TOWARDS SPEECH RECOGNITION USING PALATO-LINGUAL CONTACT PATTERNS FOR VOICE RESTORATION

Megan Jill Russell

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, June 2011

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__

_____

Megan Jill Russell

Abstract

The loss of speech following a laryngectomy presents substantial challenges, and a number of devices have been developed to assist these patients. These devices range from the electrolarynx to tracheoesophageal speech. However, all of these devices and techniques have concentrated on producing sound from the patient's vocal tract.

Research into a new type of artificial larynx is presented. This new device utilizes the measurement of dynamic tongue-palate contact patterns to infer intended speech. The dynamic tongue measurement is achieved with the use of an existing palatometer and pseudopalate. These signals are then converted to 2-D Space-Time plots and feature extraction methods (such as Principal Component Analysis, Fourier Descriptors and Generic Fourier Descriptors) are used to extract suitable features for use as input to neural network systems. Two types of neural network (Multi-layer Perceptrons and Support Vector Machines) are investigated and a voting system is formed. The final system can correctly identify fifty common English words 94.14% of the time with a rejection rate of 17.74%.

Voice morphing is investigated as a technique to match the artificially synthesized voice to the laryngectomy patient's original voice. It is successfully implemented thus creating a transfer function that can change one person's voice to sound like another's. Once the voting system has correctly identified the word said by the patient the word is then synthesized in the patient's pre-laryngectomy voice.

The final artificial larynx system solves a number of the problems inherent in previous artificial larynx designs (such as poor voice quality and invasiveness). This new artificial larynx uses current technology in a new way to produce a viable solution for alaryngeal patients.

Acknowledgements

I would like to thank the following people:

- **Prof. David Rubin and Prof. Brian Wigdorowitz:** For their guidance, help and support.
- **Prof. Tshilidzi Marwala:** For his help on the neural network aspects of this work.
- **Dr Ivan Jardine:** For helping me to understand some of the difficulties laryngectomy patients face.
- **Meir Perez:** For help on the implementation of neural networks.
- **Kyle Voster:** For reading in the male voice for voice morphing.

Contents

List of Figures

List of Tables

List of Symbols

The principal symbols used in this thesis are summarised below. The page where the symbol is first used is shown.

Nomenclature

ASR Automatic Speech Recognition

ECG Electrocardiogram

EPG Electropalatograph

FD Fourier Descriptor

GFD Generic Fourier Descriptor

GMM Gaussian Mixture Model

GUI Graphical User Interface

HSM Harmonic/Stochastic Model

IIR Infinite Impulse Response

MLP Multi-layer Perceptron

MRI Magnetic Resonance Imaging

NN Neural Network

PCA Principal Component Analysis

PET Positron Emission Tomography

SCG Scaled Conjugate Gradient

SVM Support Vector Machine

TIMIT Texas Instruments Massachusetts Institute of Technology

TTS Text-to-Speech

Chapter 1

Thesis Overview

> "Nature, as we often say, makes nothing in vain, and man is the only animal whom she has endowed with the gift of speech. And whereas mere voice is but an indication of pleasure or pain, and is therefore found in other animals, the power of speech is intended to set forth the expedient and inexpedient, and therefore likewise the just and the unjust." (Aristotle, 350 B.C.E)

Speech and the ability to communicate is one of humanity's most important tools. Thus when this faculty is taken away, either by accident or as a result of surgery, the results can be devastating both to the patients and their family members. In the USA alone 41 370 new cases of pharyngeal, laryngeal or esophageal cancer were diagnosed in 2009 (ACS, 2009). Many of these patients will face the prospect of a total laryngectomy, in which the larynx is removed and the upper part of the trachea is joined to a tracheostoma in the front of the neck (Ng et al., 1997). This leaves the patient unable to talk or produce any vocalization.

Currently the phonation restoration options available to laryngectomy patients are limited and fraught with problems. The voice output by electrolarynges is mechanical and robotic sounding. The voice quality in esophageal and tracheoesophageal speech is poor, with patients sounding gruff and harsh. Currently, research is being performed to improve the voice quality of the already existing artificial larynx devices (see for example Tack et al. (2008)). Literature directed at new techniques for solving the problem of alaryngeal speech is sparse. Fagan et al. (2008) have used magnets placed on the lips, teeth and tongue and magnetic sensors embedded in a pair of glasses to detect a small set of different words and phonemes. As the words are spoken the changing magnetic fields are picked up and a good success rate on a

small data set was achieved, however the placement and fastening of the magnets is problematic.

The new artificial larynx researched in this thesis uses the dynamic tongue movements made during speech to detect which of fifty words the user is trying to say. This is done using a palatometer system which detects tongue-palate contact signals. These signals are then formed into space-time plots which can be treated as images. Various shape features (such as Principal Component Analysis, Fourier Descriptors, Generic Fourier Descriptors and others) are extracted from the space-time plots and fed into a voting and predictive neural network system. A word identification success rate of 94.14% (of accepted words) is achieved with a rejection rate (percentage of words the system cannot classify and are rejected outright) of 17.74%. The word is then synthesized in a voice very close to the user's own voice using a technique called voice morphing. The solution proposed here is a non-invasive approach to providing laryngectomy patients with natural sounding speech.

This work is the beginning of research into a new type of artificial larynx that is intended to provide laryngectomy patients with the option of "sounding like their old-selves". Ultimately the artificial larynx will consist of an artificial palate (that would fit in the user's mouth) with touch sensors, battery and communication (bluetooth or other low-power communication module) seamlessly built in. This would then send the signals to a small microprocessor and speaker unit in the user's top pocket or clipped onto his/her clothing. This unit would then decode the tongue-palate signals and output the words articulated by the user. See Figure 1.1 for a general overview.

This thesis outlines the research and investigation done into this new type of artificial larynx and demonstrates that it circumvents many of the inherent problems of other artificial larynges.

1.1 Purpose of Research

The hypothesis for this work is that speech-free speech recognition is possible and can be used as the basis for a new approach to artificial larynges. Even though laryngectomy patients cannot utter a sound they still possess the tongue and mouth movements necessary to form words. This research is aimed at investigating and researching new techniques for a new type of artificial larynx that will provide a non invasive solution to this problem.

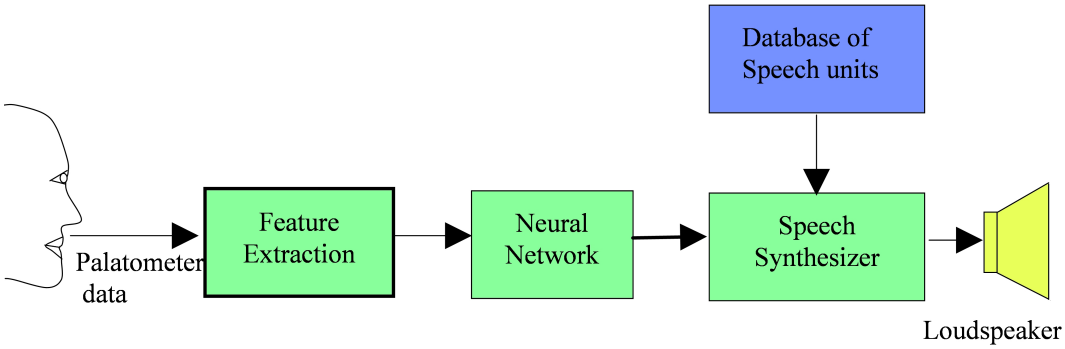


Figure 1.1: An overview of the proposed new artificial larynx

1.2 Research Questions

The following questions were investigated in this work.

1. Can speech recognition be performed on tongue-palate contact patterns from a palatometer?
2. Are standard signal processing techniques and artificial intelligence techniques sufficient to relate the data signals to the speech?
3. If the data signals can be correctly identified as speech, can the appropriate pre-recorded speech be outputted using a loudspeaker?
4. Can pre-recorded speech be altered to mimic other peoples voices?

1.3 Contribution of this Thesis

All previous artificial larynx devices have concentrated on introducing sound into the patient's vocal tract which is then modulated into words. These methods have resulted in poor voice restoration. The central idea behind the new artificial larynx research is as follows: even though laryngectomy patients have lost the ability to phonate, they still retain enough physiological movement and information in their vocal tract to allow for the recognition and classification of what they are trying to say. In this research, dynamic tongue-palate contact patterns are detected by use of a palatometer. These patterns are used in conjunction with neural networks to decipher the user's speech. Voice morphing algorithms allow the synthesizer to

output the speech in a voice very similar to the patient's pre-laryngectomy voice. This research represents a paradigm shift in the thinking behind artificial larynges and makes use of current technology to investigate a solution.

This thesis provides the research and investigation into the problems involved in this new type of artificial larynx. No previous work along these lines has been found in the literature. Speech recognition performed solely on palatometer data has not been found in the literature. By using techniques and technologies from diverse fields and applying them to new situations, inroads into this new approach to artificial larynges have been made.

A patent has been taken out on this artificial larynx (See Appendix F for more information).

1.4 Thesis Structure

The structure of this thesis is as follows:

- **Chapter 2:** This chapter contains the background to the work and the scope of the thesis. It includes an introduction to the larynx as well as currently available artificial larynges.
- **Chapter 3:** This chapter outlines the methods used in this work. The equipment is detailed and methods for feature extraction (such as Principal Component Analysis, Fourier Descriptors, Generic Fourier Descriptors) neural networks (Multi-layer Perceptrons and Support Vector Machines), voting systems and voice morphing are given.
- **Chapter 4:** The results of the methods from Chapter 3 are presented.
- **Chapter 5:** The results from Chapter 4 are discussed.
- **Chapter 6:** In this chapter conclusions are drawn and recommendations for future work are given. The contribution of this work is also highlighted.

Additional supporting material is provided in the appendices as follows:

- **Appendix A:** The ethics approval given for this work.

- **Appendix B:** The Space-Time plots of each of the fifty words used in this work.
- **Appendix C:** The grammar used to generate the second voting system.
- **Appendix D:** This appendix contains graphs showing how the success rate of the MLP changes depending upon the number of hidden nodes.
- **Appendix E:** The TIMIT sentences used for speech recording in voice morphing.
- **Appendix F:** The papers published on this work as well as an article written about this research and information on the patent filed.

A CD containing all the MATLAB code necessary to run the simulations is also provided.

Chapter 2

Background

2.1 Chapter Overview

This chapter gives an introduction to the larynx, the laryngectomy procedure as well as the current options available for phonation (the electrolarynx, esophageal speech and tracheoesophageal speech). It also outlines the current state-of-the-art in artificial larynx technology and current research into improving the voice quality of existing artificial larynx techniques. The scope of the thesis is also outlined.

2.2 The Larynx

The larynx is a continuation of the conducting tube that joins the pharynx and the trachea. It has two main functions (van de Graaff, 2002)

- to prevent food or drink from entering the trachea and lungs during swallowing;
- to produce sound

The larynx is made up of a number of different cartilages and muscles which hold it open during breathing and which close the laryngeal opening (glottis) during swallowing and speech. The epiglottis is a spoon-shaped structure that aids in closing the glottis during swallowing (see Figure 2.1). The vocal folds in the larynx are controlled by muscles and are used in sound production (van de Graaff, 2002).

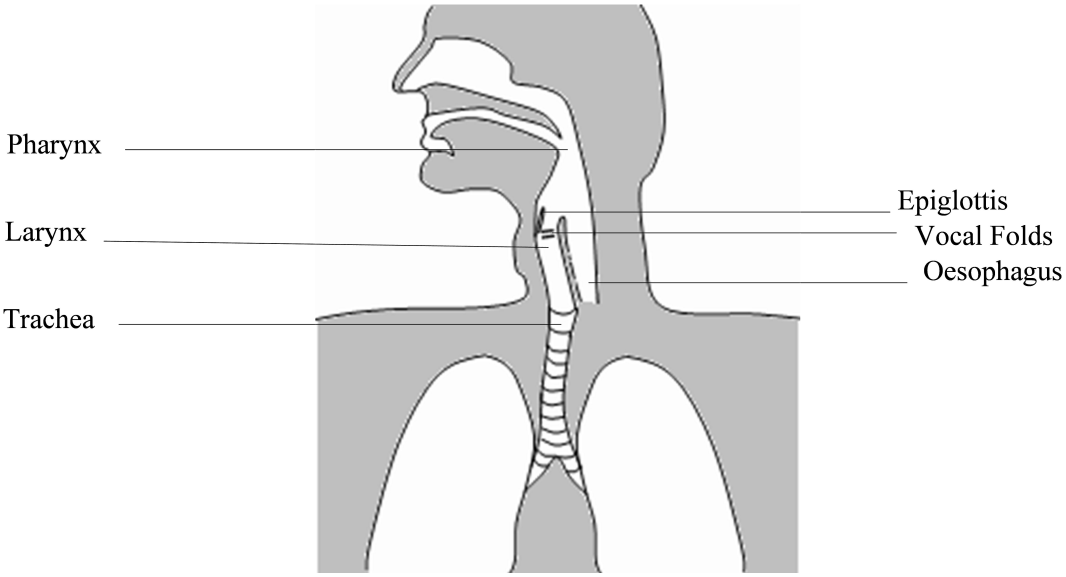


Figure 2.1: A diagram of the larynx (Russell et al., 2008)

2.2.1 The Laryngectomy Procedure

A total laryngectomy is a procedure used to remove the larynx which consists of the thyroid and cricoid cartilages, the hyoid bone, the hyoid pharynx, strap muscles, one to three rings of the trachea and possibly lobes of the thyroid gland (Shoureshi et al., 2003). On removal of the larynx the upper part of the trachea is attached to the front of the neck where a permanent opening (the tracheostoma) is created (Ng et al., 1997). The tracheostoma is mainly for breathing purposes. A total laryngectomy results in the patient being completely unable to phonate or whisper, due to the total removal of the vocal cords.

A supracricoid laryngectomy can be performed on patients in the early stages of cancer in specific sites. This type of laryngectomy preserves some of the essential functions of the larynx (such as deglutition and breathing) and the patient is still able to phonate (Torrejano & Guimaraes, 2009). However the voice quality of the patient is very poor, with absent gender distinction and poor vocal inflection and intensity variation (Torrejano & Guimaraes, 2009). In fact supracricoid laryngectomy patients regard their voice quality as worse than that of total laryngectomy patients who use a speech prosthesis (Torrejano & Guimaraes, 2009).

2.3 Artificial Speech Techniques

A number of different techniques and technologies have been developed in an attempt to restore speech to the laryngectomy patient, all of which have varying degrees of success. The three most common methods are:

2.3.1 The Electrolarynx

More than half of all laryngectomy patients use an electrolarynx (Kubert et al., 2009) as their main means of communication. Two different types of electrolarynx exist: The neck-type (see Figure 2.2) and the intra-oral type. In both kinds, an electromechanical vibrator transmits sounds waves into the oral and pharyngeal cavities where the user modulates the sound into words (Liu & Ng, 2007). Electrolaryngeal speech is associated with low intelligibility and poor listener acceptance (Liu & Ng, 2007). Some electrolarynges have a built in pitch control that the user manually controls when talking.

Figure 2.2: An electrolarynx (shown with permission from Lauder (2007))

2.3.2 Esophageal Speech

Esophageal speech is a popular voice production method, particularly in Asian countries (MacCallum et al., 2009). In this method air is gulped by the patient into the top of the esophagus which is used as an air reservoir. When the patient wishes to speak, this air (see Figure 2.3) is expelled which results in the vibration of the

pharyngeal-esophageal segment (Ng et al., 1997). This vibrational noise is modulated into words by the patient's mouth. Esophageal speech suffers from poor quality (harsh, gurgling) and low pitch, volume and duration (MacCallum et al., 2009).

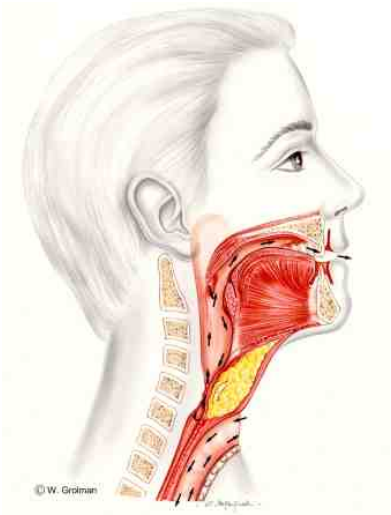


Figure 2.3: A diagram showing the path of air in esophageal speech (shown with permission from Sataloff (2007))

2.3.3 Tracheoesophageal Speech

This is considered the gold standard in voice rehabilitation of laryngectomy patients (Kazi et al., 2009). Tracheoesophageal speech requires the insertion of a one-way valve into a fistula created between the trachea and the esophagus (Ng et al., 1997). During speech the patient must close his/her tracheostoma thus forcing the air (see Figure 2.4) through the one-way valve and vibrating the pharyngoesophageal segment. Tracheoesophageal speech is the most common form of voice restoration in the USA, the UK and continental Europe, however non-English/French speakers struggle to use this method (MacCallum et al., 2009). Users also have to constantly alternate between breathing and drawing in air and covering their stoma to talk, which results in a slow speech rate (Kazi et al., 2009). Another problem is that someone listening to an unknown tracheoesophageal speaker will struggle to identify the speaker's gender (Eadie et al., 2008).

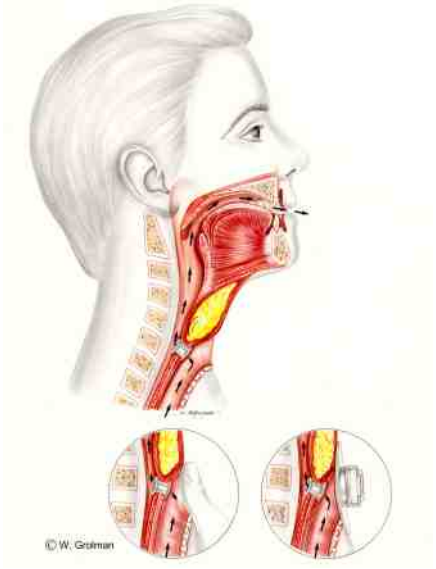


Figure 2.4: A diagram showing the path of air in tracheoesophageal speech (shown with permission from Sataloff (2007))

2.4 Current State-Of-The-Art

Generally research is being performed to improve the voice quality of the already existing artificial larynx devices. The voice quality of tracheoesophageal speech is a much researched topic (see for example Deshpande et al. (2009); Eadie et al. (2008); Most et al. (2000); Ng et al. (1997); Torrejano & Guimaraes (2009)). However, relatively little research is being undertaken into changing and developing the actual valve or voice-producing element used in tracheoesophageal speech. A new voice-producing element for female tracheoesophageal speakers has been developed that increases the pitch of the user's voice and allows for longer phonation time (Tack et al., 2008). Automatic tracheostoma valves which allow for hands-free speech have also been developed (Pawar et al., 2008) however a number of problems are associated with their use. Research is being done to increase the performance of the electrolarynx. New placement techniques, such as neck straps or intra-oral devices, as well as methods to improve the voice quality and introduce pitch control are being investigated (Kubert et al., 2009; Liu & Ng, 2007; Takahashi et al., 2005).

