaker and environment. Furthermore, even after the initial calibration, appearance of the teeth, tongue, and oral cavity during movement of the mouth can significantly alter the image histogram and affect the threshold parameters.

The approach adopted by Xinjun & Hongqiao [134] is slightly more flexible whereby they design colour filters comprising both RGB and YUV components. A typical colour filter includes arithmetic, logical, and comparison operators, for example $(R^2 + BY)Y > (Y^2 + RG)G$. The filters effectively threshold the pixels based on a colour ratio, as opposed to a fixed value. Xinjun & Hongqiao [134] cascade four

different filters to increase robustness of lip pixel detection.

Another colour-based approach uses gradient filters (e.g. Sobel, Canny-Deriche or the Prewitt operator) to extract the lip contour due to their efficacy in boundary detection [93, 135–137]. However, the segmentation of gradient-based techniques is susceptible to false boundary edges caused by shadows, skin pigmentation, and facial hair.

In more recent approaches, clustering is used to perform colour-based segmentation. Beaumesnil & Luthon [138] use k-means clustering on the U channel from LUX to classify pixels as lips or face. Skodras & Fakotakis [139] build on this by using k-means colour clustering with automatically adapted number of clusters. Hara & Chellappa [140] determine the number of clusters using Bayesian information criterion to balance the model complexity and likelihood. In a similar progression, Rohani et al. [141] use fuzzy c-means (FCM) clustering with a preassigned number of clusters; Cheung et al. [99] build on this by initialising with a superfluous number of clusters, which are then reduced by merging clusters with coincident centroids.

In [142–144], statistical models are used to estimate the lip membership map. For example, Bouvier et al. [143] estimate the distribution of skin pixel using Gaussian mixture models, which is then used to compute the membership map of the lip pixels. However, such methods tend to miscalculate the membership map due to the similarity and overlap between lip and skin pixels in colour space [145].

Colour-based approaches are computationally inexpensive and allow rapid detection of the target region [91, 136]; however, Wang et al. [146] discourage approaches that rely solely on colour information citing the low contrast between lips and skin. In addition, colour-based techniques are highly sensitive to variations in illumination and orientation of the camera, which will alter all the pixel values [81]. Some authors express concern that the resulting segmentation is often rough or noisy [93].

6.2.2 Model-based Approach

The model-based approach uses prior knowledge of the lip shape to construct a lip model. The lip model is matched to the image by optimising a cost function. Model-based techniques are usually invariant to translation, rotation, scale and illumination; however, since the model is pre-defined, variation in speaker appearance and speaker sound formation can be challenging [71]. In addition, minimising a cost function can be computationally expensive which may affect real-time performance

[91]. The three main techniques used to build a lip model are: deformable templates, active shape models, and active contour models (snakes).

Deformable Templates

The deformable template approach to lip-tracking was first described by Yuille et al. in 1989, and uses mathematically defined outlines to capture the shape and position of an object [81]. The parametrised mathematical model is based on observations and assumptions about shape and position of the object. The template is iteratively matched to the shape of the mouth by minimising a cost function using an optimisation algorithm (e.g. downhill simplex method, particle swarm, or a genetic algorithm) [81]. The cost function is based on the sum of curve and surface integrals, as well as heuristics and higher-level information [81]. The optimised model from the previous frame is used as the initial model for the following frame [81]. Figure 6.1 shows an outer-lip template used to locate the outer-contour of the mouth. The top contour is defined by two parabolas which intersect above the centre of the mouth; the bottom contour is defined by a single parabola; the top and bottom contours intersect at the corners of the mouth. Other deformable template models have been constructed using quadratic B-Splines [148] or two parabolas for the upper lip and one for the lower lip [105].

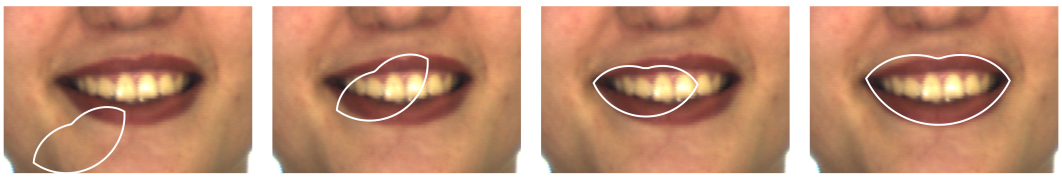


Figure 6.1: Deformable template approach – the lip template is defined by two parabolas forming the upper contour, and a single parabola forming the lower contour.

Active Shape Models

Active Shape Models (ASMs), also called Point Distribution Models (PDMs), are flexible models which represent an object by a set of labelled points [94]. The points are selected to describe the boundary or other significant parts of an object. ASMs typically involve an alignment stage whereby the lip model is aligned to a reference size (s) and rotational angle (θ), which helps reduce the effects of variation in camera zoom and speaker head position. Wang et al. [71] normalise the lip image with respect to the first image in a lip image sequence, which enables dynamic monitoring

of the scale (s) and rotational angle (θ). A statistical analysis of a labelled training dataset is performed to obtain the average shape and principal modes of variation [94]. The training data is often labelled by hand and a consistent labelling scheme must be applied to the training samples to ensure comparison of equivalent points on different samples. ASMs do not make assumptions about the shape of the object, but rather attempt to learn the legal shape deformation by examining training samples [94]. A small set of parameters is used to describe local and global deformations, which constrain the model to only deform to specific shapes. A cost function based on pixel intensity is used to measure the fit between the model and the image. An optimisation algorithm is then used to minimise the cost function and thus fit the ASM to the object (e.g. downhill simplex method, particle swarm, or a genetic algorithm), as shown in Figure 6.2.

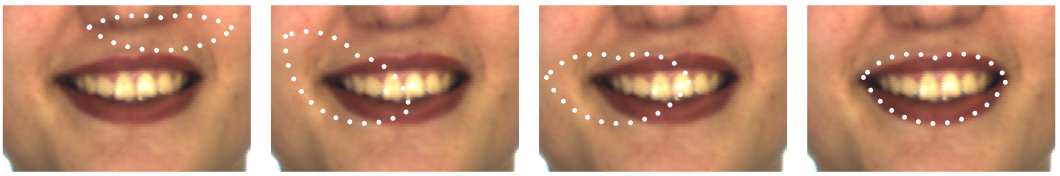


Figure 6.2: Active shape model using points to label the outer-lip contour.

Active Contour Models ("Snakes")

Active Contour Models (ACMs) (often referred to as "snakes") have been widely used for lip segmentation due to their ability to take smoothing and elasticity constraints into account [149]. A snake is an energy-minimising spline guided by internal and external energies [91, 149]. The purpose of the internal energy is to maintain the shape of the snake as regular and smooth. The simplest approach to internal energy is to assign high energy to elongated contours (elastic force) and to high curvature contours (rigid force) [91]. The purpose of the external energy is to model the edge of an object and is minimal when the snake is at the object boundary. The simplest approach to external energy uses regularised gradients as the external energy [91].

Figure 6.3 shows a snake fitting to the lower lip contour. Snakes have achieved reasonable results in lip segmentation; however, proper initialisation is crucial to successful lip tracking as snakes often converge to the incorrect result when the initial position is far from the lip edges [91, 93]. Another drawback of active contour models is that it is difficult to tune the parameters of the model. Eveno et al. [93] propose a "jumping snake" that addresses these limitations – jumping snakes can be initialised far from the lip edge and parameter adjustment is simple and intuitive [91].

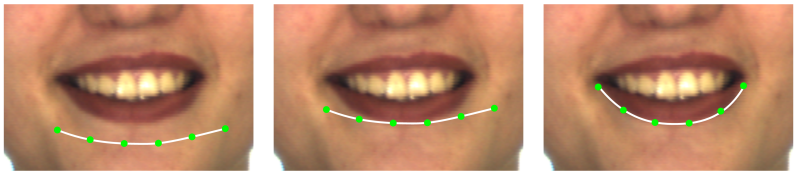


Figure 6.3: Active contour model (snakes) fitting to lower lip contour.

6.2.3 Hybrid Approach

Functional lip segmentation algorithms often combine elements from colour-based and model-based categories in so called 'hybrid techniques' [91]. Colour-based techniques are fast and not constrained by a rigid model; however, the resulting segmentation is often rough as no smoothness or shape constraint is applied. Model-based techniques are robust and accurate, but are computationally complex and limited by the flexibility of the underlying model.

In the hybrid approach, the computational complexity of model-based techniques is reduced by using colour-based techniques to obtain a quick and rough estimation of the candidate lip region. The sensitivity to illumination and rough segmentation of colour-based techniques is reduced by the smoothness and shape constraints of the model-based techniques. Finally, the cost function used by the model-based techniques is improved by enhancing the contrast between lip and non-lip pixels using a colour transformation.

Werda et al. [136] propose a hybrid technique for lip Point Of Interest (POI) localisation using colour information to locate the mouth in the first stage, and a geometric model to extract the lip contour in the second stage. Werda et al. [136] first apply a colour transform to reduce the effect of lighting, then analyse the horizontal and vertical projections to detect the corners of the mouth, and finally apply a parametrised geometric lip model.

Bouvier et al. [143] first apply a colour transform based on the red and green component values to enhance the contrast between the lips and the skin. The lip area is then estimated using expectation maximisation and a membership map. Finally, a snake is initialised based on the lip area estimation and is fitted to the upper and lower lip contours by multilevel gradient flow maximisation.

Mok et al. [150] propose a hybrid system to segment the lips by first transforming the RGB image to the CIELAB colour space, then applying a fuzzy clustering method incorporating a shape function to obtain a rough estimation. A 14-point active shape

model (ASM) is iteratively matched to the lips by deforming according to valid modes of deformation obtained from a training dataset using PCA.

Tian et al. [151] present a lip tracking method by combining colour, shape and motion. The colour information of the lips and skin is modelled as a Gaussian mixture. A multi-state deformable template model using parabolas is used to represent the different mouth states: open, relatively closed, and tightly closed. The lip motion is obtained by modified Lucas-Kanade tracking.

6.3 Lip Segmentation Algorithm

Figure 6.4 shows a high-level overview of the lip segmentation algorithm developed in this thesis, including: preprocessing, colour transformation, morphological processing, and contour smoothing. The formulation of these components is described in this section.

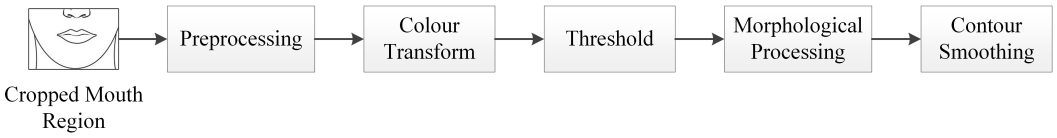


Figure 6.4: High-level overview of lip segmentation algorithm.

6.3.1 Preprocessing

The preprocessing component comprises two steps: first, a 3×3 Gaussian low pass filter (LPF) is applied to each channel R, G, and B to remove high frequency noise; second, a luminance correction is applied to each channel R, G, and B to compensate for varying illumination conditions. The luminance correction is derived from the Michaelis-Menten law which models the adaptation of human vision [114]. The correction law is shown in (6.1) and enhances chromatic information such that colours are brought out of shadowy areas. L is the luminance; a and b control the weight of luminance in the correction law ($a = 0.4, b = 0.8$).

$$X = \frac{X}{X + (b - a)L + a} \tag{6.1}$$

where

$$X = \{R, G, B\} \in [0, 1]$$
$$L = \text{Luminance} \in [0, 1]$$

6.3.2 Colour Transformation

Chapter 5 compares 33 colour transforms for lip segmentation using histogram intersection (Equation 5.14) and Otsu's discriminant (Equation 5.13). Histogram intersection simply measures the overlap between object and background histograms, but gives no indication of the complexity to perform the separation. Whereas, Otsu's discriminant measures the extent to which two histograms can be separated using a single threshold.

Histogram intersection is considered when the subsequent segmentation method involves multiple thresholds, or a model-based approach; Otsu's discriminant is considered when the segmentation method involves a single threshold.

The lip segmentation algorithm in this chapter is designed around a single histogram threshold, so Otsu's discriminant is considered in selecting the colour transform. Table 6.1 shows the top five colour transforms for lip-skin segmentation ranked by Otsu's discriminant. The top two colour transforms (YIQ-Q and MI3) are used in combination to improve contrast and reduce artefacts.

Table 6.1: Top 5 colour transforms for lip-skin segmentation according to Otsu's discriminant.

Rank	Transform	Description	Otsu
1	YIQ-Q	Q channel from YIQ	0.4666
2	MI3	Modified I3 channel (Canzler, 2002)	0.4646
3	HSV-H	Hue channel from HSV	0.4391
4	PHE	Pseudo-hue (Eveno, 2001)	0.4258
5	RE	Red exclusion (Lewis, 2002)	0.4118

YIQ-Q refers to the Q channel from the YIQ colour space. The YIQ colour space is a linear transformation of the RGB colour space, shown in (6.2). YIQ is designed to improve the transmission efficiency of the NTSC colour TV system by exploiting the colour response characteristics of the human eye [113]. Y is the luminance component, I is the orange-blue chrominance axis, and Q is the purple-green chrominance axis. Since YIQ decouples luminance and chrominance, the YIQ-Q transform provides an illumination-invariant approach (unless the illumination is so poor that the visibility is affected). On the Q axis, the lips appear more purple and the skin appears more green.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.595716 & -0.274453 & -0.321263 \\ 0.211456 & -0.522591 & 0.31135 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{6.2}$$

MI3 refers to the modified I3 transform proposed by Canzler & Dziurzyk [119]. Ohta et al. [123] generate the I3 feature by weighting the R, G and B colour components in an attempt to find an effective colour feature for region segmentation in a diverse group of images (e.g. building, seaside, home). Canzler & Dziurzyk [119] select the I3 feature for the following reasons: no singularities; good separation between brightness and colour information; simple and fast conversion from RGB. The I3 feature is modified specifically for lip segmentation by attenuating the B component, which is less prevalent in both the lips and skin.

The MI3 transform is shown in (6.3) and weights the R, G, and B components in accordance with their respective lip-to-skin separation. Table 6.2 is derived from the histogram in Figure 5.2, and shows the mean lip value, mean skin value and the difference for the R, G and B components. The G component exhibits the greatest separation (34.87), followed be the R component (28.64), while the B component exhibits the least separation (15.13). Canzler & Dziurzyk [119] construct the MI3 transform such that the G component is amplified, while the B component is attenuated.

$$I = \frac{1}{4}\left(2G - R - 0.5B\right) \tag{6.3}$$

Table 6.2: Separation of R, G and B for the lips and skin.

	Lip Mean	Skin Mean	Difference
R	132.69	161.33	28.64
G	75.95	110.82	34.87
B	64.41	79.54	15.13

Figure 6.5 shows the steps in combining the YIQ-Q and MI3 colour transforms which includes scaling, inversion, pixel multiplication, and Gaussian filtering. Several examples are shown in Figure 6.6.

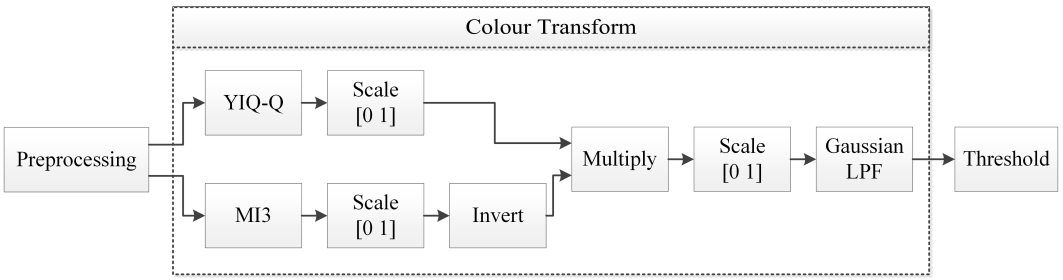
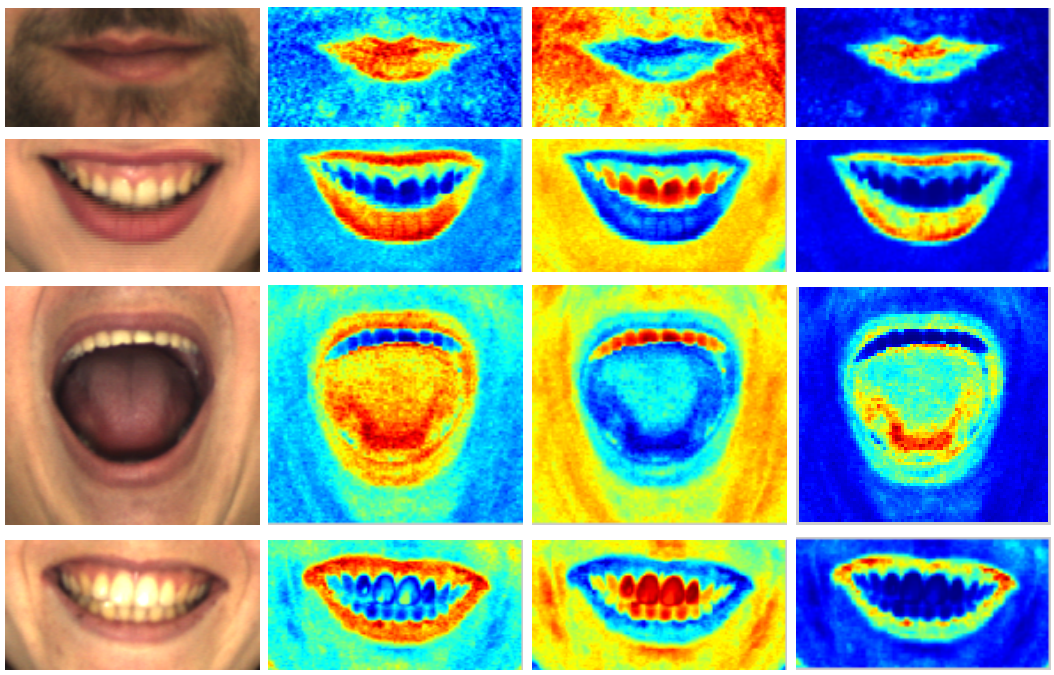


Figure 6.5: Combining the YIQ-Q and MI3 colour transforms.

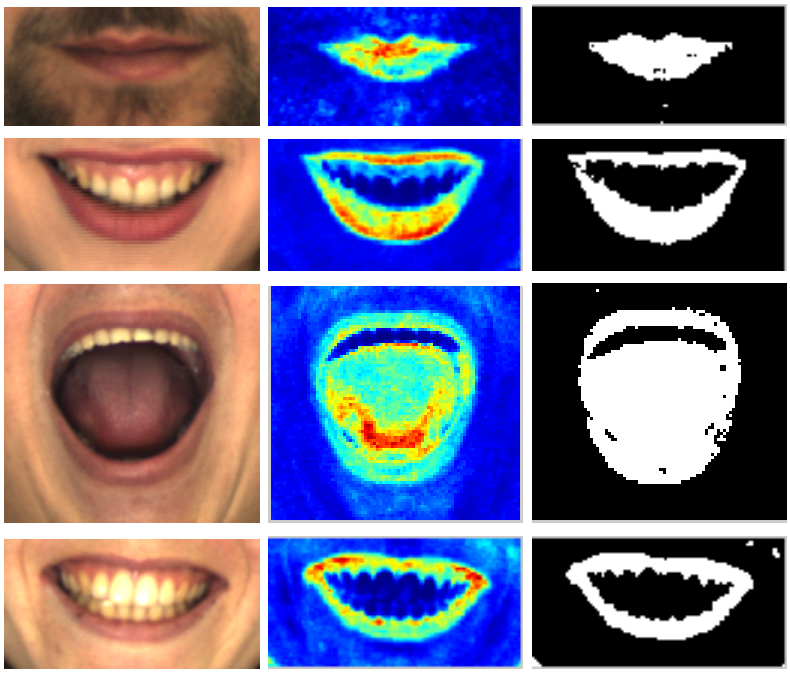


(a) Original **(b)** YIQ-Q **(c)** MI3 **(d)** Combined

Figure 6.6: Example images of the combined YIQ-MI3 colour transform.

6.3.3 Thresholding

The threshold is selected using Otsu's method [125], which parallels selection of the colour transform according to Otsu's discriminant. Otsu's method is a nonparametric and unsupervised method of automatic threshold selection, which uses the discriminant measure in Equation (5.13) to maximise the separability of the resultant classes. Figure 6.7 shows several examples of thresholding using Otsu's method.



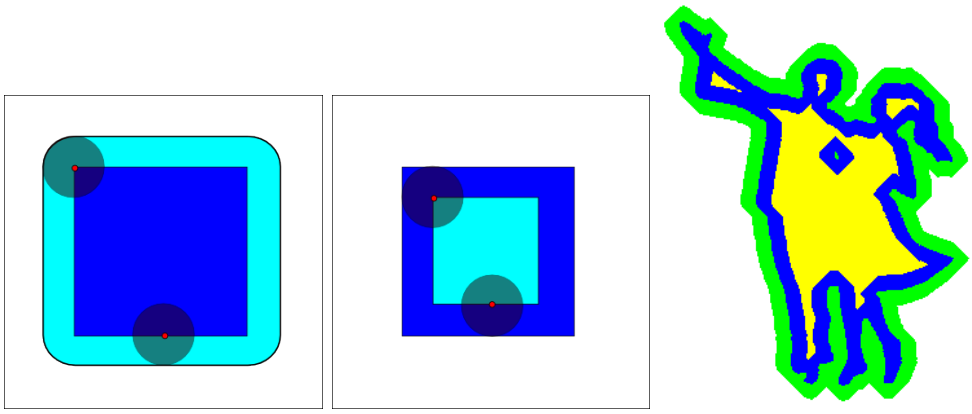
(a) Original **(b)** Colour transform **(c)** Thresholding

Figure 6.7: Example images of thresholding using Otsu's method.

6.3.4 Morphological processing

Mathematical morphology is a tool for extracting image components that are useful in the representation and description of region shape, such as boundaries [152]. The operations of *dilation* and *erosion* are fundamental to morphological image processing, and form the basis of many morphological algorithms. Dilation makes use of a structuring element to expand or "thicken" objects in a binary image. Erosion uses a structuring element to shrink or "thin" objects in a binary image. The concepts of dilation and erosion are illustrated in Figure 6.8.

After thresholding, nine morphological operations are performed to consolidate the lip region and to remove artefacts. Figure 6.9 shows the morphological processing procedure, and several examples are show in Figures 6.18 to 6.21.



(a) Dilation **(b)** Erosion **(c)** Dilation and erosion

Figure 6.8: (a) Dilation of the dark-blue square by a disk resulting in the light-blue square with rounded corners. (b) Erosion of the dark-blue square by a disk resulting in the light-blue square. (c) Original shape (blue), dilation (green) and erosion (yellow) by a diamond structuring element. Images in public domain: Keshet [153, 154, 155].

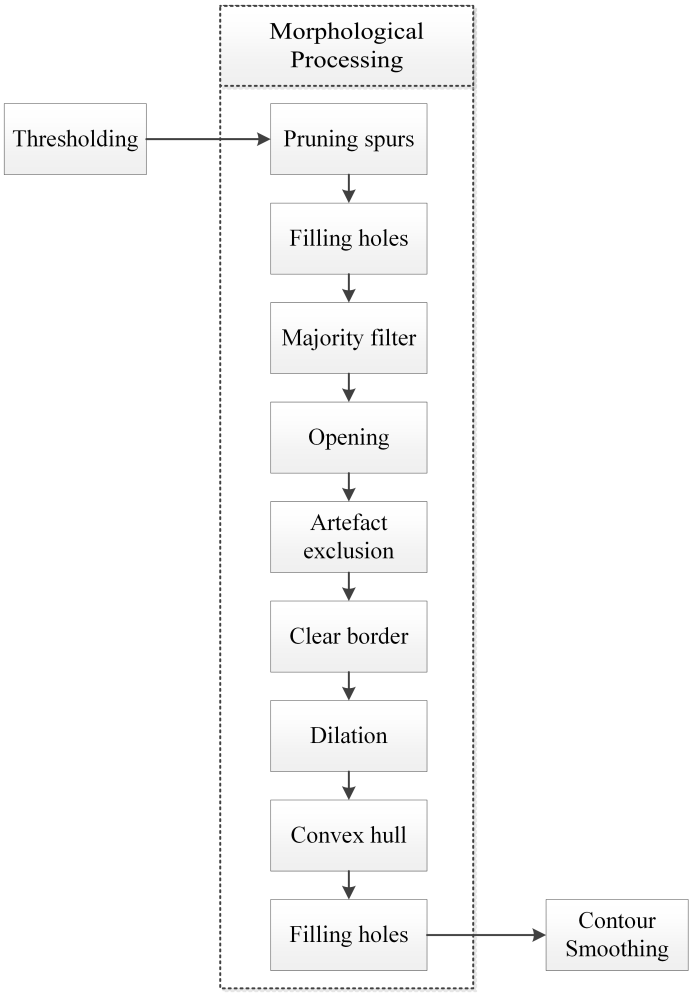


Figure 6.9: Morphological processing to consolidate lip region and remove artefacts.

1 Pruning spurs

Pruning is used to remove unwanted 'parasitic' components. The parasitic components are branches of a line which are not key to the overall shape of the line (see Figure 6.10).

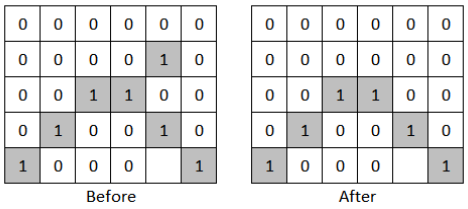


Figure 6.10: Example of pruning spurs.

2 Filling holes

The hole filling operation is a special case of the flood-fill which starts at each hole. A hole is a set of background pixels that cannot be reached by filling in the background from the edge of the image (see Figure 6.11).

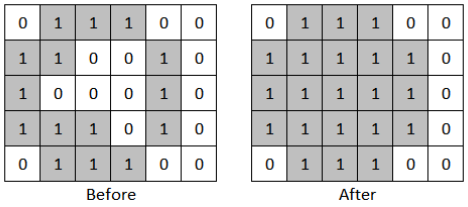


Figure 6.11: Example of hole fill.

3 Majority filter

The majority filter sets a pixel to 1 if the majority of pixels in the 3-by-3 neighbourhood are 1; otherwise, the pixel is set to 0 (see Figure 6.12). The majority filter is a shape smoother which removes small objects, holes, gaps, bays and peninsulas (both '1'-valued and '0'-valued features), but generally does not change the size of objects or background [156].

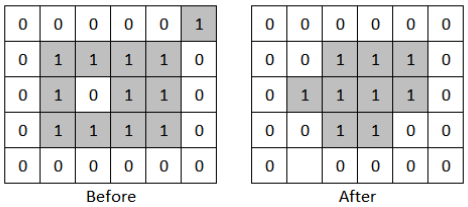


Figure 6.12: Example of majority filter.

4 Opening

Morphological opening is erosion followed by dilation, using the same structuring element. Morphological opening removes regions of an object that cannot contain the structuring element, smooths the contour of the object, breaks thin connections and removes thin protrusions [152]. The shape of the structuring

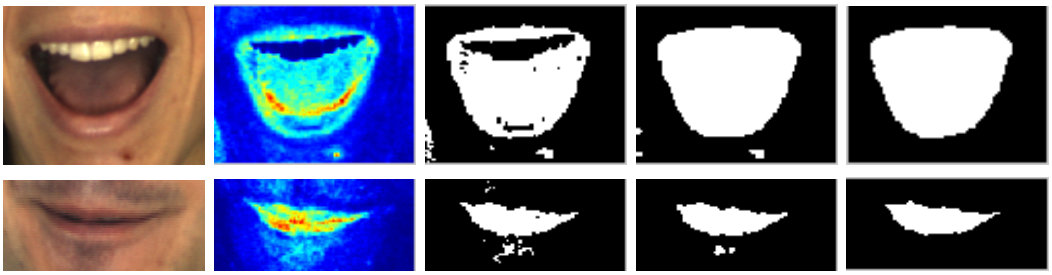
element in Figure 6.13 is similar to that of a closed mouth, which improves smoothing of the mouth region.

0	0	0	0	0	0	0	0
0	0	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	0	1	1	1	1	0	0
0	0	0	0	0	0	0	0

Figure 6.13: Structuring element for morphological opening.

5 Artefact exclusion

The preceding morphological operations (pruning spurs, filling holes, majority filter, opening) operate at the level of individual pixels or very small clusters, and serve to consolidate the regions in the image. However, these operations do not remove larger artefacts which often result from inconsistencies in skin colouration (local skin pigmentation, moles, freckles, facial hair, etc.) or illumination (reflection and shadows). These artefacts have two common characteristics: the regions are not connected to the main lip region, and the regions are significantly smaller in size than the main lip region. *Artefact exclusion* exploits these characteristics to remove the artefacts. The first step is to measure the size of all isolated regions in the image, and to determine the size of the largest connected component. Any region that is smaller than 10 % of the largest region is removed. Figure 6.14 shows the effect of artefact exclusion: in the upper row, a mole below the mouth creates an artefact which is excluded; in the lower row, a light covering of facial hair below the mouth creates an artefact which is excluded.

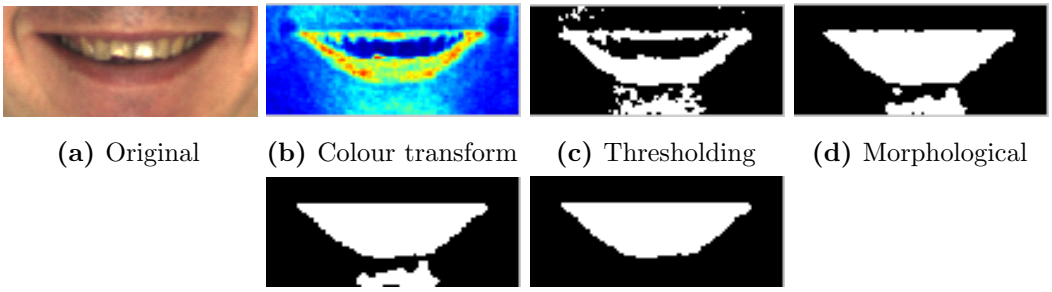


(a) Original **(b)** CT **(c)** Threshold **(d)** Morph **(e)** Artefact excl.

Figure 6.14: Example images of morphological artefact removal. In the upper row, a mole below the mouth creates an artefact which is removed; in the lower row, a light covering of facial hair below the mouth creates an artefact which is removed.

6 Clear border

Artefacts that are greater than 10 % of the largest region are not removed by *artefact exclusion*. However, these artefacts are often disconnected from the lip region and come into contact with the border of the rectangular ROI (Region Of Interest). Therefore, these artefacts can be removed by clearing the border of the image as shown in Figure 6.15.



(a) Original **(b)** Colour transform **(c)** Thresholding **(d)** Morphological

(e) Artefact exclusion **(f)** Clear border

Figure 6.15: Example images of morphological border clear.

7 Dilation

The structuring element shown in Figure 6.16 is used to expand or "thicken" the lip region. The structuring element is designed to enhance the shape of the lips by emphasising the horizontal width of the mouth and the curves of the upper lip. Dilation is used to bridge or repair gaps.

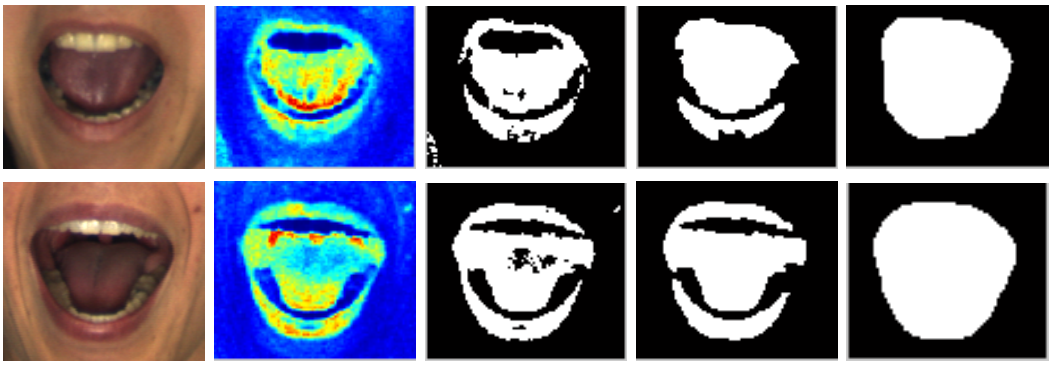
0	0	0	0	0
0	0	1	0	0
0	1	1	1	0
0	0	0	0	0

Figure 6.16: Structuring element for dilation.

8 Convex hull

In order to obtain the contour of the lip region, the image must contain only *one* cohesive region. If more than one region remains after artefact exclusion and the clear border operation, then the remaining regions must be joined before the contour can be extracted. The convex hull of a set of points in two dimensions is the smallest convex region enclosing all points in the set. Consider a set of nails knocked into a wooden board, where each nail represents the coordinates of a specific point. The convex hull of the set would be formed by placing an elastic band around the set of nails [157]. Figure 6.17 shows the use of the convex hull in joining isolated regions before the lip contour is

extracted. The colour transform and thresholding operations often segment the teeth and skin in the same category; therefore, the convex hull can be very useful in restoring the outer contour of the mouth by grouping together the lips, teeth and oral cavity.



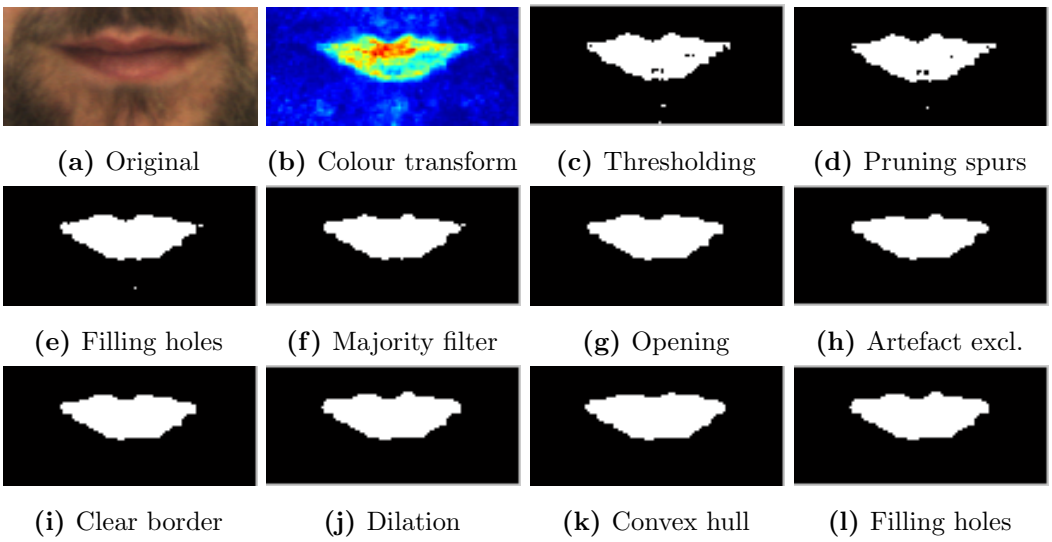
(a) Original **(b)** CT **(c)** Threshold **(d)** Morph **(e)** Convex hull

Figure 6.17: Example images of convex hull operation.

9 Filling holes

The preceding morphological operations attempt to consolidate the lip region and remove artefacts. Hole filling is first performed in stage two of morphological processing, however morphological opening and dilation may join existing regions thereby creating new holes. The second hole filling operation is delayed until the artefacts have been removed.

Figure 6.18 to Figure 6.21 show examples of the full morphological processing procedure.

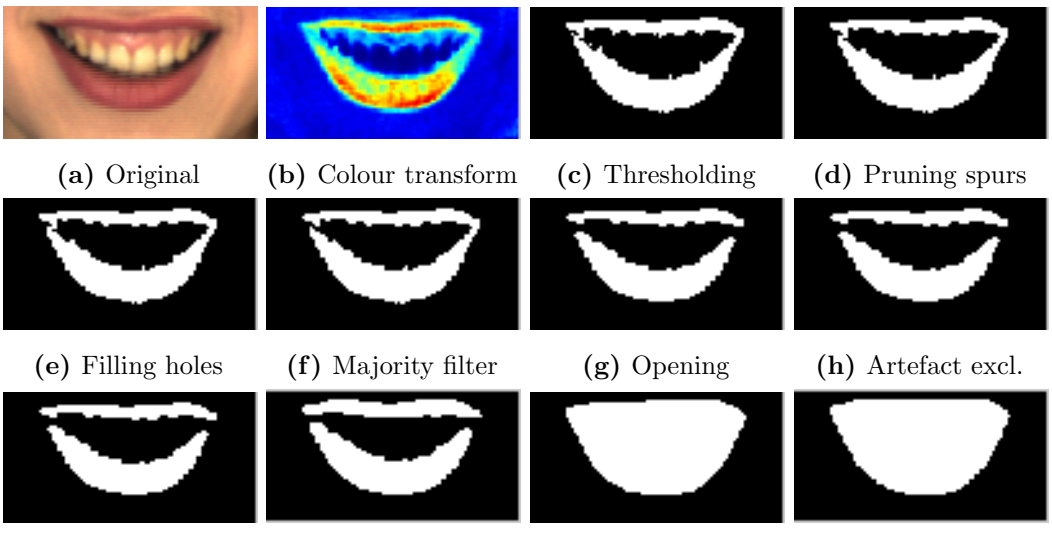


(a) Original **(b)** Colour transform **(c)** Thresholding **(d)** Pruning spurs

(e) Filling holes **(f)** Majority filter **(g)** Opening **(h)** Artefact excl.

(i) Clear border **(j)** Dilation **(k)** Convex hull **(l)** Filling holes

Figure 6.18: Example 1 – all stages in morphological processing.

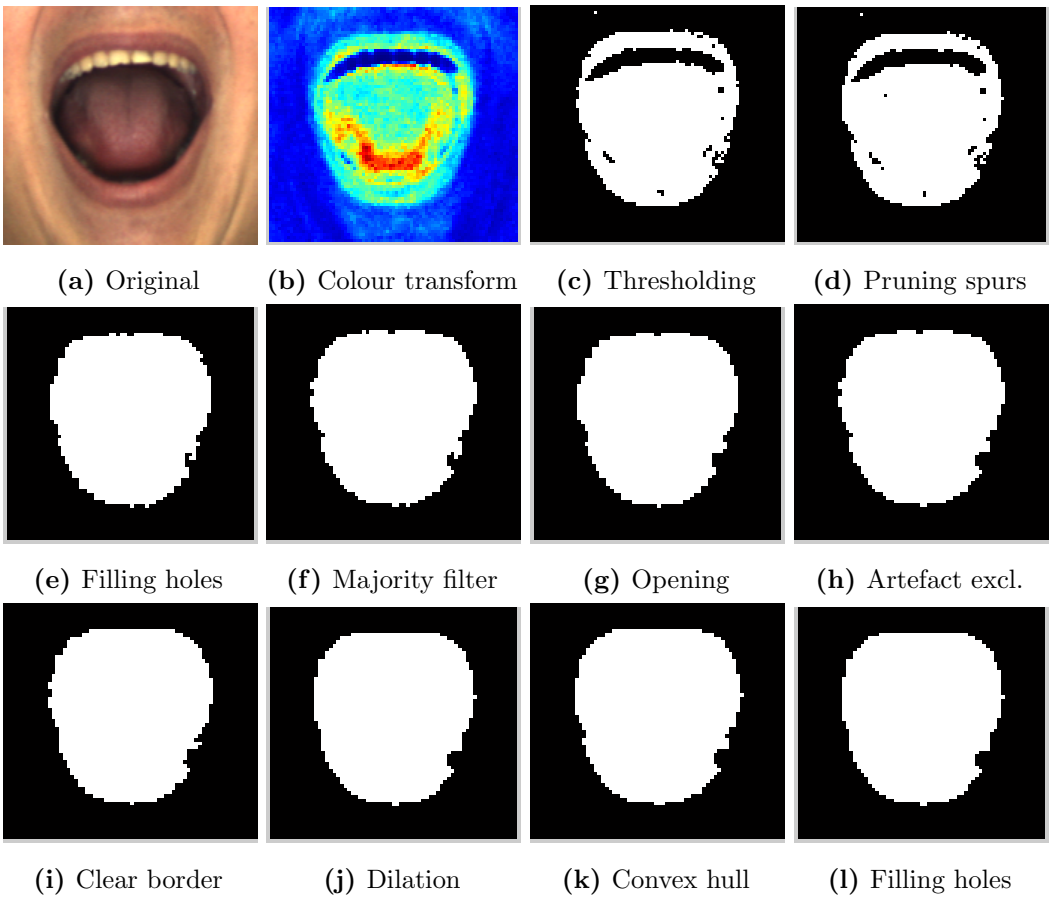


(a) Original **(b)** Colour transform **(c)** Thresholding **(d)** Pruning spurs

(e) Filling holes **(f)** Majority filter **(g)** Opening **(h)** Artefact excl.

(i) Clear border **(j)** Dilation **(k)** Convex hull **(l)** Filling holes

Figure 6.19: Example 2 – all stages in morphological processing.

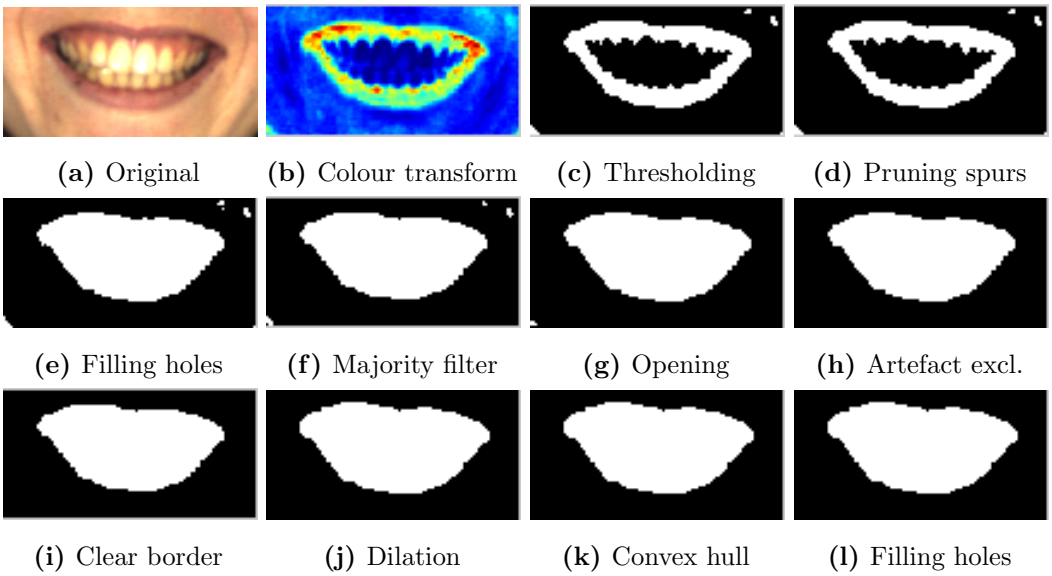


(a) Original **(b)** Colour transform **(c)** Thresholding **(d)** Pruning spurs

(e) Filling holes **(f)** Majority filter **(g)** Opening **(h)** Artefact excl.

(i) Clear border **(j)** Dilation **(k)** Convex hull **(l)** Filling holes

Figure 6.20: Example 3 – all stages in morphological processing.



(a) Original **(b)** Colour transform **(c)** Thresholding **(d)** Pruning spurs

(e) Filling holes **(f)** Majority filter **(g)** Opening **(h)** Artefact excl.

(i) Clear border **(j)** Dilation **(k)** Convex hull **(l)** Filling holes

Figure 6.21: Example 4 – all stages in morphological processing.

6.3.5 Contour Smoothing

After the morphological processing, the lip contour is extracted from the outer boundary of the lip region. However, the morphological processing does not incorporate any shape or smoothness constraints, so the resulting segmentation is often rough. Thus, the lip segmentation algorithm uses cubic splines to smooth the lip contour.

A spline function is a curve constructed from polynomial segments that are subject to continuity and smoothness conditions at the joints [158]. Given a set of co-ordinates $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, the object of spline interpolation is to bridge the gap between adjacent points $(x_i, y_i), (x_{i+1}, y_{i+1})$ using a smooth polynomial function that is piecewise defined:

$$S(x) = \begin{cases} S_1(x) & \text{if } x_1 \leq x < x_2 \\ S_2(x) & \text{if } x_2 \leq x < x_3 \\ & \vdots \\ S_{n-1}(x) & \text{if } x_{n-1} \leq x < x_n \end{cases} \tag{6.4}$$

In the case of cubic spline interpolation, $S_i(x)$ is a third degree polynomial defined by

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \tag{6.5}$$

for $i = 1, 2, ..., n - 1$

To ensure that the resulting curve is continuous and smooth, three conditions are imposed at the joints or "knots" between adjacent segments:

1. The adjacent functions S_{i-1} and S_i for $i = 2, ..., n - 1$ should meet at the point (x_i, y_i):
$$S_{i-1}(x_i) = S_i(x_i) = y_i \tag{6.6}$$
2. The first derivatives of adjacent functions should be equal:
$$S'_{i-1}(x_i) = S'_i(x_i) \tag{6.7}$$
3. The second derivatives of adjacent functions should be equal:
$$S''_{i-1}(x_i) = S''_i(x_i) \tag{6.8}$$

The interpolating spline is useful in approximating a smooth function that passes through each data point; however, if the data includes random fluctuations, then it is necessary to allow the spline to depart from the data points to approximate the underlying function. The smoothing spline S minimises the spline objective function:

$$L = \lambda \sum_{i=1}^{n} (y_i - S_i)^2 + (1 - \lambda) \int_{x_1}^{x_n} \left(S''(x)\right)^2 dx \tag{6.9}$$

where

$$S_i = S(x_i)$$
$$\lambda = \text{smoothing parameter, } \lambda \in [0, 1]$$

The first term is simply the mean squared error of using the curve $S(x)$ to predict y. This quantity measures how closely $S(x)$ adheres to the data points.

In the second term, $S''(x)$ is the second derivative of S with respect to x, which measures the curvature of S at x. The sign of $S''(x)$ indicates whether the curvature is concave (negative) or convex (positive). If $S''(x)$ is zero, then the curve is linear at x. The quantity of interest is the total curvature of $S(x)$, therefore the second derivative is squared and integrated over the length of the curve.

The parameter λ reflects the relative importance given to the conflicting objectives of remaining close to the data on the one hand, and of obtaining a smooth curve on the other hand. At the one extreme, if $\lambda = 0$ and only smoothness matters, then $S(x)$ will become a straight line. At the other extreme, if $\lambda = 1$ and closeness to the data is the only concern, then $S(x)$ will be an interpolating spine which passes exactly through the data points. Increasing λ will increase the adherence to the data points, while decreasing λ will increase the smoothness of the curve.

The lip segmentation algorithm uses two cubic smoothing splines to smooth the top lip contour and the bottom lip contour respectively. To ensure that the top and bottom smoothing splines meet at the corners of the mouth, it is necessary to weight the error measure to ensure that the curve passes through these points. The spline objective function is modified by specifying the weight w of each error measure, as shown in (6.10). The first and last points correspond to the corners of the mouth, and thus are assigned a weight 100 times greater than the intervening points.

$$L = \lambda \sum_{i=1}^{n} w_i (y_i - S_i)^2 + (1 - \lambda) \int_{x_1}^{x_n} \left(S''(x)\right)^2 dx \tag{6.10}$$

where

$$w = [100, 1, 1, ..., 1, 100]$$

Figure 6.22 shows smoothing of the top and bottom lip contours using cubic splines. For the top lip contour $\lambda_{top} = 0.003$, while for the bottom lip contour $\lambda_{bottom} = 0.006$.

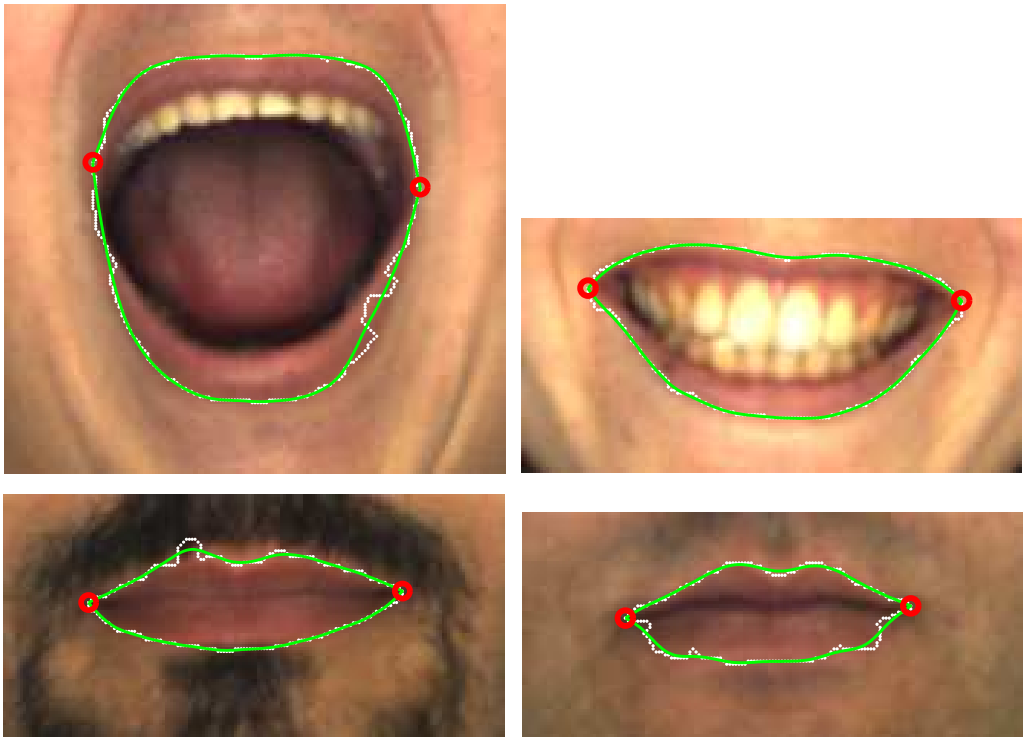


Figure 6.22: Example images of contour smoothing using cubic splines. The top lip contour is smoothed using a cubic spline with $\lambda_{top} = 0.003$, and the bottom lip contour is smoothed using a cubic spline with $\lambda_{bottom} = 0.006$. The splines for the top and bottom lip contours meet at the corners of the mouth.

6.4 Conclusion

The lip segmentation algorithm begins by filtering the pre-cropped mouth region, and applying luminance correction based on the Michaelis-Menten law. The two best transforms from the comparison in Chapter 5 (YIQ-Q and MI3) are combined to enhance the contrast between the lips and the skin. Otsu's method is used to select the threshold, followed by morphological processing to consolidate the lip region and to remove artefacts. Finally, the lip segmentation algorithm uses cubic splines to smooth the lip contour.

Chapter 7

Testing and Analysis of Threshold-based Segmentation Algorithm

7.1 Introduction

This chapter presents the results and analysis of the lip segmentation algorithm detailed in Chapter 6. The algorithm is tested on 895 mouth region images from the AR Face Database, using percentage overlap (OL) and segmentation error (SE) to quantify performance. The discussion presents examples of both successful and unsuccessful segmentation results, which leads to an understanding of the strengths and the weaknesses of the algorithm.

7.2 Dataset

The dataset comprises 895 pre-cropped mouth regions from the AR Face Database [100]. The manual markings for the lips are interpolated using two cubic smoothing splines (see Section 6.3.5 for details). The interpolated lip contour is used to generate the binary ground truth as shown in Figure 7.1.

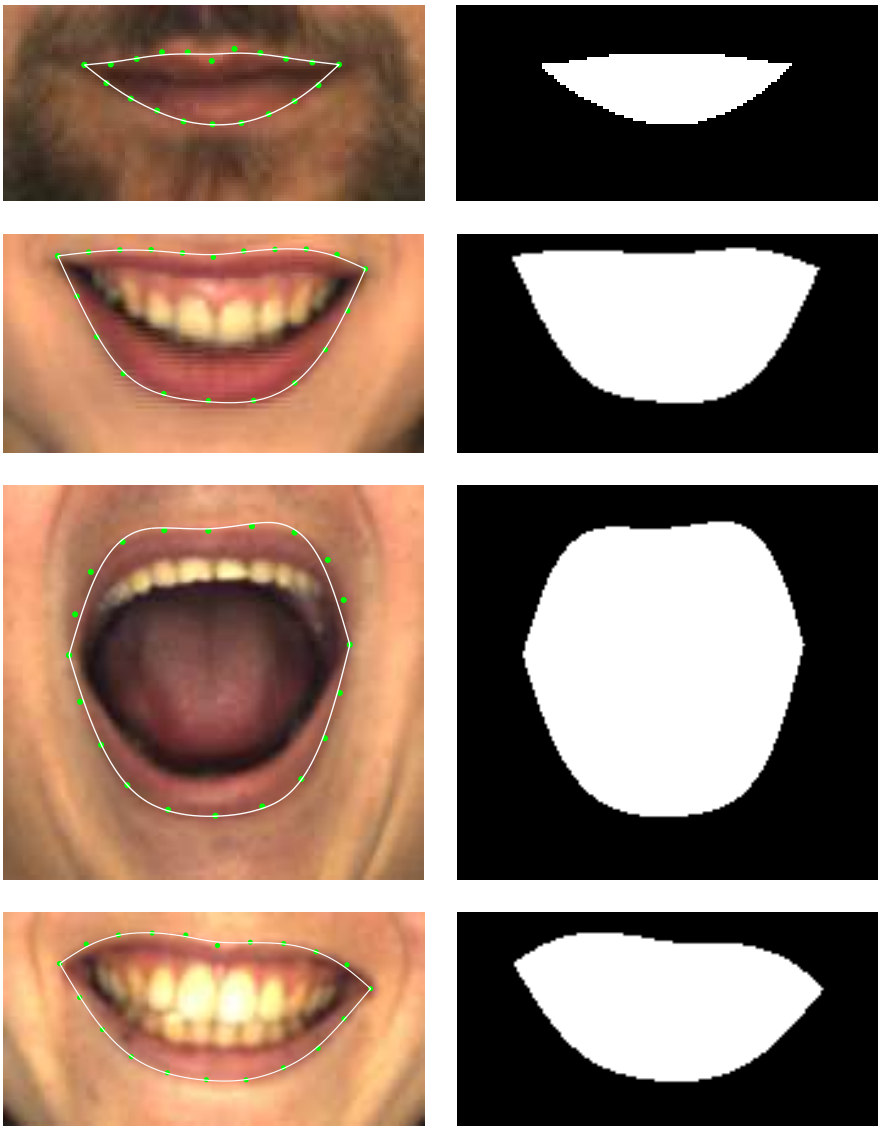


Figure 7.1: The full-face images from the AR Face Database are pre-cropped to the rectangular mouth region. The manual markings from Ding & Martinez [101] are interpolated using cubic smoothing splines to generate the ground truth.

7.3 Metrics

Wang et al. [159] defined two types of error to quantify the quality of lip segmentation: *Outer Lip Error (OLE)* is the number of non-lip pixels classified as lip pixels; and *Inner Lip Error (ILE)* is the number of lip-pixels classified as non-lip pixels. Liew et al. [160] used these error measures to develop two metrics to quantify the overall lip segmentation accuracy: *Percentage Overlap (OL)* and *Segmentation Error (SE)*. OL measures the percentage overlap between the segmented lip region (A) and the ground truth (A_G), as shown in (7.1). Total agreement between the ground truth (A_G) and the segmented region (A) has an overlap of 100 %. SE measures the segmentation error, and is 0 % for total agreement, as shown in (7.2). Equation (7.3) and (7.4) show OL and SE in terms of the binary classification metrics (TP, TN, FP, FN). OL and SE have been widely adopted to quantify lip segmentation accuracy: Saeed & Dugelay [91], Cheung et al. [99], Guan [121], Cheung et al. [145], Liew et al. [160], Pan et al. [161], Cheung & Li [162], Chin et al. [163], Saeed [164], Guan [165].

$$OL = \frac{2 \times (A \cap A_G)}{A + A_G} \times 100\,\% \tag{7.1}$$

$$SE = \frac{OLE + ILE}{2 \times TL} \times 100\,\% \tag{7.2}$$

where

A = segmented lip region
A_G = ground truth
OLE = outer lip error – the number of non-lip pixels classified as lip pixels
ILE = inner lip error – the number of lip pixels classified as non-lip pixels
TL = number of lip pixels in the ground truth

$$OL = \frac{2 \times TP}{TP + TP + FP + FN} \times 100\,\% \tag{7.3}$$

$$SE = \frac{FP + FN}{2 \times (TP + FN)} \times 100\,\% \tag{7.4}$$

where

TP = true positive – number of lip pixels classified as lip pixels
TN = true negative – number of skin pixels classified as skin pixels
FP = false positive – number of skin pixels classified as lip pixels
FN = false negative – number of lip pixels classified as skin pixels

7.4 Results

The lip segmentation algorithm is tested on the dataset of 895 mouth region images, and the performance is analysed in terms of percentage overlap (OL) and segmentation error (SE).

Figure 7.2 shows a histogram of percentage overlap (OL) for all images in the dataset. The values for OL range from a minimum of 50.46 % to a maximum of 98.05 %. The percentage of images above 90 % OL is 78.2 %, and only 4.2 % of images fall below 80 % OL. The peak of the histogram is at 96-97 % with 132 images. The mean OL for all 895 mouth region images is 92.23 %.

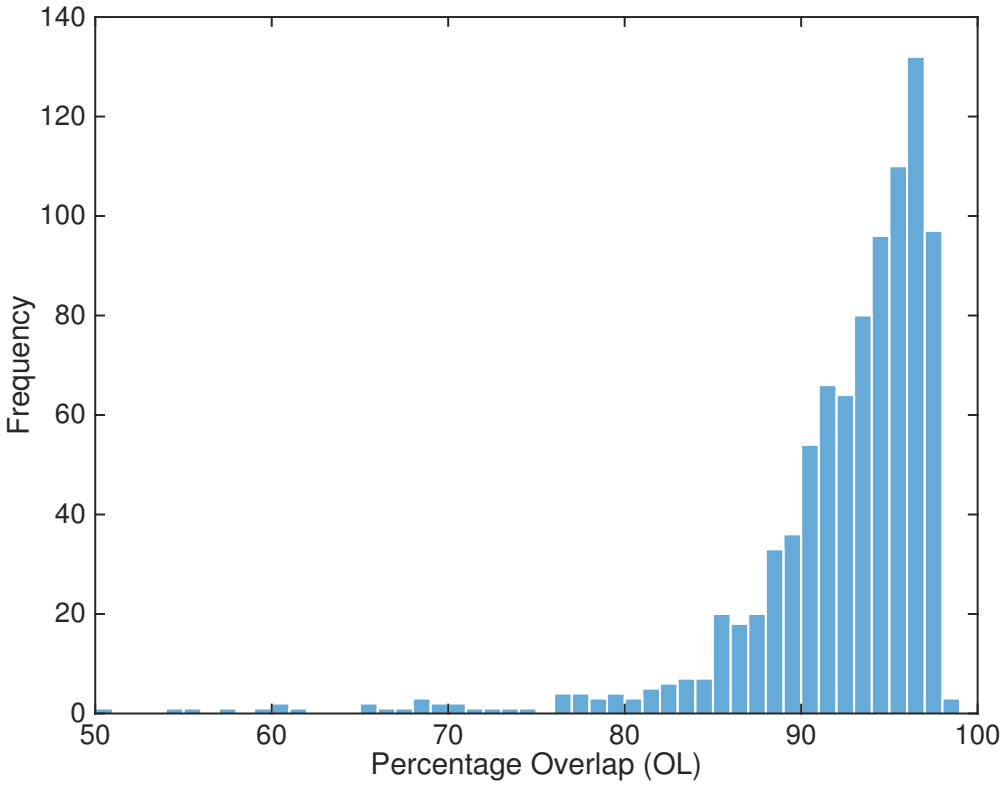


Figure 7.2: Histogram showing percentage overlap (OL) between segmentation result and ground truth.