Completely new approaches to the problem of alaryngeal speech are few and far between. In the literature, the only research found that used a novel approach was Fagan et al. (2008). They used magnets placed on the lips, teeth and tongue and magnetic sensors embedded in a pair of glasses to detect a small set of different words and phonemes (13 phonemes and 9 words). As the words are spoken the changing

magnetic fields are picked up, and by using a dynamic time-warping algorithm the words are classified. A good success rate on a small data set was achieved, however the placement of the magnets currently uses surgical glue and long time use of the device would require surgical implantation of the magnets. No speaker outputs are produced in this research.

2.5 Scope of the Thesis

This research is mainly focused on speech-free speech detection that could be used for a new type of artificial larynx. A limited set of common words were chosen to test the idea. The objective of this work was to explore a new approach to the problem of alaryngeal speech. The development of new neural network techniques for this application is a topic for future research. Extensive clinical testing on the new artificial larynx will need to be done in future to prove its viability.

Chapter 3

Methods

3.1 Chapter Overview

Automatic Speech Recognition (ASR) is a much researched topic in the field of signal processing. However the main challenge in the design of the new artificial larynx was how to perform speech recognition and synthesis without any input speech. The equipment used in this research is detailed, and data normalization and pre-processing are explained.

The two different artificial intelligence machines used to analyze the data, namely the multilayer perceptron neural network and the support vector machine are explained. Raw palatometer data cannot be fed into the machines, thus it had the salient features extracted from it and these used as input to the machines. The various feature selection techniques are explained in this chapter as well.

How the neural networks are assembled into voting systems is explained and speech synthesis and the artificial larynx simulator are introduced. Thus this chapter gives an overview of the equipment and methods used to create the artificial larynx.

3.2 Equipment

A challenge in this work was the finding of appropriate non invasive instrumentation. Microphones are unsuitable as laryngectomy patients cannot make any type of sound. They cannot whisper, as the white noise sound used in whispering is also created by the larynx and vocal tract.

A palatometer system is used to provide data to allow for speech-free speech recognition. This device has been used since the mid twentieth century for analyzing tongue palate contact patterns (see Figure 3.1). It is primarily used by speech therapists to help correct speech problems in individuals. It is also used by linguists and in investigations into speech production (see Figure 3.2 for an example of a typical stream of information from a palatometer used in speech studies). Literature using palatometer data alone for speech recognition has not been found.

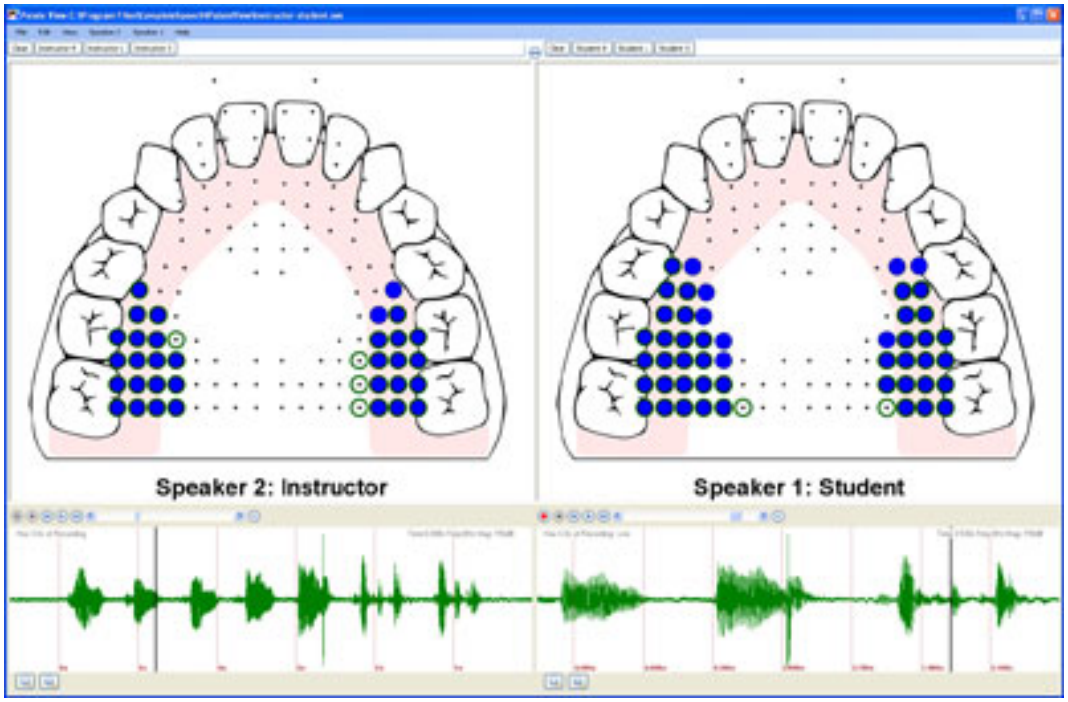


Figure 3.1: A screen shot of the palatometer system showing the real time position of the teacher's tongue and the student's tongue (May (2010)).

In this research, a palatometer system (also known as an Electropalatograph (EPG)) developed by CompleteSpeech (formerly known as LogoMetrix), Arizona, was used to detect the tongue-palate contact patterns made during speech. The palatometer system (see Figure 3.3) consists of a custom made and fitted pseudopalate which contains 118 gold contact sensors (see Figure 3.4) and a basic signal processing unit (CompleteSpeech, 2008). The pseudopalate is sampled at a rate of 100Hz.

Dromey & Sanders (2009) recently used the CompleteSpeech palatometer system to measure inter-speaker variability in consonant production. Cho & Keating (2009) used palatometer data for consonantal measures in their study of the effects of three prosodic factors. Toutios & Margaritis (2008) mapped acoustic signals to electropalatograph data. Palatometers have been used for many years in phonetics

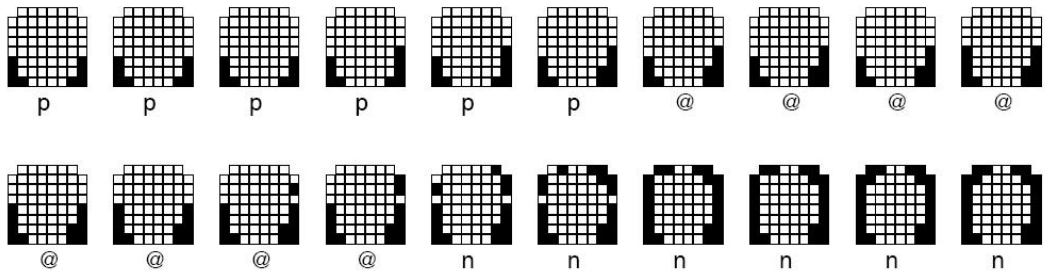


Figure 3.2: Typical EPG sequence. Black squares indicate a contact between the tongue and the palate. Segment is from the word "opens" (Toutios & Margaritis (2006)).

research (for example see Christensen et al. (1992)) and speech therapy (for example see Dagenais (1995)). EPG recordings have also been used to augment speech recognition systems (Soquet et al., 1999) but have never been used by themselves to recognize speech and never in the development of an artificial larynx.

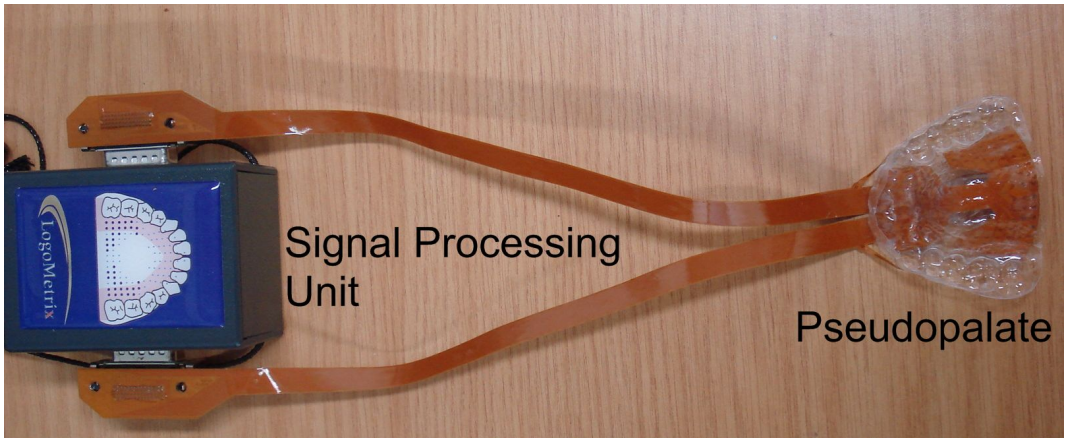
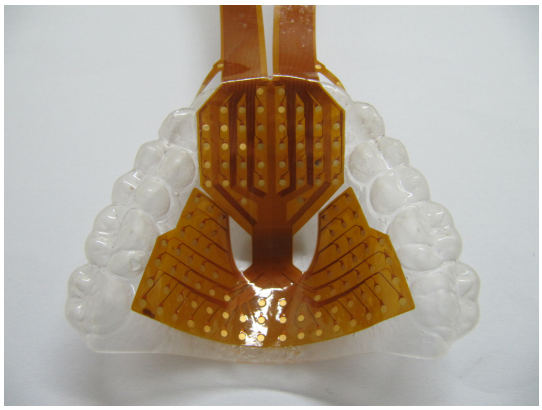
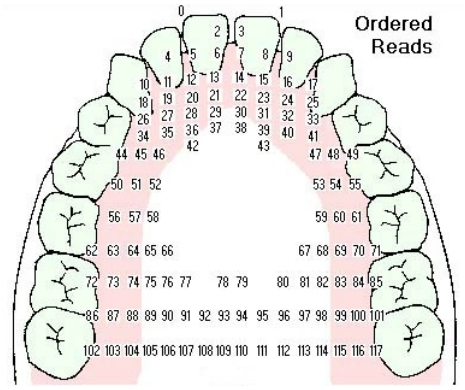


Figure 3.3: The CompleteSpeech System showing the pseudopalate and the signal processing unit

A computer with an Intel Core Duo 2.66GHz CPU and 3.23GB of RAM was used. MATLAB Version 7.6.0.324 (R2008a) as well as NETLAB (free downloadable MATLAB software for data analysis and neural network implementation (Nabney, 2003)) was used. For the speech synthesis a MATLAB toolbox for voice morphing from the Universitat Politecnica de Catalunya Speech Processing Group (Erro, 2008) was used as well as a speech processing program SFS (Release 4.7/Windows) (Huckvale, 2008). To separate the audio information from the palatometer data and to allow it to be imported into MATLAB a program called LogoCracker (provided by



(a) A close-up of the pseudopalate



(b) The placement of the gold electrodes on the pseudopalate (CompleteSpeech, 2008)

Figure 3.4: The pseudopalate.

CompleteSpeech) was used.

A Plantronics .AUDIO 625 USB stereo headset and ProRec speech recording software (Huckvale, 2007) were used to record the word library used in the speech synthesis. The microphone had a frequency response that was 100Hz-10KHz.

Ethics approval for human testing on the principal investigators was granted for this work (see Appendix A).

3.3 Data Capture and Display

3.3.1 The Datasets

Fifty common English words were chosen (see Table 3.1). Palatometer recordings of these words were made 20 times each, to create a dataset of 1000 words. This was called the training dataset. The words were then read another four times each, to create the testing dataset, which contained 200 words. All the words were read by the same person.

To create a word library the words were recorded using the microphone (sampling at a rate of 16KHz). They were read in by both a male and a female speaker.

Table 3.1: The 50 words chosen to test the larynx design (from Russell et al. (2009*a*))

1. the	11. if	21. day	31. take	41. small
2. and	12. will	22. come	32. place	42. large
3. is	13. about	23. did	33. live	43. spell
4. that	14. many	24. sound	34. through	44. big
5. was	15. then	25. number	35. just	45. change
6. for	16. them	26. call	36. form	46. kind
7. I	17. write	27. first	37. great	47. picture
8. they	18. like	28. down	38. same	48. animal
9. have	19. long	29. side	39. sentence	49. head
10. them	20. make	30. been	40. three	50. stand

3.3.2 Word Choice

The fifty words used in this work were all randomly chosen from the top 200 of a list of 500 of the most commonly used words in British, American and Australian English (WorldEnglish, 2009). As this research focuses on a proof of concept, it was thought that commonly used words would be the most useful in the testing of the functionality of this device.

3.3.3 Phonetic Response of the Palatometer

According to Hardcastle et al. (1989) a palatometer system can detect all phonemes and sounds except back vowels (for example /ɒ, ɔː, ʊ, uː/) and relatively open vowels (for example /ɑː, ə, ʌ/). The fact that the palatometer cannot detect some vowels is why the decision was made to focus on word recognition as opposed to phoneme recognition. The meaning of the phonetic symbols is given in Figure 3.5.

3.3.4 Space-Time Plots and Standardization of Data

The palatometer data, once separated from the audio data, can be viewed as a Space-Time plot (Russell et al., 2008) (see Figure 3.6). The palatometer sensors

Vowels

Short Vowels	p**i**t	p**e**t	p**a**t	p**u**tt	p**o**t	p**u**t	**a**noth**er**		
	ɪ	e	æ	ʌ	ɒ	ʊ	ə ə		
Long Vowels	b**ea**n	b**ar**n	b**or**n	b**oo**n	b**ur**n				
	iː	aː	ɔː	uː	ɜː				
Diphthongs	b**ay**	b**uy**	b**oy**	n**o**	n**ow**	p**eer**	p**air**	p**oor**	
	eɪ	ai	ɔɪ	əʊ	aʊ	ɪə	eə	ʊə	

Consonants

Plosives	**p**in	**b**in	**t**in	**d**in	**k**in	**g**um			
	p	b	t	d	k	g			
Affricates	**ch**ain	**J**ane							
	ʧ	ʤ							
Fricatives	**f**ine	**v**ine	**th**ink	**th**is	**s**eal	**z**eal	**sh**eep	mea**s**ure	**h**ow
	f	v	θ	ð	s	z	ʃ	ʒ	h
Nasals	su**m**	su**n**	su**ng**						
	m	n	ŋ						
Approximants	**l**ight	**r**ight	**w**et	**y**et					
	l	r	w	j					

Figure 3.5: Symbols for English transcription (adapted from Roach (2002))

run along the x-axis and the time epochs run along the y-axis from the top to the bottom. The Space-Time plots for all fifty words can be found in Appendix B.

The palatometer data was standardized by ensuring that each word recorded started when the first tongue-palate contact was made. The length of time of the recordings was also standardized, with all words being zero-padded to be the same size as the largest word in all the datasets. Thus all the words are 118x140 pixels (118 being the number of palatometer sensors, 140 being the number of time epochs).

The order of the palatometer sensors was investigated, as a number of the feature extraction algorithms tend to perform better on images that consist of only one "'object"' rather than a number of fragmented "'objects"'. However, it was found that the ordering as devised by CompleteSpeech was as close to optimal as could be expected. By reordering the sensors, the space-time plots tended to become more fragmented, thus increasing the difficulty in feature extraction.

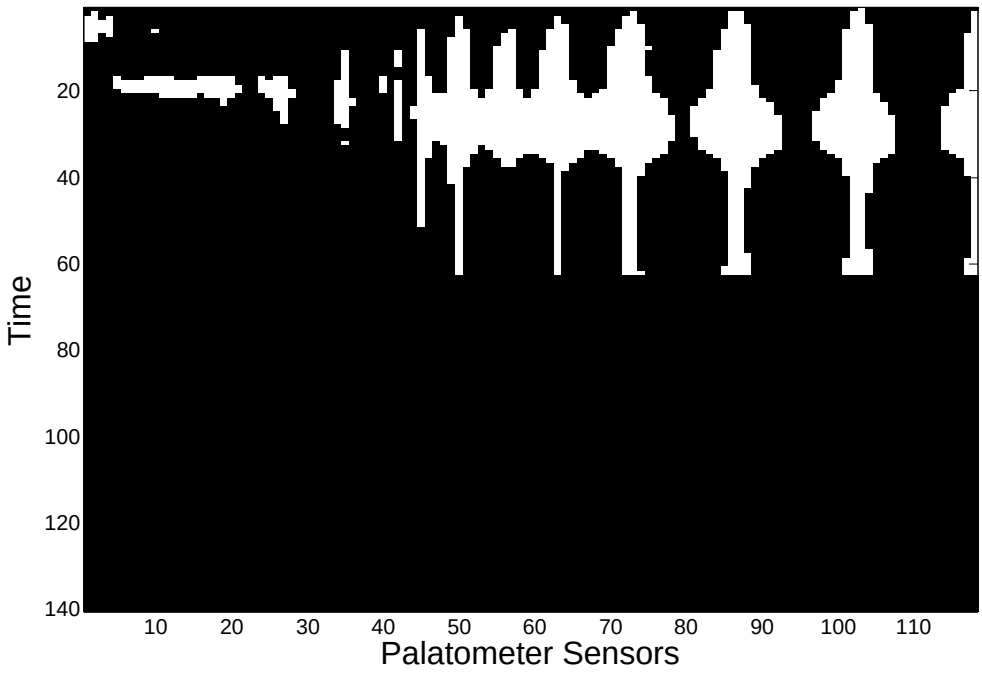


Figure 3.6: Standardized Space-Time plot for the word "many"

3.3.5 Word Variance

The cases of each word are not identical. Even though the tongue palate contact pattern is fairly standard for each word, there is still variability in the different cases of the same word. This is due to how quickly the word is said as well as how precisely it is enunciated. The variance for the word 'many' can be seen in Figure 3.7. The white pixels show which pixels (and thus sensors) never vary between cases of the word, while the varying shades of gray show how often other sensors are activated (the closer to white they are, the more frequently they are activated). This variance increases the difficulty of accurate word recognition.

3.3.6 Audiovisual Synchronization

A potential problem with the newly proposed larynx is that the user will say a word, a noticeable amount of time will elapse and only then will the word be emitted. However this does not take account of the plasticity of the listener's brain. It has

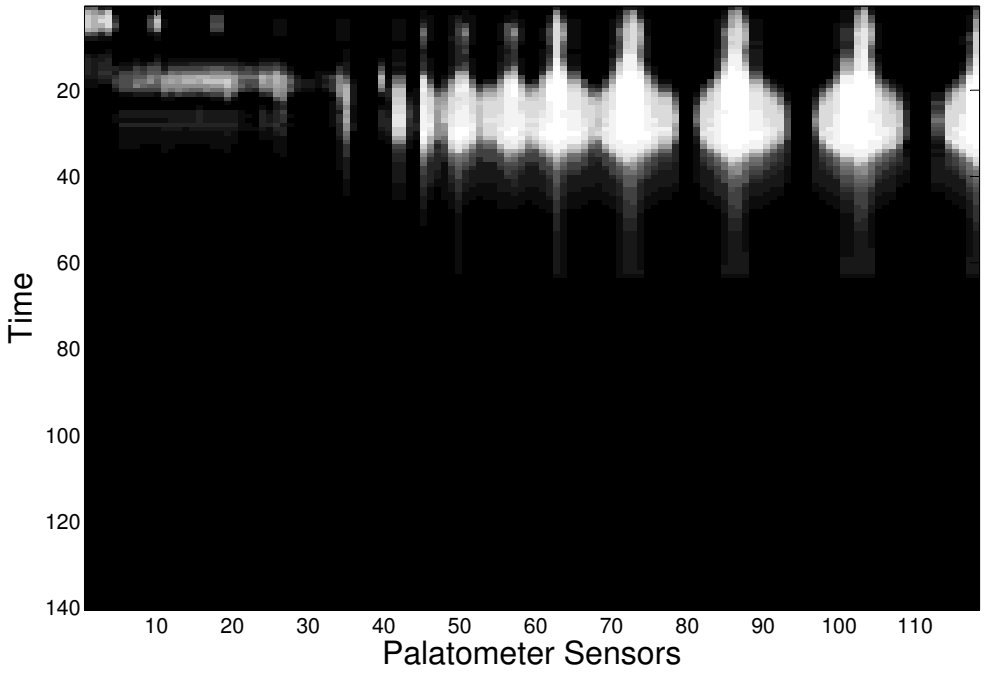


Figure 3.7: Space-Time plot showing the variance in sensor activation for the word "many"

been shown by a number of researchers (for example McGrath & Summerfield (1985); Navarra et al. (2009, 2005); van Wassenhove et al. (2007)) that a window of up to 300ms exists wherein the listener will not realize that the visual and auditory signals are not perfectly synchronous. It is hoped that by refining the new artificial larynx (possibly using hardware to implement the neural networks), the processing time can be decreased to well within this window.

However, if the word takes longer than 300ms to say there still might be a perceptible gap between the listener seeing the start of the word and hearing the start of the word (as in the current method the system can only recognize the word once it has been spoken in full).

3.4 Machine Learning for Word Identification

Pattern recognition is a vast topic with many different techniques and methods for achieving results. Machine learning is just one such section, however it is unique

in that the model for predicting the output 'learns' from a set of training data. It tweaks the parameters of the model until the model associates the training data with the correct output (Bishop, 2006). Two different types of machine learning were used, namely the Multilayer Perceptron Neural Networks and Support Vector Machines.

3.4.1 The Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP) is a popular neural network for classification problems because of its simplicity, scalability and adaptivity (Trenn, 2008). According to Li & Meng (2009) the MLP has many advantages over other classifiers due to its generalization ability, robust performance and need for less training data. An MLP is a feedforward neural network with an input layer, a hidden layer and an output layer (see Figure 3.8). There is full connectivity between the input and hidden layers and the hidden and output layers (Nabney, 2002).

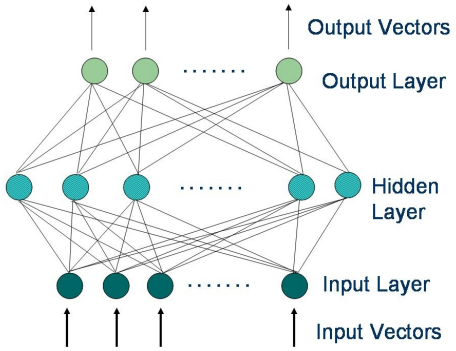


Figure 3.8: The basic layout of an MLP neural network (Adapted from Morgan & Bourlard (1995))

Recent use of MLPs for image classification can be seen in Li & Meng (2009) who use an MLP and SVM to classify colour and texture features in endoscopy images. The MLP provided very good results with an accuracy of around 90%, however the SVM provided such poor results that the authors did not even include them in the article. Garcia et al. (2009) used neural networks (a multilayer perceptron, a radial basis function and a support vector machine) in their classification of retinal images of patients with diabetic retinopathy. All three of the neural networks achieved a mean sensitivity of 100% (proportion of test data having diabetic retinopathy classified as having diabetic retinopathy). The multilayer perceptron acheived a mean specificity (proportion of test data without diabetic retinopathy classified as not having diabetic

retinopathy) of 97.01% while the radial basis function and support vector machines acheived 81.48% and 77.78% respectively. de Albuquerque et al. (2009) used MLPs and self-organizing map topologies to analyze and segment micro-structural elements in metallographic images. The MLP consistently outperformed the self-organizing map in its ability to correctly segment images and its robustness to errors. Diaz et al. (2009) used an MLP and a SVM to classify erythrocytes infected with malaria in light microscopy images. The SVM was the best performing classifier and identifier of the infection stage (although the MLP also performed well with an effectiveness of over 92%). Automatic identification of infected erythrocytes showed a specificity of 99.7% and a sensitivity of 94. The infection stage was determined with an average sensitivity of 78.8% and average specificity of 91.2%. Selver et al. (2008) use an MLP and a K-means based classification system to automatically identify the liver in computed-tomography angiography images. These are combined with a data-dependent and automated switching mechanism that decides which method to apply to which image (the MLP is used for atypical liver shapes). The MLP outperforms the K-means based system, however it takes a long time to process, thus the authors created a system which combines the fast running but less accurate K-means system with the slower but more accurate MLP.

The MLP is capable of universal approximation provided that there are enough hidden units and that the weights and biases are chosen effectively Nabney (2002).

The definition of a two-layer MLP can be given as follows (Nabney, 2002):
The input to the network is given as x_i where $i = 1, \ldots, d$. The input layer forms M linear combinations of the input values to produce a set of intermediate activation functions a_j^1 with one of these variables associated with each hidden layer unit:

$$a_j^1 = \sum_{i=1}^{d} w_{ji}^{(1)} x_i + b_j^{(1)} \quad j = 1, \ldots, M \tag{3.1}$$

$w_{ji}^{(1)}$ are the elements of the input layer's weight matrix and $b_j^{(1)}$ are the bias parameters of the hidden layer's units. A non-linear activation function (for example, the *tanh* function) is then applied to a_j^1. The *tanh* function is given by:

$$z_j = tanh(a_j^1) \quad j = 1, \ldots, M \tag{3.2}$$

The second layer weights and biases then transform z_j to give the second layer

activation function a_k^2:

$$a_k^2 = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + b_k^{(2)} \quad k = 1, \ldots, c \tag{3.3}$$

where c is the total number of outputs. Finally the values of a_k^2 are passed through to the output activation function to give output values y_k where $k = 1, \ldots, c$. For classification problems a logistic sigmoidal activation function is used. It is applied to each of the outputs independently, such that:

$$y_k = \frac{1}{1 + exp(-a_k^{(2)})} \tag{3.4}$$

There are many learning algorithms to train feed-forward networks. During training the network is taught to associate an input vector with a particular output vector. The goal of training is to model the underlying process that produced the data in order to make accurate predictions when new data is presented to the network (Nabney, 2002). In back-propagation training, when an input vector is fed into the neural network the generated output vector is compared to the desired output vector. All the relevant weights are then adjusted to ensure that the next time the same input vector is shown to the network the generated output vector will be closer to the ideal one. In this way the error between what is desired and what is generated is slowly decreased until a minimum error is reached or the desired number of training epochs has passed (Orozco & Garcia, 2003). A popular back-propagation method in classification problems is called Scaled Conjugate Gradient (SCG), developed by Moller (1993). In terms of optimization, training a network is equivalent to minimizing an error function (Moller, 1993). If training is successful the network's errors will decrease with each iteration and the network will converge to a stable set of weights (Gonzalez & Woods, 2002). SCG has been shown to be considerably faster than other conjugate gradient methods as well as quasi-Newton methods and standard back-propagation (Moller, 1993).

Regularization

When training an MLP it is essential that the network is not over-trained. When a network is over-trained it fits the training data precisely (often by employing very large weights) but will give poor output for any other data (Bishop, 2006). Thus a

fine balance must be created between training a network enough to get good results and over-training it. This can be done by using a regularization parameter (λ) in the error function used to train the network. The sum-of-squares error function is given by (Bishop, 2006):

$$E = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, \mathbf{w}) - t_n)^2 + \frac{\lambda}{2}||\mathbf{w}|| \tag{3.5}$$

where (x, t) are the training data (input values and target values respectively) and $\mathbf{w}$ are the weights. λ dictates how important the regularization term is as compared with the sum-of-squares error term. As can be seen from the above equation, if the weights of the network get too large (indicating over training) then the error function increases even though the sum-of-squares error term might be very small. Thus the network cannot be over-trained, as once it has reached an optimal weight set it will remain there (if λ is large it will settle at a higher value of error after fewer iterations and vice versa). The choice of value for λ is attained by trial and error. See Figures 3.9, 3.10, 3.11 and 3.12 for illustrations of how as λ increases the error increases but the number of iterations taken to reach a stable error falls. Figure 3.13 shows how the accuracy of the neural network decreases as λ gets too large.

Implementation

Using the Netlab toolbox for MATLAB the MLPs were implemented with the following features:

- **Network Structure:** The Neural Network was a 2-layer feed-forward network. Weights are drawn from a zero-mean, isotropic Gaussian function. The regularization coefficient λ was set as 0.001. The output activation function was logistic.
- **Units:** The number of input units depends on the type of input. The number of hidden units varies depending on the number of input units. The hidden units use a *tanh* activation function. See Appendix D for graphs illustrating how the number of hidden units changes the success rate of the neural network. Six binary outputs are used, as in order to represent fifty different outputs 2^6 binary units are needed. According to a study done by Francis & Kucera (1982) a vocabulary size of 1000 words would allow a reader to understand 72% of all text he/she would read. The basic English introduced by Ogden

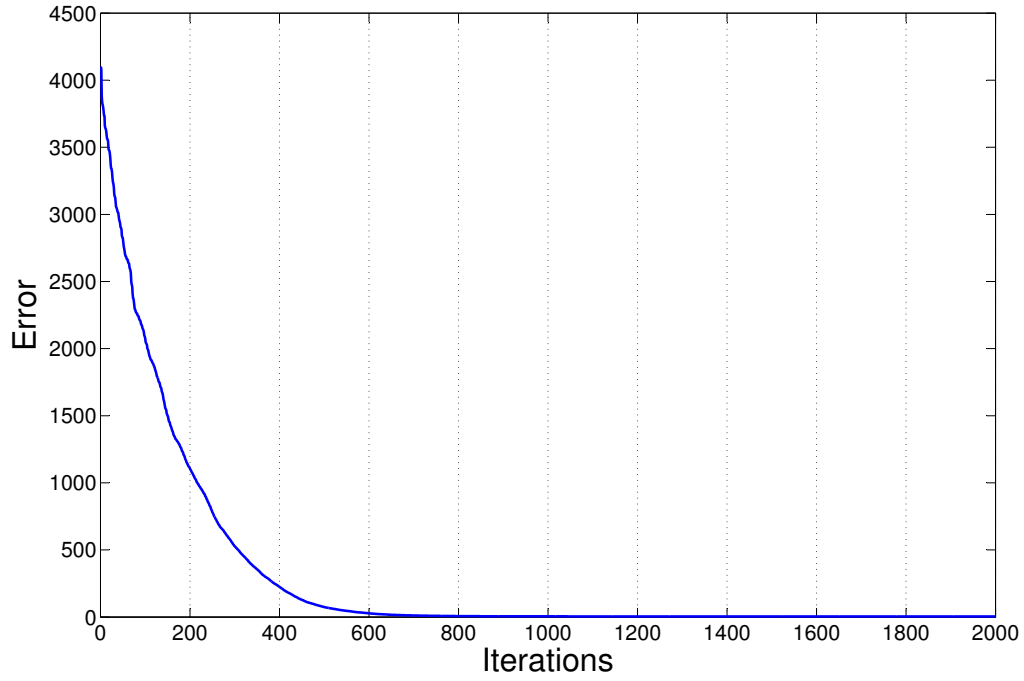


Figure 3.9: Graph showing Error Rate vs. Iterations when $\lambda = 0$

(1937) for foreign speakers contains just 850 core words. Thus a vocabulary of about 2000 words should be sufficient in the final version of the artificial larynx. 2000 words can be encoded by 11 binary outputs which is a feasible output size.

- **Training:** A Scaled Conjugate Gradient optimization algorithm is used for training with a maximum number of iterations of 10 000 to 20 0000.

3.4.2 Support Vector Machines (SVM)

Support Vector Machines are tools for data classification and regression (Maglogiannis & Zafiropoulos, 2004); they map the input vectors into a high-dimensional feature-space using a non-linear transformation (Cortes & Vapnik, 1995). In this space a linear surface can be created that separates the input vectors into two classes (Cortes & Vapnik, 1995). SVMs are an estimation algorithm that separates data into two classes, however they can be adapted for multi-class use (Bishop, 2006; Maglogiannis & Zafiropoulos, 2004). A *one-versus-the-rest* approach was used,

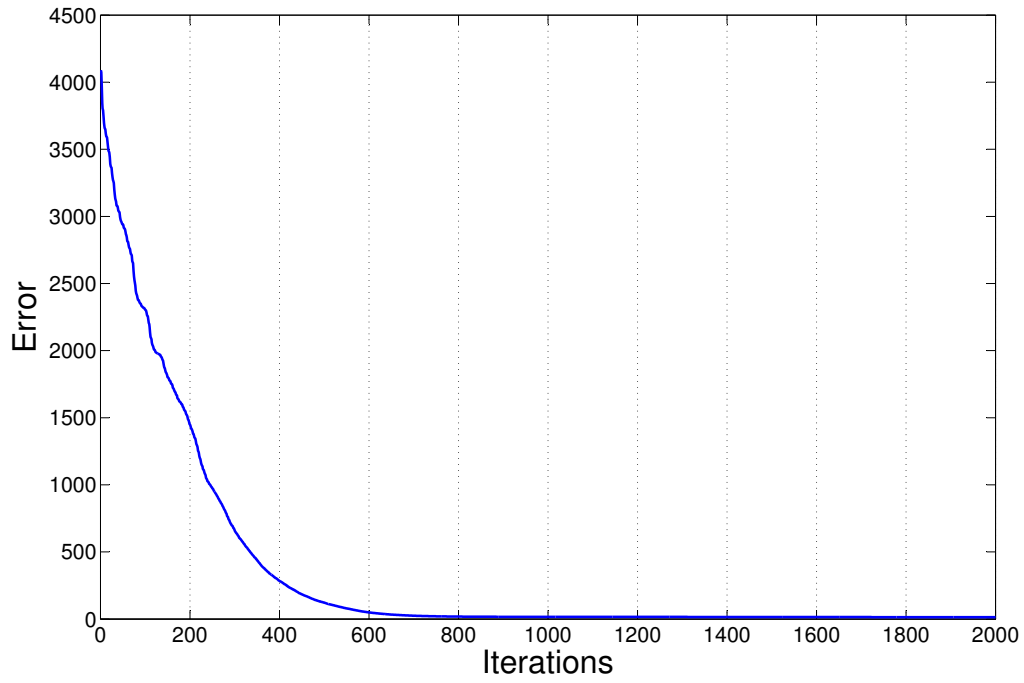


Figure 3.10: Graph showing Error Rate vs. Iterations when $\lambda = 0.001$

where if there are K classes the k^{th} model is trained using the data from its class as the positive training data and the $K-1$ data as the negative training data (Bishop, 2006). When SVMs are compared to other learning techniques they perform well on high dimensional input data and have a high generalization performance (Zhang et al., 2001).