Figure 7.3 shows a histogram of the segmentation error (SE) for all images in the dataset. The values for SE range from a minimum of 1.93 % to a maximum of 69.26 %. The percentage of images below 10 % SE is 81.5 %, and the mean SE for all 895 images is 7.39 %.

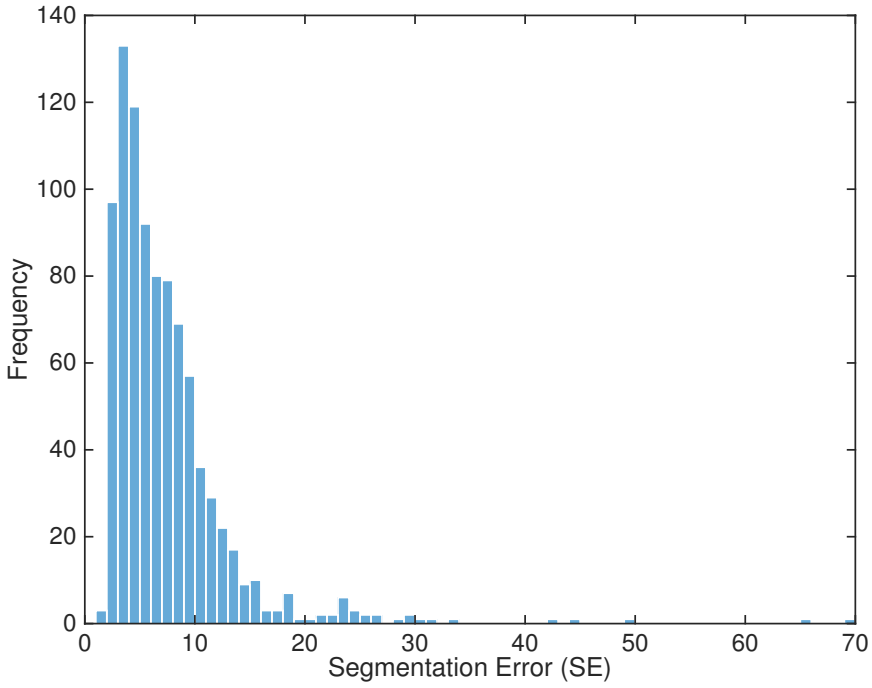


Figure 7.3: Histogram showing segmentation error (SE) between segmentation result and ground truth.

7.4.1 Discussion

It is difficult to evaluate the success of the lip segmentation algorithm from the histograms alone. For example, the mean OL is 92.23 % – is this high enough? Or, the mean SE is 7.39 % – is this acceptable? In order to understand the results and evaluate the algorithm, it is necessary to examine sample images to determine what values of OL and SE correspond to acceptable segmentation.

Table 7.1 shows a summary of the OL histogram of Figure 7.2, in which OL has been grouped into discrete bins. Similarly, Table 7.2 shows a summary of the SE histogram of Figure 7.3.

Of the 895 images in the dataset, the segmentation algorithm obtained an OL of 95-100 % for 342 images. Figure 7.4 shows several examples of images in the 95-100 % bin. In these images, the segmentation adheres almost exactly to the contour of the lips. The examples in this bin include male and female, different skin tones, beards, and even orthodontic braces. SE for this bin ranges from 1.93-5.13 % with an average of 3.51 %.

The lip segmentation algorithm achieved an OL of 90-95 % for 360 images. Figure 7.5

Table 7.1: OL in discrete bins.

OL Range	Num	%	Mean OL (%)
<70	17	1.90	62.93
70-80	21	2.35	76.21
80-90	155	17.32	87.00
90-95	360	40.22	92.77
95-100	342	38.21	96.47
Total	895	100	92.23

Table 7.2: SE in discrete bins.

SE Range	N	%	Mean SE (%)
0-5	352	39.33	3.56
5-10	377	42.12	7.24
10-20	137	15.31	12.77
20-30	21	2.35	24.44
>30	8	0.89	45.83
Total	895	100	7.39

shows several examples of images in the 90-95 % bin. The segmentation appears to occasionally divert slightly from the contour of the lips, but still remains accurate to the human eye. The examples again include variation in gender, skin tone, and facial hair. SE for this bin ranges from 4.82-10.19 % with an average of 7.00 %.

Images with OL above 90 % seem acceptable for lip segmentation applications, which amounts to 78 % of the 895 images.

Images in the 80-90 % OL bin (Figure 7.6) show more significant deviation from the lip contour. The algorithm identifies about three quarters of the lip contour, but deviates substantially in the remaining quarter.

Figure 7.4: Example images of segmentation with OL above 95 %.

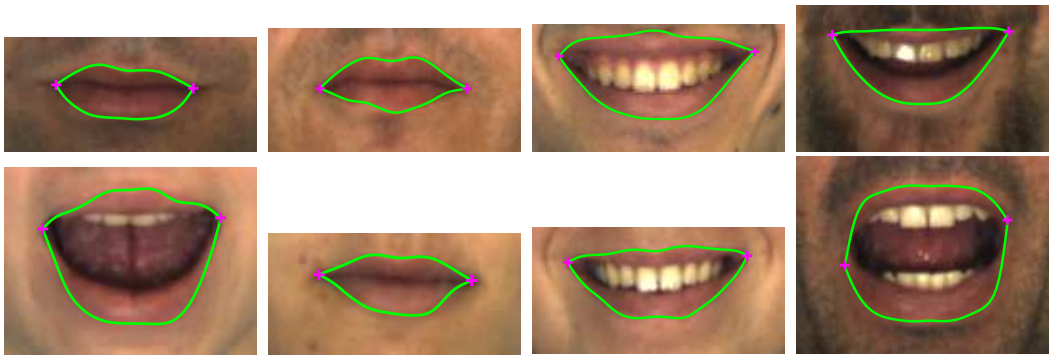


Figure 7.5: Example images of segmentation with OL from 90 % to 95 %.

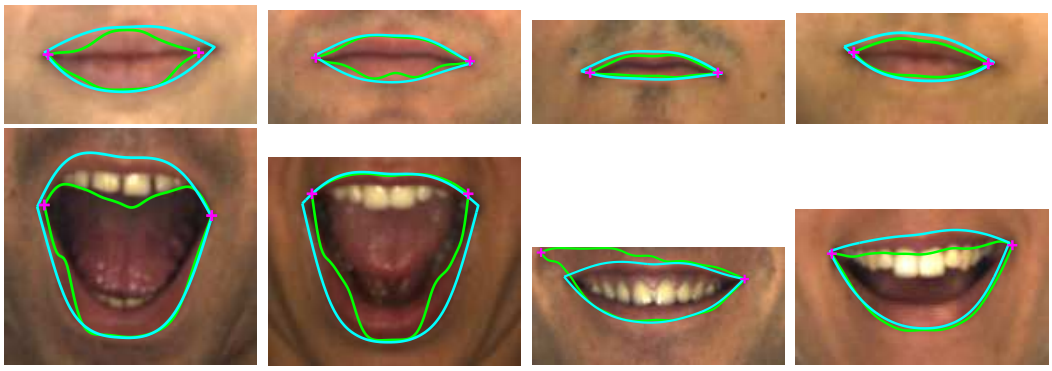


Figure 7.6: Example images of segmentation with OL from 80 % to 90 %.

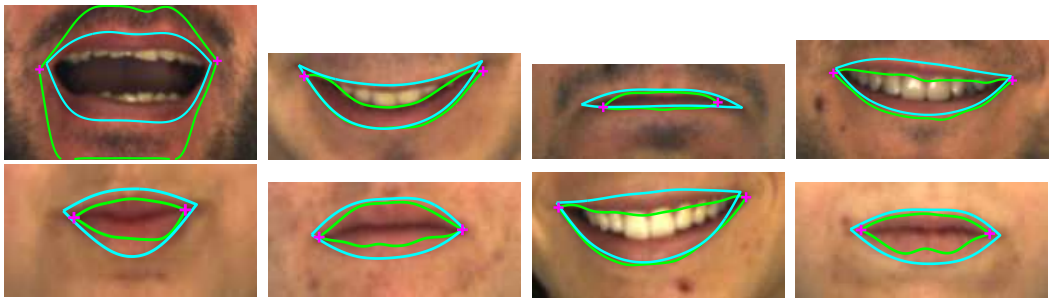


Figure 7.7: Example images of segmentation with OL from 70 % to 80 %.

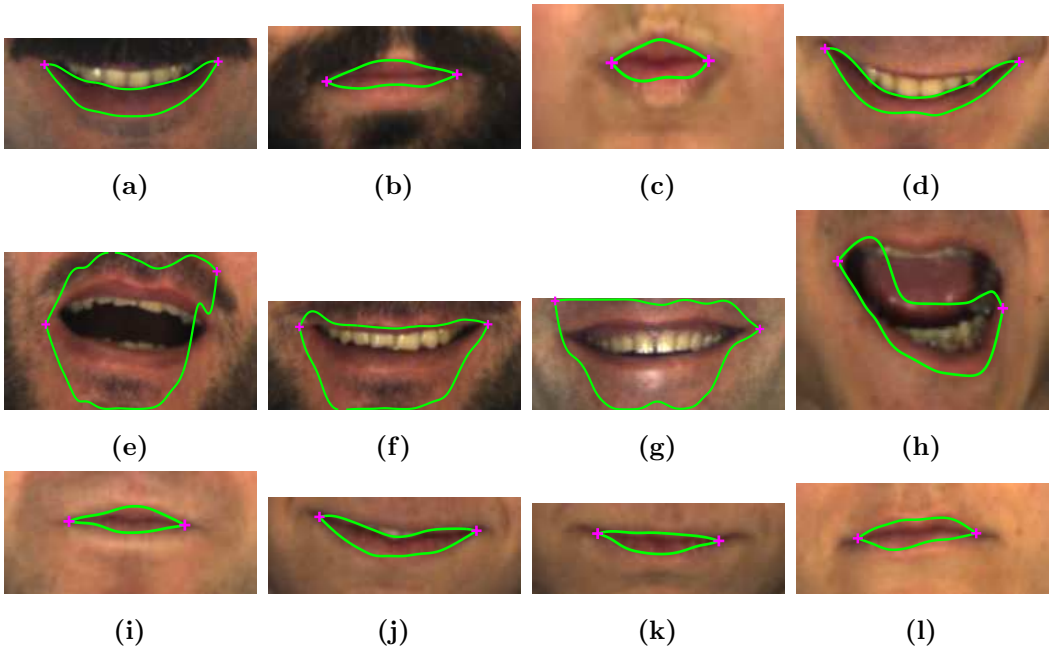


Figure 7.8: Example images of poor lip segmentation: OL below 70 %.

Only 1.9 % of images have a OL below 70 %. The examples in this bin highlight some of the shortcomings of the lip segmentation algorithm, which are analysed in the notes below (see Figure 7.8).

Figure 7.8 (a): Facial hair does not necessarily present a problem, as shown in Figure 7.4 where the algorithm obtains an accuracy in excess 96 % for some images with thick facial hair. However, in (a) the moustache causes a problem because it occludes the top lip contour. This presents a difficult challenge for lip segmentation as the algorithm cannot segment the lips when they are not visible. One possible solution is to augment the lip segmentation algorithm with a teeth detection algorithm. Once the teeth have been identified, the algorithm should ensure that the teeth are contained within the lips.

Figure 7.8 (b-c): At first glance, the issue with image (b) seems to be the dark, thick beard. However, upon closer inspection, the algorithm successfully identifies the upper lip contour, despite the beard. The problem is in fact very low contrast between the lower lip and the skin just below. The problem is the same in image (c). Accurate segmentation of the lips in these images is challenging to even a human, and therefore this case is not of particular concern.

Figure 7.8 (d): In this image, the top lip is extremely thin, and thus the algorithm only identifies the bottom lip. This case would benefit from a teeth detection algorithm which ensures that the teeth are contained within the lips.

Figure 7.8 (e-f): The presence of a light covering of facial hair results in three peaks on the histogram: lips, skin and facial hair. As a result, Otsu's method selects a poor threshold which results in poor segmentation. This type of problem is an ideal candidate for improvement with ATO (see Chapter 9).

Figure 7.8 (g): A slight amount of moisture on the skin surrounding the lips coupled with the angle of the light source creates a reflection. As a result, the histogram is comprised of three peaks (lips, skin and reflection), which results in a poor threshold selection. This type of problem is another candidate for improvement with ATO (see Chapter 9).

Figure 7.8 (h): The asymmetry of the expression exposes the teeth on one side of the mouth, but not the other. The colour transform causes teeth to appear similar to skin, which breaks the lip contour at the corner of the mouth. The subsequent morphological and smoothing operations result in the segmentation as shown. This image would benefit from first identifying and masking the teeth.

Figure 7.8 (i-l): These images are from the *neutral* or *anger* expressions. The segmentation does not appear to be entirely inaccurate, and seems to be significantly better than other images in the <70 % bin (a-h). In fact, the segmentation appears to follow the lip contours with reasonable accuracy, and only appears to deviate by a small number of pixels. To understand this result, it is necessary to consider the OL and SE metrics shown in equations (7.1) and (7.2). The denominator in SE is $2 \times TL$, where TL is the number of lip pixels in the ground truth. An image with the mouth wide open will have a large TL value, and thus any slight deviation from the lip contour will have little effect on SE. In contrast, an image with the mouth closed will have a small TL value, and thus any slight deviation from the lip contour will significantly increase SE. Similar logic applies to the OL metric. Thus, images with a small mouth area will be sensitive to small deviations from the lip contour, while images with a large mouth area will be insensitive to small deviations from the lip contour. OL and SE measure the segmentation accuracy for a region, but do not measure the adherence/deviation from a contour. However, when humans assess the accuracy of a lip segmentation result, the natural focus is on how well the segmented contour matches the ground truth contour. This suggests that although the OL and SE are widely used in the literature to evaluate lip segmentation algorithms, perhaps it would be better to use metrics that measure adherence/deviation from the lip contour. In other words, it may be more appropriate to measure the difference between the segmented contour and the ground truth contour, as opposed to the difference between the segmented region and the ground truth region.

7.4.2 Analysis of Expressions

Figure 7.9 and Figure 7.10 show the results for OL and SE for each expression. These results are summarised in the form of mean OL and mean SE in Table 7.3. *Scream* obtained the greatest OL of 95.06 %, followed by *smile* at 93.22 %. *Neutral* and *anger* obtained the lowest OL of 91.09 % and 89.54 % respectively. It is interesting to note that the percentage overlap seems to correlate with the degree of mouth opening: in the *scream* expression, the mouth is wide open (95.06 %); in the *smile* expression, the mouth is moderately open (93.22 %); in the *neutral* expression, the mouth is closed (91.09 %); and, in the *anger* expression, the mouth is tightly closed (89.54 %). There are two main factors which contribute to higher overlap and lower segmentation error when the mouth is open.

Table 7.3: Results for OL and SE for each expression.

	Expression	Mean OL (%)	Mean SE (%)	No. Images
1	Scream	95.06	4.91	224
2	Smile	93.22	6.95	224
3	Neutral	91.09	8.24	223
4	Anger	89.54	9.44	224
	Total	92.23	7.39	895

The first factor is the size of the mouth in each expression. When the mouth is wide open, as in the *scream* expression, the area of the mouth occupies close to 50 % of the rectangular mouth region; when the mouth is closed, the area of the mouth occupies less than 20 % of the rectangular mouth region. Otsu's method is used to threshold the image, which performs best on bimodal histograms having two distinct peaks (see Section 6.3.3). An image with 50 % object and 50 % background is more likely to have two distinct peaks than an image with 20 % object and 80 % background; therefore, Otsu's method performs better on open mouth histograms.

The second factor contributing to better performance when the mouth is open is the visibility of the lips in each expression. In the *neutral* expression, the mouth is closed and the lips are lightly pressed together, which causes part of the lip surface to be occluded by the other lip. Less surface area of lips is visible, and as a result, it is more difficult to identify and segment the lips. The expression for *anger* is typically formed by frowning with the eyebrows and clamping the lips tightly together, which results in even less visible lip area and further decreases the segmentation accuracy.

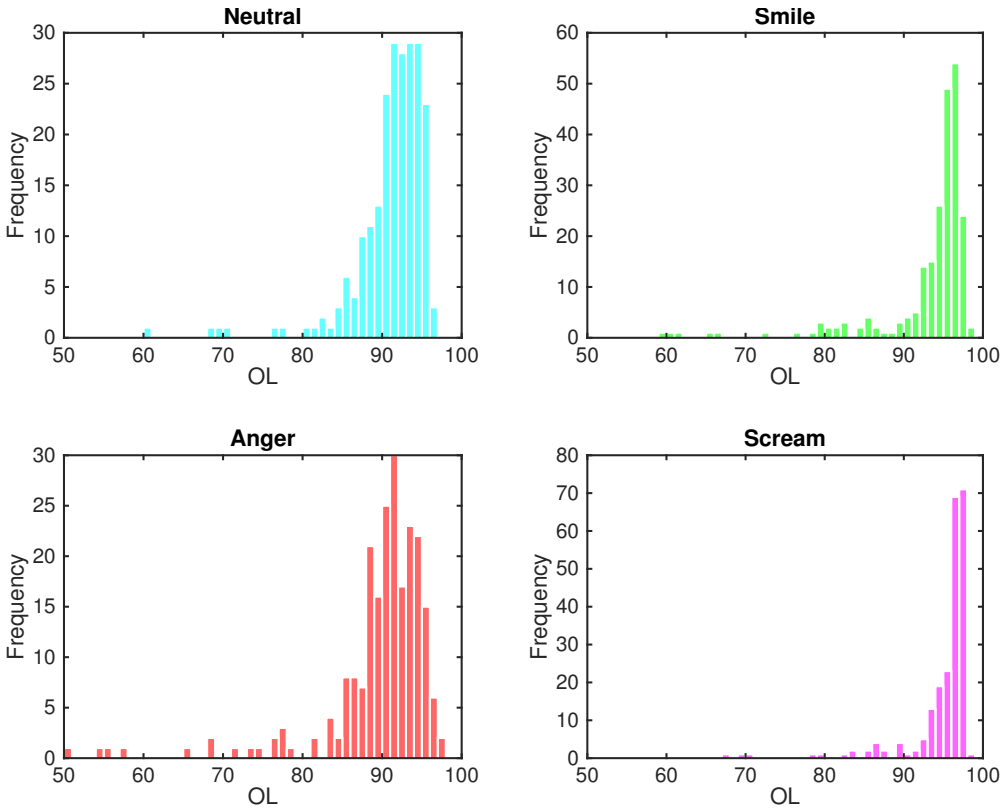


Figure 7.9: Histogram showing percentage overlap (OL) by expression.

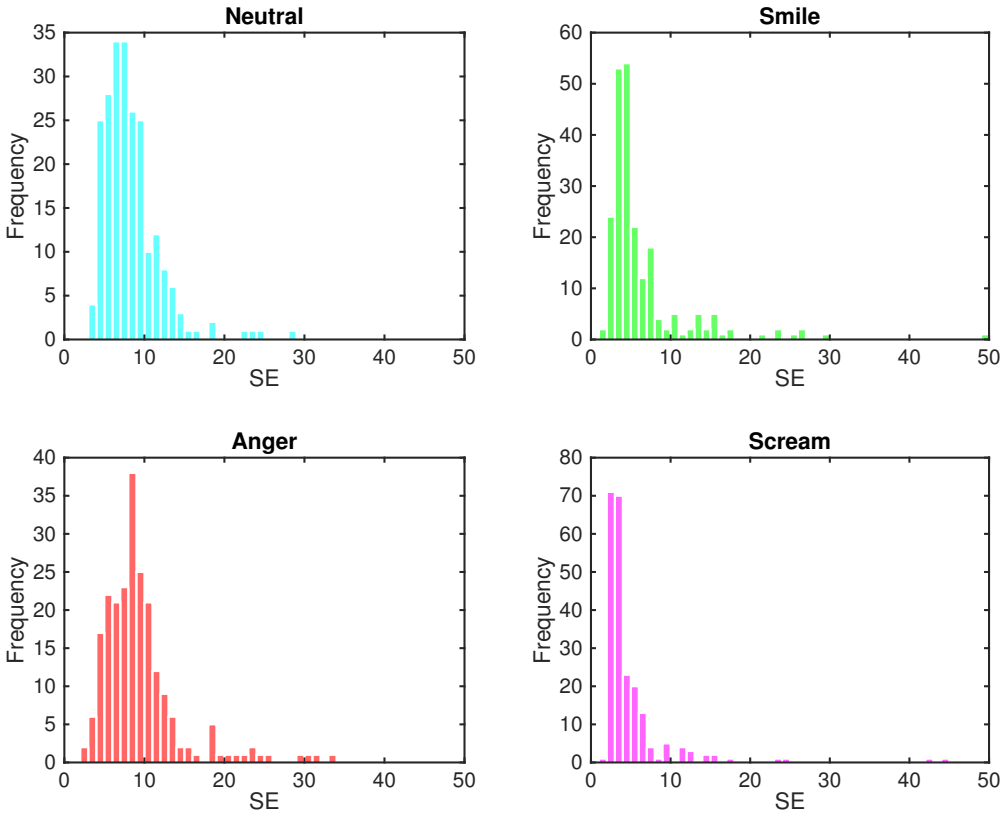


Figure 7.10: Histogram showing segmentation error (SE) by expression.

However, when the mouth is moderately open as in the *smile* expression, or wide open as in the *scream* expression, more of the lip surface is visible resulting in better segmentation of the lips.

7.5 Conclusion

The lip segmentation algorithm was tested on 895 mouth region images from the AR Face Database, using percentage overlap (OL) and segmentation error (SE) to quantify performance. The mean OL was 92.23 % and the mean SE was 7.39 %. Of the 895 images in the dataset, 78 % obtained an OL above 90 %, which seems acceptable for lip segmentation applications. The images in this bin include variation in gender, skin tone, and facial hair. Only 1.90 % of images obtained an OL below 70 %.

The following scenarios presented a challenge to the algorithm:

- facial hair obscuring the lips
- low contrast between the lips and skin
- thin lips
- light covering of facial hair causing poor threshold selection
- reflections caused by moisture on the skin

Several images from the neutral or anger expression appear to follow the lip contours with reasonable accuracy, however perform poorly according to the OL metric. This result is explained by considering the sensitivity of OL and SE to the size of the lip area. Although widely used in lip segmentation literature, OL and SE metrics are region-based metrics and may not be ideal for quantifying lip segmentation accuracy where the focus is on adherence/deviation from a contour.

Chapter 8

The Challenge of Threshold Selection

8.1 Introduction

Threshold-based segmentation methods provide a simple and efficient way to implement lip segmentation. However, automatic computation of robust thresholds presents a major challenge [103]. Fixed thresholds cannot be generalised due to variability in speaker appearance and lighting conditions, hence the threshold parameters must be calibrated for the specific speaker and environment. Furthermore, even after the initial calibration, appearance of the teeth, tongue, and oral cavity during movement of the mouth can significantly alter the image histogram and affect the threshold parameters.

The *base algorithm* described in Chapter 6 represents a typical threshold-based lip segmentation algorithm, and incorporates preprocessing, colour transform, thresholding, morphological processing, and contour smoothing. The default threshold in the base algorithm is selected using Otsu's method, which chooses the threshold to minimise the intraclass variance of black and white pixels [125]. Otsu's method is effective when the histogram is comprised of two compact and distinct peaks, one for lip pixels and one for skin pixels. However, Otsu's method may not select an adequate threshold in cases where the lip and skin components overlap considerably, or where the histogram includes additional peaks caused by facial hair, skin colouration, and illumination artefacts.

The base algorithm is analysed in Chapter 7, and results show that the algorithm performs well on 78 % of images obtaining OL above 90 %. However, the segmentation produced by the base algorithm on the remaining 22 % falls below the desired accuracy.

The images in this 22 % generally comprise more challenging scenarios including facial hair, low lip-skin contrast, shadows, and reflections.

This chapter illustrates the challenge of threshold selection by analysing two examples where Otsu's method fails to select a suitable threshold. Thereafter, this chapter quantifies the improvement in segmentation accuracy that can be obtained by adjusting the threshold value.

8.2 Qualitative Examples

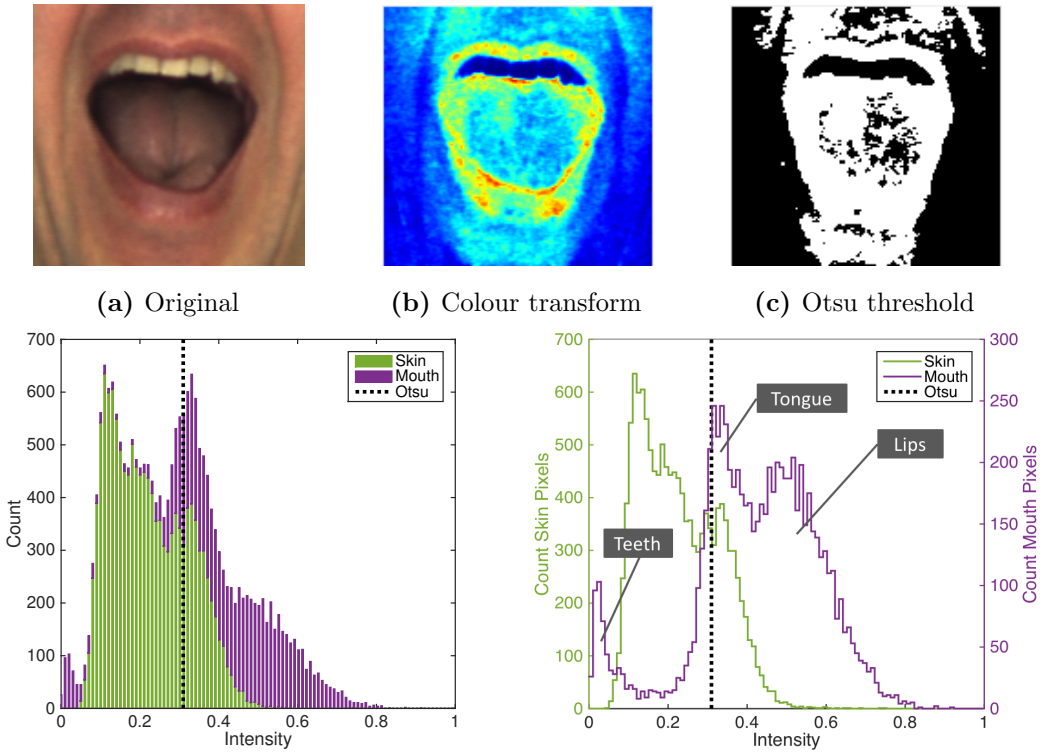
To illustrate some of the challenges of threshold selection, two segmentation examples are discussed below, using Otsu's threshold as the starting point.

Example 1

Figure 8.1 presents the first segmentation example using Otsu's method to select the threshold, which leads to poor lip segmentation. The figure contains the following subfigures: (a) original image; (b) intensity image after colour transform; (c) binary image after applying Otsu's threshold; (d) histogram of the intensity image; (e) decomposed histogram. The decomposed histogram is created by first classifying pixels as either *mouth* or *skin* according to the ground truth. Separate histograms are then computed for *mouth* and *skin* pixels. Note the different scales on the left and right y-axes.

The histogram in 8.1(d) contains two clear peaks at 0.1 and 0.35 respectively, with a moderate roll-off from 0.4 to 0.8. Otsu's method selects the threshold at 0.3, partway up the second peak. Considering only the shape of histogram 8.1(d), intuitively it seems preferable to select the threshold at the trough between the two peaks. In other words, it appears that the threshold selected by Otsu is too high, and should be decreased.

Subfigure 8.1(e) shows the histogram decomposed into mouth and skin components. By analysing the decomposed histogram in 8.1(e) relative to the intensity image in 8.1(b), the mouth pixels can be broken down further into lips, tongue, and teeth. It can be seen from the decomposed histogram in 8.1(e) that the ideal threshold to segment the lips and skin would then be around 0.4. Subfigure 8.1(c) shows the binary image after applying Otsu's threshold. The high number of false positive lip



(a) Original **(b)** Colour transform **(c)** Otsu threshold

(d) Histogram **(e)** Decomposed histogram

Figure 8.1: Example 1 – shortcomings of thresholding using Otsu's method. Otsu's threshold is too low, which results in a high number of skin pixels classified erroneously as lip pixels.

pixels indicates that the threshold is too lenient, and substantiates the assertion that the threshold needs to be increased.

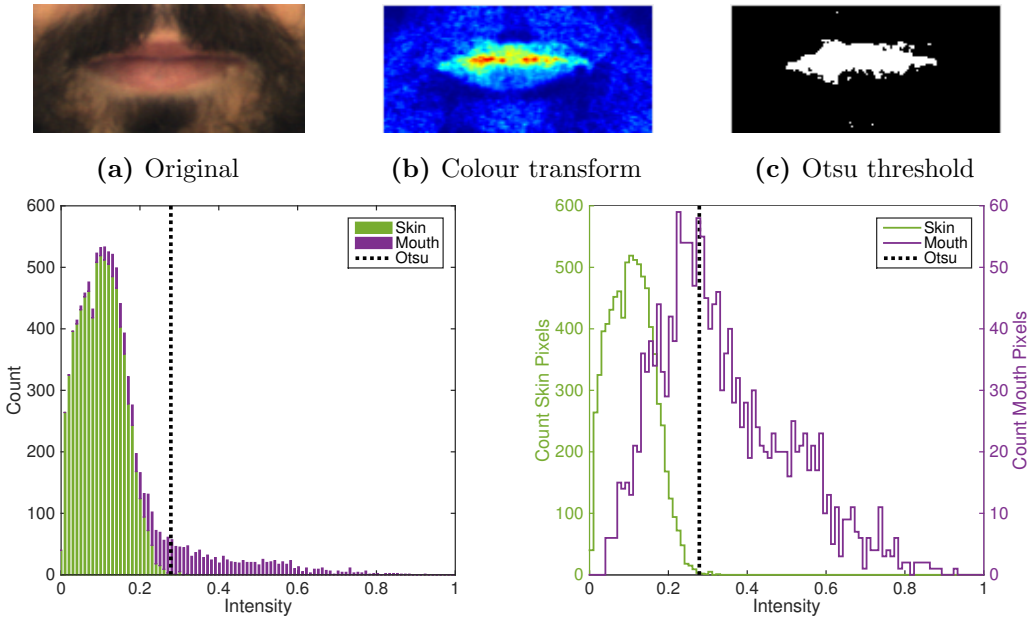
Considering only the shape of histogram 8.1(d), it initially seems logical to select the threshold between the two peaks; however, after analysing the decomposed histogram in 8.1(e) it becomes clear that the threshold should actually be increased. This example highlights the issue that in certain cases it is very difficult to compute a suitable threshold considering only the histogram.