Krishnan et al. (2009) successfully used SVMs to classify normal and pre-cancerous connective tissue cells. Various image features such as eccentricity, compactness, orientation and area were used as input to the SVM for the detection of oral submucous fibrosis (a pre-cancerous condition). The system achieved a sensitivity of 90.47% and a specificity of 87.54%. Tsantis et al. (2009) achieved a good classification rate of malignancy risk using SVMs in ultrasound images of thyroid nodules. Morphological shape properties such as area, smoothness, concavity and symmetry as well as wavelets were used in this investigation. The shape and wavelet features were used as input to two different neural networks (a support vector machine and a probabilistic neural network). The authors then looked at the effect different amounts of speckle had on the classification ability of the neural networks. Before

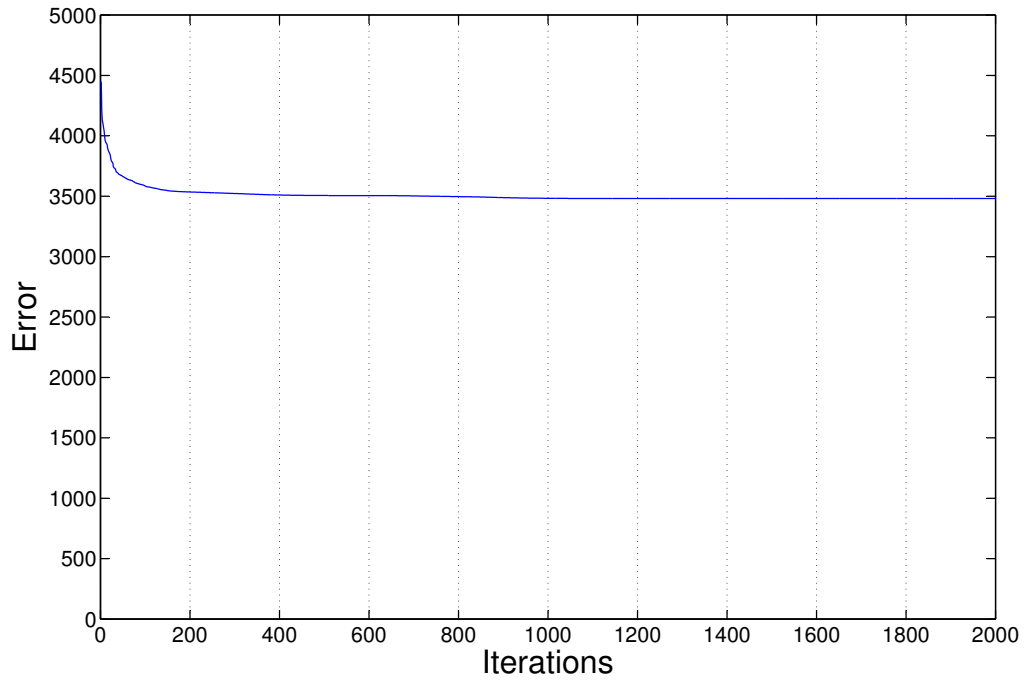


Figure 3.11: Graph showing Error Rate vs. Iterations when $\lambda = 10$

speckle was added the support vector machine had sensitivity and specificity values of 93% and 98% and the probabilistic neural network had sensitivity and specificity values of 96% and 94%. After speckle was added the support vector machine had sensitivity and specificity values of 93% and 96% and the probabilistic neural network had sensitivity and specificity values of 84% and 88%. Thus showing that SVMs are more robust to noise than probabilistic neural networks. da Silva Sousa et al. (2009) used SVMs in their system for automatic lung nodule detection in CT (Computerized Tomography) images, the SVMs were used to reduce the number of false-positives diagnoses produced by the system. The SVMs were used with a number of image features such as geometry, texture, histogram, gradient and spatial. The system had a sensitivity of 84.84% and a specificity of 96.15%. Hotta (2008) attained a higher rate with their system that used an SVM with a local Gaussian summation kernel for face recognition with partial occlusion. In this approach, local kernels are arranged at all local regions of a recognition target and are used in conjunction with an SVM to create a robust face recognition system. Even though this system performed better than systems with global kernels it still did not produce outstanding results, with occluded face recognition being between 70 and 80%.

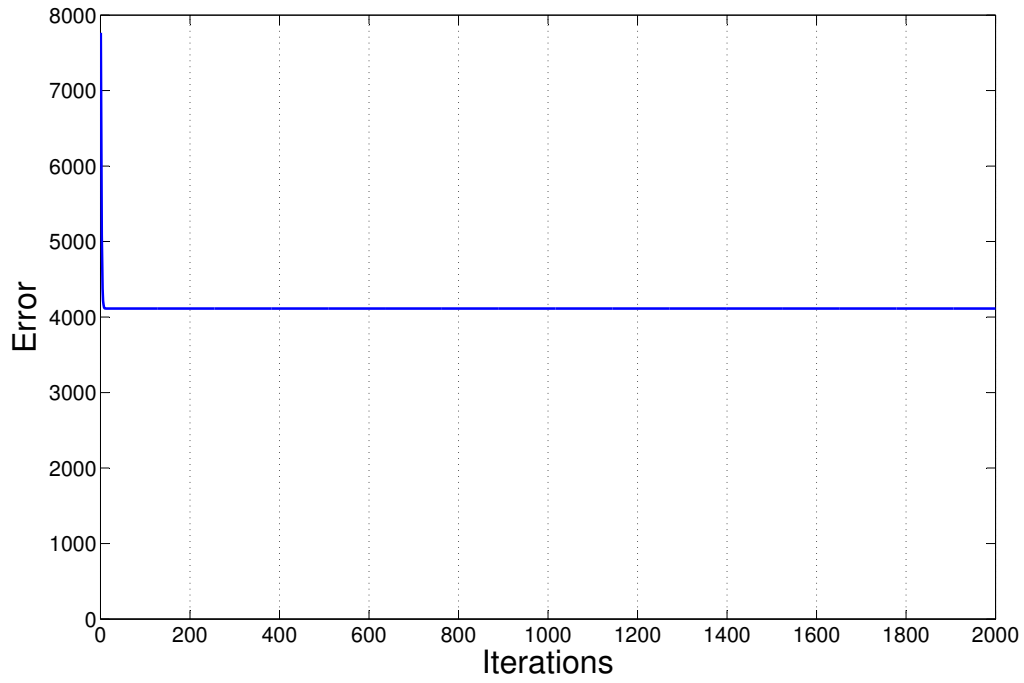


Figure 3.12: Graph showing Error Rate vs. Iterations when $\lambda = 100$

The working of a SVM in a linearly separable case is as follows (Zhang et al., 2001): The general form of a linear classification function is $g(\mathbf{x}) = \mathbf{h} \cdot \mathbf{x} + b$ which corresponds to a separating hyperplane (see Figure 3.14) $\mathbf{h} \cdot \mathbf{x} + b = 0$ where $(\mathbf{x}_i, y_i)_{1 \leq i \leq N}$, $\mathbf{x}_i \in R^d$ are the set of linearly separable training samples and $y_i \in \{-1, 1\}$ is the class label to which $\mathbf{x}_i$ belongs. Normalize $g(\mathbf{x}_i)$ to satisfy $|g(\mathbf{x}_i)| \geq 1$ for all of $\mathbf{x}_i$ so that the distance from the hyperplane to the closest point is $1/||\mathbf{h}||$. To find the optimal separating hyperplane minimize $||\mathbf{h}||$ since the distance to the closest point is $1/||\mathbf{h}||$ thus the objective function becomes:

$$min\phi(\mathbf{h}) = \frac{1}{2}||\mathbf{h}||^2 = \frac{1}{2}(\mathbf{h} \cdot \mathbf{h}) \quad y_i(\mathbf{h} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots, N \tag{3.6}$$

The N non-negative Lagrange multipliers $(\alpha_1, \ldots, \alpha_N)$ associated with the constraints of the equation above allow for the Optimal Separating Hyperplane to be constructed by solving a constrained quadratic programming problem. The solution $\mathbf{h}$ can be expanded as $\mathbf{h} = \sum_i \alpha_i y_i \mathbf{x}_i$ in terms of support vectors (ie. training patterns which lie on the margin). Thus the classification function can be written

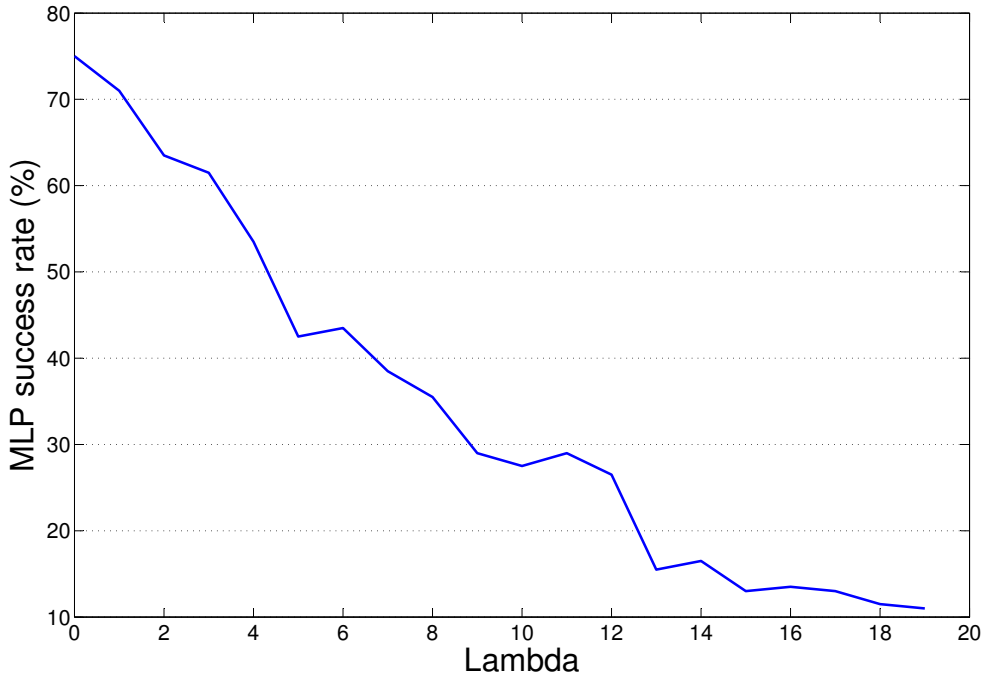


Figure 3.13: Graph showing MLP Success Rate vs. λ

as:

$$f(\mathbf{x}) = sign\Big(\sum_i \alpha_i y_i \mathbf{x}_i \cdot \mathbf{x} + b\Big) \tag{3.7}$$

If the input data is not linearly separable it can be non-linearly mapped into a high-dimensional feature space where the construction of the Optimal Separating Hyperplane takes place. When the dot product satisfies Mercer's condition (i.e. that the Kernel k is a positive definite function (Tian et al., 2007)) it can then be represented by $k(\mathbf{x}, \mathbf{y}) := (\phi(\mathbf{x}) \cdot \phi(\mathbf{y}))$. The final classification function is thus:

$$f(\mathbf{x}) = sign\Big(\sum_i \alpha_i y_i \cdot k(\mathbf{x}_i \cdot \mathbf{x}) + b\Big) \tag{3.8}$$

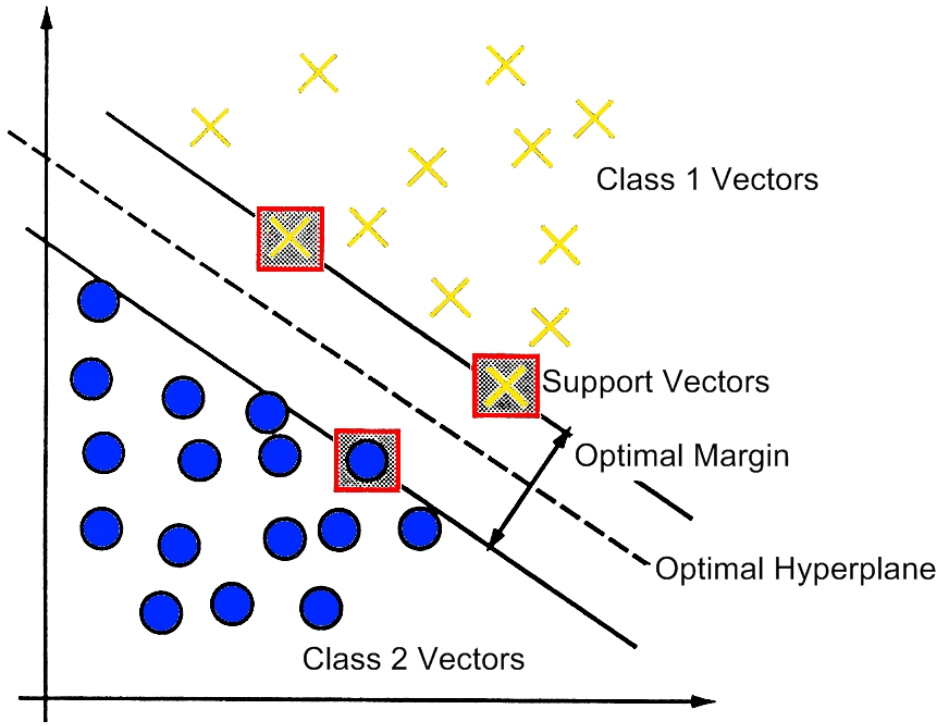


Figure 3.14: A Support Vector Machine with two linearly separable classes (Adapted from Cortes & Vapnik (1995))

Implementation

The Support Vectors Machines were created using the Bioinformatics toolbox in MATLAB. They have the following properties:

- **Kernel Function:** A linear (or dot product) kernel function was used to map the training data into kernel space.
- **Optimal Hyperplane Method:** Quadratic Programming was used thus the classifier is a soft-margin (allows for overlapping class distributions (Bishop, 2006)), two-norm Support Vector Machine.

3.5 Feature Selection to Generate MLP and SVM Inputs

The data from the palatometer takes the form of a stream of binary data containing space and time information (Carreira-Perpinan & Renals, 1998). This large amount

of high dimensionality data cannot be inputted straight into the MLP or SVM. As the data from the palatometer can be viewed as a Space-Time image, image processing techniques can be applied to it to extract the salient features which can then be used as input to the MLP and SVM.

Visually, shape is a discerning feature between the 50 words therefore shape descriptors were investigated. According to Zhang & Lu (2004) shape descriptors can be divided into two main groups, Contour-based and Region-based; these are illustrated in Figure 3.15. The classification is based on how the algorithm extracts data from the shape, i.e. whether it uses only the circumference (Contour-based) or if it uses the whole shape (Region-based). Data reduction techniques were also investigated. Choosing appropriate techniques to investigate was a difficult task as many shape descriptors are designed to be immune to rotational, translation, scale and affine transformations, however in this application these transformations all provide valuable information (Russell et al., 2009*a*). Thus many common shape description techniques were found to be unsuitable.

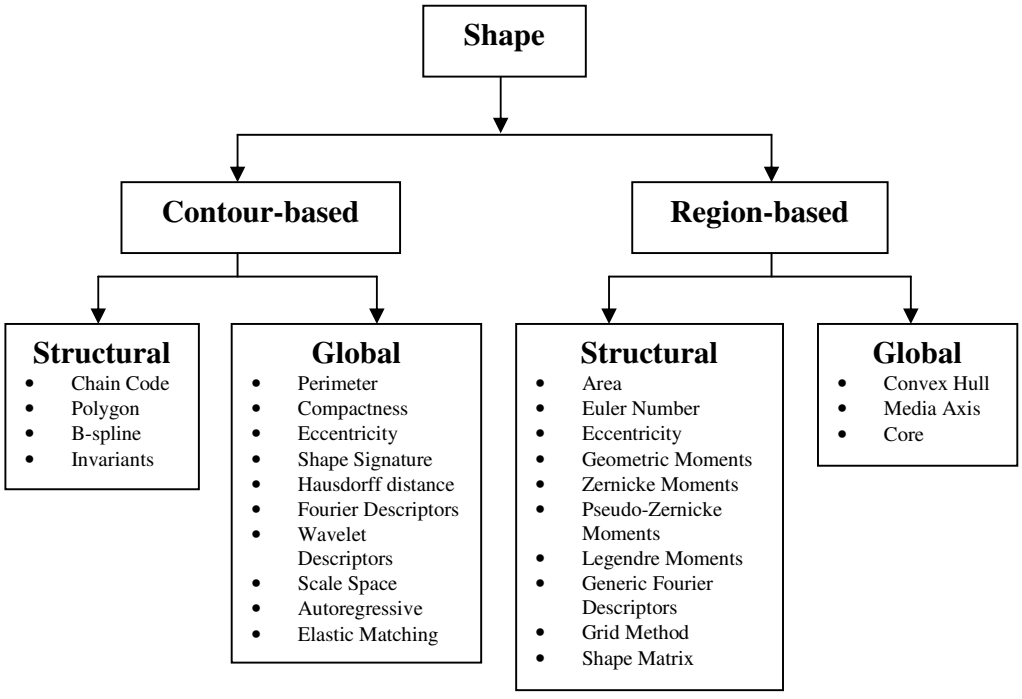


Figure 3.15: Summary of shape classification techniques (Adapted from Zhang & Lu (2004))

3.5.1 Principal Component Analysis (PCA)

Principal Component Analysis is a data reduction technique that is commonly used to find patterns (particularly similarities and differences) in high dimensionality data (Smith, 2006). It provides a method for identifying the important data in a dataset.

Recent usage of PCA is varied across a number of different applications. Babaoglu et al. (2009) successfully used PCA to decrease their dataset of coronary artery disease features from 23 features to 18 features for use with SVMs. They found that the SVMs produced better results when trained with the PCA features than with the original 23 features. Using the PCA features also decreased the training error and the time to run. Lee et al. (2009) used PCA to analyze motion capture data from loaded and unloaded walking. They successfully used PCA to quantify changes in gait on loaded and unloaded subjects. The use of PCA allowed quantification of significant differences that were not apparent in standard temporal analysis. Monasterio et al. (2009) used PCA to analyze multilead ECG (Electrocardiogram) signals in their analysis of T-wave Alternans which is associated with a high incidence of sudden cardiac death. By performing PCA on the ECG signals spatial redundancy was eliminated before feeding the data into a generalized likelihood test. This technique allowed for the detection of T-wave Alternans with a signal-to-noise-ratio of 30dB lower than detected with single lead ECG. Malagon-Borja & Fuentes (2009) used PCA in an interesting technique to reduce the number of false positives in their pedestrian image recognition system. By using the knowledge that PCA can only optimally compress the kinds of pictures that were used to compute the principal components it can then be inferred that images not of a certain type will be poorly compressed using only a few principal components. By quantifying how well the image is reconstructed from a few principal components the image can be assigned to a class. This system reaches classification rates of up to 99.02%. Wagner (2005) performed PCA on palatometer data. However a stream of palatometer data was not considered, this PCA was performed on the maximal contact frame in the phoneme (when the most electrodes were activated) combined with six lip shape measurements. It was found that two principal components were sufficient to describe the data.

To perform PCA the following steps are followed (Smith, 2006):

1. Ensure the variables in the data set have a zero mean.

2. Calculate the covariance matrix for an $n \times n$ matrix C using:

$$C^{n \times n} = cov(Dim_i, Dim_j) \tag{3.9}$$

where Dim_x is the xth dimension of C. Covariance is given by:

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \tag{3.10}$$

where $\bar{X}$ and $\bar{Y}$ are the mean of the dataset in the x or y dimension. The covariance matrix illustrates how much the dimensions in the dataset vary from the mean with respect to each other.

3. Find the eigenvalues and the eigenvectors of the covariance matrix and order them so that the eigenvector with the largest eigenvalue comes first and the eigenvector with the second largest eigenvalue comes second, and so on. This gives a set of the principal components from most important to least important.

4. Choose the first p eigenvectors to form the feature vector.

$$FeatureVector = (eigenvector_1, eigenvector_2, \ldots, eigenvector_p) \tag{3.11}$$

5. The new dataset can now be derived as follows:

$$NewDataset = FeatureVector^T \times dataset^T \tag{3.12}$$

The $NewDataset$ now gives the original data solely in terms of the chosen principal components, with data items in columns and the dimensions in rows.

As the data from the palatometer is binary it can be converted to decimal, thus keeping all the information but reducing the size of the image. PCA was performed on the pure, uncompressed data and on the decimal converted data (14-bit binary strings were converted). For the uncompressed data the first 21 vectors were used. For the decimal compressed data, the first 20 vectors (providing 80% of the variance) were used.

3.5.2 Correlation

Correlation is a measure of how similar an object is to another object. The formula for correlation between an image $f(x, y)$ of size $M \times N$ and a subimage $w(x, y)$ of size $J \times K$ can be given by (Gonzalez & Woods, 2002):

$$corr(x, y) = \sum_s \sum_t f(s, t)w(x + s, y + t) \tag{3.13}$$

for $x = 0, 1, 2, \ldots, M - 1$, $y = 0, 1, 2, \ldots, N - 1$ and the summation is taken over the region where f and w overlap and $J \leq M$ and $K \leq N$.

Gershikov et al. (2007) used correlation to implement a new method for color image compression. This method uses the inter-correlation of the colors in an RGB image to approximate two of the colors as a function of the third. This new correlation based method outperforms the common JPEG image compression method.

Avants et al. (2008) used correlation in their technique to quantify the amount of neural degeneration in brain MRI (Magnetic Resonance Imaging) images in patients with neurological degeneration. This technique uses cross-correlation to normalize and register the images before segmenting out various regions of interest. Correlation techniques were chosen due to correlation's robustness to inconsistent illumination, reflectance and MRI inhomogeneity. This neural degeneration quantification technique performs favorably when compared to other segmenting and quantification techniques. Bing et al. (2009) and Moerman et al. (2009) used correlation to determine changes in images. Bing et al. (2009)found the coefficient of thermal expansion in polymer film by correlating images taken of the film at different temperatures, where as Moerman et al. (2009) used correlation and finite element modeling to determine the mechanical properties of human soft tissue.

In this work templates were made of each of the 50 words. These were created by aligning and averaging the Space-Time images for each of the fifty words (Russell et al., 2009*a*). An unknown word could then be correlated with the templates and either the maximum correlation coefficient from each template could be used as input to the neural network or the number of the template that had the highest correlation coefficient could be used as input.

Classification using correlation and template matching by itself and not as input to the neural networks can also be used.

3.5.3 Fourier Descriptors (FD)

Fourier descriptors are a contour-based technique (Zhang & Lu, 2005). This means that just the boundary of the shape is taken into consideration. Fourier descriptors are a popular choice of shape descriptor due to their stability, clarity of meaning and coarse to fine description ability (Zhang & Lu, 2005).

Menesatti et al. (2009) explored the use of shape matching and FDs to track identification tags in video images. The FD technique performed faster (if slightly less accurately) than the shape matching technique which makes it useful for real time applications. Chen et al. (2009) designed a real time image alignment system for industry using FDs. Using the magnitude and phase information in the FDs an image can be matched to a template image (as FDs allow for rotation invariant matching), then by using a novel phase shift technique the object can be aligned to the same orientation as the object in the template image. Chen et al. (2009) achieved results comparable to commercialized methods but with faster processing times, again showing FD's applicability to real time applications.According to Lestrel et al. (2009) FDs provide an excellent way to precisely quantify the boundary of shapes. Previous to the Lestrel et al. (2009) study gender differences in human skull remains were mainly determined by size however, Lestrel et al. (2009) used FDs to describe six elements of the human skull in order to distinguish gender. FDs were chosen due to their ability to outperform simple angle and ratio morphological features and their ability to normalize for size.

To calculate the Fourier descriptor of a shape, the shape signature must be calculated. A shape signature is a 1-D representation of a 2-D area that captures the perceptual features of the shape (Zhang & Lu, 2005). The shape signature $u(t)$ used in this application is called the complex coordinate and it is created from the boundary coordinates $(x(t), y(t))$, $t = 0, 1, \ldots, N-1$ of the shape (Zhang & Lu, 2005):

$$u(t) = [x(t) - x_c] + i[y(t) - y_c] \tag{3.14}$$

where (x_c, y_c) is the centroid of the shape which is the average of the boundary coordinates:

$$x_c = \frac{1}{N}\sum_{i=0}^{N-1} x(t), \quad y_c = \frac{1}{N}\sum_{i=0}^{N-1} y(t) \tag{3.15}$$

For any 1-D shape signature the discrete Fourier transform $q(n)$ is given by (Zhang & Lu, 2005):

$$q(n) = \frac{1}{N} \sum_{i=0}^{N-1} u(t) exp(-j2\pi nt/N) \quad n = 0, 1, \ldots, N-1 \tag{3.16}$$

This gives a set of Fourier coefficients that describe the shape. The general shape of the object is described by the low frequency coefficients, whereas the high frequency coefficients describe the fine details of the shape (Kunttu et al., 2006). A common technique is to use a subset of the low frequency coefficients for image description. It has been found that 10 Fourier descriptors are sufficient to provide for generic shape description (Zhang & Lu, 2005).

3.5.4 Generic Fourier Descriptors (GFD)

The generic Fourier descriptor can be applied to more general images than the Fourier descriptor detailed above. In most cases GFD are used in image classification and database image retrieval applications. Yadav et al. (2008),Ohbuchi et al. (2003) and Saykol et al. (2005) used GFDs for image retrieval and classification in image databases. While Yadav et al. (2008) found GFDs outperformed by wavelet Zernike moment descriptors in their image retrieval application, Ohbuchi et al. (2003) found that GFD produced the best results in complex 3D shape recognition and retrieval. Saykol et al. (2005) found their histogram based approach to color and shape queries in an image and video database, though faster, did not perform better than GFD. Yu et al. (2007) used Fourier descriptors and generic Fourier descriptors in a different imaging application. They used it to describe particle morphology which is used in the monitoring and control of particulate processes. Neither method outperformed the other, which led Yu et al. (2007) to recommend using both methods in their application.

The GFD $F(u_p, v_q)$ is calculated from the spectral domain by applying a Fourier transform to the polar-raster sampled (see Figure 3.16) image (Yadav et al., 2008):

$$GFD(\rho, \phi) = \sum_r \sum_i f(r, \theta_i) exp[-j2\pi(\frac{r}{R}\rho + \frac{\phi}{T})] \tag{3.17}$$

where $0 \leq r < R$ and $\theta_i = i(2\pi/T)(0 \leq i < T)$, $0 \leq \rho < R$, $0 \leq \phi < T$. R and T are the radial and angular resolutions and $f(x, y)$ is the binary shape function. GFD

has been shown to outperform other shape descriptors such as Zernike moments, geometric moments and Fourier descriptors (Zhang & Lu, 2004).

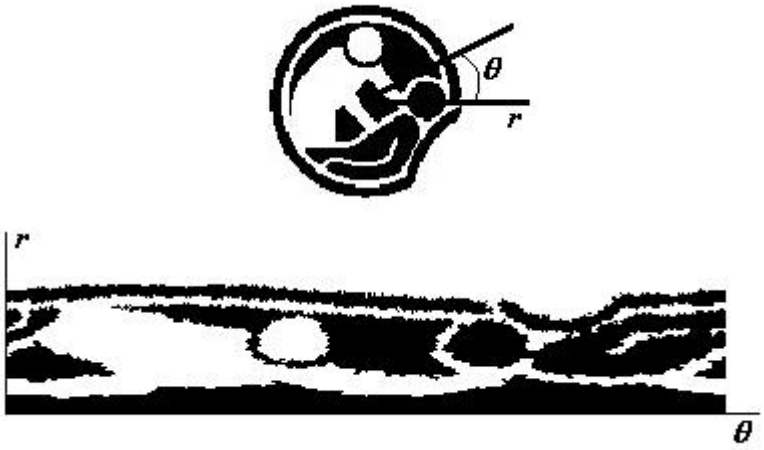


Figure 3.16: An example of an image (top) which has been polar-raster sampled (bottom) (Adapted from Zhang & Lu (2003))

3.5.5 Image Properties

There are a large number of morphological image properties that can be used for image classification. The variety of these properties means that they are suitable for describing images and objects in a large number of unrelated fields. Naqa et al. (2009) used shape properties such as eccentricity, Euler number, solidity and extent as part of their investigation of a system for predicting cancer treatment outcomes from PET images. They found that the shape properties provided relevant information in the case of head and neck cancer, while other features such as texture proved more relevant in the case of cervical cancer. They thus recommended a multi modal system. According to Helmuth et al. (2009), the size and shape of cell organelles give invaluable information about their function. Thus accurate shape information is of great interest to researchers. Helmuth et al. (2009) used area, eccentricity, and concavity in their work on the shape reconstruction of cell organelles from microscopy images. Their method worked well, even in images with a low signal to noise ration, however the curvature of outlines was consistently underestimated.

A number of investigations into systems that will diagnose various human ailments also use morphological image properties. Garcia et al. (2009) used a number of color

and shape features (such as region size and compactness) as input to various neural networks (a multilayer perceptron, a radial basis function and a support vector machine) in their classification of retinal images of patients with diabetic retinopathy. All three of the neural networks achieved a mean sensitivity of 100% (proportion of test data having diabetic retinopathy classified as having diabetic retinopathy). The multilayer perceptron acheived a mean specificity (proportion of test data without diabetic retinopathy classified as not having diabetic retinopathy) of 97.01% while the radial basis function and support vector machines acheived 81.48% and 77.78% respectively.Krishnan et al. (2009) used eccentricity, compactness, orientation and area as image features for the detection of oral sub-mucous fibrosis (a pre-cancerous condition). These image features were used as input to a support vector machine. The system achieved a sensitivity of 90.47% and a specificity of 87.54%. Tsantis et al. (2009) used morphological shape properties such as area, smoothness, concavity and symmetry as well as wavelets in a thyroid nodule evaluation system. The shape and wavelet features were used as input to two different neural networks (a support vector machine and a probabilistic neural network). The authors then looked at the effect different amounts of speckle had on the classification ability of the neural networks. Before speckle was added the support vector machine had sensitivity and specificity values of 93% and 98% and the probabilistic neural network had sensitivity and specificity values of 96% and 94%. After speckle was added the support vector machine had sensitivity and specificity values of 93% and 96% and the probabilistic neural network had sensitivity and specificity values of 84% and 88%.

The following image properties were used on the Space-Time images (Russell et al., 2009*a*):

- **Area:** The area of an image is a measure of how many pixels are in the region (Gonzalez & Woods, 2002).
- **Euler Number:** This is defined as the number of connected regions in an area subtracted by the number of holes in the area (Gonzalez & Woods, 2002).
- **Centroid:** This refers to the coordinates for the center of mass of the object. Thus there is a horizontal (x-axis) centroid and a vertical (y-axis) centroid.
- **Major and Minor Axis Length:** The major axis length is the pixel length of the major axis of the ellipse that has the same normalized second central moments as the region. The minor axis length is the minor axis length of the same ellipse (Mathworks, 2008).

- **Eccentricity:** This is the ratio of the major axis length to the minor axis length (Gonzalez & Woods, 2002).
- **Orientation:** This is the angle (in degrees) between the x-axis and the major axis (Mathworks, 2008).
- **Convex Area:** This is the number of pixels of the smallest convex polygon the region can fit into (this is known as the convex hull) (Mathworks, 2008).
- **Filled Area:** This is the number of pixels in the image with all the holes filled (Mathworks, 2008).
- **Equivalent Diameter:** This is the diameter of a circle with the same area as the region (Mathworks, 2008).
- **Solidity:** This is the number of pixels in the convex hull that are also in the region. It is calculated by dividing the Area by the Convex Area (Mathworks, 2008).
- **Extent:** This is calculated by dividing the Area of the region by the area of the bounding box (the smallest box which contains the whole region) (Mathworks, 2008).

The above image properties were used in various combinations or all together to provide input to the Neural Networks (Russell et al., 2009*a*).

3.5.6 SVM Specific Input

Due to the ability of Support Vector Machines to employ high-dimensional data for classification the following techniques were also implemented as input:

- **Sum of Space-Time image columns:** How long each sensor was activated.
- **Sum of Space-Time image rows and columns:** How long each sensor was activated and how many sensors were activated at the same time.
- **Sum of Space-Time image columns and the length of the word:** How long each sensor was activated and how long the word took to say (number of rows).

3.5.7 Combinations of Inputs used with MLP and SVM

The features described above are used in a number of different combinations and permutations as inputs to either the MLP, SVM or both (see Table 3.2). Note that when 13 image properties are used, the full list as described in 3.5.5 is used. However when four image properties are used this means that only the Area, Euler number, X center of mass and Y center of mass are used.

Table 3.2: Which image features are used as input to the MLP and SVM

Image Feature (Input to MLP or SVM)	MLP	SVM
Principal Component Analysis	✓	✓
PCA and correlation number	✓	✓
Fourier descriptors	✓	✓
Fourier descriptors and correlation number	✓	✓
Fourier descriptors and 4 Image Properties and Correlation Number	✓	✓
Correlation against templates (input -1 to 1)	✓	✓
Correlation against templates (input 0 to 1)	✓	✗
Correlation against templates (abs input)	✓	✗
4 Image Properties and Correlation Number	✓	✓
13 image properties	✓	✗
13 image properties and correlation number	✓	✗
Generic Fourier Descriptors - centered at center of mass	✓	✗
GFD not centered	✓	✗
Column Sums of each word	✗	✓
Column and row sums of each word	✗	✓
Column sums and length of each word	✗	✓

3.6 Voting Systems to Increase Word Classification Rate

Team decisions are usually better than individual decisions, and the same applies to neural networks (Battiti & Colla, 1994). It is possible to increase the performance of a classification system by creating a voting system, which will perform better than the best classifier in the system (Battiti & Colla, 1994). In classification, the outputs of the different neural networks are often combined in a voting scheme or committee. The committee (of M members) assigns the pattern to the class that has the maximum number of votes (Tresp et al., 2001):

$$\widehat{class}(x) = arg\ max_j \sum_{i=1}^{M} g_i f_{i,class=j(x)} \tag{3.18}$$

where $f_{i,class=j(x)}$ is the output of the classifier i for class j. By using a committee approach it has been shown that the performance of the committee will be better than, or equal to the average performer in the committee (Tresp et al., 2001). Thus the squared error between the committee prediction $\hat{t}$ and the true but unknown target t for a committee member f_i, is (Tresp et al., 2001):

$$(\hat{t} - t)^2 \leq \frac{1}{M} \sum_{i=1}^{M} (f_i - t)^2 \tag{3.19}$$

A committee of neural networks can be applied to many different scenarios and situations. Nanni et al. (2010) used a fusion of two neural network designs (a LevenbergMarquardt neural network and a variant of the Ada-Boost) on a cell phenotype database and achieved excellent classification results (97.5% accuracy). Reddy & Buch (2003) used five of the best performing neural networks in a committee for use in speaker verification. Using a winner-takes-all approach Reddy & Buch (2003) achieved a 100% success rate on their datasets even though in the majority of the decisions the result was not unanimous. Bogdanov (2008) used a committee consisting of neural (attractor dynamics algorithm) and non-neural (classifier masking algorithm) based classifiers to classify sea-ice imagery. The results of Bogdanov (2008) show that training and combining of the individual classifier outputs in the committee significantly improve the robustness and the error tolerance of the classification system as compared to a single classifier.

Research on different types of committees as well as the properties of the committees is ongoing. For example Muhlbaier et al. (2009) introduced a new algorithm (called

Learn++.NC) for a committee to learn new classes with fewer classifiers. In this new algorithm individual classifiers consulted with each other to determine which classifier was the most qualified to classify a given instance as well as how important each classifiers decision is.

In this application, a voting system allows for the rejection of unknown words thus increasing the accuracy of the system.

3.6.1 Voting System 1 (Winner-takes-all)

To create a higher success rate, a voting system was introduced into the artificial larynx system. This system combines the top three classification performers into a winner-takes-all panel which rejects outputs that have a very low probability of being correct (Russell et al., 2009*b*). An unknown word is presented to the three top performing classifiers; if two or more of the systems agree with each other about the output, the word classification is accepted, otherwise the word classification is rejected (Russell et al., 2009*b*). See Figure 3.17 for the outline of this system.

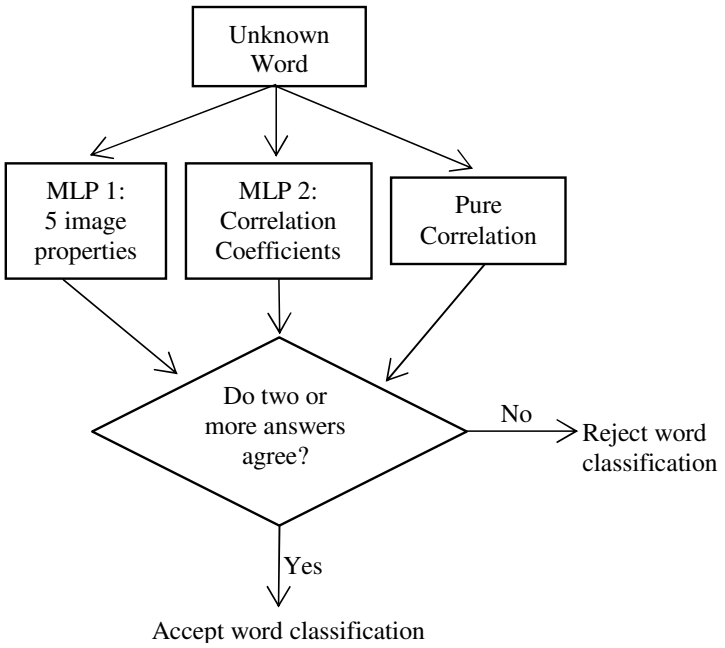


Figure 3.17: The Winner-takes-all Voting System (Russell et al., 2009*b*)

3.6.2 Voting System 2 (Grammar Prediction)

The application of this system deals with classifying unknown words (rather than pictures or numbers); thus predictions can be introduced into whether a certain word can follow another word. A unique grammar (see Figure 3.18 and Appendix C) was generated for the 50 words used in this application, and this was used to further increase the accuracy of the voting system (see Figure 3.19 (Russell et al., 2009*b*)). If the word order does not follow the grammar rules then the word is rejected.