Example 2

Figure 8.2 presents the second example using Otsu's method to select the threshold, which results in poor lip segmentation. The lips in 8.2(a) are surrounded by a dark, thick beard, and only a small area of skin is visible in the ROI. The histogram in 8.2(d) comprises one broad peak from 0 to 0.2. Otsu's method selects the threshold at 0.2, separating the broad peak from the roll-off. Considering only the shape of

the histogram in 8.2(d), it is difficult to comment on whether Otsu's threshold will produce a reasonable segmentation.

The decomposed histogram in 8.2(e) shows that Otsu's threshold excludes a significant number of lip pixels, and as a result, the entire bottom lip is classified as skin. The binary image after Otsu's threshold in 8.2(c) contains a high number of false negative lip pixels. In this case Otsu's threshold is too stringent, and should be decreased to produce better lip segmentation.



(a) Original **(b)** Colour transform **(c)** Otsu threshold

(d) Histogram **(e)** Decomposed histogram

Figure 8.2: Example 2 – shortcomings of thresholding using Otsu's method. Otsu's threshold is too high, which results in a high number of lip pixels classified erroneously as skin pixels.

The two examples illustrate some of the challenges of threshold selection. In certain cases it is very difficult to manually select an appropriate threshold, which implies that it is even more difficult to find an automatic method for this task (e.g. Otsu's method). However, it is clear from the examples above that the problem is not necessarily with the segmentation algorithm itself, and that simply adjusting the threshold value may produce accurate segmentation.

8.3 Quantifying the Improvement that can be Achieved by Optimising the Threshold

In threshold-based lip segmentation, such as the *base algorithm* of Chapter 6, the threshold value is usually the most sensitive parameter. In cases where the algorithm produces poor segmentation results, a slight adjustment of the threshold value may result in significantly better segmentation. This section quantifies the improvement that can be achieved by adjusting the threshold. In other words, this section compares the segmentation accuracy obtained when using Otsu's threshold to segment an image, versus the segmentation accuracy obtained using the optimal threshold. The algorithm itself remains unchanged, and only the threshold value is altered.

Equations (8.1) and (8.2) define the absolute and relative improvement, applicable to OL and SE defined in Section 7.3. In this section, $V_{reference}$ is the value obtained using Otsu's threshold, and V is the value obtained using the optimal threshold. The unit of measurement for absolute improvement between two percentages is *percentage point* (pp), while the unit of measurement for relative improvement is *percent* (%).

$$absolute\ improvement = V - V_{reference} \tag{8.1}$$

$$relative\ improvement = \frac{V - V_{reference}}{V_{reference}} \tag{8.2}$$

8.3.1 Method to Determine the Optimal Threshold

In order to quantify the improvement that can be achieved by adjusting the threshold, it is necessary to first determine the optimal threshold for an image. The optimal threshold is the value that results in the best segmentation accuracy. In order to obtain this value, a linear threshold search is performed for the base algorithm.

The threshold is incremented in steps of 0.01 from an initial value of 0 to a final value of 1. Setting the threshold value to 0 results in the all pixels in the ROI classified as lip pixels, while setting the threshold value to 1 results in the all the pixels classified as skin pixels. Setting the threshold value between 0 and 1 results in lip-skin segmentation with varying degrees of accuracy.

Figure 8.3 presents two examples of the linear search for the optimal threshold value. In Example 1, Subfigure 8.3(a) shows the segmentation error (SE) as a function of the threshold value. Below a threshold value of 0.11, all the pixels are classified as

lip pixels, and as a result the SE remains constant at the maximum 133 %. From 0.11 to the minimum value at 0.52 the SE decreases, thereafter the SE increases until 0.85 from where it remains constant. At the default threshold of 0.31, the SE is 44.8 %. The minimum SE is 6.09 % at a threshold of 0.52, which is the optimal threshold. The absolute improvement between the default threshold and the optimal threshold is 38.7 pp, and the relative improvement is 86 %.

Subfigure 8.3(b) shows the percentage overlap (OL) as a function of threshold. The pattern is inverse to that of the SE curve, with the optimal threshold occurring again at 0.52. Subfigure 8.3(c) shows the segmentation produced by the base algorithm using the default threshold, while 8.3(d) shows the segmentation at the optimal threshold.

Example 2 is shown in (e – h). The SE and OL curves follow a similar trend to Example 1, and the optimal threshold occurs at 0.25. The absolute improvement in SE between the default threshold and the optimal threshold is 6.7 pp, and the relative improvement is 69.1 %.

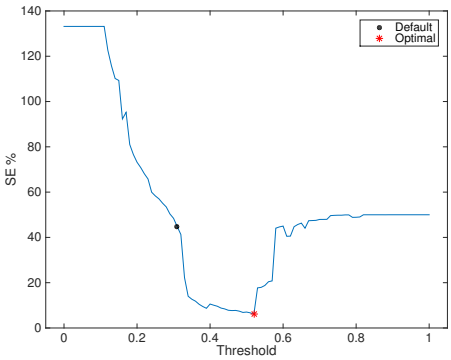
8.3.2 Quantifying the Accuracy Limit of the Base Algorithm

The accuracy limit of the base algorithm can be obtained by evaluating the segmentation accuracy at the optimal threshold. This represents the maximum accuracy that can be achieved by manipulating the threshold, without changing the algorithm itself.

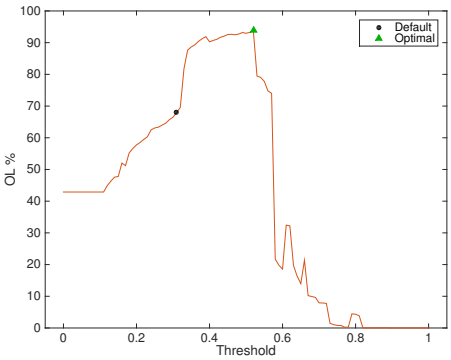
To quantify the potential for improvement, it is necessary to compare the performance of the base algorithm using the default threshold (Otsu) against the performance using the optimal threshold. The overall results of this comparison on the 895 AR Face images are presented in Table 8.1. The maximum accuracy limit of the base algorithm is shown under the *optimal threshold*: SE 4.75 % and OL 95.2 %. SE at the optimal threshold reflects a 2.64 pp absolute improvement and 35.8 % relative improvement over the default threshold. In terms of OL, the optimal threshold results in a 3.00 pp absolute improvement and 3.25 % relative improvement.

Figure 8.4 shows the cumulative distribution function (CDF) comparing the default threshold and the optimal threshold. In 8.4(a), the cumulative number of images is plotted against SE. A perfect segmentation algorithm would produce a curve that runs along the y-axis at $SE = 0\%$. The optimal curve shows the maximum accuracy that can be achieved by manipulating the threshold; any further improvement would

Example 1: Threshold Search



(a) Segmentation error (SE)

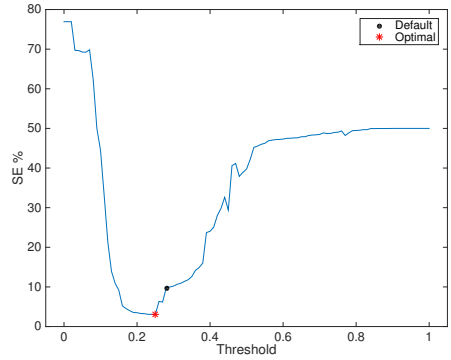


(b) Percentage overlap (OL)

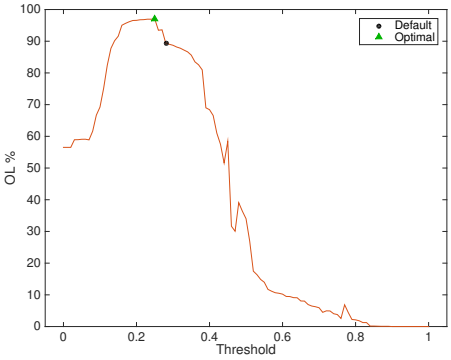
(c) Default threshold

(d) Optimal threshold

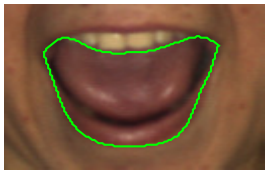
Example 2: Threshold Search



(e) Segmentation error (SE)



(f) Percentage overlap (OL)



(g) Default threshold

(h) Optimal threshold

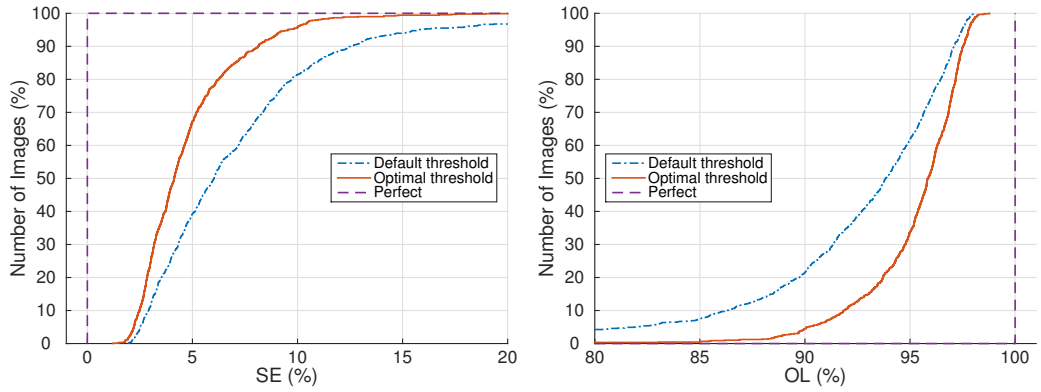
Figure 8.3: Linear threshold search to find the optimal values for Example 1 (a – d) and Example 2 (e – h). The optimal threshold corresponds to the minimum SE value, or the maximum OL value. Subfigures (c) and (g) show the segmentation produced by the base algorithm using the default threshold, while (d) and (h) show the segmentation at the optimal threshold.

Table 8.1: Comparing the performance of the base algorithm using the default threshold versus the optimal threshold. The segmentation error (SE) and percentage overlap (OL) reflect the mean values across 895 AR Face images.

	Default threshold (%)	Optimal threshold (%)	Improvement absolute (pp)	Improvement relative (%)
SE	7.39	4.75	2.64	35.8
OL	92.2	95.2	3.00	3.25

require changing the actual algorithm. Considering the optimal threshold, 67 % of segmentations obtain SE below 5 %, and 96 % of segmentations obtain SE below 10 %. By comparison, with the default threshold only 39 % of images obtain SE below 5 %, and 81 % of segmentations obtain SE below 10 %.

In terms of OL shown in Figure 8.4(b), a perfect segmentation algorithm would produce a curve that runs parallel to the y-axis at $OL = 100\%$. The default threshold produces 22 % of segmentations with OL below 90 %, while the optimal threshold produces just 4.4 % of segmentations below 90 %.

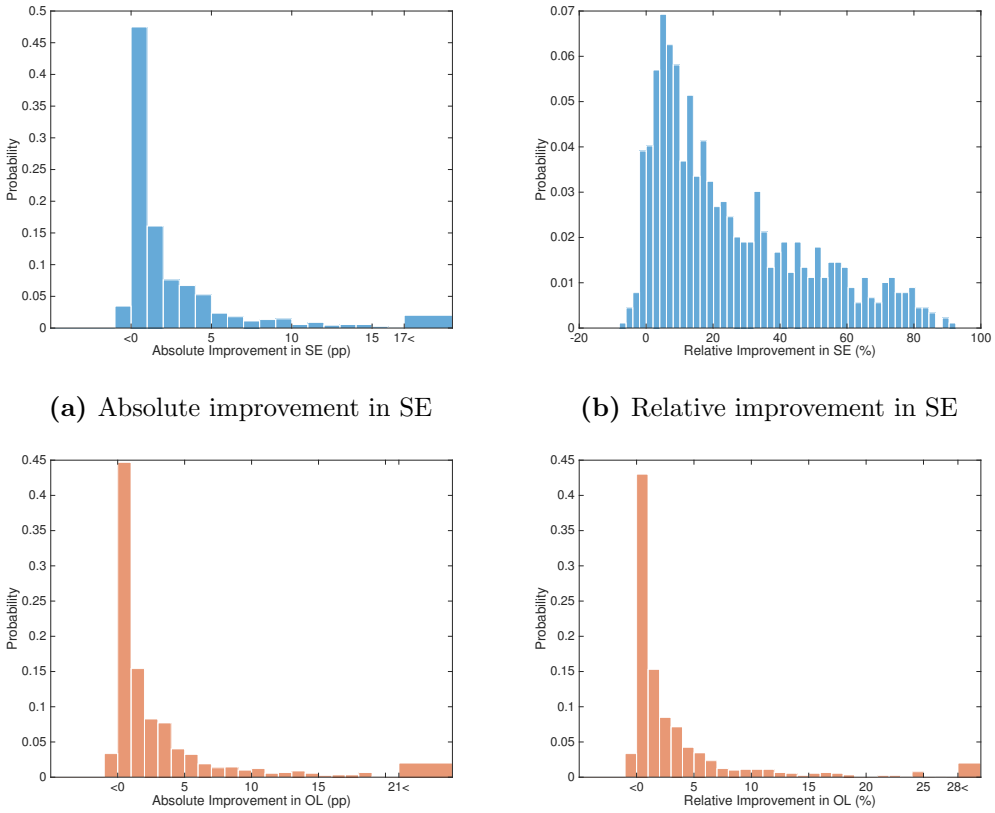


(a) Segmentation error (SE) CDF **(b)** Percentage overlap (OL) CDF

Figure 8.4: Cumulative distribution function (CDF) comparing the default threshold versus the optimal threshold.

Figure 8.5 shows a breakdown of the improvement between the default threshold and the optimal threshold. The relative improvement in SE is shown in 8.5(b). Half of the segmentations improve moderately by 0 – 20 %, while 37 % of the segmentations improve substantially by 20 – 40 %. In cases where the Otsu's method selected a particularly poor threshold value, the segmentation improved by over 60 %, which occurred in 10 % of images.

In a small number of images, the optimal threshold actually produced slightly worse segmentation than the default threshold. The reason for this is that the linear threshold search is incremented in steps of 0.01, whereas Otsu's method operates on a continuous range. As an example, consider a case where Otsu's method selects a very good threshold, say 0.034, resulting in SE of 3.6 %. The linear threshold search only searches the threshold values 0.03 and 0.04 which may result in slightly less accurate segmentation.



(a) Absolute improvement in SE

(b) Relative improvement in SE

(c) Absolute improvement in OL

(d) Relative improvement in OL

Figure 8.5: Histograms of improvement in segmentation accuracy between the default threshold (Otsu) and the optimal threshold obtained from the linear threshold search.

8.4 Conclusion

In the field of image segmentation, poor segmentation accuracy indicates a shortcoming of the algorithm. In the case of threshold-based segmentation, the problem may reside in the general approach of the algorithm, or the problem may be specifically in the method of threshold selection.

The *base algorithm* described in Chapter 6 is investigated from both qualitative and quantitative perspectives, to determine the potential impact of improving the threshold selection method.

The quantitative analysis reports that the accuracy limit of the base algorithm with the optimal threshold is SE 4.75 %, which represents a 35.8 % relative improvement over the default threshold. It is clear from this analysis that the segmentation produced by the base algorithm can be improved significantly by implementing a better threshold selection technique.

Chapter 9

Adaptive Threshold Optimisation (ATO) Algorithm

9.1 Introduction

Threshold-based segmentation methods provide a simple and efficient way to implement lip segmentation. However, automatic computation of robust thresholds presents a major challenge. Chapter 8 describes the challenge of threshold selection, and concludes that the *base algorithm* from Chapter 6 can be improved by up to 25 % relative SE.

This chapter describes an adaptive algorithm for selecting the histogram threshold, based on feedback of shape information. The algorithm reduces unnecessary overhead by first comparing the initial segmentation to a reference lip shape model to decide if optimisation is required ('validation stage'). In cases where optimisation is required, the algorithm iteratively adjusts the threshold to reduce the segmentation error ('optimisation stage'). This novel technique for threshold selection is referred to as *Adaptive Threshold Optimisation (ATO)*.

The dataset to train and test ATO comprises 895 images from the AR Face Database [100], which includes 112 different subjects. The 112 subjects are randomly divided into two groups: 60 % in the training group, and 40 % in the test group. The training dataset comprises 544 images from the subjects in the training group, and the test dataset comprises 351 images from the subjects in the test group. The training dataset is used to build the lip shape model (LSM) in this chapter, and the test dataset is used to conduct the three tests in Chapter 10.

The sections that follow detail the components of ATO including construction of the lip shape model (LSM), the validation stage, and the optimisation stage.

9.2 Algorithm Overview

At a high-level, the algorithm consists of two components: the base algorithm which represents a typical threshold-based segmentation algorithm; and the *adaptive threshold optimisation (ATO)* algorithm. Figure 9.1 presents an overview of the base algorithm and the ATO algorithm shown in the shaded blocks. ATO is not a standalone segmentation algorithm; rather it is designed to augment an existing threshold-based algorithm. ATO aims to improve the segmentation of the base algorithm by optimising the threshold parameter.

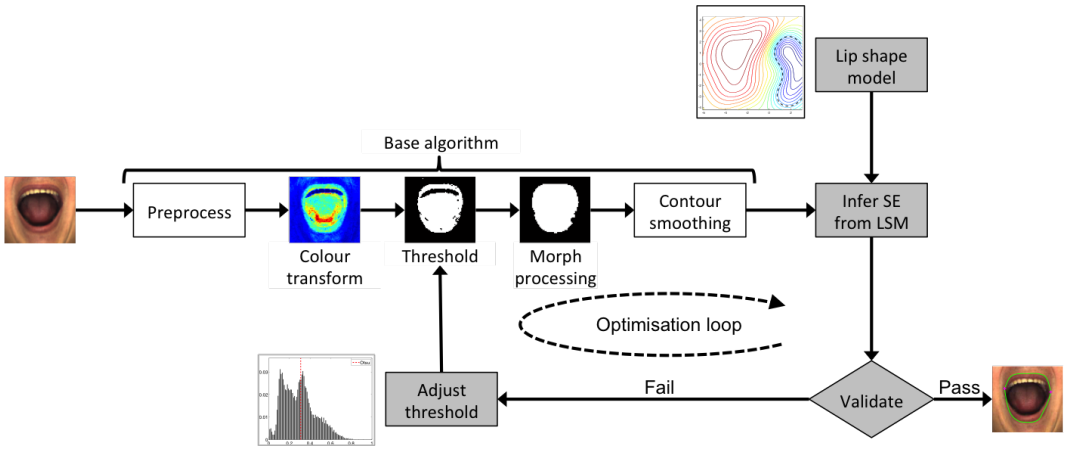


Figure 9.1: Block diagram of the base algorithm with adaptive threshold optimisation (ATO) shown in the shaded blocks.

The threshold-based segmentation algorithm described in Chapter 6 is used as the base algorithm in this research. The base algorithm from Chapter 6 comprises the following components: colour transform, thresholding, morphological processing, and contour smoothing. The base algorithm uses Otsu's method [125] to select the default threshold value.

It is important to note that ATO does not depend on the specific make-up of the base algorithm, or the method of computing the default threshold. The base algorithm from Chapter 6 may be replaced by various different threshold-based algorithms, for example: Coianiz et al. [105], Wark et al. [132], Chiou & Hwang [133] or Zhang & Mersereau [109].

The principle of ATO is to use feedback of shape information to drive selection of the threshold (see Figure 9.1). ATO incorporates shape information by constructing a lip shape model (LSM) from the training images.

In simple cases, the segmentation produced by the base algorithm using the default threshold is adequate, and it is not necessary to optimise the threshold. Therefore, to preserve the overall efficiency, ATO first determines whether or not the initial segmentation is satisfactory in the *validation stage*. Only if the initial segmentation is deemed unsatisfactory, then ATO proceeds to recompute the threshold in the *optimisation stage*.

The validation stage uses the LSM to infer the segmentation error, then determines whether to accept or reject the segmentation. If the segmentation is rejected, then the optimisation stage iteratively adjusts the threshold value to minimise the inferred error.

9.3 Lip Shape Model (LSM)

Figure 9.2 shows the procedure to construct a lip shape model (LSM) from the training dataset. Samples of both good and bad segmentation from the base algorithm are combined with the ground truth images to create the discriminant dataset. The discriminant dataset is used to train a regression SVM model to estimate the segmentation error. The parameters of the SVM model are optimised using a grid search approach with cross validation.

Since the LSM is trained from samples produced by the base algorithm, the resulting model effectively characterises the segmentation errors of the underlying algorithm. If a different base algorithm is chosen, then it is necessary to train a new LSM.

The subsections that follow explain the process to construct the LSM in more detail.

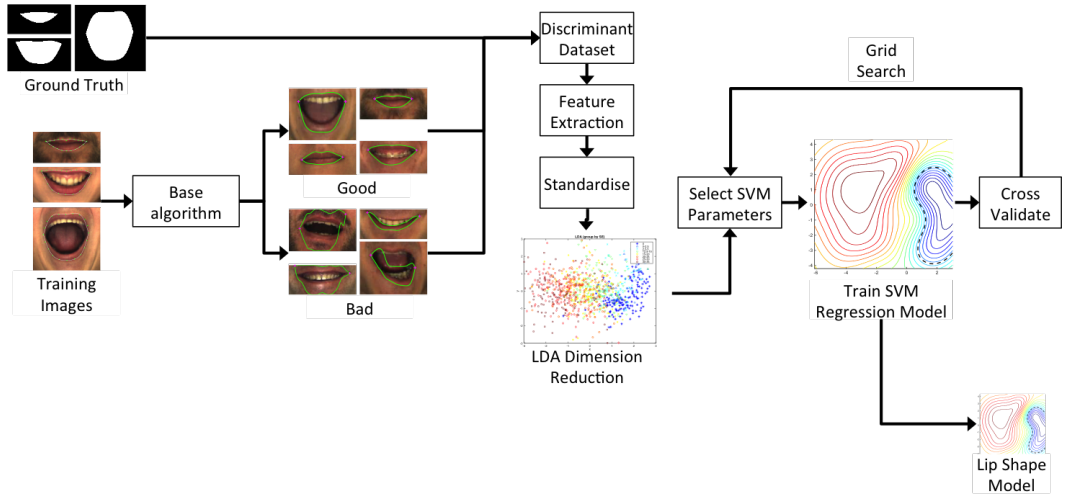


Figure 9.2: Process to construct the lip shape model (LSM) using ground truth images and output from the base algorithm. Samples of both *good* and *bad* segmentation from the base algorithm are combined with the *ground truth* images to create the discriminant dataset. The discriminant dataset is used to train a regression SVM model to estimate the segmentation error.

9.3.1 Discriminant Dataset

The 544 images from the training dataset are used to generate three subsets which comprise the discriminant dataset:

1. Ground truth images ($SE = 0\%$)
2. Samples of good segmentation from the base algorithm ($SE < 5\%$)
3. Samples of bad segmentation from the base algorithm ($SE > 12.5\%$)

The *good* and *bad* samples are generated by varying the threshold of the base algorithm to produce output with a range of segmentation errors. For example, applying four different thresholds to the same image (e.g. 0.1, 0.25, 0.8, 0.9), will result in four different segmentations, with different segmentation errors (e.g. 21 %, 15 %, 5.5 %, 14 %). Using this approach, the discriminant dataset is expanded to three times larger than the training dataset.

The output of the base algorithm is classified as *'good'* if the segmentation error is below 5 %, and *'bad'* if the segmentation error is above 12.5 %. To create a clear distinction between *good* and *bad*, output with segmentation error from 5 % to 12.5 % is not included in the discriminant dataset. Table 9.1 shows a breakdown of the discriminant dataset by segmentation error. *Ground truth* and *good* segmentation together make up 27 % of the discriminant dataset, while *bad* segmentation accounts

for the remaining 73 %.

Table 9.1: Breakdown of discriminant dataset by SE.

	SE	Number	Subtotal
Ground Truth	0	447	447
Good	0-2.5	49	
	2.5-5	290	339
Bad	12.5-15	414	
	15-20	441	
	20-25	429	
	25-30	427	
	30-35	423	2134
		2920	2920

Figure 9.3 shows examples from the discriminant dataset including *ground truth*, *good*, and *bad* segmentation. The threshold (TH) used to generate the segmentation is shown along with the corresponding segmentation error (SE). The left column of *bad* examples shows segmentation produced with a lenient threshold, resulting in a high number of false positive lip pixels. The right column of *bad* examples shows segmentation produced with a strict threshold, resulting in a high number of false negative lip pixels.

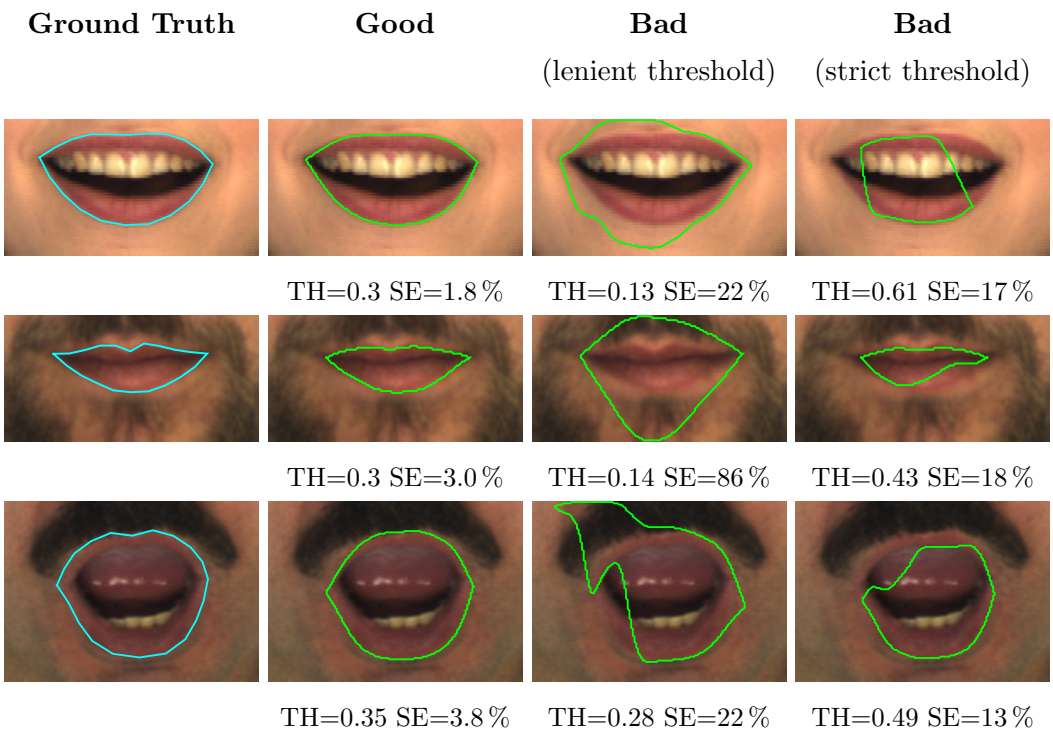


Figure 9.3: Segmentation examples from the discriminant dataset, which comprises three subsets: *ground truth*, *good*, and *bad*. The *good* and *bad* samples are generated by varying the threshold (TH) of the base algorithm to produce output with a range of segmentation errors (SE).

9.3.2 Feature Extraction

The output of the base algorithm is a binary image containing the lip region and the skin region. The feature vector for the lip shape model comprises fourteen geometric features shown in Table 9.2. The metrics relating to size and distance are dependant on the physical characteristics of the subject (e.g. size of mouth), the proximity to the camera, and the zoom of the camera; therefore, it is necessary to normalise these measures relative to the mouth region of interest (ROI). Table 9.2 shows the fourteen features and the reference quantities used to normalise the features. For example, the area of the lips is measured relative to the area of the ROI. The feature vector is standardised (centred and scaled) by calculating the z-score shown in (9.1).

$$z = \frac{x - \mu}{\sigma} \tag{9.1}$$

where

x = raw score

μ = mean of the population

σ = standard deviation of the population

Table 9.2: Feature vector for lip shape model.

No.	Feature	Reference
1	Area	ROI area
2	Perimeter	ROI perimeter
3	Centroid X	ROI width
4	Centroid Y	ROI height
5	Major axis length	ROI diagonal
6	Minor axis length	ROI diagonal
7	Eccentricity	
8	Orientation	
9	Bounding box X	ROI width
10	Bounding box Y	ROI height
11	Width	ROI width
12	Height	ROI height
13	Distance transform mean	ROI diagonal
14	Distance transform std dev	ROI diagonal

9.3.3 LDA Dimensionality Reduction

Linear discriminant analysis (LDA) is a statistical method to find a linear combination of features that characterises or separates two or more classes. The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory information as possible.

In building the lip shape model, LDA is used to find the eigenvectors that separate the three subsets in the discriminant dataset – *ground truth*, *good*, and *bad*. Since there are three classes ($N = 3$), LDA returns two eigenvectors ($N - 1 = 2$).

Table 9.3 shows the two eigenvectors obtained from LDA. Eigenvector 1 accounts for 94.8 % of the discriminating power, while Eigenvector 2 only accounts for 5.2 % of the discriminating power.

The features in Table 9.3 are ranked according to the absolute value of Eigenvector 1. Top ranked features exhibit the greatest distinguishing power, while lower ranked features are less effective in discriminating between *good* and *bad* segmentation. Two features dominate the distinguishing power of Eigenvector 1: *distance mean* and *perimeter*.

Table 9.3: Linear discriminant analysis (LDA) eigenvectors and eigenvalues with the features ranked by absolute value of Eigenvector 1.