Word 1 can be followed by words:
10, 12, 14, 19, 21, 24, 25, 26, 27, 29, 32, 36, 37, 38, 39,
40, 41, 42, 44, 45, 46, 47, 48, 49, 50

Word 2 can be followed by words:
1, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 30, 21,
22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37,
39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

Word 3 can be followed by words:
1, 2, 4, 6, 10, 13, 14, 18, 19, 21, 24, 25, 27, 38, 34, 35, 37,
40, 41, 42, 44, 46
45, 46, 47, 48, 49, 50

.....etc

Figure 3.18: An example of the grammar used (Russell et al., 2009*b*)

3.6.3 Voting System 3 (Bit-by-bit)

The previous voting systems looked at the entire 6-bit output from each of the three classifiers as a whole. Another approach is for the voting to occur on a bit-by-bit basis. This is illustrated in Figure 3.20. Each of the classifiers outputs a 6-bit result, each of the bits is then compared and voted upon.

3.7 Speech Synthesis

Speech synthesis is the automatic generation of speech and is often the final step in a Text-to-Speech (TTS) application (O'Shaughnessy, 2003). Most speech synthesizers are focused on creating a 'reader' which can accurately read out text with correct intonation and prosody. However, in this application, it would be optimal if the new voice created for the laryngectomy patient was identical (or at least very similar) to his/her pre-laryngectomy voice. Thus a technique called voice morphing (or voice conversion) was used, in which a subject's voice can be transformed into a target

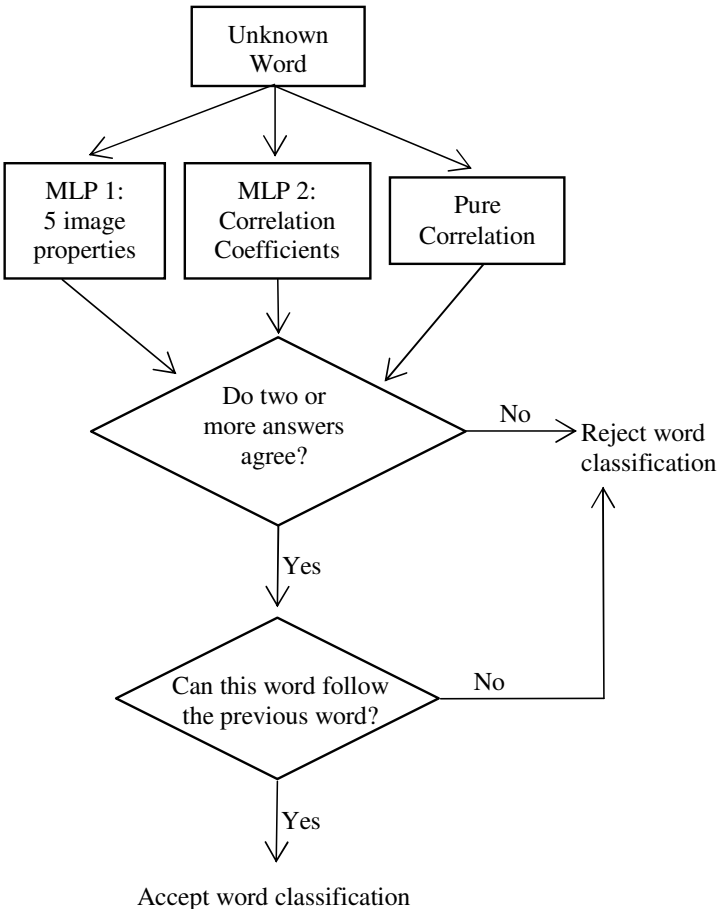


Figure 3.19: The grammar prediction voting system (Russell et al., 2009*b*)

voice. If the patient has a recording of his/her voice made before the laryngectomy or preferably before their voice quality deteriorates due to disease, application of this technique should be able to restore their pre-laryngectomy voice.

3.7.1 Voice Morphing

Voice morphing has many applications, such as dubbing movies into different languages while still retaining the original actors voices, as well as personalization of Text-to-Speech systems in email and language teaching software (Duxans & Bonafonte, 2003). Felps et al. (2009) used a voice morphing system to help with foreign language pronunciation. In this application the voice the learner imitates is the

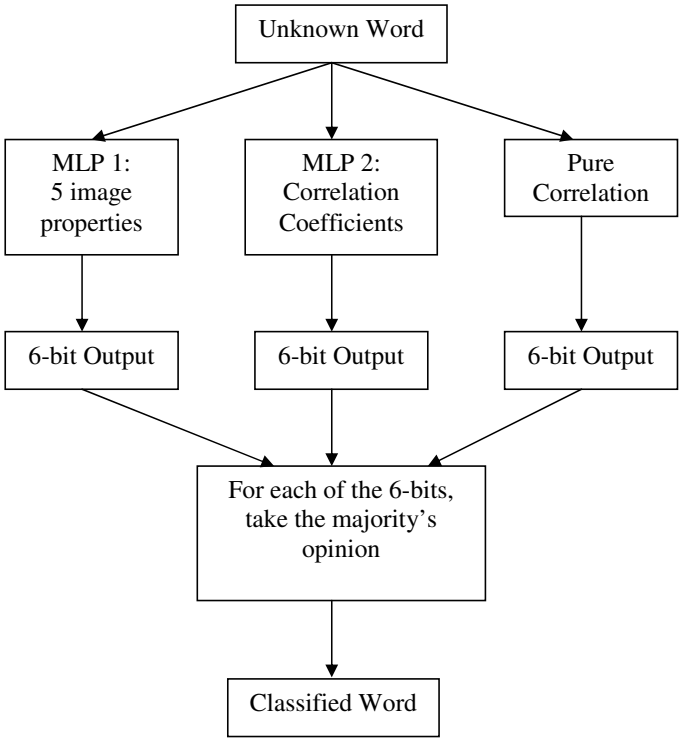


Figure 3.20: The Bit-by-Bit voting system

learner's own voice with a foreign accent. Rao (2009) used neural network techniques to develop a new type of voice morphing algorithm. Neural networks were developed to make mapping functions for each level of speaker characteristics introduced along the vocal tract. For example the shape of the glottal pulse (which has features of the excitation source), the shape of the vocal tract (which introduces vocal tract characteristics) and long-term characteristics (such as prosody). This method outperformed prior methods of vocal characterization

The use of voice morphing in an artificial larynx has not been explored before this research.

A person's voice contains a lot of information that is unique to that person, such as spectral characteristics, prosodic characteristics (phone duration, pitch), lexical and syntactic properties (Duxans & Bonafonte, 2003). One of the main problems in voice morphing is to find the minimum set of features that describe a voice.

Speech Capture and Processing

To perform voice conversion a speech library is needed. This was created by recording a female and a male, each saying the 50 words of the dataset as well as 24 phonetically rich TIMIT sentences (UPenn, 2008) (The TIMIT sentences can be found in Appendix E). The TIMIT sentences were created to contain all the English phonemes. These were recorded using a program called *ProRec - Prompt and Record Version 1.2* (Huckvale, 2007) which allows the user to create applications that will display and record text.

Once the recordings were made, each of the 50 words and the TIMIT sentences had to be individually processed using the SFS program (Huckvale, 2008) to extract the timing of the pitch marks (also known as pitch epochs). Pitch marks are related to the instants of glottal closure in the speech cycle, which is a moment of significant excitation of the vocal-tract system during production of speech (Murty & Yegnanarayana, 2008). Pitch markings are used to calculate a number of characteristics of the speech waveform, such as the fundamental frequency and the frequency response of the vocal tract (Murty & Yegnanarayana, 2008). An example of a speech waveform with its associated pitch markings is shown in Figure 3.21.

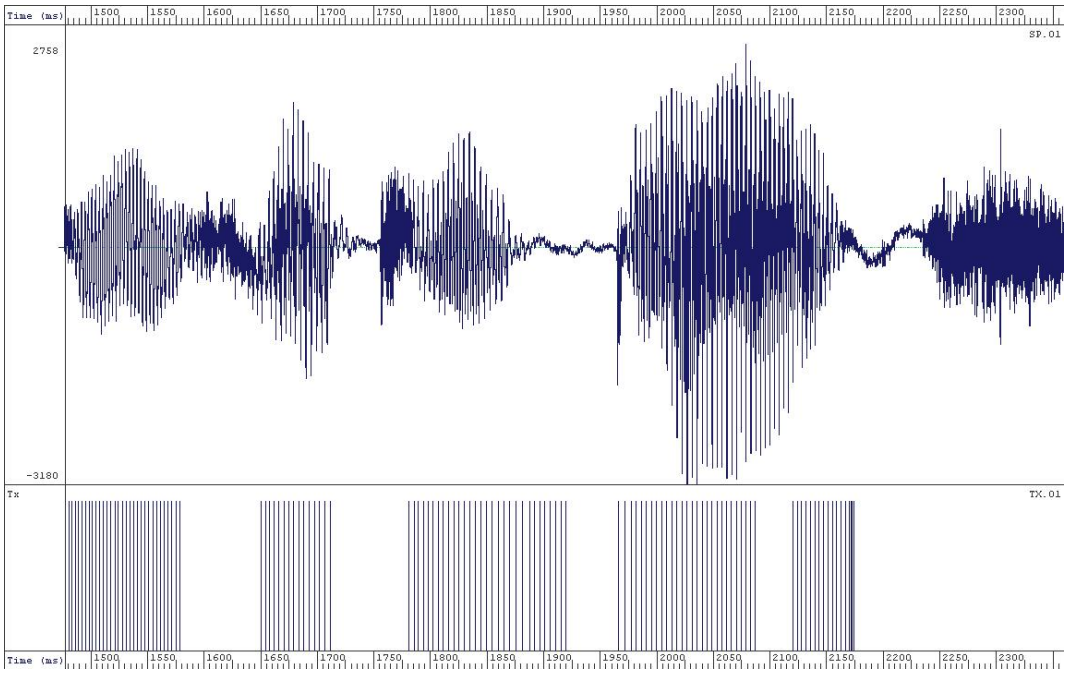


Figure 3.21: Screen shot from the SFS software showing a speech waveform on top and its' associated pitch marks on the bottom

The speech file in a .wav format, as well as a text file containing the timing of the pitch marks, is required for each of the 50 words and TIMIT sentences. These are needed in order to use the MATLAB toolbox for voice morphing (Erro, 2008).

Voice Morphing Process

The MATLAB voice morphing toolbox (Erro, 2008) was used to perform the following:

1. **Harmonic/Stochastic Model (HSM):** First the audio files are analyzed according to a Harmonic/Stochastic Model. Harmonic/stochastic models of speech represent speech as a sum of harmonically related sinusoids (with parameters that vary over time), as well as a noise-like component (Banos et al., 2008). Voiced speech segments contain harmonic components whereas unvoiced segments (such as breath sounds) are represented by stochastic components (Banos et al., 2008). The voiced segments can be broken down frame by frame into the fundamental frequency and phases and amplitudes of the harmonics (Banos et al., 2008). Unvoiced segments are represented by the coefficients of an IIR (Infinite Impulse Response) filter (Banos et al., 2008). The pitch marks files are necessary for this section of the voice morphing process.

2. **Training of the Conversion Function:** An option is present in the toolbox to use parallel or non-parallel recordings to train the conversion function. The method for non-parallel recordings was chosen for the application (thus allowing for any sample of the user's pre-laryngectomy voice to be used). The voice conversion function is found using Gaussian Mixture Models (GMM). GMMs are probability density functions built as the weighted sum of n Gaussian components (Eslava, 2008). The whole acoustic space of the vocal tract can be represented by GMMs. The transformation function is calculated by minimizing the acoustic distance between the source voice and the target voice (Erro & Moreno, 2007).

Once the transformation function has been calculated, any utterance by the source speaker can be converted into a format that will make it sound like it came from the target speaker. A full description of the techniques and algorithms used for voice morphing can be found in Banos et al. (2008); Duxans & Bonafonte (2003); Duxans et al. (2006); Erro & Moreno (2007); Eslava (2008).

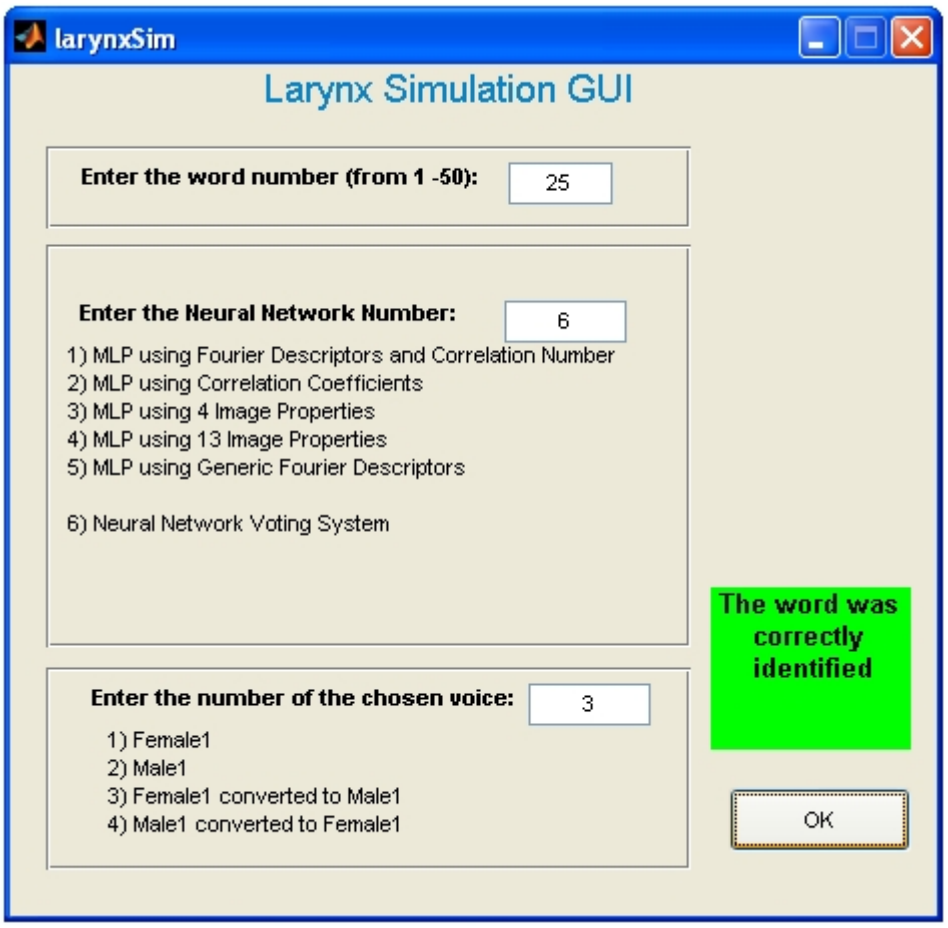
3.8 The Artificial Larynx Simulator

The aim of the simulator (see Figure 3.22) is to allow for easy demonstration of the work done in this research. As the palatometer cannot directly interface with MATLAB, a live demonstration of the artificial larynx cannot be done. Thus the simulator brings together all the parts of the artificial larynx (the feature extraction, neural network classification, voice morphing and word synthesis) in a real time situation. It allows the user to chose a word as the input and then hear the word outputted in the voice of their choice. The top performing MLPs as well as the top performing voting system perform classification of the word chosen by the user, thereby allowing the user to get a sense of how accurate each system is and how long each system takes. The option of using one of the SVMs to classify the word is not given as only the top performing systems were chosen and all of the SVM systems performed poorly.

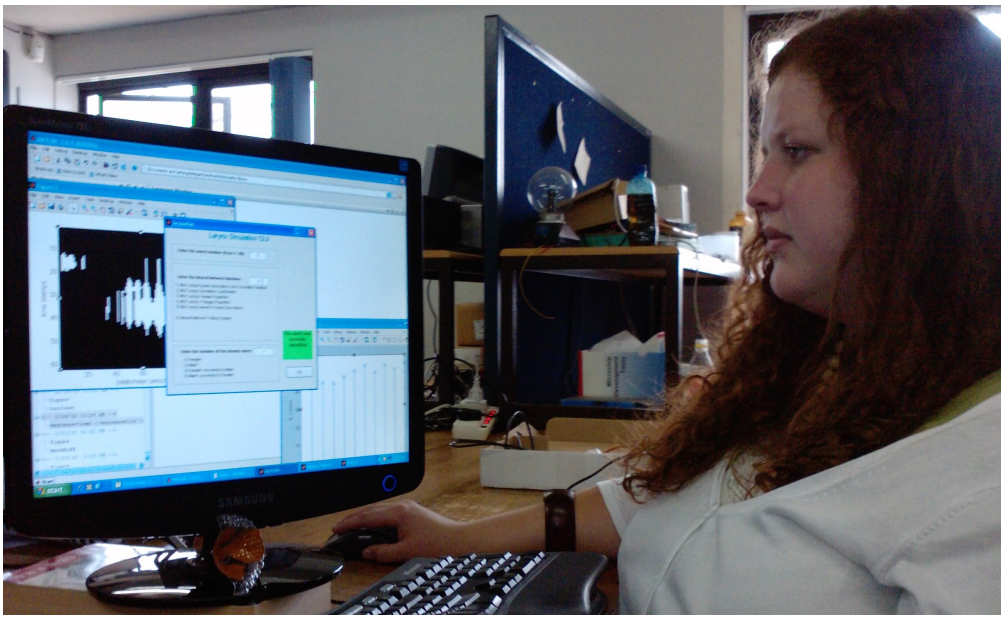
The simulator is used in the following way:

1. The user first enters a number from one to fifty, representing the word spoken by the patient. The palatometer information for this word is randomly chosen from the four cases of each word in the testing database.
2. The user then chooses which image features to use as input to an MLP. The top performing voting system is also an option.
3. The user chooses the output voice. Female1 and Male1 are unchanged, recorded voices. Voice morphing options are also given (Female1 converted to Male1 and Male1 converted to Female1).
4. The output from the Larynx Simulator is the word spoken in the chosen voice. An indicator panel shows whether or not the word has been correctly identified.

An outline of the final system can be seen in Figure 3.23.



(a) Screen shot of the GUI



(b) The simulator in use

Figure 3.22: The artificial larynx simulator.