No.	Feature	Eigenvector 1	Eigenvector 2
1	Distance Mean	-4.498	-0.130
2	Perimeter	-3.552	1.104
3	Location X	-1.374	-3.823
4	Height	-1.338	0.099
5	Width	-1.308	-3.709
6	Distance Std	1.164	0.112
7	Minor Axis	1.135	-1.185
8	Major Axis	0.860	-0.293
9	Location Y	0.526	0.435
10	Area	0.451	0.164
11	Centroid Y	-0.401	-0.179
12	Centroid X	0.381	0.767
13	Orientation	0.028	0.196
14	Eccentricity	-0.025	0.085
	Eigenvalue	1.88	0.103
	Eigenvalue percent	94.8 %	5.2 %

Figure 9.4 shows examples of the *perimeter* and *distance mean* features. The distance transform is shown in the intensity images, which measures the euclidean distance between each pixel and the nearest nonzero pixel in a binary image. Pixels within the segmentation boundary have zero distance (blue), while pixels further from the boundary have increasing distance (red). As the *perimeter* of the region increases, the corresponding *distance mean* tends to decrease. *Bad* segmentation with a lenient threshold is characterised by high *perimeter* and low *distance mean*, while *bad* segmentation with a strict threshold is the inverse, low *perimeter* and high *distance mean*. *Good* segmentation tends to be somewhere in-between.

In Table 9.3, *distance mean* is ranked first, having a weight of -4.5 in Eigenvector 1. The magnitude of *distance mean* is the largest of the fourteen features, which indicates that this feature contains the greatest discriminating power. The negative weighting of *distance mean* implies that images in the *bad* subset have a high value of *distance mean*, while images in the *good* subset have a low value of *distance mean*.

Perimeter is ranked second, having a weight of -3.5 in Eigenvector 1. The negative weighting of *perimeter* indicates that large *perimeter* corresponds to *bad* segmentation,

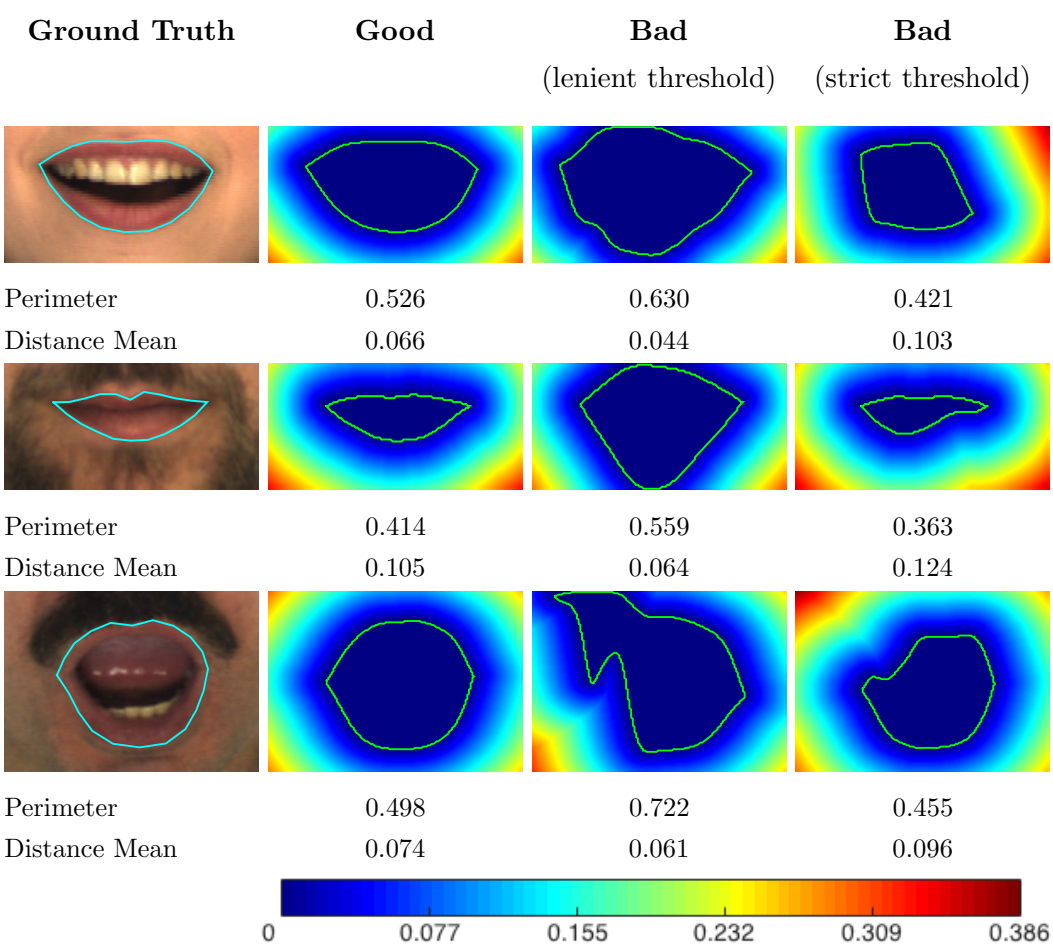
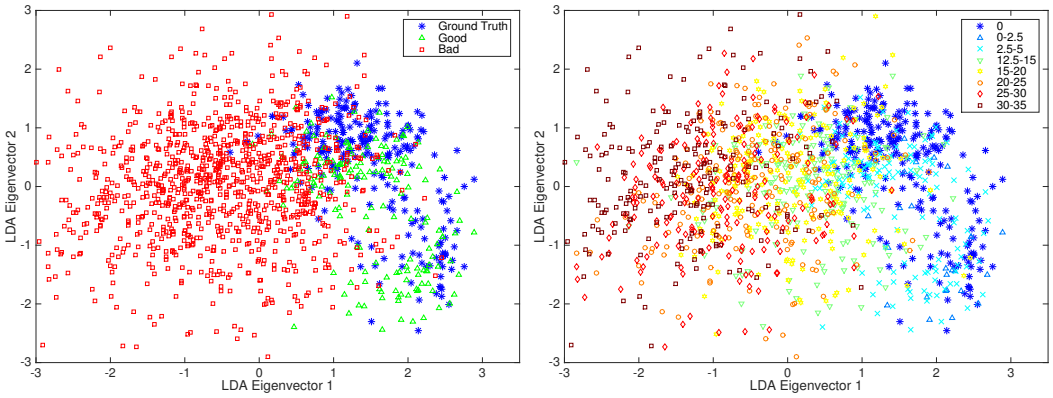


Figure 9.4: Examples of the *Perimeter* and *Distance Mean* features, which dominate the distinguishing power of Eigenvector 1. The distance transform is shown in the intensity images, which measures the euclidean distance between each pixel and the nearest nonzero pixel in a binary image. The *Distance Mean* and *Perimeter* for each segmentation is shown below the image, and the corresponding threshold (TH) and segmentation error (SE) can be found in Figure 9.3.

while small *perimeter* corresponds to *good* segmentation.

Figure 9.5 shows the training images projected in the two-dimensional plane comprising LDA Eigenvector 1 and LDA Eigenvector 2. In (a) the feature vectors are coloured by subset, and in (b) the feature vectors are coloured by segmentation error (SE). The discriminating power of Eigenvector 1 can be seen clearly as *ground truth* and *good* segmentation tend to have high values of Eigenvector 1, while *bad* segmentation tends to have low values of Eigenvector 1. Figure 9.5(b) shows the trend in segmentation error (SE) – as Eigenvector 1 increases the segmentation error decreases.



(a) Grouped by subset **(b)** Grouped by SE

Figure 9.5: Scatter plot of feature vectors from training dataset after LDA to reduce dimensionality. In (a) the feature vectors are coloured by subset, and in (b) the feature vectors are coloured by segmentation error (SE).

9.3.4 ϵ-Support Vector Regression (ϵ-SVR)

The LSM uses *ϵ-Support Vector Regression (ϵ-SVR)* to fit a model which approximates the relation between the shape features and the corresponding segmentation error (SE). Once the model has been trained, the shape features of an unknown region can then be used to infer the segmentation error.

The fundamental principles of ϵ-SVR are reviewed in this section, and the reader is referred to [166] for a comprehensive tutorial on support vector regression.

Basic Principles of ϵ-SVR

Consider a set of training points, $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_m}, y_m)$, where $\mathbf{x_i} \in \mathbb{R}^n$ is a feature vector and $y_i \in \mathbb{R}^1$ is the target output. In the case of the lip shape model, $\mathbf{x_i}$ is the reduced two-dimensional feature vector ($n = 2$), and y_i is the corresponding segmentation error.

In ϵ-SVR [167], the goal is to find a function $f(\mathbf{x})$ that satisfies two objectives:

1. The deviation of $f(\mathbf{x})$ from the actually obtained targets y_i must be less than ϵ for all training data
2. The function $f(\mathbf{x})$ must be as flat as possible

For the case of linear functions $f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b$, the objectives can be written as a convex optimisation problem where $\langle \cdot, \cdot \rangle$ denotes the dot product:

$$\begin{aligned} &\underset{w,b}{\text{minimise}} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to} \quad \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x_i} \rangle - b & \leq \epsilon \\ \langle \mathbf{w}, \mathbf{x_i} \rangle + b - y_i & \leq \epsilon \end{cases} \end{aligned} \tag{9.2}$$

where $\epsilon > 0$ is a predefined constant which controls the noise tolerance. Minimising the norm $\|\mathbf{w}\|^2 = \langle \mathbf{w}, \mathbf{w} \rangle$ ensures the flatness of $f(x)$.

Equation (9.2) assumes that the optimisation problem is feasible, in other words, that such a function f actually exists that approximates all pairs $(\mathbf{x_i}, y_i)$ with precision ϵ. In some cases, this assumption does not hold, and no function f exists that satisfies the ϵ precision constraint. It may be beneficial to allow for some errors whereby creating a 'soft margin' loss function [168], which was used in support vector machines by Cortes & Vapnik [169]. The error is introduced by way of slack variables ξ_i, ξ_i^* to arrive at the standard form of support vector regression [167]:

$$\begin{aligned} &\underset{w,b,\xi,\xi^*}{\text{minimise}} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ &\text{subject to} \quad \begin{cases} y_i - \langle \mathbf{w}, \mathbf{x_i} \rangle - b & \leq \epsilon + \xi_i \\ \langle \mathbf{w}, \mathbf{x_i} \rangle + b - y_i & \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0, \text{for } i = 1, 2, ..., m \end{cases} \end{aligned} \tag{9.3}$$

The constraint $C > 0$ determines the trade-off between flatness of f and the tolerance of deviations larger that ϵ. Larger values of C favour minimising the error, at the expense of a more complex model. Smaller values of C favour a more flat model, at the expense of larger error.

Figure 9.6 is a graphical representation of (9.3). Points within the shaded region $\pm\epsilon$ do not contribute to the cost, while points outside the region are penalised in a linear fashion.

In most cases the optimisation problem 9.3 can be solved more easily in its dual formulation. Furthermore, the dual formulation facilitates extending the support vector machine to non linear functions. The reader is referred to [166] for more details on the dual problem and quadratic programs.

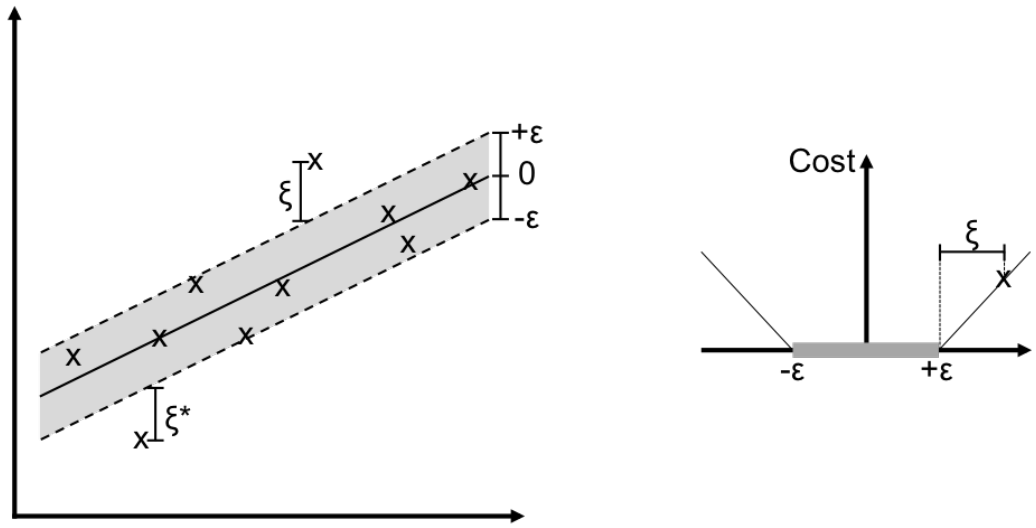


Figure 9.6: Graphical representation of Eq (9.3) soft margin for support vector regression [170], with ϵ-insensitive loss function [167].

Kernel Selection

The radial basis function (RBF) kernel, also known as the Gaussian kernel, is used in the lip shape model. The RBF kernel nonlinearly maps samples into a higher dimensional space, which enables the kernel to handle cases where the relation between the features and the target is nonlinear. The RBF kernel is shown in (9.4).

$$K(\mathbf{x_i}, \mathbf{x_j}) = e^{-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2} \tag{9.4}$$

where

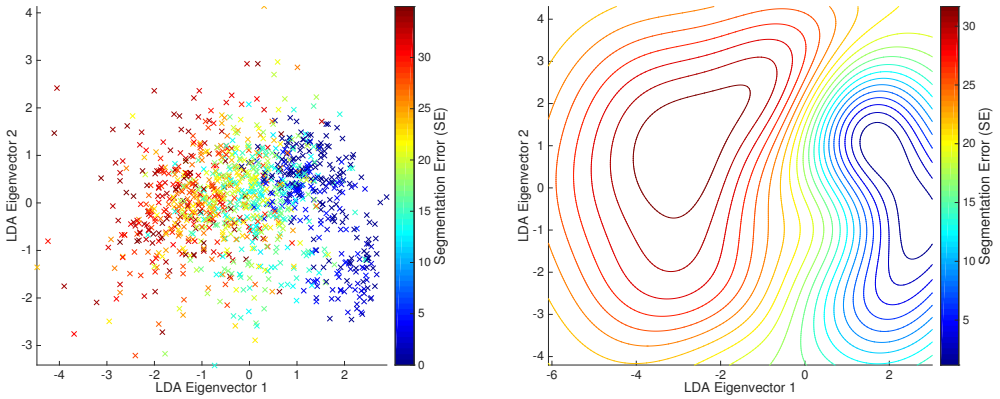
$$\gamma = 1/2\sigma^2$$

$\|\mathbf{x_i} - \mathbf{x_j}\|^2$ is the squared Euclidean distance between the support vector $\mathbf{x_i}$ and the data point $\mathbf{x_j}$. The support vector $\mathbf{x_i}$ is the centre of the RBF, and γ determines the area of influence of the support vector over the feature space. For small values of γ, the area of influence is large, which results in a smooth decision surface with fewer support vectors. For large values of γ, the area of influence is small, which results in a more complex model.

Applying ϵ-SVR to Training Data

The ϵ-SVR model is trained with the 2D reduced feature vectors $\mathbf{x_i}$ and corresponding segmentation error y_i. The LIBSVM library is used to implement ϵ-SVR [162]. Figure 9.7 shows an example of the training data and the resulting ϵ-SVR model.

The model is dependant of the selection of two parameters: C which controls the trade-off between flatness and tolerance of deviations, and γ which controls the influence of the support vectors. Selection of the parameters C and γ is discussed in Section 9.3.5.



(a) Training data **(b)** ϵ-SVR model

Figure 9.7: ϵ-SVR model trained from 2D reduced feature vectors $\mathbf{x_i}$ and corresponding segmentation error y_i. Figure (a) shows a scatter plot of the training data, and Figure (b) shows the ϵ-SVR model. The model is dependant on selection of two parameters: C which controls the flatness of the model, and γ which controls the support vector influence in the kernel.

9.3.5 SVM Parameter Selection

The ϵ-SVR model is dependant two parameters: C from (9.3) which controls the trade-off between flatness and tolerance of deviations larger that ϵ; and the kernel parameter γ from (9.4) which controls the influence of the support vectors. The goal is to select (C, γ) such that the model can accurately predict the lip segmentation error of an unknown region.

In order to prevent overfitting the model to the training data, k-fold cross-validation is used whereby the training set is divided into k subsets of equal size. Sequentially, the model is trained on $k-1$ subsets, and tested on the remaining one subset. The mean squared error (MSE) is calculated from model performance on the test subsets. The value of k is set to 10.

The grid search technique is used to tune the values of C and γ. The motivation for using the grid-search approach is two-fold. First, the grid search approach increases the researcher's confidence that the input space has been adequately covered [171].

Second, the grid search approach does not require significantly more computational power than more advance methods since there are only two parameters [171].

Since a complete grid search can be time consuming, the grid search starts with a coarse grid which is then narrowed to find the best (C, γ). Furthermore, to cover a larger input space, the values C and γ are tried in an exponentially growing sequence:

$$C = 2^{-15}, 2^{-14}, ..., 2^{14}, 2^{15}$$

$$\gamma = 2^{-15}, 2^{-14}, ..., 2^{14}, 2^{15}$$

Figure 9.8 shows the results of the grid search starting with the coarse grid, progressing to the medium grid, and ending with the fine grid. For each pair (C, γ) the value of $log_{10}(MSE)$ is plotted on an intensity scale where the maximum of $log_{10}(MSE)$ is red, and the minimum of $log_{10}(MSE)$ is blue. The intensity bar is rescaled to the range of values in the corresponding grid. The white asterisk * shows the minimum MSE at each scale. In the fine grid search, the minimum $log_{10}(MSE) = 1.587$ i.e. $MSE = 38.65$. The minimum MSE occurs at:

$$\begin{array}{llllll} log_2(C) & = 1 & \text{i.e.} & C & = 2 \\ log_2(\gamma) & = -2 & \text{i.e.} & \gamma & = 0.25 \end{array}$$

9.3.6 Final Lip Shape Model

The final lip shape model (LSM) is trained using ϵ-SVR with the parameters $C = 2$ and $\gamma = 0.25$. The final LSM is shown in Figure 9.9. The model infers the lip segmentation error (SE) of an unknown region from the reduced feature vector of the region.

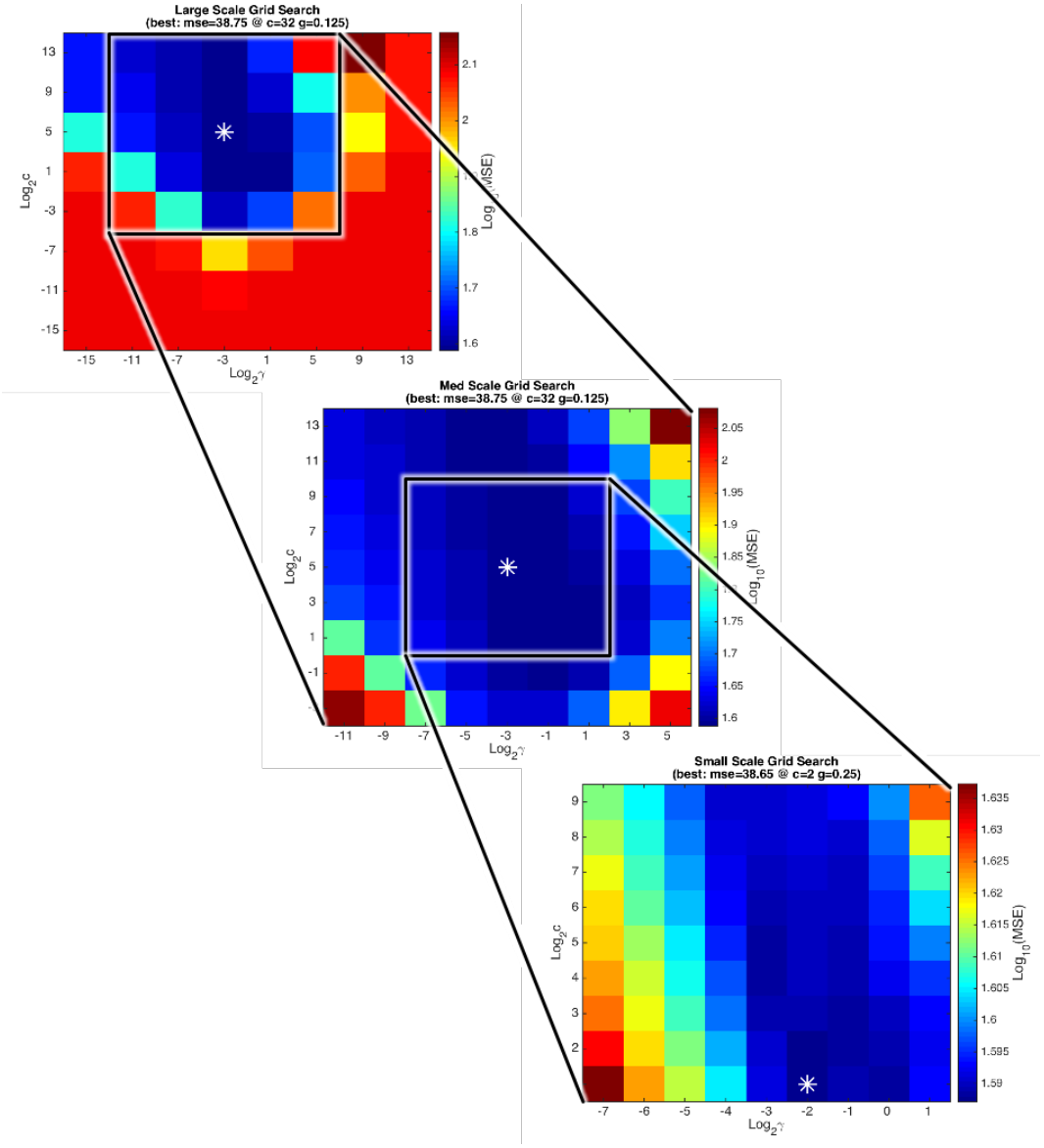


Figure 9.8: Grid search for ϵ-SVR parameters C and γ starting with the coarse grid, progressing to the medium grid, and ending with the fine grid. For each pair (C, γ) the value of $log_{10}(MSE)$ is plotted on an intensity scale where the maximum $log_{10}(MSE)$ is red, and the minimum $log_{10}(MSE)$ is blue. The intensity bar is rescaled to the range of values in the corresponding grid. The white asterisk * shows the minimum MSE at each scale.

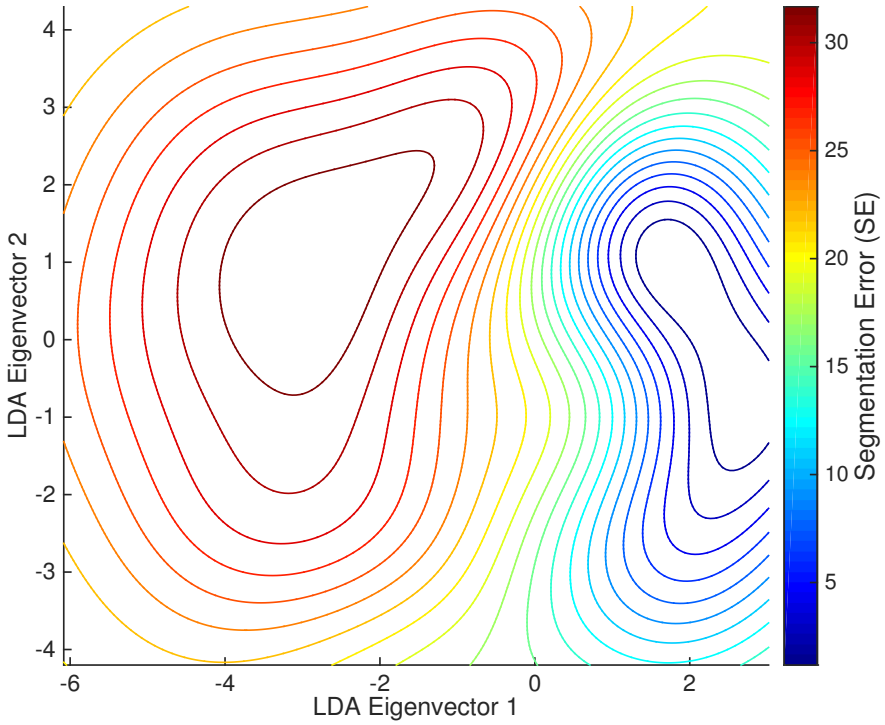


Figure 9.9: Final lip shape model (LSM) trained using ϵ-SVR. The model infers the lip segmentation error (SE) of an unknown region from the shape features of the region.

9.4 Validation Stage

The objective of the validation stage is to determine whether the output from the base algorithm represents an acceptable lip segmentation. The validation stage uses the shape model to infer the SE of the candidate lip region. If the inferred SE is below the SE boundary, then the candidate region is accepted and becomes the final segmentation. If the inferred SE is above the SE boundary, then the candidate region is rejected, and ATO proceeds to the optimisation stage.

The value of the SE boundary is subject to the trade-off between segmentation accuracy and processing time. If the candidate region fails validation then the threshold is optimised to improve segmentation accuracy, however the optimisation process is computationally expensive. Selecting a low SE boundary results in a high number of images that are optimised, while selecting a high SE boundary results in a low number of images that are optimised.

In order to select the SE boundary, it is useful to consider the number of images from the training dataset that would require optimisation. After segmentation by the base algorithm, the LSM is used to infer the SE from the candidate region. In Figure 9.10 the x-axis represents the SE inferred from the LSM, and ranges from $-0.5\,\%$ to $31\,\%$. The blue histogram (left y-axis) shows the probability distribution of the inferred SE. The peak of the probability distribution is at 0–$2.5\,\%$ SE which accounts for $20\,\%$ of the images.

The orange line plot (right y-axis) in Figure 9.10 shows the cumulative distribution function (CDF) starting from -31% and ending at -0.5%. The inferred SE is above $5\,\%$ for $64\,\%$ of images, and the inferred SE is above $10\,\%$ for $43\,\%$ of images. In other words, if the SE boundary is selected at $5\,\%$ then $64\,\%$ of images would be optimised, while if the SE boundary is selected at $10\,\%$ then $43\,\%$ of images would be optimised. The SE boundary is set at $10\,\%$, which is shown in Figure 9.11.

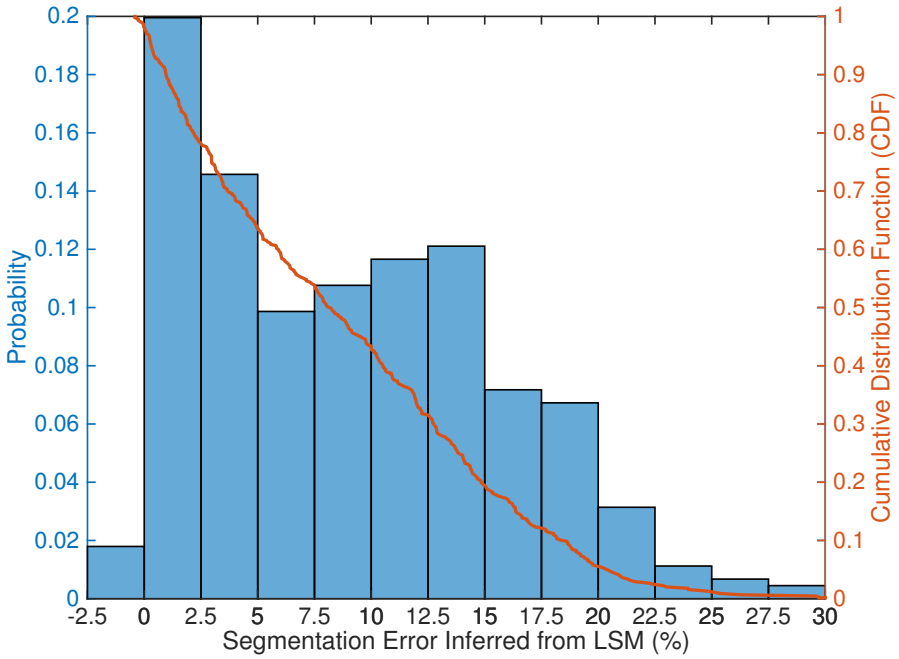


Figure 9.10: Distribution of segmentation error (SE) inferred from the lip segmentation model (LSM).

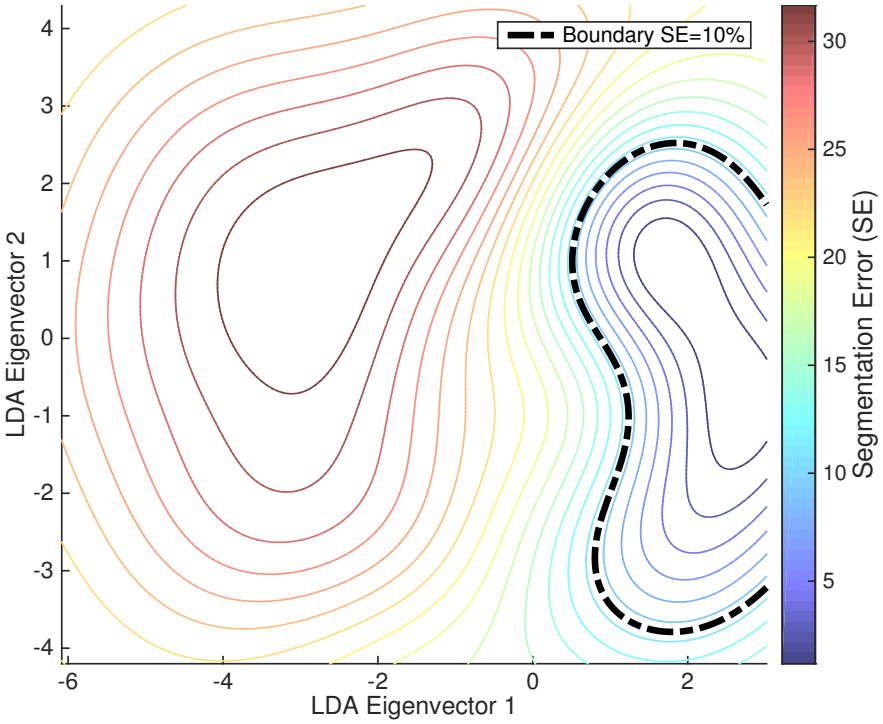
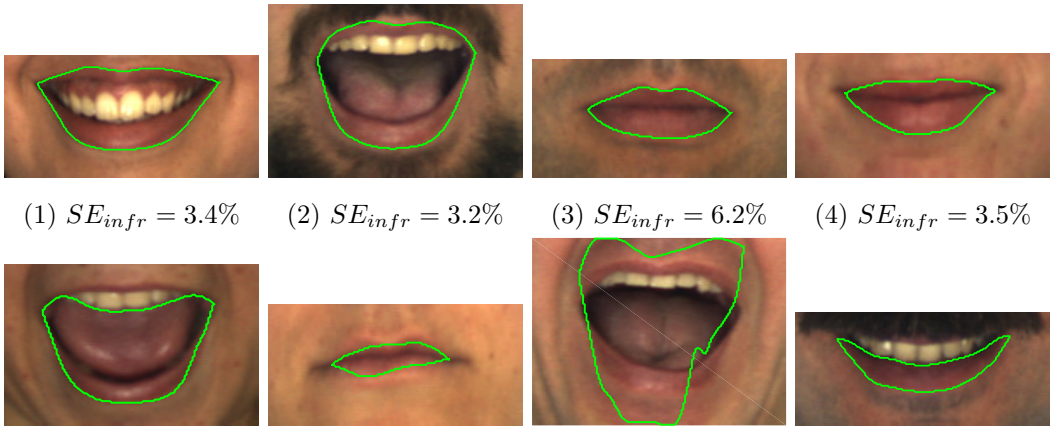


Figure 9.11: LSM with validation boundary set at SE = 10 %.