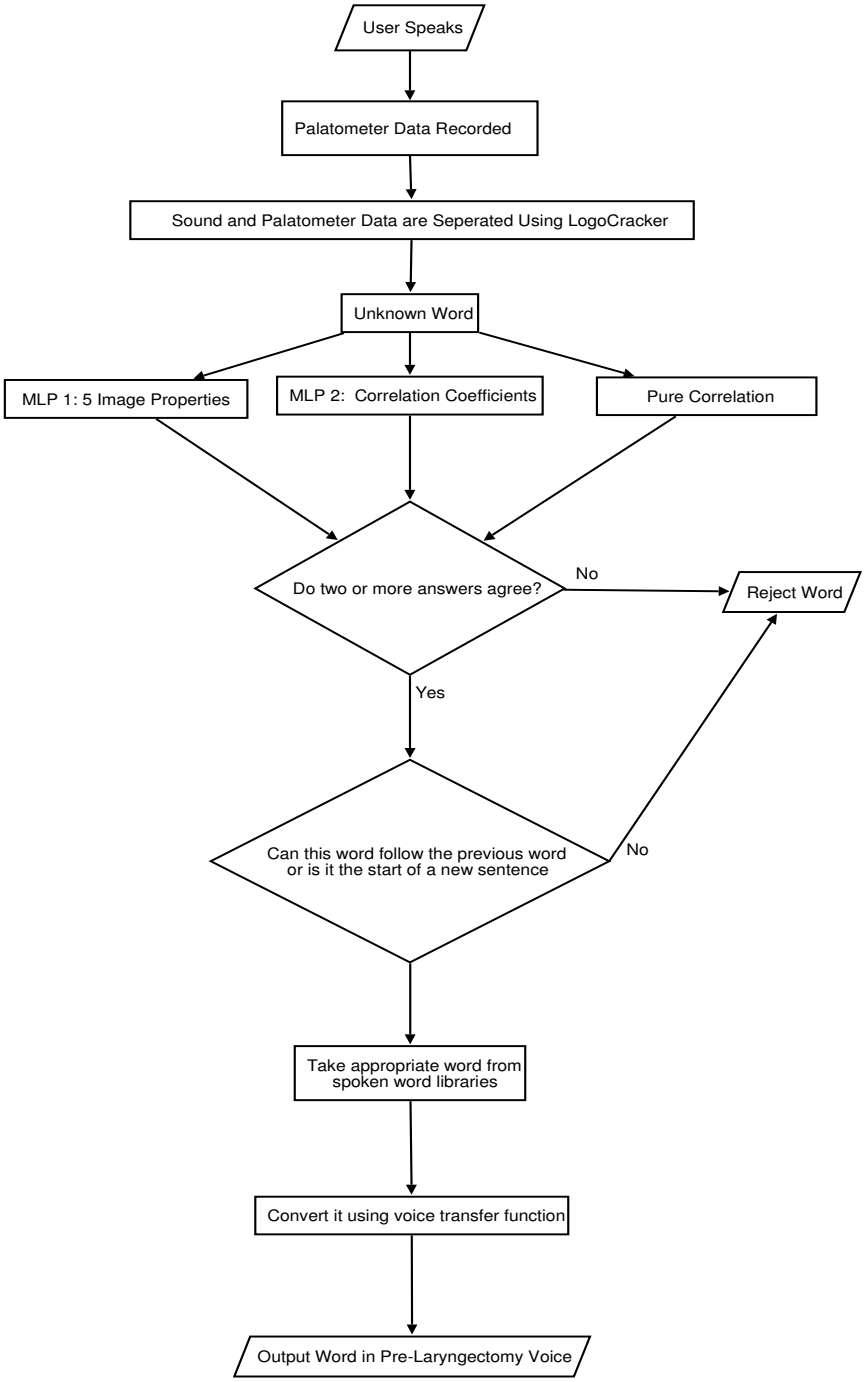


Figure 3.23: Outline of the whole artificial larynx

Chapter 4

Results

4.1 Chapter Overview

This chapter presents and discusses the results for the different combinations of inputs to the MLP and SVM. It also highlights the improvement in performance (an increase of 9.14%) created by using voting systems and discusses voice morphing. It shows that the most successful system uses a voting and predicting scheme consisting of the three most successful classification networks alongside a rudimentary grammar. This system has a correct classification rate of 94.14% and has a rejection rate of 17.74%.

4.2 Results

The results from the various inputs to the MLP and SVM as well as the voting schemes and voice morphing are detailed below. For a complete set of the MLP and SVM results with different inputs see Table 4.1 and see appendix D for how the number of hidden nodes changes the success rate of the MLP.

4.2.1 Results of MLP using Image Features

The results of using the image features described in Section 3.5 as input to the MLP are summarized in Table 4.2 and Figure 4.1. By chance alone, there is a one-in-fifty probability (2% chance) of randomly generating the correct result.

Table 4.1: The complete results of using different image features as input to the MLP and SVM

Image Feature (Input to MLP or SVM)	Classification Success rate using the MLP (Percent)	Classification Success rate using the SVM (Percent)
Principal Component Analysis	8	4
PCA and correlation number	8	2.66
Fourier descriptors	22.5	1.5
Fourier descriptors and correlation number	50	0.5
Fourier descriptors and 4 Image Properties and Correlation Number	60	1
Correlation against templates (input -1 to 1)	76.5	1.5
Correlation against templates (input 0 to 1)	76	
Correlation against templates (abs input)	78	
4 Image Properties and Correlation Number	71.5	1.5
13 image properties	45	
13 image properties and correlation number	67	
Generic Fourier Descriptors - centered at center of mass	38	
GFD not centered	27	
Column Sums of each word		47
Column and row sums of each word		45.5
Column sums and length of each word		45

The top performing MLP used correlation coefficients as its input. The worst performing MLP used Principal Components.

Table 4.2: The results of using different image features as input to the MLP (from Russell et al. (2009*a*))

Image Feature	Classification Success rate using the MLP (Percent)
PCA	8
Fourier Descriptors	22.5
Generic Fourier Descriptors	38
4 Image Properties	31
13 Image Properties	45
Correlation Coefficients	78
4 Image Properties and Correlation Number	71.5
13 Image Properties and Correlation Template Number	67
Fourier Descriptors and Correlation Number	50
Fourier Descriptors and 4 Image Properties and Correlation number	60

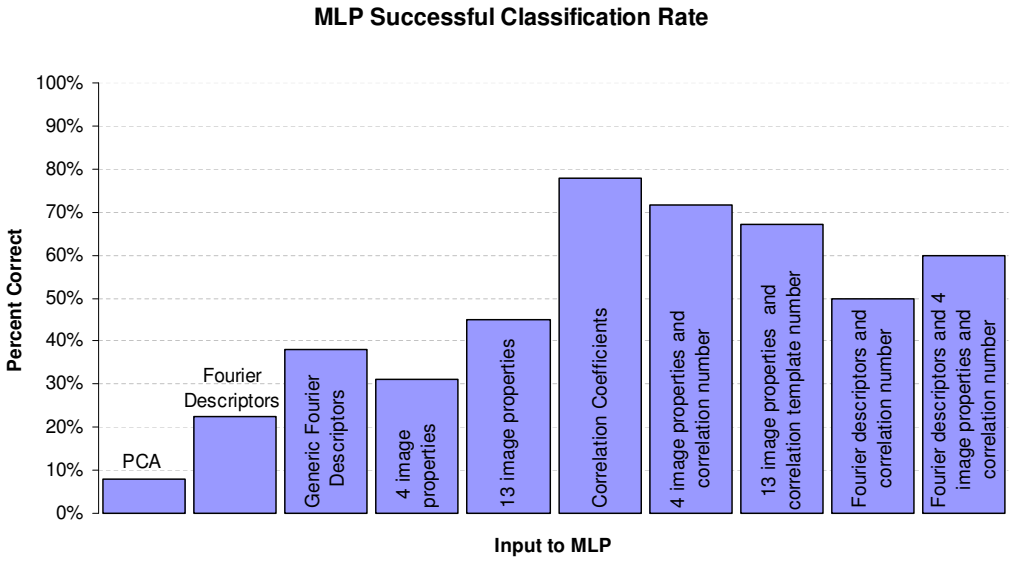


Figure 4.1: A graphical representation of the success rate of the MLP using different features to classify each of the 50 words

4.2.2 Results of SVM using Image Features

Support Vector Machines are good at classifying high dimensional data. Thus the image properties which produced high dimensional data (PCA, Fourier Descriptors) were used as input to the SVM. Some other image properties (summing of the columns and rows of the Space-Time images) were also high dimensional and used as input to the SVMs. The results can be seen in Table 4.3.

Table 4.3: The results of using different Image Features as input to the SVM

Image Feature	Classification Success rate using the SVM (Percent)
PCA	1.5
Fourier Descriptors	1.5
Sum of Space-Time Image's Columns	47
Sum of Space-Time Image's Rows and Columns	45.5
Sum of Space-Time Image's Columns and the Length of the Word (Number of Rows)	45

The top performing SVM used the sum of the Space-Time image's columns

4.2.3 Results of Classifying Using Correlation Alone

Using no neural networks, the results of classification using correlation template matching alone are very good. An 85% success rate was achieved. Four unseen cases of each word were presented to the system and the number of times each word was correctly classified can be seen in Figure 4.2.

4.2.4 Results of Voting System 1 (Winner-takes-all)

Using the top three classifiers (pure correlation, MLP using 5 image properties and MLP using correlation coefficients), this voting system correctly classifies non-rejected words 93.5% of the time, and has a rejection rate of 17.36% (Russell et

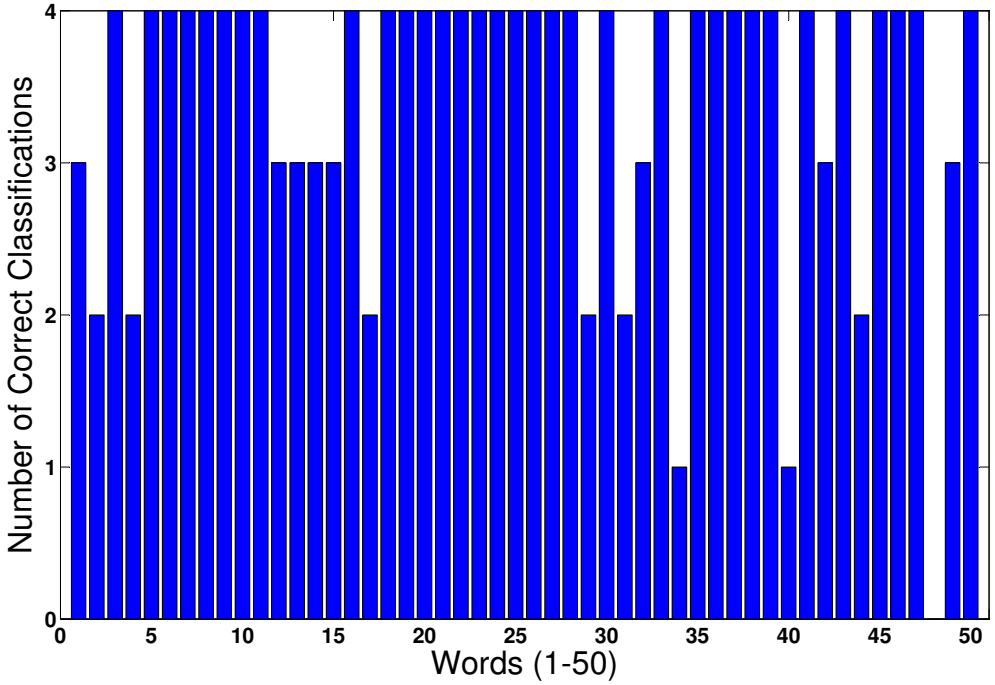


Figure 4.2: The success rate of using pure correlation to classify each of the 50 words

al., 2009*b*). Four unseen cases of each word were presented to the system and the number of times each word was correctly classified are shown in Figure 4.3.

4.2.5 Results of Voting System 2 (Grammar Prediction)

This system identifies an unknown word correctly 94.14% of the time, although the system now has a slightly increased rejection rate of 17.74% of the input words (Russell et al., 2009*b*). 1000 sentences of five words each were randomly created using the grammar rules and unseen cases of each word. These were then presented to the system and the number of times each word is correctly identified can be seen in Figure 4.4.

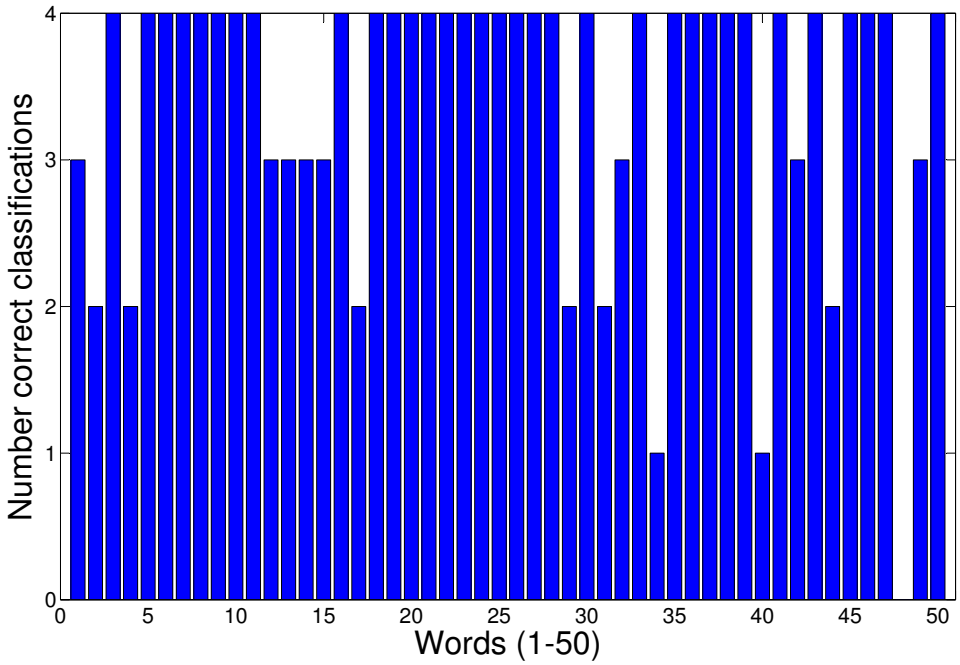


Figure 4.3: The success rate of Voting System 1 to classify the test dataset

4.2.6 Results of Voting System 3 (Bit-by-bit)

This system correctly identifies 82.5% of words, however it is not able to reject incorrect classifications as the other voting systems were able to. Four unseen cases of each word were presented to the system and the number of times each word was correctly classified can be seen in Figure 4.5.

4.2.7 Processing Time

The average times to process different sections of the MATLAB code are given in Table 4.4. As can be seen from Table 4.4, the complete system (i.e. classifying and outputting the synthesized voice) takes significantly longer than the voting systems by themselves.

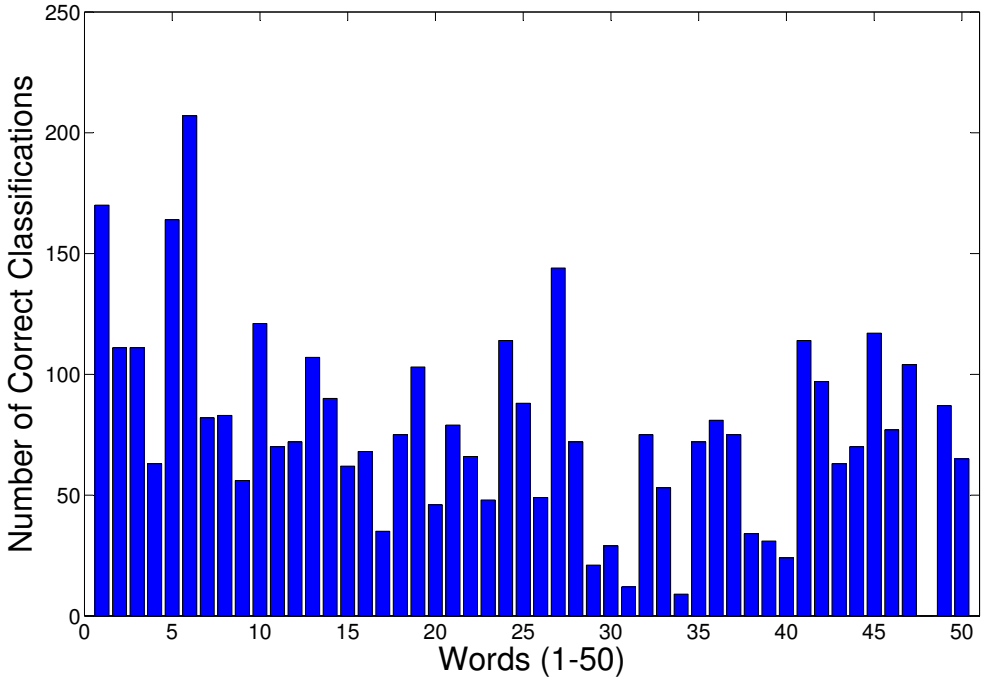


Figure 4.4: The success rate of the Voting System 2 to classify each of the 50 words when used in sentences

Table 4.4: The average processing times of the different systems

System	Time (seconds)
Voting System 1 (Winner-takes-all)	0.2462
Voting System 2 (Grammar Prediction)	0.2406
Voting System 3 (Bit-by-bit)	0.2462
Loading and Synthesizing a Word	0.7921
Complete System with Voice Output	1.1283

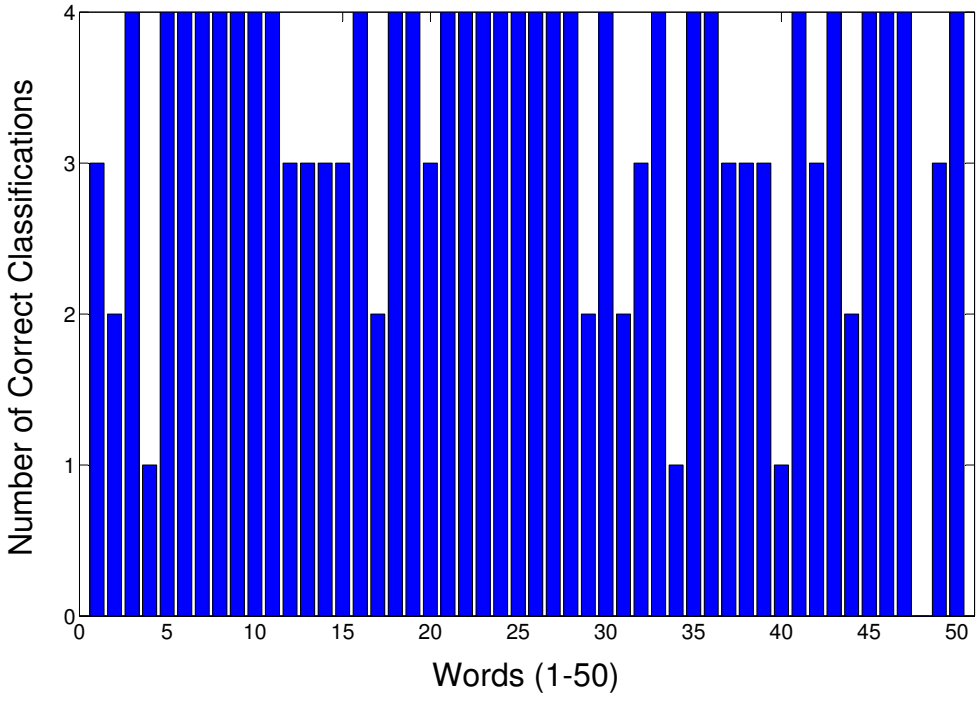


Figure 4.5: The success rate of the Bit-by-Bit Voting System to classify the test dataset

Chapter 5

Discussion

5.1 Chapter Overview

This chapter discusses the results obtained from the different neural networks and the voting systems.