Figure 9.12 shows eight examples of the candidate lip region produced by the base algorithm. The segmentation in these images is produced using the default threshold, before any further optimisation of the threshold. Each candidate region is plotted on the lip shape model (LSM), where the image numbers correspond to the labels on the contour plot. The LSM is used to infer the SE of the candidate region, which is shown below each image (SE_{infr}). The candidate regions (1) – (4) pass validation, while the candidate regions (5) – (8) fail validation.



(1) $SE_{infr} = 3.4\%$ (2) $SE_{infr} = 3.2\%$ (3) $SE_{infr} = 6.2\%$ (4) $SE_{infr} = 3.5\%$

(5) $SE_{infr} = 12\%$ (6) $SE_{infr} = 23\%$ (7) $SE_{infr} = 30\%$ (8) $SE_{infr} = 30\%$

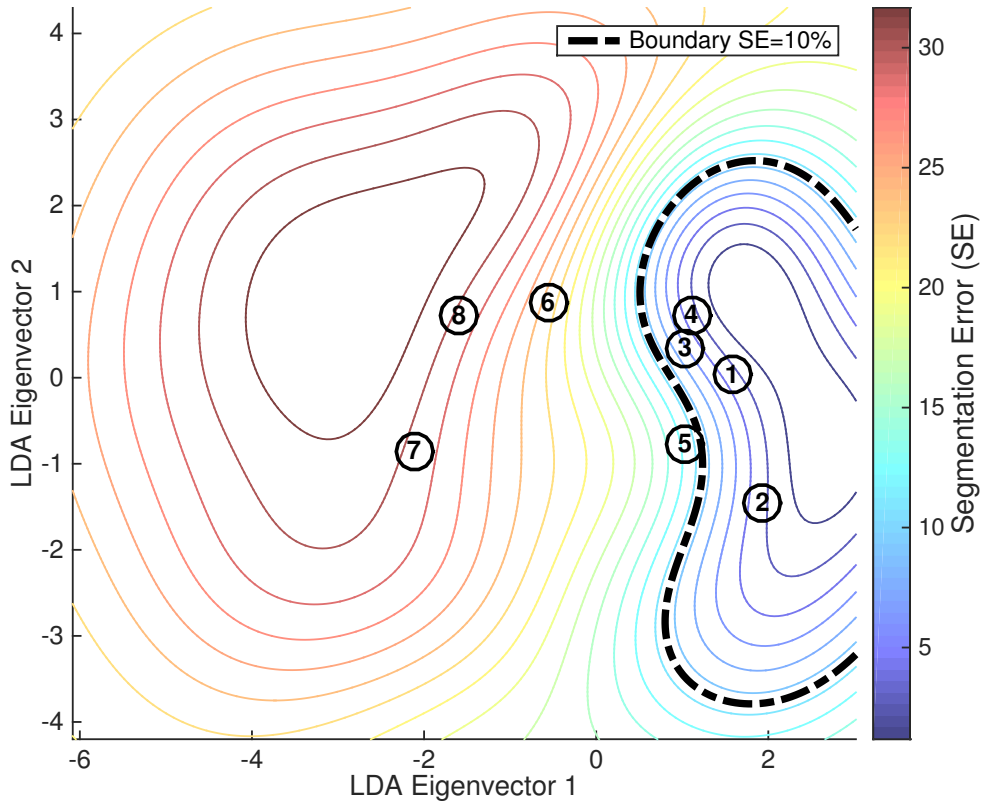


Figure 9.12: Examples of the validation stage. Images (1) – (8) show the examples of candidate lip region produced by the base algorithm. Each candidate region is plotted on the lip shape model (LSM), the image numbers correspond to the labels on the contour plot. The LSM is used to infer the SE of the candidate region, which is shown below each image (SE_{infr}). The candidate regions (1) – (4) pass validation, while the candidate regions (5) – (8) fail validation.

9.5 Optimisation Stage

The base algorithm uses the default threshold (Otsu) to produce the candidate lip region. If the candidate region is rejected by the validation stage, then ATO proceeds to the optimisation stage.

Figure 9.13 shows the objective function, which comprises thresholding, morphological processing, smoothing, feature extraction, and inferring the SE from the LSM. The input to the objective function is the threshold, and the output is the inferred SE (SE_{infr}). The optimisation stage aims to minimise the inferred segmentation error by changing the threshold.

The objective function is minimised using *Direct Search*, which polls a set of points, called a *mesh*, around the current point [172, 173]. If the poll is *successful*, i.e. *Direct Search* finds a point in the mesh that improves the objective function at the current point, then the new point becomes the current point at the next step of the algorithm. If the pole is *unsuccessful*, i.e. *Direct Search* does not find a point in the mesh that improves the objective function at the current point, then the current point remains unchanged and the mesh is refined. See [172, 173] for more details on the *Direct Search* algorithm.

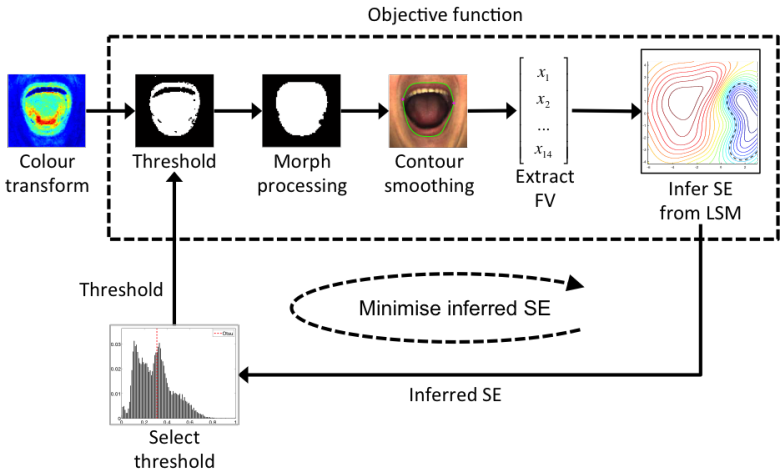


Figure 9.13: Block diagram of ATO optimisation stage, which aims to minimise the inferred segmentation error (SE_{infr}) by changing the threshold.

Figure 9.14 shows an example of the ATO algorithm, which aims to minimise the inferred SE (SE_{infr}) by changing the threshold value. In Figure 9.14(a), images (1) – (10) show the candidate lip region at each poll. SE_a is the actual segmentation error at the beginning and end of the optimisation. Figure 9.14(b) shows the *Inferred*

SE on the left y-axis (blue) and the *Threshold* on the right y-axis (orange) at each poll. Successful polls are indicated by the green markers. In Figure 9.14(c), the feature vector at each poll is plotted on the LSM.

The example image in Figure 9.14 presents a challenging case for the base algorithm to segment. The mouth is open which exposes the black oral cavity, as well as the white teeth. Furthermore, the image is from a male who has a light covering of black facial hair surrounding the mouth. The base algorithm using the default threshold produces the candidate lip segmentation shown in poll (1). The segmentation is very poor, and covers a region much larger than the lip contour. The high number of false positive lip pixels indicates that the threshold is too lenient, and needs to be increased. In Figure 9.14(c), the feature vector for poll (1) is plotted on the LSM, and corresponds to the labelled data point 1. The inferred SE at this point is $SE_{infr} = 26.1\%$, while the actual SE is $SE_a = 44.8\%$.

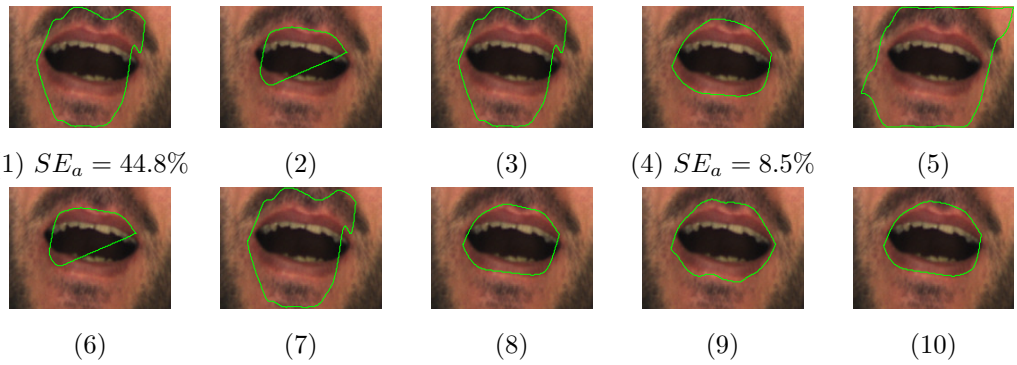
At poll (2), the threshold is increased from 0.31 to 0.56, which results in the segmentation excluding part of the mouth region. At this poll the threshold is too stringent, which results in a high number of false negative lip pixels. The inferred SE improves slightly compared to poll (1), and decreases from 26.1 to 19.8 %, corresponding to a *successful poll*.

At poll (3), the inferred SE increases from 19.8 % to 26.1 %, corresponding to an *unsuccessful poll*.

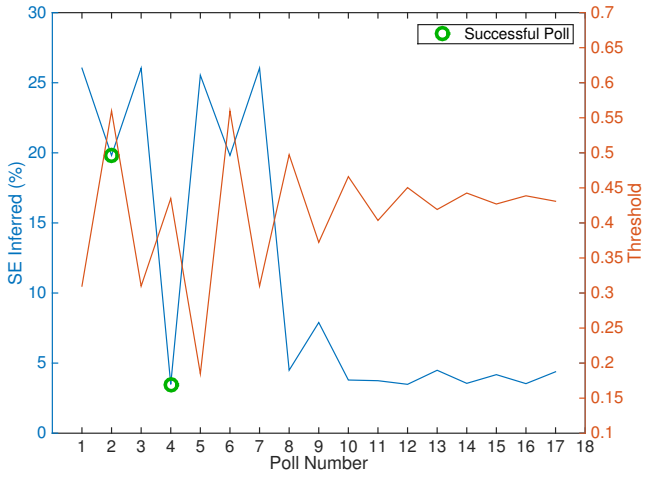
At poll (4), the threshold is set at 0.435, and the inferred SE drops dramatically to 3.49 %, corresponding to a *successful poll*. The segmentation in poll (4) adheres well to the lower lip contour, with only minor deviations from the upper lip contour.

From poll (5) onwards, *Direct Search* attempts to improve on the inferred SE of poll (4) by refining the mesh. The process of refining the mesh can be seen from the *Threshold* curve of Figure 9.14(b), which oscillates around the value of 0.435 from poll (4). However, *Direct Search* is unsuccessful in finding a threshold value that improves on the inferred SE of 3.49 % from poll (4), and therefore the threshold of 0.435 from poll (4) is selected as the final threshold value.

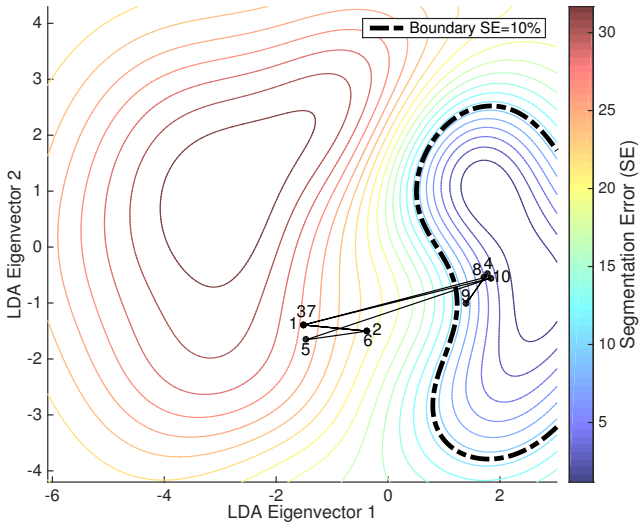
The actual SE improves from 44.8 % before ATO to 8.5 % after ATO, which is an absolute improvement of 36.3 pp, or a relative improvement of 81 %.



(a) Segmentation at each poll



(b) Inferred SE and Threshold at each poll



(c) Feature vector at each poll plotted on LSM

Figure 9.14: Graphical explanation of the ATO optimisation stage, which aims to minimise the Inferred SE by changing the threshold value. In (a), images (1) – (10) show the candidate lip region produced by the base algorithm at each poll. SE_a is the actual segmentation error at the beginning and end of the optimisation. Subfig (b) shows the Inferred SE (left y-axis) and Threshold (right y-axis) at each poll. In (c), the feature vector at each poll is plotted on the LSM.

9.6 Conclusion

Adaptive Threshold Optimisation (ATO) is a novel algorithm to select the histogram threshold based on feedback of shape information. ATO incorporates three stages: construction of a lip shape model (LSM), validation stage, and optimisation stage.

The LSM uses ϵ-support vector regression (ϵ-SVR) to model the relation between the shape of a lip region, and the corresponding segmentation error. Once the model has been trained, the shape features of an unknown region can then be used to infer the segmentation error.

The validation stage determines whether the output from the base algorithm constitutes an acceptable lip segmentation. The validation stage uses the shape model to infer the segmentation error (SE) of the candidate lip region. If the inferred SE is below the SE boundary, then the candidate region is accepted and becomes the final segmentation. If the inferred SE is above the SE boundary, then the candidate region is rejected, and ATO proceeds to the optimisation stage.

The optimisation stage aims to minimise the inferred segmentation error by iteratively changing the threshold.

Chapter 10

Testing and Analysis of ATO

10.1 Introduction

The ATO algorithm presented in Chapter 9 tackles the challenge of threshold selection in threshold-based segmentation methods. The ATO algorithm operates in two stages: the validation stage identifies poor segmentations produced using a default threshold, and the optimisation stage adjusts the threshold to improve the segmentation.

This chapter analyses the performance of the ATO algorithm by conducting three tests. In the first test, the validation stage is evaluated as a binary classifier which aims to identify poor segmentation. In the second test, the optimisation stage is evaluated by comparing the lip segmentation accuracy of the default threshold, versus the ATO threshold. In the final test, the performance of the overall algorithm is evaluated, which includes images that are optimised as well as images that are not optimised.

The 112 subjects in the AR Face Database are assigned 60 % to the training dataset and 40 % to the test dataset. The 544 images of the training dataset are used to build the lip shape model (LSM) in Chapter 9, and the 351 images from the test dataset are used to conduct the three tests in this chapter.

To note, the aggregated results from testing the base algorithm in Chapter 7 differ slightly from the results presented in this chapter. The differences stem from the images comprising the test dataset: in Chapter 7 the test dataset comprises all 895 images; whereas, in this chapter the test dataset comprises a subset of 351 images.

10.2 Test 1: Evaluation of the Validation Stage

The *candidate lip region* is segmentation output produced by the base algorithm, using the default threshold. The purpose of the validation stage is to determine whether the candidate lip region represents an acceptable lip segmentation. The validation stage uses the lip shape model (LSM) to infer the SE of the candidate lip region. If the inferred SE is below the predefined SE boundary, then the candidate region passes validation; however, if the inferred SE is above the SE boundary, then the candidate region fails validation. The SE boundary is set at 10 %.

The validation stage is essentially a binary classifier which infers whether SE is greater than 10 % (fail validation) or less than 10 % (pass validation) for a particular segmentation. As such, the validation stage is tested by comparing the inferred pass/fail label to the actual pass/fail label.

The confusion matrix for the validation stage is shown in Table 10.1. The validation stage correctly identifies 17.4 % of images that fail validation, and require further processing in the optimisation stage. The validation stage also correctly identifies 51.9 % of images that pass validation, in which case optimisation is not required.

Table 10.1: Confusion matrix for ATO validation stage where *pass* corresponds to $SE <= 10\%$ and *fail* corresponds to $SE > 10\%$.

	Infer fail	Infer pass
Actual fail	17.4%	2.3%
Actual pass	28.5%	51.9%

Since the objective of the validation stage is to correctly identify poor segmentation, the number 17.4 % represents true positive (TP) classification, and the number 51.9 % represents true negative (TN) classification. The calculations of precision and recall follow in (10.1) and (10.2) respectively. *Precision* is the fraction of retrieved instances that are relevant, i.e. of the images that are selected for optimisation, how many actually need optimisation. *Recall*, or sensitivity, is the fraction of relevant instances that are retrieved, i.e. of the images that need optimisation, how many are actually selected for optimisation.

$$precision = \frac{TP}{TP + FP} = \frac{17.4}{17.4 + 28.5} = 0.379 \tag{10.1}$$

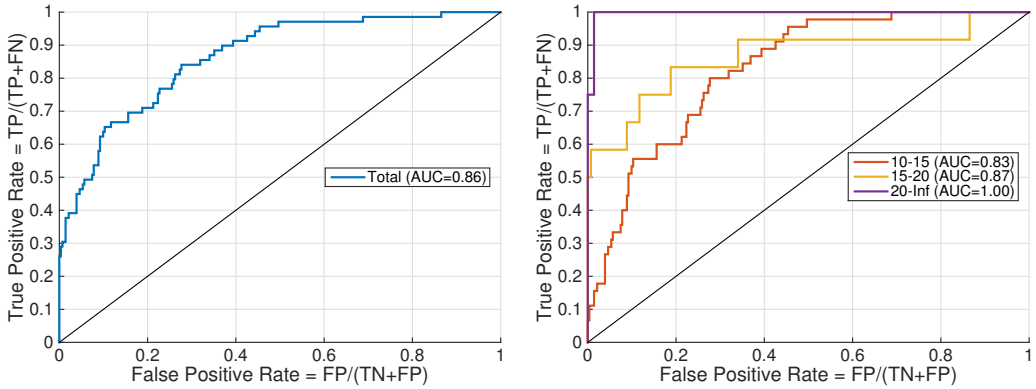
$$recall\ (sensitivity) = \frac{TP}{TP + FN} = \frac{17.4}{17.4 + 2.3} = 0.883 \tag{10.2}$$

The precision of the validation stage is 37.9 %, which implies that of the images selected for optimisation, only 37.9 % actually need optimisation. The low precision value does not necessarily affect the accuracy of the ATO algorithm, since optimising images unnecessarily may or may not further improve the segmentation accuracy. However, the low precision value does translate to a computational cost, whereby good segmentations still undergo the optimisation procedure.

The recall of the validation stage is 88.3 %, which implies that of the poor segmentations, 88.3 % are selected for optimisation. In contrast to the precision value, the recall value directly impacts the accuracy of the ATO algorithm. The recall of 88.3 % corresponds to the *miss rate* of $(1 - recall) = 11.7\%$. The miss rate refers to poor segmentations that do not continue to the optimisation stage, and hence remain poor segmentations.

Figure 10.1 presents the ROC curves of the validation stage, showing the accuracy in identifying poor segmentations. Subfigure (a) shows the total ROC curve, with the area under curve (AUC) of 0.86. At false positive rate (FPR) of 20 %, the true positive rate (TPR) is 70 %. In other words, the validation stage correctly identifies 70 % of poor segmentations, at a cost of incorrectly identifying 20 % of the good segmentations.

Subfigure (b) shows the ROC curve split by SE group. The AUC for moderate-poor segmentation in the 10 – 15 % group is 0.83, while the AUC for very poor segmentation in the 20 %+ group rounds up to 1.00. The validation stage is better at identifying very poor segmentation, as compared to moderate-poor segmentations.



(a) Total ROC **(b)** Split by SE groups

Figure 10.1: ROC curves showing the accuracy of the validation stage in identifying lip segmentation with SE greater than 10%. (a) shows the total ROC curve, and (b) shows the ROC curve split into three SE groups: 10 – 15%, 15 – 20%, and 20%+. The area under curve (AUC) is shown for each ROC curve.

Finally, Table 10.2 compares the mean SE and OL for images that pass and fail validation. The mean SE for images that pass validation is 4.89 %, while the mean SE for images that fail validation is 10.9 %, more than two times higher.

Table 10.2: Mean segmentation error (SE) and percentage overlap (OL) for segmentations that pass and fail validation.

	Number	SE	OL
Pass	190	4.89	95.0
Fail	161	10.9	88.5
Total	351	7.65	92.0

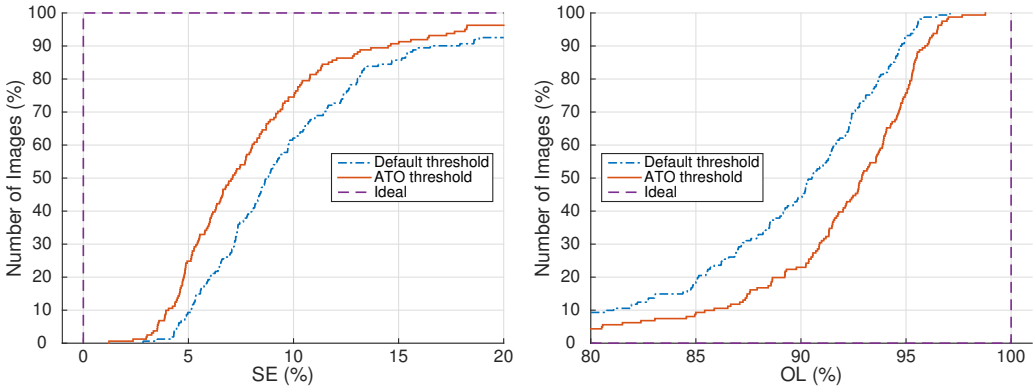
10.3 Test 2: Evaluation of the Optimisation Stage

Of the 351 images in the test dataset, 161 images fail validation and continue to the optimisation stage. The optimisation stage computes the ATO threshold, and uses this threshold to segment the image. Test 2 evaluates the performance of the optimisation stage on the 161 images that fail validation.

Figure 10.2 presents the cumulative distribution function (CDF) comparing the default threshold versus the ATO threshold, and Table 10.3 lists specific CDF values. Subfigure 10.2(a) shows the improvement in SE between the default threshold and the

ATO threshold. The percentage of segmentations below 5 % SE improves from 11.3 % to 27.4 %, corresponding to an increase of 2.4 times. The percentage of segmentations below 10 % SE improves by 11.9 percentage points from 61.9 % to 73.8 %.

The CDF for OL in Subfigure (b) shows similar results. The percentage of segmentations above 95 % OL improves by a factor of 2.7, from 8.93 % to 24.4 %. The percentage of segmentations above 90 % improves by 19 percentage points from 56.0 % to 75.0 %.



(a) Segmentation error (SE) CDF **(b)** Percentage overlap (OL) CDF

Figure 10.2: Cumulative distribution function (CDF) comparing the default threshold versus the ATO threshold for images that are selected for optimisation (i.e. fail validation). (a) shows the CDF for segmentation error (SE) and (b) shows the CDF for percentage overlap (OL).

Table 10.3: Specific SE and OL values from the CDF for images that are selected for optimisation.

(a) % of segmentations below SE value

	% of segmentations	
SE	Default	ATO
< 5 %	11.3	27.4
< 10 %	61.9	73.8
< 15 %	86.3	90.5
< 20 %	92.9	96.4
< 25 %	94.6	97.6

(b) % of segmentations above OL value

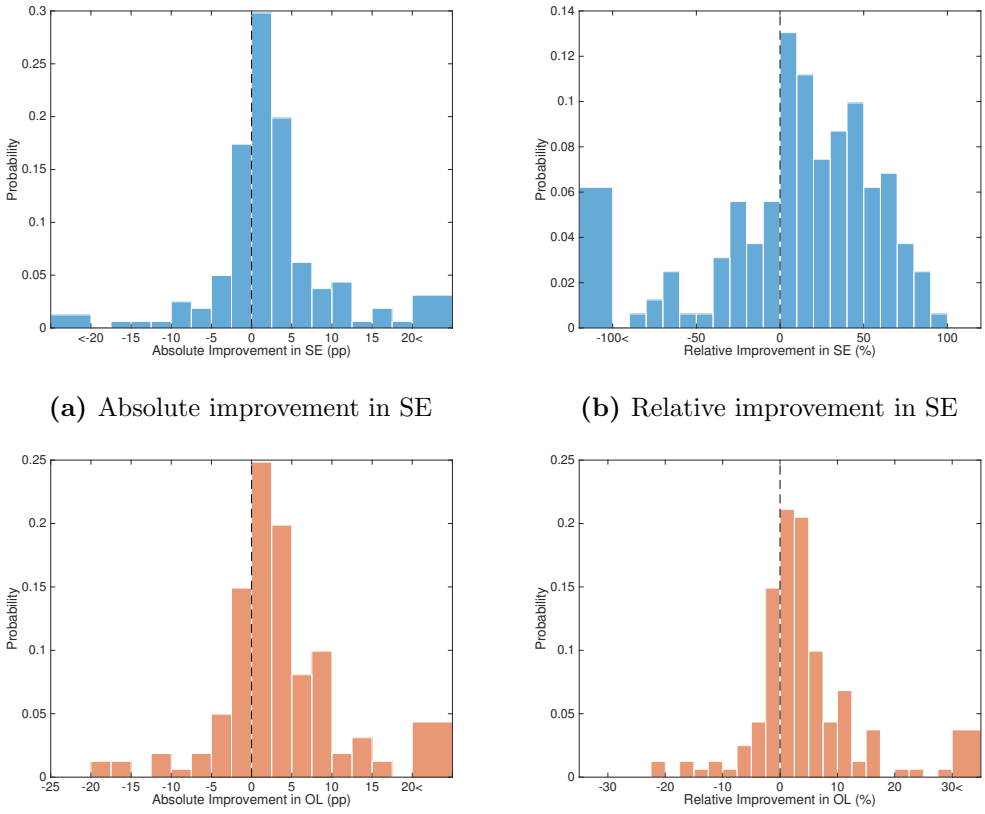
	% of segmentations	
OL	Default	ATO
75 % <	94.0	98.2
80 % <	91.1	95.8
85 % <	82.7	90.5
90 % <	56.0	75.0
95 % <	8.93	24.4

Table 10.4 shows the average improvement for the 161 images that are optimised. The absolute improvement in SE is 2.52 pp, corresponding to a relative improvement of 23.1 %. The absolute improvement in OL is 3.12 pp.

Table 10.4: Improvement in segmentation error (SE) and percentage overlap (OL) for images that are selected for optimisation (i.e. fail validation).

	Default (%)	ATO (%)	Improvement absolute (pp)	Improvement relative (%)
SE	10.9	8.39	2.52	23.1
OL	88.5	91.6	3.12	3.52

The histograms in Figure 10.3 present a breakdown of the absolute and relative improvement: (a) and (b) show the improvement in SE, (c) and (d) show the improvement in OL. The histograms show that ATO does not improve the segmentation accuracy for all images, and in some cases ATO reduces the accuracy. However, considering the relative SE improvement in (b) for example, it is clear that the positive improvement significantly outweighs the negative.



(a) Absolute improvement in SE

(b) Relative improvement in SE

(c) Absolute improvement in OL

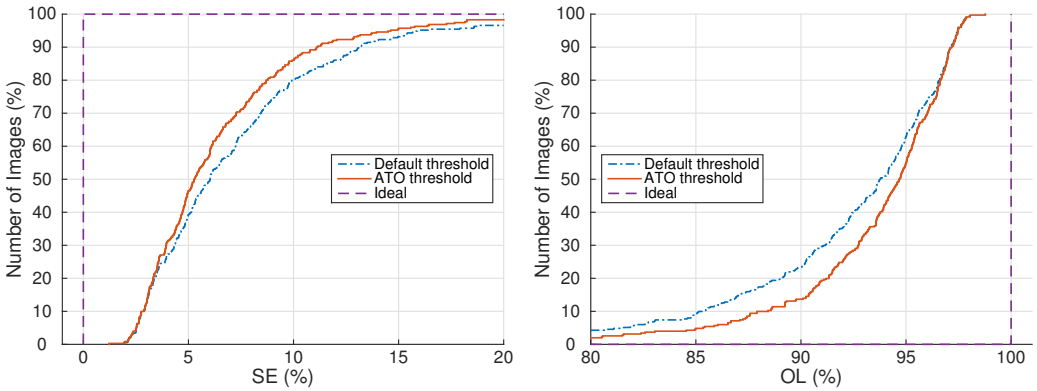
(d) Relative improvement in OL

Figure 10.3: Histograms showing the improvement in segmentation error (SE) and percentage overlap (OL) for images that are selected for optimisation (i.e. fail validation).

10.4 Test 3: Evaluation of the Overall Algorithm

The overall algorithm is evaluated by considering the segmentation accuracy across all 351 test images, including images that are selected for optimisation by the validation stage, as well as images that are not optimised.

Figure 10.4 presents the CDF for the overall algorithm, and Table 10.5 lists specific CDF values. After ATO, 46.4 % of segmentations achieve SE below 5 %, and 86.0 % of segmentations obtain SE below 10 %. Considering OL before and after ATO, the percentage of segmentations above 95 % improves by 21.6 % relative from 37.0 % to 45.0 %.



(a) Segmentation error (SE) CDF **(b)** Percentage overlap (OL) CDF

Figure 10.4: Cumulative distribution function (CDF) showing the overall performance of the ATO algorithm, including images that pass and fail validation.

Table 10.5: Specific SE and OL values from the overall CDFs.

(a) % of segmentations below SE value

	% of segmentations	
SE	Default	ATO
< 5 %	39.3	46.4
< 10 %	80.3	86.0
< 15 %	93.2	95.4
< 20 %	96.6	98.3
< 25 %	97.4	99.1

(b) % of segmentations above OL value

	% of segmentations	
OL	Default	ATO
75 % <	97.2	99.1
80 % <	95.7	98
85 % <	91.2	95.2
90 % <	76.6	86.3
95 % <	37.0	45.0

Table 10.6 shows the mean improvement for the overall ATO algorithm on the 351 test images. The overall SE after ATO decreases from 7.65 % to 6.50 %, corresponding to an absolute improvement of 1.16 pp and relative improvement of 15.1 %. The mean OL improved from 92.0 % to 93.4 %.

Figure 10.5 shows several examples of the lip segmentation results before and after optimising the threshold with ATO. The threshold selected by ATO significantly improves the segmentation in several scenarios in which the base algorithm with the default threshold does not perform well, including: facial hair, low contrast between the lips and skin, and inconsistent illumination.

Table 10.6: Improvement in segmentation error (SE) and percentage overlap (OL) for overall ATO algorithm, including images that pass and fail validation.

	Default (%)	ATO (%)	Improvement absolute (pp)	Improvement relative (%)
SE	7.65	6.50	1.16	15.1
OL	92.0	93.4	1.43	1.55

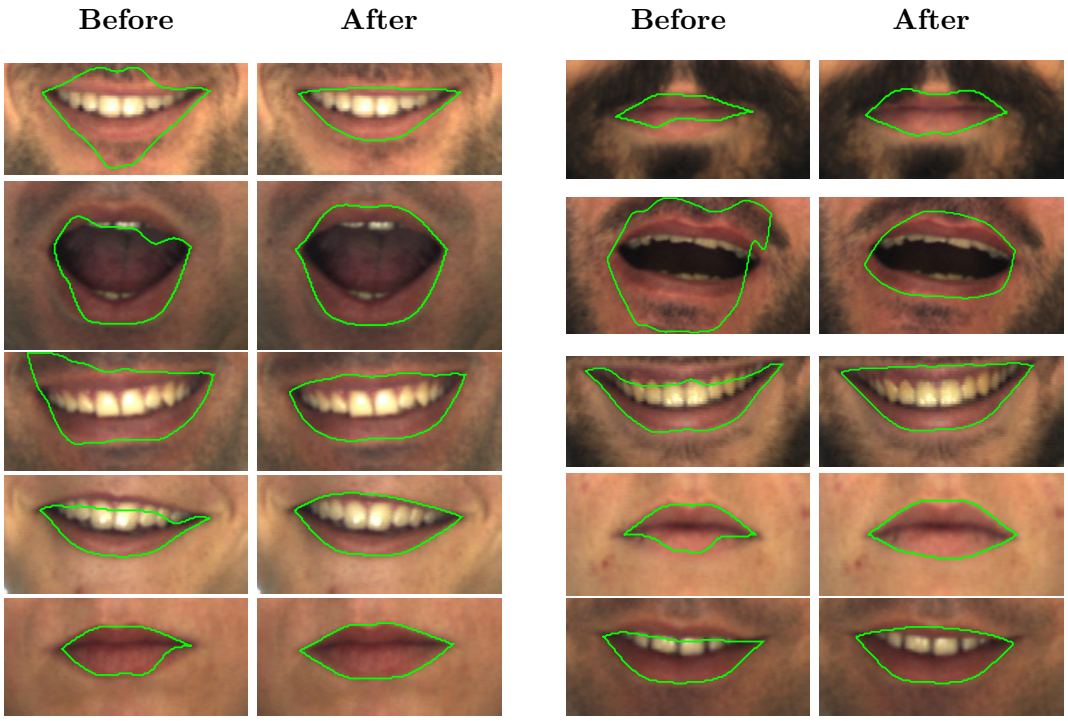


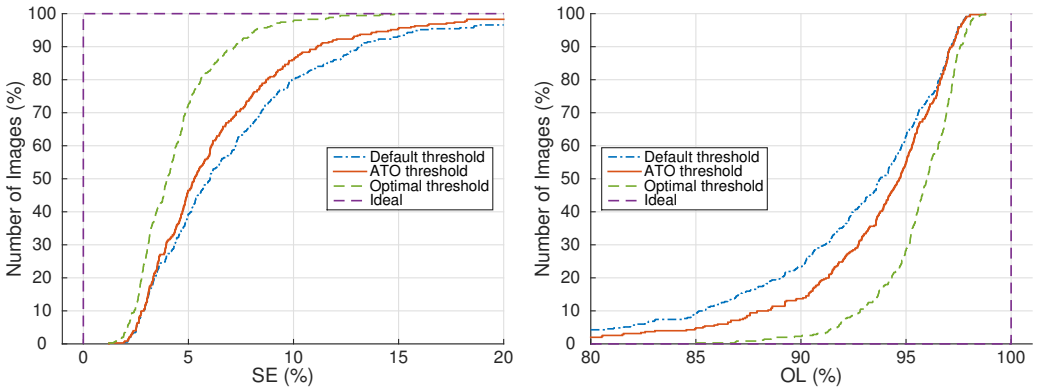
Figure 10.5: Examples of the segmentation produced by the base algorithm before and after optimising the threshold with ATO.

10.5 Scope for Further Improvement

The new ATO algorithm is a technique to adjust the histogram threshold to improve the resulting segmentation accuracy. While the ATO results presented in this chapter show significant improvement in the segmentation accuracy, the approach of optimising the threshold still has scope for further improvement.

Figure 10.6 shows the CDF comparing the ATO algorithm, the default threshold, and the optimal threshold. As a reminder, the optimal threshold described in Chapter 8 is found by scanning across all threshold values, and selecting the best threshold value for a particular image. Comparing the *optimal threshold* to the *ATO threshold* in (a) and (b), it is clear that there is certainly scope for further improvement. Obviously,

the challenge remains developing a robust technique to compute the correct threshold.



(a) Segmentation error (SE) CDF **(b)** Percentage overlap (OL) CDF

Figure 10.6: Cumulative distribution function (CDF) comparing the ATO algorithm, default threshold, and optimal threshold.

10.6 Comparison to Existing Methods

The most recent work in the field of lip segmentation was published by Cheung et al. [99] in November 2015, six months prior to submission of this thesis. In this work, Cheung et al. present a fuzzy clustering-based approach to lip segmentation, which does not depend on prior knowledge of the number of segments. The novelty of the approach arises from the technique to merge coincident cluster centroids, which allows the algorithm to be robust against the preassigned number of clusters.

Cheung et al. [99] compare the performance of their method with four existing methods: Liew2003 [160], Leung2004 [174], Wang2007 [175], and classical fuzzy c-means clustering (FCM). Liew2003 and Leung2004 propose fuzzy clustering taking into account both colour and spatial information. Wang2007 improves on this approach by adding automatic selection of the number of segments. An important difference to note is that aside from parameter selection, these techniques are essentially unsupervised, whereas ATO is a supervised technique.

For a direct comparison between ATO and the methods analysed in Cheung et al. [99], it is necessary to evaluate all algorithms according to the same experimental method. Table 10.7 shows a comparison between the experimental method used in the current research against that of Cheung et al. [99]. The source database and the metrics are the same, but there are differences in the images selected, clipping, and ground truth. It is not possible to reproduce the Cheung2015 test images in a

consistent manner, since the article does not provide a list of selected images, and the ground truth has not been made publicly available. Several attempts to contact the authors to request the test images were unsuccessful, so it is necessary to resort to a somewhat indirect comparison.

Table 10.7: Comparison of experimental methods between the *Current Research* and *Cheung2015*.

	Current Research	Cheung2015
Database	AR Face Database	AR Face Database
Selection	Randomly partition subjects 60 % for training, 40 % for testing. Test dataset comprises 351 images from 44 subjects.	Randomly select 50 images
Clipping	Size of image cropped according to size of mouth	128×128 pixels
Ground truth	Manual markings from Ding & Martinez [101] are interpolated using cubic smoothing splines	Manually segmented by authors
Metrics	OL and SE	OL and SE

To assess the validity of the comparison, it is necessary to consider the potential implications of the differences in experimental methods. First, regarding the selection of the test images, even though the test images differ between Cheung2015 and the current research, in both cases the selection is random, so no bias is introduced. Second, the preparation of the ground truth in both experiments involves manual markings by human judges. Ding & Martinez [101] report that the within-judge variability for facial key points was 3.8 pixels or 1.2 %, so variability in the ground truth is not expected to have a significant impact the comparison.

The final difference to consider is the size of the clipped images. Figure 10.7 shows the width and height of images used for testing in the current research versus Cheung2015. All 50 images in Cheung2015 are 128×128 in size (16 384 pixels). Observing the *Area Boundary*, 20 % of images in the current research are larger than Cheung2015 images, while 80 % are smaller. While it is difficult to quantify the implication of image size on the segmentation accuracy, it is likely that the larger images in Cheung2015 would be more challenging to segment. However, the author argues that current techniques, such as the Viola-Jones detector [89], adapt the region dimensions to

the size of the mouth, and can easily detect regions smaller than 128×128. Some cursory experimentation with the Viola-Jones detector [89] validate this assertion.

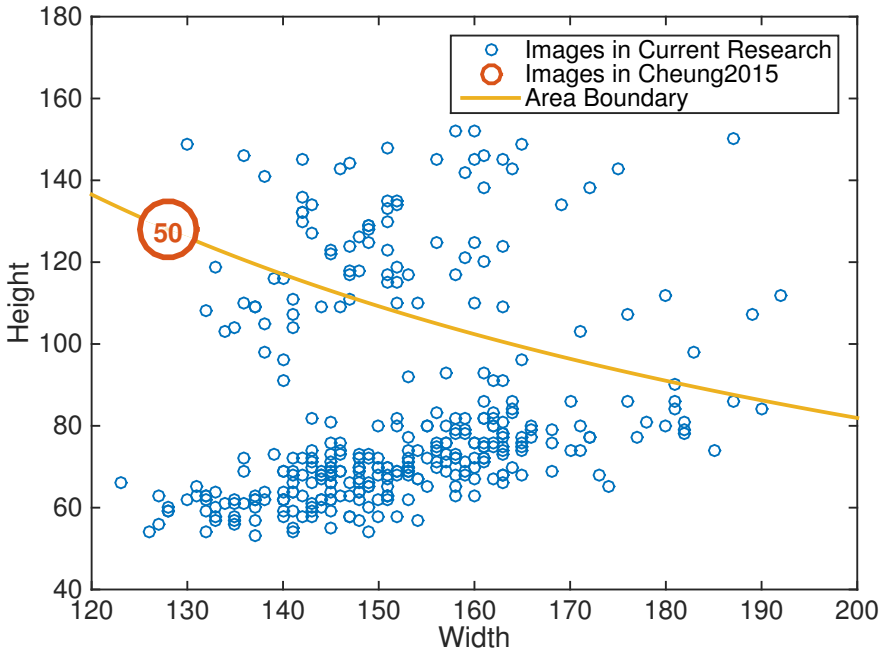


Figure 10.7: Width and height of images used for testing in the Current Research versus Cheung2015 [99]. The test dataset in Cheung2015 comprises 50 images from the AR Face Database, all clipped to 128×128 (16 384 pixels). Images above the *Area Boundary* are larger in number of pixels than the Cheung2015 images, while those below the *Area Boundary* are smaller.

To summarise the discussion on the validity of the comparison: image selection and the ground truth are not likely to bias the comparison, however the size of images may have an effect. Nevertheless, the experimental methods appear to be sufficiently compatible to draw informative conclusions.

Table 10.8 presents the comparison of ATO to the lip segmentation methods analysed in Cheung et al. [99], acknowledging the limitations discussed above. With respect to segmentation error (SE), ATO outperforms the nearest rival, Cheung2015, by 1.9 pp absolute or 22.6 % relative. Aside from Cheung2015, the other techniques are not competitive with SE ranging from 13.25 % to 42.82 %.

Cheung2015 and Wang2007 place a large emphasis on automatically determining the number of segments; by contrast, ATO assumes that the image comprises only two segments. The superior performance of ATO suggests that it is more important to correctly select the threshold value, than the number of segments. If the boundary

between the lips and skin is clearly differentiated by the threshold, then the subsequent post-processing can consolidate the lip and skin segments.

Table 10.8: Comparison of ATO to lip segmentation methods analysed in Cheung et al. [99]. The limitations of the comparison arising from differences in the experimental methods are acknowledged in Table 10.7 and the accompanying discussion.

	OL	SE
FCM	65.25%	42.82%
Liew2003 [160]	83.50%	17.31%
Leung2004 [174]	89.92%	13.25%
Wang2007 [175]	87.35%	21.13%
Cheung2015 [99]	90.80%	8.40%
Ato2016	**93.45%**	**6.50%**

10.7 Conclusion

The ATO algorithm is tested by evaluating the validation stage and the optimisation stage individually in Test 1 and Test 2 respectively. Thereafter, the overall ATO algorithm is evaluated in Test 3.

Test 1 evaluates the validation stage as a binary classifier with the objective of identifying poor segmentation. The precision of the validation stage is 37.9 %, while the recall of the validation stage is 88.3 %. The low precision value does necessarily affect the accuracy of the ATO algorithm, since optimising images unnecessarily may or may not further improve the segmentation accuracy. The high recall value is essential to the segmentation accuracy, as poor segmentations missed by the validation stage will not proceed to the optimisation stage, and will remain poor.

Test 2 evaluates the optimisation stage by comparing the lip segmentation accuracy of the default threshold versus the ATO threshold. The percentage of segmentations above 95 % OL improves by a factor of 2.7 from 8.93 % to 24.4 %. The mean SE for the optimised images improves by 2.52 pp absolute, or 23.1 % relative.

Test 3 evaluates the performance of the overall algorithm by considering the segmentation accuracy across all test images, including images that are selected for optimisation by the validation stage, as well as images that are not optimised. After ATO, 46.4 % of segmentations achieve SE below 5 %, and 86.0 % of segmentations

obtain SE below 10 %. The overall SE after ATO decreases from 7.65 % to 6.50 %, corresponding to an absolute improvement of 1.16 pp absolute or relative improvement of 15.1 %.

Comparing the optimal threshold from Chapter 8 to the ATO threshold shows that the approach of optimising the threshold may still offer potential for further improvement.

Finally, although the comparison is somewhat indirect, ATO seems to outperform existing lip segmentation methods.

Chapter 11

Conclusion

11.1 Overview

A *laryngectomy* is the partial or complete surgical removal of the larynx (voice box), which leaves the patient unable to speak. Restoration of some degree of voice is crucial to the laryngectomy patient's morale, self esteem and reintegration into society. The conventional methods to restore verbal communication (electrolarynx, tracheo-oesophageal speech, and oesophageal speech) have achieved some success in restoring speech; however, there are various limitations and drawbacks associated with each technique.

Modern advances in electronic miniaturisation and portable computing have paved the way for a computer-based solution. A silent speech interface (SSI) is a system enabling speech communication in the absence of an intelligible acoustic signal. SSIs improve on some of the limitations of conventional techniques, while giving rise to a new set of challenges: sensor positioning and robustness, speaker independence, vocabulary size, cost, and practicality for routine use.

The revolution in mobile computing has provided the platform for optical SSI devices – a standard smartphone possesses all the hardware requirements including a camera, processing power, and audio/video output. The technique of retrieving speech content from visual clues, such as the movement of the lips, tongue and teeth, is known as automatic lip-reading (ALR). The two main challenges of ALR are lip segmentation and recognition. This thesis addresses the image processing challenge of lip segmentation, and focuses on three specific contributions:

1. Comparison of 33 colour transforms used in lip segmentation algorithms
2. Development of new threshold-based lip segmentation algorithm
3. Development of a novel threshold selection technique called Adaptive Threshold Optimisation (ATO)

11.2 Summary of Contributions

11.2.1 Comparison of Colour Transforms

The first stage in lip segmentation typically involves applying a suitable colour transform to enhance the contrast between the lips and the surrounding skin; however, no consensus exists among researchers as to the best colour transform for this task. Chapter 5 presents the most comprehensive study to date by evaluating and comparing 33 different colour transforms: 21 channels from 7 colour space models (RGB, HSV, YCbCr, YIQ, CIEXYZ, CIELUV, CIELAB); and 12 additional transforms, 8 of which are designed specifically for lip segmentation. The contrast between the lips and the skin is used to obtain the outer-lip contour, while the contrast between the lips and the oral cavity is used to obtain the inner-lip contour. As such, this thesis identifies the transforms most appropriate for lip-skin segmentation and for lip-oral cavity segmentation.

The 33 colour transforms are compared based on two metrics: histogram intersection which indicates the maximum segmentation accuracy, and Otsu's discriminant which measures the separability attainable using a single threshold.

Results for lip-skin segmentation validate the experimental approach, as 11 of the top 12 transforms have been used in lip segmentation algorithms in the literature. The necessity of selecting the correct transform is demonstrated by an increase in segmentation accuracy of up to three times. Hue-based transforms (including pseudo-hue and hue domain filtering) perform the best for lip-skin segmentation, with the hue component of HSV achieving the greatest accuracy of 93.85 %. The a^* component of CIELAB performs the best for lip-oral cavity segmentation, while pseudo-hue and the LUX transform perform reasonably well for both lip-skin segmentation and lip-oral cavity segmentation.

This work has been published in *Signal, Image and Video Processing* [14].

11.2.2 Threshold-based Lip Segmentation Algorithm

Lip segmentation is a fundamental system component in a range of applications including: automatic lip-reading (ALR), virtual face animation, biometric speaker identification, and emotion recognition. Lip segmentation presents a challenging image processing problem arising from the variability associated with the speaker profile, movement of the lips, and environmental conditions.

The second contribution of this thesis is the development of a new threshold-based lip segmentation algorithm in Chapter 6. The lip segmentation algorithm begins by filtering the pre-cropped mouth region, and applying luminance correction based on the Michaelis-Menten law. The two best colour transforms from the comparison in Chapter 5 (YIQ-Q and MI3) are combined to enhance the contrast between the lips and the skin. Otsu's method is used to select the histogram threshold, which is followed by nine morphological operations to consolidate the lip region and to remove artefacts. Finally, the lip segmentation algorithm uses cubic splines to smooth the lip contour.

In Chapter 7, the lip segmentation algorithm was tested on 895 mouth region images from the AR Face Database, using percentage overlap (OL) and segmentation error (SE) to quantify performance. The mean OL was 92.23 % and the mean SE was 7.39 %. Of the 895 images in the dataset, 78 % obtained OL above 90 %, which appears to be acceptable for lip segmentation applications. The images in this category include variation in gender, age, skin colour, make-up, and facial hair. Only 1.90 % of images obtained OL below 70 %.

The following scenarios presented a challenge to the algorithm:

- facial hair obscuring the lips
- low contrast between the lips and skin
- thin lips
- light covering of facial hair causing poor threshold selection
- reflections caused by moisture on the skin

The ATO algorithm, which forms the third contribution, is designed to improve the segmentation accuracy in these challenging scenarios.

11.2.3 Adaptive Threshold Optimisation (ATO)

Threshold-based segmentation methods provide a simple and efficient way to implement lip segmentation. However, automatic computation of robust thresholds presents a significant challenge. The segmentation algorithm developed in the second contribution represents a typical threshold-based lip segmentation algorithm, which uses Otsu's method to compute the threshold value. This algorithm is referred to as the *base algorithm*, and forms the platform to develop and test a novel threshold selection technique called ATO.

In Chapter 8, the significance of the threshold value in the base algorithm is interrogated by performing a linear search to find the optimal threshold. The *optimal threshold* is the value that results in the best segmentation accuracy. This value is compared to the *default threshold* (Otsu) to quantify the improvement in segmentation accuracy that can be obtained by adjusting the threshold value. The analysis reports that the accuracy limit of the base algorithm with the optimal threshold is SE 4.75 %, which represents a 35.8 % relative improvement over the default threshold. It is clear that the segmentation produced by the base algorithm can be improved significantly by implementing a better threshold selection technique.

Chapter 9 presents *Adaptive Threshold Optimisation (ATO)*, which is a novel technique to select the histogram threshold based on feedback of shape information. ATO incorporates three stages: construction of a lip shape model (LSM), the validation stage, and the optimisation stage. The LSM uses ϵ-support vector regression (ϵ-SVR) to model the relation between the shape of a lip region, and the corresponding segmentation error. Once the model has been trained, the shape features of an unknown region can then be used to infer the segmentation error (SE). The *validation stage* uses the inferred SE to determine whether the output from the base algorithm constitutes an acceptable lip segmentation. If the segmentation is rejected, then ATO proceeds to the *optimisation stage*, which minimises the inferred SE by iteratively adjusting the threshold.

The ATO algorithm is tested in Chapter 10 by individually evaluating the validation stage in Test 1, and the optimisation stage in Test 2. Thereafter, the overall ATO algorithm is evaluated in Test 3.

Test 1 evaluates the validation stage as a binary classifier with the objective of identifying poor segmentation. The precision of the validation stage is 37.9 %, while the recall of the validation stage is 88.3 %. The low precision value does necessarily

affect the accuracy of the ATO algorithm, since optimising images unnecessarily may or may not further improve the segmentation accuracy. The high recall value is essential to the segmentation accuracy, as poor segmentations missed by the validation stage will not proceed to the optimisation stage, and will remain poor.

Test 2 evaluates the optimisation stage by comparing the lip segmentation accuracy of the default threshold versus the ATO threshold. The percentage of segmentations above 95 % OL improves by a factor of 2.7 from 8.93 % to 24.4 %. The mean SE for the optimised images improves by 2.52 pp absolute, or 23.1 % relative.

Test 3 evaluates the performance of the overall algorithm by considering the segmentation accuracy across all test images, including images that are selected for optimisation, as well as images that are not optimised. After ATO, 46.4 % of segmentations achieve SE below 5 %, and 86.0 % of segmentations obtain SE below 10 %. The overall SE after ATO decreases from 7.65 % to 6.50 %, corresponding to an absolute improvement of 1.16 pp or relative improvement of 15.1 %.

To note, the results from testing the base algorithm in the second contribution (Chapter 7) differ slightly from the results presented in the third contribution (Chapter 10). The differences stem from the images comprising the test dataset: in Chapter 7 the test dataset comprises all 895 images; whereas, in Chapter 10 the test dataset comprises a subset of 351 images.

Comparing the optimal threshold from Chapter 8 to the ATO threshold shows that the approach of optimising the threshold may still offer potential for further improvement.

Finally, although the comparison is somewhat indirect, ATO seems to outperform existing lip segmentation methods.

This work has been published in the journal of *Signal, Image and Video Processing* [15], and *Proceedings of the 2015 Conference on Facial Analysis and Animation (FAA2015)* [16].

11.3 Recommendations for Future Research

The major area for future research concerns adaptive threshold optimisation (ATO), the novel threshold selection technique introduced in this thesis. ATO uses feedback of shape information to select the threshold in the underlying segmentation algorithm.

11.3.1 Generalise to Other Applications

In this thesis, ATO was used to augment the lip segmentation algorithm; however, ATO is by no means restricted to lip segmentation. In fact, ATO may be used as a general image processing technique to improve threshold selection when prior shape information is available. ATO can be applied to different segmentation challenges by simply building shape models appropriate for the new task. In future research, ATO should be tested on different segmentation problems to determine if similar benefits can be achieved.

11.3.2 Multi-Object Segmentation

ATO can be extended to multiple object segmentation by training multiclass shape models. In this work, the ϵ-SVR model is trained with lip segmentation samples. The model can be expanded by training with samples of another object, the eyes for example. The threshold is still adjusted with the objective of minimising the inferred SE, but in the multi-object case the segmentation result iteratively moves towards either the eyes or the mouth.

11.3.3 From Static Segmentation to Dynamic Tracking

In this thesis, ATO was developed around segmentation in static images. However ATO lends itself to easily be adapted to dynamic tracking applications by simply using the threshold selected in the previous frame to initialise the threshold in the current frame.

11.3.4 Further Improving the Accuracy

Regarding the general approach of adjusting the threshold to improve the segmentation accuracy, Chapter 10 compares ATO to the optimal threshold (from Chapter 8) to show that there is scope for further improvement.

In supervised learning tasks, the performance of the system depends on three components: the data quantity and quality, the features, and the learning technique.

With reference to ATO, the ϵ-SVR model is trained with 544 images from the dataset. Increasing the size of training dataset will most likely lead to an improvement in

the performance of ATO. Furthermore, in subject-specific applications such as voice restoration for laryngectomy patients, the performance of ATO could be improved by training subject-dependant models.

Considering the features used in the ATO algorithm, the current feature set comprises region-based geometric features, such as area and height. The performance of ATO may be improved by adding point distribution features [71, 94], or appearance features which include both shape and colour [176, 177].

11.3.5 Improving the Evaluation Metrics

It is necessary to raise one final concern regarding the metrics used to quantify the segmentation accuracy of the algorithm (see Section 7.3). In this thesis, percentage overlap (OL) and segmentation error (SE) are used to quantify segmentation accuracy, in line with the lip segmentation literature [91, 121, 160–165]. In several images from the neutral and anger expressions, the segmented contour appears to follow the true lip contour with reasonable accuracy, however the segmentation performs poorly according to OL and SE (see Section 7.4.1). This result is explained by considering the sensitivity of OL and SE to the size of the lip area. When the accuracy of a lip segmentation result is assessed with the human eye, the focus is on how well the segmented contour adheres to the ground truth contour. OL and SE are region-based metrics, and despite wide adoption in the lip segmentation literature, it may be preferable to construct a metric that measures the adherence/deviation from the lip contour.

References

[1] Aristotle (350 B.C.E.), *Politics*, Digireads.com Publishing, 2005 edn., translated by Benjamin Jowett.

[2] Matyja, A., Matyja, G., Tarnowska, C., Rogowska, D. & Horodnicki, J. (2005), 'Anxiety and depression in patients with laryngeal and hypopharyngeal cancer (preliminary report)].', *Polski Merkuriusz Lekarski: Organ Polskiego Towarzystwa Lekarskiego*, vol. 19, no. 111, p. 390.

[3] Cady, J. (2002), 'Laryngectomy: Beyond loss of voice - caring for the patient as a whole', *Clinical Journal of Oncology Nursing*, vol. 6, no. 6, pp. 347–351.

[4] Byrne, A., Walsh, M., Farrelly, M. & O'Driscoll, K. (1993), 'Depression following laryngectomy. A pilot study.', *The British Journal of Psychiatry*, vol. 163, no. 2, pp. 173–176.

[5] Shapiro, P.A., Kornfeld, D.S. et al. (1987), 'Psychiatric aspects of head and neck cancer surgery.', *The Psychiatric Clinics of North America*, vol. 10, no. 1, p. 87.

[6] A.D.A.M. Medical Encyclopedia (2015), 'Laryngectomy', Online, URL `http://www.nlm.nih.gov/medlineplus/ency/article/007398.htm`, accessed: 5 May 2016.

[7] American Cancer Society (2015), *Cancer facts & figures*, The Society.

[8] WellCare (2015), 'Voice prosthesis for voice rehabilitation following total laryngectomy', *Clinical Coverage Guideline*, , no. HS-083.

[9] InHealth Technologies (2013), 'What is a laryngectomy', Online, URL `http://store.inhealth.com/category_s/60.htm`, accessed: 6 May 2013.

[10] Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. & Brumberg, J. (2010), 'Silent speech interfaces', *Speech Communication*, vol. 52, no. 4, pp. 270–287.

[11] AudiSoft Technologies Inc. (2016), URL `www.audiosoft.net`.

[12] Bernstein, L.E., Tucker, P.E. & Demorest, M.E. (2000), 'Speech perception without hearing', *Perception & Psychophysics*, vol. 62, no. 2, pp. 233–252.

[13] Henrey, E. (2007), 'Recent developments in the use of experts and the admissibility of expert evidence - an international perspective', in 'Second National Forum on Expert Evidence in Criminal Proceedings: Strategies for Avoiding Wrongful Convictions and Acquittals', .

[14] Gritzman, A.D., Rubin, D.M. & Pantanowitz, A. (2015), 'Comparison of colour transforms used in lip segmentation algorithms', *Signal, Image and Video Processing*, vol. 9, no. 4, pp. 947–957.

[15] Gritzman, A.D., Aharonson, V., Rubin, D.M. & Pantanowitz, A. (2015), 'Automatic computation of histogram threshold for lip segmentation using feedback of shape information', *Signal, Image and Video Processing*, pp. 1–8.

[16] Gritzman, A.D., Aharonson, V., Rubin, D. & Pantanowitz, A. (2015), 'Threshold-based lip segmentation using feedback of shape information', in 'Proceedings of the Conference on Facial Analysis and Animation 2015 (FAA2015)', ACM, p. 4.

[17] Levelt, W.J. (1993), *Speaking: From intention to articulation*, vol. 1, MIT press.

[18] Krames StayWell (2013), 'Parts of the throat and neck', Online, URL `http://www.uofmchildrenshospital.org/healthlibrary/Article/84507`, accessed: 30 July 2013.

[19] Deng, L. & Dang, J. (2007), 'Speech analysis: The production-perception perspective', *Advances in Chinese Spoken Language Processing*, pp. 3–32.

[20] Hall, J.E. & Guyton, A.C. (2006), *Textbook of medical physiology*, Saunders.

[21] Mannell, R. (2008), 'Sound sources in the vocal tract', Online, URL `http://clas.mq.edu.au/acoustics/frequency/source.html`, accessed: 7 May 2013.

[22] Hoofring, A. (2003), 'Larynx and nearby structures', Online, URL `http://commons.wikimedia.org/wiki/File:Larynx_and_nearby_structures.jpg`, accessed: 31 July 2013.

[23] Hoofring, A. (2003), 'Larynx (top view)', Online, URL `http://commons.wikimedia.org/wiki/File:Larynx_(top_view).jpg`, accessed: 31 July 2013.

[24] Jarvis, J.F. (1978), *An Introduction to the Anatomy and Physiology of Speech and Hearing*, Juta.

[25] Dickson, D.R. & Maue-Dickson, W. (1982), *Anatomical and physiological bases of speech*, Little, Brown Boston.

[26] Clark, J.E., Yallop, C. & Fletcher, J. (1995), 'An introduction to phonetics and phonology', .

[27] Kubert, H.L., Stepp, C.E., Zeitels, S.M., Gooey, J.E., Walsh, M.J., Prakash, S. & Hillman, R.E. (2009), 'Electromyographic control of a hands-free electrolarynx using neck strap muscles', *Journal of Communication Disorders*, vol. 42, no. 3, pp. 211–225.

[28] Liu, H. & Ng, M.L. (2007), 'Electrolarynx in voice rehabilitation', *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332.

[29] Fagan, M., Ell, S., Gilbert, J., Sarrazin, E. & Chapman, P. (2008), 'Development of a (silent) speech recognition system for patients following laryngectomy', *Medical Engineering & Physics*, vol. 30, no. 4, pp. 419–425.

[30] Atos Medical AB (2013), 'Atos medical image bank', Online, URL `http://www.atosmedical.com/Corporate/Media/Image_Bank`, accessed: 1 August 2013.

[31] Kazi, R., Kanagalingam, J., Venkitaraman, R., Prasad, V., Clarke, P., Nutting, C.M., Rhys-Evans, P. & Harrington, K.J. (2009), 'Electroglottographic and perceptual evaluation of tracheoesophageal speech', *Journal of Voice*, vol. 23, no. 2, pp. 247–254.

[32] Fagan, M.J., Chapman, P.M. & Gilbert, J.M. (2006), 'Generation of data from speech or voiceless mouth speech', .

[33] Heaton, J. & Parker, A. (1994), 'Indwelling tracheo-oesophageal voice prostheses post-laryngectomy in Sheffield, UK: a 6-year review', *Acta Oto-Laryngologica*, vol. 114, no. 6, pp. 675–678.

[34] Ell, S. (2000), 'A retrieval study to investigate the failure of silastic speaking valves used post-laryngectomy', *MD Thesis, Leeds.*

[35] Ell, S. (1996), 'Candida 'the cancer of silastic", *The Journal of Laryngology & Otology*, vol. 110, no. 03, pp. 240–242.

[36] Ell, S., Mitchell, A. & Parker, A. (1995), 'Microbial colonization of the Groningen speaking valve and its relationship to valve failure', *Clinical Otolaryngology & Allied Sciences*, vol. 20, no. 6, pp. 555–556.

[37] Eerenstein, S.E.J., Schouwenburg, P.F., Velden, L.A.V.D. & Boer, M.F.D. (2001), 'First results of the VoiceMaster prosthesis in three centres in the Netherlands', *Clinical Otolaryngology & Allied Sciences*, vol. 26, no. 2, pp. 99–103.

[38] Everaert, E.P.J.M., Mahieu, H., Chung, R.P.W., Verkerke, G., der Mei, H.V. & Busscher, H. (1997), 'A new method for in vivo evaluation of biofilms on surface-modified silicone rubber voice prostheses', *European Archives of Oto-Rhino-Laryngology*, vol. 254, no. 6, pp. 261–263.

[39] Hilgers, F.J.M., Ackerstaff, A.H., Balm, A.J.M., Brekel, M.W.M.V.D., Tan, I.B. & Persson, J. (2003), 'A new problem-solving indwelling voice prosthesis, eliminating the need for frequent Candida-and "underpressure"-related replacements: Provox ActiValve', *Acta Oto-Laryngologica*, vol. 123, no. 8, pp. 972–979.

[40] Jacobson, M., Franssen, E., Birt, B.D., Davidson, M.J. & Gilbert, R.W. (1997), 'Predicting postlaryngectomy voice outcome in an era of primary tracheoesophageal fistulization: A retrospective evaluation', *Journal of Otolaryngology*, vol. 26, no. 3, pp. 171–179.

[41] MacCallum, J.K., Cai, L., Zhou, L., Zhang, Y. & Jiang, J.J. (2009), 'Acoustic analysis of aperiodic voice: perturbation and nonlinear dynamic properties in esophageal phonation', *Journal of Voice*, vol. 23, no. 3, pp. 283–290.

[42] Eadie, T.L., Doyle, P.C., Hansen, K. & Beaudin, P.G. (2008), 'Influence of speaker gender on listener judgments of tracheoesophageal speech', *Journal of Voice*, vol. 22, no. 1, pp. 43–57.

[43] Petajan, E. (1984), 'Automatic lipreading to enhance speech recognition', in 'IEEE Communications Society Global Telecommunications Conference', Atlanta, USA, pp. 265–272.

[44] Yuhas, B.P., Goldstein Jr, M.H. & Sejnowski, T.J. (1989), 'Integration of acoustic and visual speech signals using neural networks', *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71.

[45] Chan, M.T., Zhang, Y. & Huang, T.S. (1998), 'Real-time lip tracking and bimodal continuous speech recognition', in '1998 IEEE Second Workshop on Multimedia Signal Processing', IEEE, pp. 65–70.

[46] Yoshinaga, T., Tamura, S., Iwano, K. & Furui, S. (2003), 'Audio-visual speech recognition using lip movement extracted from side-face images', in 'International Conference on Audio-Visual Speech Processing (AVSP 2003)', .

[47] Motlíček, P. & Černockỳ, J. (2004), 'Multimodal phoneme recognition of meeting data', in 'Seventh International Conference on Text, Speech and Dialogue (TSD 2004)', Springer, pp. 379–384.

[48] Kubanek, M. (2005), 'Technique of video features extraction for audio-video speach recognition system', *Computing, Multimedia and Intelligent Techniques*, vol. 1, no. 1, pp. 181–190.

[49] Lucey, S., Chen, T., Sridharan, S. & Chandran, V. (2005), 'Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition', *IEEE Transactions on Multimedia*, vol. 7, no. 3, pp. 495–506.

[50] Denby, B., Oussar, Y., Dreyfus, G. & Stone, M. (2006), 'Prospects for a silent speech interface using ultrasound imaging', in '2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)', IEEE, vol. 1, pp. I–I.

[51] Hueber, T., Chollet, G., Denby, B., Stone, M. & Zouari, L. (2007), 'Ouisper: corpus based synthesis driven by articulatory data', in '16th International Congress of Phonetic Sciences', .

[52] Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G. & Stone, M. (2010), 'Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips', *Speech Communication*, vol. 52, no. 4, pp. 288–300.

[53] Russell, M.J., Rubin, D.M., Marwala, T. & Wigdorowitz, B. (2010), 'Pattern recognition and feature selection for the development of a new artificial larynx', in 'World Congress on Medical Physics and Biomedical Engineering 2009', Springer, Munich, Germany, pp. 736–739.

[54] Sugie, N. & Tsunoda, K. (1985), 'A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production', *IEEE Transactions on Biomedical Engineering*, , no. 7, pp. 485–490.

[55] Schultz, T. & Wand, M. (2010), 'Modeling coarticulation in EMG-based continuous speech recognition', *Speech Communication*, vol. 52, no. 4, pp. 341–353.

[56] Jou, S.C.S., Schultz, T., Walliczek, M., Kraft, F. & Waibel, A. (2006), 'Towards continuous speech recognition using surface electromyography.', in 'Interspeech 2006 – ICSLP', .

[57] Wand, M. & Schultz, T. (2009), 'Towards speaker-adaptive speech recognition based on surface electromyography.', in 'International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2009)', pp. 155–162.

[58] Suppes, P., Lu, Z.L. & Han, B. (1997), 'Brain wave recognition of words', *Proceedings of the National Academy of Sciences*, vol. 94, no. 26, pp. 14965–14969.

[59] Neuper, C., Müller, G., Kübler, A., Birbaumer, N. & Pfurtscheller, G. (2003), 'Clinical application of an eeg-based brain–computer interface: a case study in a patient with severe motor impairment', *Clinical Neurophysiology*, vol. 114, no. 3, pp. 399–409.

[60] Wester, M. & Schultz, T. (2006), *Unspoken speech-speech recognition based on electroencephalography*, Master's thesis, Universität Karlsruhe (TH), Karlsruhe, Germany.

[61] Brumberg, J.S., Nieto-Castanon, A., Kennedy, P.R. & Guenther, F.H. (2010), 'Brain-computer interfaces for speech communication', *Speech Communication*, vol. 52, no. 4, pp. 367–379.

[62] Brumberg, J.S., Wright, E.J., Andreasen, D.S., Guenther, F.H. & Kennedy, P.R. (2011), 'Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex', *Frontiers in Neuroscience*, vol. 5.

[63] Talea, H. & Yaghmaie, K. (2011), 'Automatic visual speech segmentation', in '2011 IEEE 3rd International Conference on Communication Software and Networks (ICCSN 2011)', IEEE, pp. 184–188.

[64] McGurk, H. & MacDonald, J. (1976), 'Hearing lips and seeing voices', .

[65] Lucey, P., Martin, T. & Sridharan, S. (2004), 'Confusability of phonemes grouped according to their viseme classes in noisy environments', in 'Tenth Australian International Conference on Speech Science and Technology', .

[66] Luettin, J., Thacker, N.A. & Beet, S.W. (1996), 'Speechreading using shape and intensity information', in 'Fourth International Conference on Spoken Language (ICSLP 1996)', IEEE, vol. 1, pp. 58–61 vol. 1.

[67] Chen, T. & Rao, R.R. (1998), 'Audio-visual integration in multimodal communication', *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852.

[68] Potamianos, G., Neti, C., Gravier, G., Garg, A. & Senior, A.W. (2003), 'Recent advances in the automatic recognition of audiovisual speech', *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326.

[69] Jeffers, J. & Barley, M. (1971), *Speechreading (lipreading)*, Thomas Springfield.

[70] Nitchie, E.B. (1930), 'Principles and practice of lip reading', .

[71] Wang, S., Lau, W., Leung, S. & Yan, H. (2004), 'A real-time automatic lipreading system', in 'International Symposium on Circuits and Systems (ISCAS 2004)', IEEE, vol. 2, pp. II–101–104.

[72] International Phonetic Association (1999), *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press.

[73] Alothmany, N., Boston, R., Li, C., Shaiman, S. & Durrant, J. (2010), 'Classification of visemes using visual cues', in '52nd International Symposium ELMAR-2010', pp. 345–349.

[74] Cappelletta, L. & Harte, N. (2011), 'Viseme definitions comparison for visual-only speech recognition', in '19th European Signal Processing Conference (EUSIPCO 2011)', pp. 2109–2113.

[75] Pandzic, I., Escher, M. & Thalmann, N.M. (1998), 'Facial deformations for MPEG-4', in 'Computer Animation 1998 (CA 1998)', IEEE, pp. 56–62.

[76] Hilder, S., Theobald, B.J. & Harvey, R. (2010), 'In pursuit of visemes', in '9th International Conference on Auditory-Visual Speech Processing (AVSP 2010)', .

[77] Bozkurt, E., Erdem, C., Erzin, E., Erdem, T. & Ozkan, M. (2007), 'Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation', in '2007 IEEE 15th Signal Processing and Communications Applications (SIU 2007)', IEEE, pp. 1–4.

[78] Hazen, T.J., Saenko, K., La, C.H. & Glass, J.R. (2004), 'A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments', in '6th International Conference on Multimodal Interfaces (ICMI 2004)', ACM, pp. 235–242.

[79] Mattheyses, W., Latacz, L. & Verhelst, W. (2011), 'Automatic viseme clustering for audiovisual speech synthesis', in 'Twelfth Annual Conference of the International Speech Communication Association', .

[80] Goldschen, A.J. (1993), 'Continuous automatic speech recognition by lipreading', .

[81] Stork, D.G. & Hennecke, M.E. (1996), 'Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques', in 'Second International Conference on Automatic Face and Gesture Recognition', IEEE, pp. XVI–XXVI.

[82] WenJuan, Y., YaLing, L. & MingHui, D. (2010), 'A real-time lip localization and tacking for lip reading', in '2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE 2010)', IEEE, vol. 6, pp. V6–363–V6–366.

[83] Cappelletta, L. & Harte, N. (2010), 'Nostril detection for robust mouth tracking', in 'Signals and Systems Conference (ISSC 2010)', IET, pp. 239–244.

[84] Pietruch, R. & Grzanka, A. (2010), 'Combining acoustic and visual modalities in vowel recognition system for laryngectomees', in '2010 10th Symposium on Neural Network Applications in Electrical Engineering (NEUREL 2010)', IEEE, pp. 175–179.

[85] Koo, H. & Song, H. (2010), 'Facial feature extraction for face modeling program', *International Journal of Circuits, Systems and Signal Processing*, vol. 4, no. 4, pp. 169–176.

[86] Shiell, D.J., Terry, L.H., Aleksic, P.S. & Katsaggelos, A.K. (2009), 'Audio-visual and visual-only speech and speaker recognition: Issues about theory, system design', in A.W.C. Liew & S. Wang (editors), 'Visual Speech Recognition: Lip Segmentation and Mapping', pp. 1–38.

[87] Chollet, G., Landais, R., Bredin, H., Hueber, T., Mokbel, C., Perrot, P. & Zouari, L. (2007), 'Some experiments in audio-visual speech processing, in non-linear speech processing', *Progress in Non-Linear Speech Processing*.

[88] Wilson, P. & Fernandez, J. (2006), 'Facial feature detection using Haar classifiers', *Journal of Computing Sciences in Colleges*, vol. 21, no. 4, pp. 127–133.

[89] Viola, P. & Jones, M. (2001), 'Rapid object detection using a boosted cascade of simple features', in '2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)', IEEE, vol. 1, pp. I–511.

[90] Hewitt, R. (2007), 'Seeing with OpenCV part 2: Finding faces in images', *SERVO Magazine*.

[91] Saeed, U. & Dugelay, J.L. (2010), 'Combining edge detection and region segmentation for lip contour extraction', *Articulated Motion and Deformable Objects*, pp. 11–20.

[92] Aleksic, P.S. & Katsaggelos, A.K. (2009), 'Lip feature extraction and feature evaluation in the context of speech and speaker recognition', in A.W.C. Liew & S. Wang (editors), 'Visual Speech Recognition: Lip Segmentation and Mapping', pp. 39–69.

[93] Eveno, N., Caplier, A. & Coulon, P. (2004), 'Accurate and quasi-automatic lip tracking', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 706–715, iD: 1.

[94] Luettin, J. & Thacker, N.A. (1997), 'Speechreading using probabilistic models', *Computer Vision and Image Understanding*, vol. 65, no. 2, pp. 163–178.

[95] Huang, X., Ariki, Y. & Jack, M. (1990), *Hidden Markov Models for Speech Recognition*, Edinburgh, UK, Edinburgh University Press.

[96] Lewis, T.W. & Powers, D.M.W. (2002), 'Audio-visual speech recognition using red exclusion and neural networks', in 'Australian Computer Science Communications', Australian Computer Society, Inc., vol. 24, pp. 149–156.

[97] Baldwin, J.F., Martin, T.P. & Saeed, M. (1999), 'Automatic computer lip-reading using fuzzy set theory', in 'International Conference on Auditory-Visual Speech Processing (AVSP 1999)', .

[98] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D. & Povey, D. (2006), 'The HTK book (for HTK version 3.4)', .

[99] Cheung, Y., Li, M., Peng, Q. & Chen, C. (2015), 'A cooperative learning-based clustering approach to lip segmentation without knowing segment number.', *IEEE Transactions on Neural Networks and Learning Systems*.

[100] Martinez, A. (1998), 'The AR face database', *CVC Technical Report*, vol. 24.

[101] Ding, L. & Martinez, A. (2010), 'Features versus context: An approach for precise and detailed detection and delineation of faces and facial features', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 11, pp. 2022–2038.

[102] Chellappa, R., Wilson, C.L. & Sirohey, S. (1995), 'Human and machine recognition of faces: A survey', *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–741.

[103] Caplier, A., Stillittano, S., Bouvier, C. & Coulon, P. (2009), 'Lip modelling and segmentation', in A.W.C. Liew & S. Wang (editors), 'Visual Speech Recognition: Lip Segmentation and Mapping', pp. 70–127.

[104] Lievin, M. & Luthon, F. (2004), 'Nonlinear color space and spatiotemporal mrf for hierarchical segmentation of face features in video', *IEEE Transactions on Image Processing*, vol. 13, no. 1, pp. 63–71.

[105] Coianiz, T., Torresani, L. & Caprile, B. (1996), '2D deformable models for visual speech analysis', in G. Stork & M.E. Hennecke (editors), 'NATO Advanced Study Institute: Speechreading by Man and Machine', Springer-Verlag, pp. 391–398.

[106] Vogt, M. (1996), 'Fast matching of a dynamic lip model to color video sequences under regular illumination conditions', *Nato ASI Subseries F: Computer and Systems Sciences*, vol. 150, pp. 399–408.

[107] Liang, Y.L. & Du, M.H. (2011), 'Lip extraction method based on a component of lab color space', *Computer Engineering*, vol. 37, no. 3.

[108] Eveno, N., Caplier, A. & Coulon, P.Y. (2002), 'Key points based segmentation of lips', in '2002 IEEE International Conference on Multimedia and Expo', IEEE, vol. 2, pp. 125–128 vol. 2.

[109] Zhang, X. & Mersereau, R. (2000), 'Lip feature extraction towards an automatic speechreading system', in '2000 International Conference on Image Processing (ICIP 2000)', IEEE, vol. 3, pp. 226–229.

[110] Hurlbert, A.C. & Poggio, T.A. (1988), 'Synthesizing a color algorithm from examples', *Science*, vol. 239, no. 4839, pp. 482–485.

[111] Goldschen, A.J., Garcia, O.N. & Petajan, E. (1994), 'Continuous optical automatic speech recognition by lipreading', in '28th Asilomar Conference on Signals, Systems and Computers', IEEE, vol. 1, pp. 572–577 vol. 1.

[112] Dahlman, E., Parkvall, S., Beming, P., Bovik, A.C., Fette, B.A., Jack, K., Skold, J., Dowla, F., Chou, P.A. & DeCusatis, C. (2009), *Communications engineering desk reference*, Academic Pr.

[113] Ford, A. & Roberts, A. (1998), 'Colour space conversions', *Westminster University, London*, pp. 1–31.

[114] Eveno, N., Caplier, A. & Coulon, P.Y. (2001), 'New color transformation for lips segmentation', in '2001 IEEE Fourth Workshop on Multimedia Signal Processing', pp. 3–8, iD: 1.

[115] McClain, M., Brady, K., Brandstein, M. & Quatieri, T. (2004), 'Automated lip-reading for improved speech intelligibility', in '2004 IEEE International Conference on Acoustics, Speech, and Signal Processing', IEEE, vol. 1, pp. I–701.

[116] Hsu, R.L., Abdel-Mottaleb, M. & Jain, A.K. (2002), 'Face detection in color images', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696–706.

[117] Thejaswi, N.S. & Sengupta, S. (2008), 'Lip localization and viseme recognition from video sequences', in '14th National Conference on Communications (NCC 2008)', .

[118] Hamilton, J. (1992), 'Color space conversion', *Green Harbor Publications*.

[119] Canzler, U. & Dziurzyk, T. (2002), 'Extraction of non manual features for videobased sign language recognition', in 'IAPR Workshop on Machine Vision Applications (IAPR MVA 2002)', pp. 318–321.

[120] Gong, Y. & Sakauchi, M. (1995), 'Detection of regions matching specified chromatic features', *Computer Vision and Image Understanding*, vol. 61, no. 2, pp. 263–269.

[121] Guan, Y.P. (2008), 'Automatic extraction of lips based on multi-scale wavelet edge detection', *IET Computer Vision*, vol. 2, no. 1, pp. 23–33.

[122] Zhang, J., Tao, H., Wang, L., Zhan, Y. & Song, S. (2004), 'A real-time approach to the lip-motion extraction in video sequence', in '2004 IEEE International Conference on Systems, Man and Cybernetics', IEEE, vol. 7, pp. 6423–6428 vol. 7.

[123] Ohta, Y., Kanade, T. & Sakai, T. (1980), 'Color information for region segmentation', *Computer Graphics and Image Processing*, vol. 13, no. 3, pp. 222–241.

[124] Watson, A.B. & Poirson, A. (1986), 'Separable two-dimensional discrete Hartley transform', *Journal of the Optical Society of America A*, vol. 3, no. 12, pp. 2001–2004.

[125] Otsu, N. (1975), 'A threshold selection method from gray-level histograms', *Automatica*, vol. 11, no. 285-296, pp. 23–27.

[126] Demirkaya, O. & Asyali, M.H. (2004), 'Determination of image bimodality thresholds for different intensity distributions', *Signal Processing: Image Communication*, vol. 19, no. 6, pp. 507–516.

[127] The MathWorks Inc. (1998), 'MATLAB user guide', vol. 4.

[128] Fu, K.S. & Mui, J. (1981), 'A survey on image segmentation', *Pattern recognition*, vol. 13, no. 1, pp. 3–16.

[129] Haralick, R.M. & Shapiro, L.G. (1985), 'Image segmentation techniques', *Computer vision, graphics, and image processing*, vol. 29, no. 1, pp. 100–132.

[130] Pal, N.R. & Pal, S.K. (1993), 'A review on image segmentation techniques', *Pattern recognition*, vol. 26, no. 9, pp. 1277–1294.

[131] Zhang, Y.J. (1996), 'A survey on evaluation methods for image segmentation', *Pattern recognition*, vol. 29, no. 8, pp. 1335–1346.

[132] Wark, T., Sridharan, S. & Chandran, V. (1998), 'An approach to statistical lip modelling for speaker identification via chromatic feature extraction', in 'Fourteenth International Conference on Pattern Recognition (ICPR 1998)', IEEE, vol. 1, pp. 123–125.

[133] Chiou, G.I. & Hwang, J.N. (1997), 'Lipreading from color video', *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1192–1195.

[134] Xinjun, M. & Hongqiao, Z. (2015), 'Lip segmentation algorithm based on bi-color space', in '34th Chinese Control Conference (CCC2015)', IEEE, pp. 3776–3779.

[135] Pardàs, M. & Sayrol, E. (2001), 'Motion estimation based tracking of active contours', *Pattern Recognition Letters*, vol. 22, no. 13, pp. 1447–1456.

[136] Werda, S., Mahdi, W. & Hamadou, A.B. (2007), 'Colour and geometric based model for lip localisation: Application for lip-reading system', in '14th International Conference on Image Analysis and Processing (ICIAP 2007)', IEEE, pp. 9–14.

[137] Delmas, P., Eveno, N. & Lievin, M. (2002), 'Towards robust lip tracking', in '16th International Conference on Pattern Recognition', IEEE, vol. 2, pp. 528–531.

[138] Beaumesnil, B. & Luthon, F. (2006), 'Real time tracking for 3d realistic lip animation', in 'Pattern Recognition, 2006. ICPR 2006. 18th International Conference on', IEEE, vol. 1, pp. 219–222.

[139] Skodras, E. & Fakotakis, N. (2011), 'An unconstrained method for lip detection in color images', in 'Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on', IEEE, pp. 1013–1016.

[140] Hara, K. & Chellappa, R. (2014), 'Growing regression forests by classification: Applications to object pose estimation', in 'European Conference on Computer Vision', Springer, pp. 552–567.

[141] Rohani, R., Alizadeh, S., Sobhanmanesh, F. & Boostani, R. (2008), 'Lip segmentation in color images', in 'Innovations in Information Technology, 2008. IIT 2008. International Conference on', IEEE, pp. 747–750.

[142] Gacon, P., Coulon, P.Y. & Bailly, G. (2005), 'Non-linear active model for mouth inner and outer contours detection', in 'Signal Processing Conference, 2005 13th European', IEEE, pp. 1–4.

[143] Bouvier, C., Coulon, P.Y. & Maldague, X. (2007), 'Unsupervised lips segmentation based on ROI optimisation and parametric model', in '2007 IEEE International Conference on Image Processing', IEEE, vol. 4, pp. IV–301–IV–304.

[144] Li, M. & Cheung, Y.m. (2010), 'Automatic segmentation of color lip images based on morphological filter', in 'Artificial Neural Networks–ICANN 2010', Springer, pp. 384–387.

[145] Cheung, Y.m., Li, M., Cao, X. & You, X. (2014), 'Lip segmentation under map-mrf framework with automatic selection of local observation scale and number of segments', *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3397–3411.

[146] Wang, S., Liew, A., Lau, W.H. & Leung, S.H. (2009), 'Lip region segmentation with complex background', *Visual Speech Recognition: Lip Segmentation and Mapping: Lip Segmentation and Mapping*, p. 150.

[147] Yuille, A.L., Hallinan, P.W. & Cohen, D.S. (1992), 'Feature extraction from faces using deformable templates', *International Journal of Computer Vision*, vol. 8, no. 2, pp. 99–111.

[148] Kaucic, R., Dalton, B. & Blake, A. (1996), 'Real-time lip tracking for audio-visual speech recognition applications', *Computer Vision – ECCV 1996*, pp. 376–387.

[149] Kass, M., Witkin, A. & Terzopoulos, D. (1988), 'Snakes: Active contour models', *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331.

[150] Mok, L.L., Lau, W.H., Leung, S.H., Wang, S.L. & Yan, H. (2004), 'Person authentication using ASM based lip shape and intensity information', in '2004

International Conference on Image Processing (ICIP 2004)', IEEE, vol. 1, pp. 561–564 Vol. 1.

[151] Tian, Y., Kanade, T. & Cohn, J. (2000), 'Robust lip tracking by combining shape, color and motion', in '4th Asian Conference on Computer Vision (ACCV 2000)', pp. 1040–1045.

[152] Gonzalez, R.C., Woods, R.E. & Eddins, S.L. (2009), *Digital image processing using MATLAB*, vol. 2, Gatesmark Publishing Knoxville.

[153] Keshet, R. (2008), 'Dilation', Online, URL `http://en.wikipedia.org/wiki/File:Dilation.png`, accessed: 24 October 2013.

[154] Keshet, R. (2008), 'Erosion', Online, URL `http://en.wikipedia.org/wiki/File:Erosion.png`, accessed: 24 October 2013.

[155] Keshet, R. (2008), 'Dilation and erosion', Online, URL `http://en.wikipedia.org/wiki/File:DilationErosion.png`, accessed: 24 October 2013.

[156] Bovik, A.C. (2009), *The essential guide to image processing*, Academic Press.

[157] De Berg, M., Van Kreveld, M., Overmars, M. & Schwarzkopf, O.C. (2000), *Computational geometry*, Springer.

[158] Pollock, D. & Mary, Q. (1993), *Smoothing with cubic splines*, Department of Economics, Queen Mary and Westfield College.

[159] Wang, S., Leung, S. & Lau, W. (2002), 'Lip segmentation by fuzzy clustering incorporating with shape function', in '2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002)', IEEE, vol. 1, pp. I–1077.

[160] Liew, A.C., Leung, S.H. & Lau, W.H. (2003), 'Segmentation of color lip images by spatial fuzzy clustering', *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, pp. 542–549.

[161] Pan, J., Guan, Y. & Wang, S. (2012), 'A new color transformation based fast outer lip contour extraction', *Journal of Information and Computational Science*, vol. 9, no. 9, pp. 2505–2514.

[162] Cheung, Y.m. & Li, M. (2011), 'Map-MRF based lip segmentation without true segment number', in '2011 18th IEEE International Conference on Image Processing (ICIP 2011)', IEEE, pp. 769–772.

[163] Chin, S.W., Seng, K.P. & Ang, L.M. (2010), 'Enhanced snake model and modified H-infinity for lips contour detection and tracking', in '2010 International Conference on Computer Applications and Industrial Electronics (ICCAIE 2010)', IEEE, pp. 659–664.

[164] Saeed, U. (2010), 'Comparative analysis of lip features for person identification', in '8th International Conference on Frontiers of Information Technology', ACM, New York, NY, USA, FIT '10, ISBN 978-1-4503-0342-2, pp. 20:1–20:6, doi: 10.1145/1943628.1943648, URL `http://doi.acm.org/10.1145/1943628.1943648`.

[165] Guan, Y.P. (2006), 'Automatic extraction of lip based on wavelet edge detection', in 'Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC 2006)', IEEE, pp. 125–132.

[166] Smola, A.J. & Schölkopf, B. (2004), 'A tutorial on support vector regression', *Statistics and Computing*, vol. 14, no. 3, pp. 199–222.

[167] Vapnik, V. (1998), *Statistical learning theory. 1998*, Wiley, New York.

[168] Bennett, K.P. & Mangasarian, O.L. (1992), 'Robust linear programming discrimination of two linearly inseparable sets', *Optimization methods and software*, vol. 1, no. 1, pp. 23–34.

[169] Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine Learning*, vol. 20, no. 3, pp. 273–297.

[170] Schölkopf, B. & Smola, A.J. (2002), *Learning with kernels: Support vector machines, regularization, optimization, and beyond*, MIT press.

[171] Hsu, C.W., Chang, C.C., Lin, C.J. et al. (2003), 'A practical guide to support vector classification', .

[172] MathWorks (2016), 'How pattern search polling works', Online, URL `http://www.mathworks.com/help/gads/how-pattern-search-polling-works.html`, accessed: 30 August 2016.

[173] MathWorks (2016), 'What is direct search?', Online, URL `http://www.mathworks.com/help/gads/what-is-direct-search.html`, accessed: 30 August 2016.

[174] Leung, S.H., Wang, S.L. & Lau, W.H. (2004), 'Lip image segmentation using fuzzy clustering incorporating an elliptic shape function', *IEEE Transactions on Image Processing*, vol. 13, no. 1, pp. 51–62.

[175] Wang, S.L., Lau, W.H., Liew, A.W.C. & Leung, S.H. (2007), 'Robust lip region segmentation for lip images with complex background', *Pattern Recognition*, vol. 40, no. 12, pp. 3481–3491.

[176] Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S. & Harvey, R. (2002), 'Extraction of visual features for lipreading', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213.

[177] Zheng, Z., Jiong, J., Chunjiang, D., Liu, X. & Yang, J. (2008), 'Facial feature localization based on an improved active shape model', *Information Sciences*, vol. 178, no. 9, pp. 2215–2223.