5.2 The Multi-Layer Perceptron

The results of using Principal Component Analysis to decrease the size of the dataset for input into the MLP were poor. When the principal components of each word in the training set (50 words, 20 cases of each word) were correlated against each other, many unrelated words had a high correlation (greater than 90% correlation) (see Figure 5.1). The diagonal line running from the top left corner to the bottom right corner shows the correlation of each word's Principal Components with themselves (i.e. these should have a correlation value of 1). All the other white areas in the image show words correlating with other words. This suggests that PCA is unsuitable as a data reduction technique of palatometer data (Russell et al., 2009*a*) due to the fact that the Principal Components of the words are all too similar to eachother. This may be due to the data being non-linear and therefore not being linearly decomposable (Russell et al., 2009*a*).

Fourier Descriptors alone as input to the MLP also performed poorly. This is probably due to the fact that the shapes in the Space-Time plots are very complex and contain many parts (Russell et al., 2009*a*); and Fourier Descriptors tend to be used on simple images (i.e. images that contain only one "'object"' or area of interest).

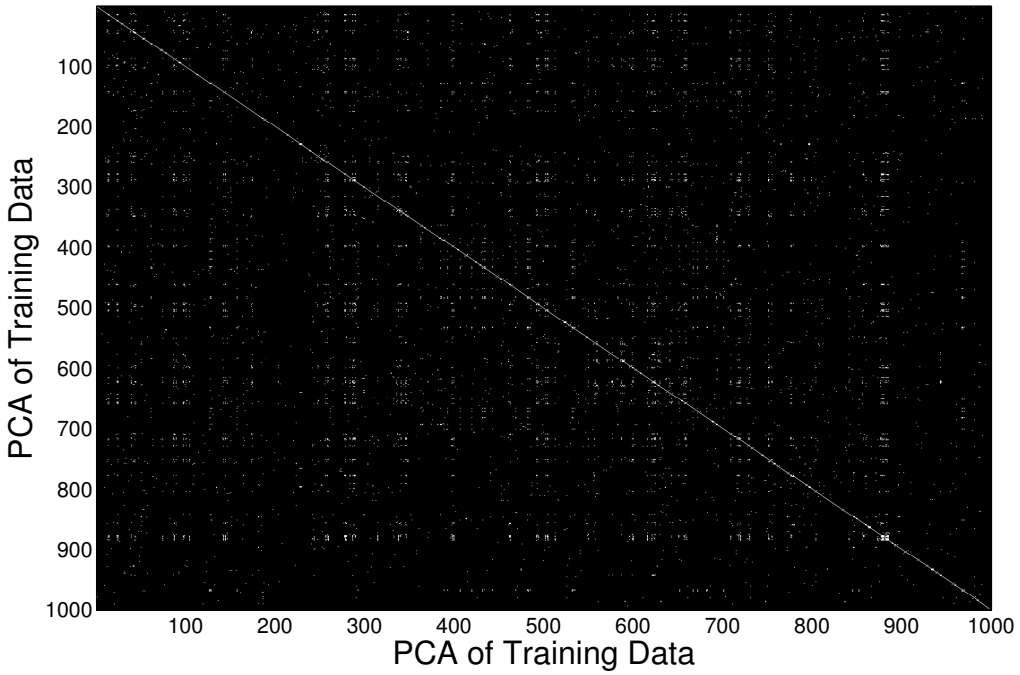


Figure 5.1: Correlation of the PCA vectors of the training set against each other.

As the space-time plots are binary (and thus have extremely sharply defined edges), the Fourier Descriptors would have an infinite number of coefficients, which means that the image would not be well described with only a small portion of the Fourier coefficients. Generic Fourier Descriptors also performed poorly for similar reasons, as well as the fact that the shape has to be positioned at the center of mass, which destroys any spatial information contained in the image (Russell et al., 2009*a*).

The four image descriptors used as input to the MLP were: area, Euler number, center of mass on the X axis and center of mass on the Y axis. These performed poorly, however, as soon as the correlation number is added the performance of the MLP increases (Russell et al., 2009*a*). The results of the 13 image properties mirror those of the four image properties with their performance increasing as the correlation number is added (Russell et al., 2009*a*).

The MLP produces the best results (71.5% correct identification) when using Correlation Coefficients as the input (obtained by convolving each test word against the templates) (Russell et al., 2009*a*). Combinations of inputs were also tried with

varying degrees of success (see Table 4.1).

5.3 Support Vector Machines

Due to the fact that a one-versus-the-rest approach was used, the training set can be seen as unbalanced (i.e. only 20 of the 1000 words in the training set are in class 1 and 980 words are in class 2) (Bishop, 2006). This contributes significantly to the poor performance of the SVMs. Also, if more than one SVM claims the unknown word as belonging to it, there is no way of deciding which to class it actually belongs.

5.4 Correlation Alone

Correlation alone performed very well. It is successfully used in the voting systems to increase the overall classification rate. As can be seen in Figure 5.2 most cases of each word correlate well with each other.

5.5 Voting Systems

The voting systems substantially increase the success of the system. Using the grammar voting system there is an increase of 9.14% in the correct classification of words as compared to the leading individual classifier.

5.6 Undetected Words

Word 48 ("animal") was never correctly identified. This seems to be due to inconsistency in the data and thus the recordings. When all the words were correlated against each other, it was expected that different cases of the same word would strongly correlate with each other. Figure 5.2 and Figure 5.3 show the results of correlating each case of each word against every single other word in the training set. Most words correlate strongly with other cases of the same word (i.e. form square regions in Figure 5.2 and Figure 5.3). However, as can be seen in Figure 5.2 and Figure 5.3, the cases for word 48 do not correlate with each other at all. This was due to inconsistencies in how the word was pronounced during recording. It is

thus recommended that before training of the neural networks is begun, it is ensured that all cases of the same word correlate highly with each other.

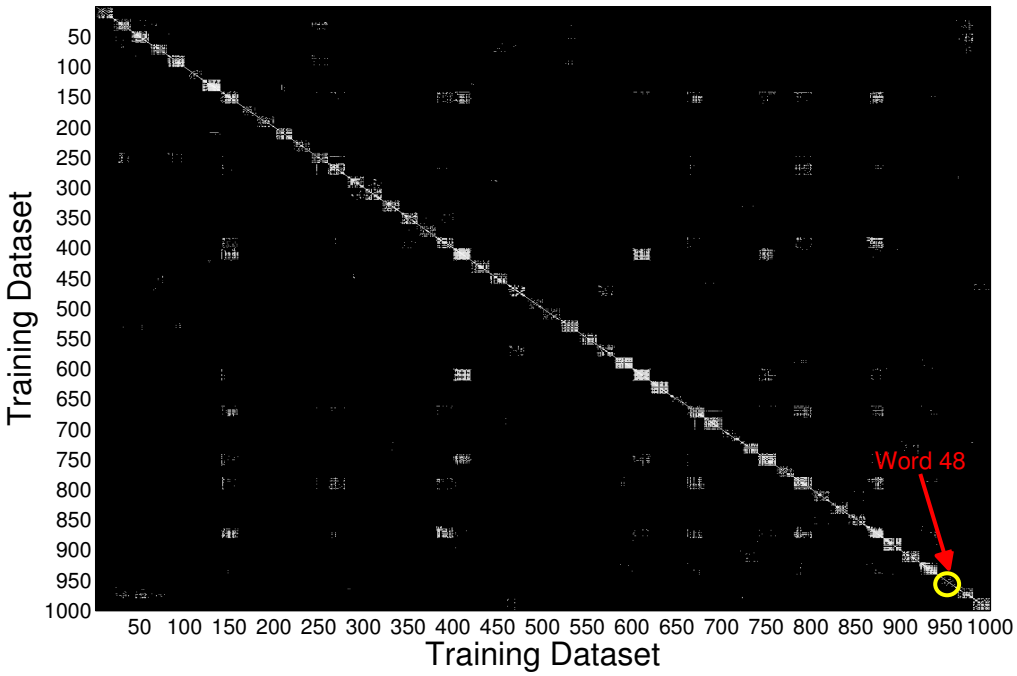


Figure 5.2: Correlating 20 cases of each word for all of the 50 words. Only correlation values of over 0.75 are shown. The arrow indicates where word 48 is situated

5.7 Voice Morphing

Transfer functions that can successfully take one person's voice and output another person's words were created. These transfer functions can be used to make any sentence spoken by the first speaker sound like it was spoken by the second speaker.

5.7.1 Processing Time

As can be seen from Table 4.4 the process that takes the longest is the voice morphing and synthesizing. The neural networks, once they are trained, can easily classify an unknown word in under the required time. The only way to increase the speed of

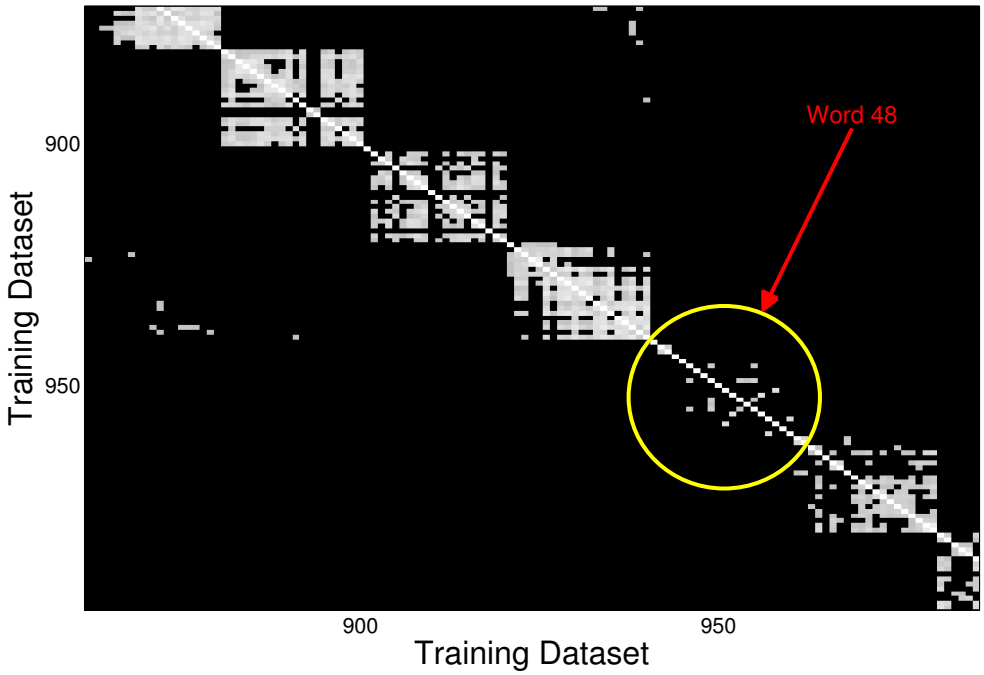


Figure 5.3: Closeup on word 48 from Figure 5.2 showing the lack of correlation in the cases of the same word

the voice synthesizing system would be to implement it in different software and this is beyond the scope of this study.

5.8 Results and Research Questions

The following research questions were introduced in the beginning of this work and are discussed below:

1. **Can speech recognition be performed on tongue-palate contact patterns from a palatometer?**
   Speech recognition can be performed on palatometer data. However, the palatometer does present some limitations in that it cannot detect some vowels. However, most of this problem can be solved by detecting and recognizing whole words and not just phonemes. The use of space-time plots to represent the palatometer data means that even if a word contains a phoneme that is

not detected the rest of the word will be detected and have a unique pattern to it, thus allowing for identification.

2. **Are standard signal processing techniques and artificial intelligence techniques sufficient to relate the data signals to the speech?**
   The techniques used to identify the palatometer data perform fairly well (Voting System 2 identifies an unknown word correctly 94.14% of the time with a rejection rate of 17.74%). By using voting systems, incorrect classifications can be recognized and thus outputting the wrong word can be avoided. However many of the image feature extraction algorithms performed poorly on the space-time images. This may be due to the images being binary in nature or may be due to the fact that the images tend to be comprised of a number of discrete parts. Also many of the feature extraction algorithms do not take into account position information in the image, thus data about when each sensor is activated is lost.

3. **If the data signals can be correctly identified as speech, can the appropriate pre-recorded speech be outputted using a loudspeaker?**
   The appropriate pre-recorded speech can be outputted and by using Voting System 2 94.14% or the outputted words will be correct. The processing times also lead to the conclusion that once the neural networks are implemented in hardware the processing time will be fast enough to allow for audio visual synchronisation.

4. **Can pre-recorded speech be altered to mimic other peoples voices?**
   This is a technique much researched for use in the movie industry. By using the Voice Morphing toolbox the pre-recorded speech could be altered to sound like another person's voice.

The answers to the research questions indicate that the hypothesis (that speech-free speech recognition is possible) is correct and that it can be used as a basis for a new type of artificial larynx.

Chapter 6

Conclusion and Recommendations

Research into speech-free speech recognition and the use of voice morphing for a new type of artificial larynx has been done. Words can be successfully classified by using data from the palatometer and these words can then be synthesized in the user's pre-laryngectomy voice. Using the voting system with grammar a satisfactory recognition success rate of 94.14% is achieved with a rejection rate of 17.74% and a voice morphing system has been implemented. A simulator has been developed to allow for easy testing.

6.1 Contribution to Knowledge

Research toward the design and implementation of a new type of artificial larynx has been demonstrated. No previous work along these lines has been found in the literature. Speech recognition performed solely on palatometer data has not been performed before. Techniques and technologies from diverse fields have been applied to new situations and the feasibility of creating a new type of artificial larynx has been shown.

6.2 Recommendations for Further Research

A number of different aspects can be considered for future work on the artificial larynx:

- **Increase Word Database Size:** The word recognition system should be expanded to increase the functionality of the device. As discussed in Section 3.4.1

2000 words should provide the user with sufficient functionality.

- **Palatometer Technology:** The palatometer could be improved by removing the ribbon cables that currently project from the user's mouth and replacing the connectivity with Blue Tooth or possibly some other low-power wireless connection. This would make the artificial palate almost unnoticeable to observers.

- **Unidentified Words:** Some additional hardware is required to aid in the recognition of words unable to be identified by the palatometer (e.g. "I"). Other physiological measurements such as force measurement, jaw opening and lip shape detection should be investigated.

- **Feature Selection:** It is unlikely that the optimal features to describe palatometer recordings have been found. A technique similar to the "eigenfaces" algorithm (see for example Kim et al. (2002) and Wang et al. (2005)) could be investigated. Other options include wavelets, data compression techniques and new ways of representing the palatometer data.

- **Predictive Neural Networks:** A concern with the artificial larynx is that the words must be recognized and synthesized in real time (or within 300ms). The likelihood of this happening could be increased by employing Hidden Markov Models (see for example Rabiner (1989) and Bishop (2006)) to predict the outcome of the word before the user has finished saying it.

- **Training of Voting System:** Introducing the use of Bagging and Boosting (see for example Tresp et al. (2001), Schapire (2002) and Nanni et al. (2010)) to the voting schemes training algorithms could increase the success rate of the classification.

- **Speed:** The speed of the total processing would need to fall below 300ms as discussed in Section 3.3.6. This could be done by implementing the code in C++ or by implementing the neural network architecture in hardware. The system that has the slowest processing time is the voice loading and synthesizing thus possibly more sophisticated speech synthesizing software than MATLAB could be used to solve this problem.

- **Microprocessors:** The neural network as well as the various libraries and templates could be stored and computed using a microprocessor system. This would allow the user to carry the artificial larynx in a pocket, allowing for portability.

- **Emotion in Speech:** Advances in speech technology are happening at a fast rate. These, including adding emotion to speech, could all be implemented. Other sensors such as heart rate monitors and skin conductivity sensors which detect the physiological changes due to emotions could be used as additional input to the speech synthesizer thus allowing for even more authentic sounding speech.

6.3 Additional Outcomes

A patent for this artificial larynx has been applied for. This work has been featured in the MIT Technology Review online magazine. (For further details on the patent and the article see Appendix F.)

References

ACS (2009), *Cancer Facts and Figures*, Atlanta: American Cancer Society.

Aristotle (350 B.C.E), *Politics*, Vol. Translated by Benjamin Jowett, 2005 edn, Digireads.com Publishing.

Avants, B., Epstein, C., Grossman, M. & Gee, J. (2008), 'Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain', *Medical Image Analysis*, vol. 12, pp. 26–41.

Babaoglu, I., Findik, O. & Bayrak, M. (2009), 'Effects of principle component analysis on assessment of coronary artery diseases using support vector machine', *Expert Systems with Applications*.

Banos, E., Erro, D., Bonafonte, A. & Moreno, A. (2008), 'Flexible harmonic/stochastic modeling for hmm-based speech synthesis', *Actas de las V Jornadas en Tecnologias del Habla*, pp. 145–148.

Battiti, R. & Colla, A. M. (1994), 'Democracy in neural nets: Voting schemes for classification', *Neural Networks*, vol. 7, no. 4, pp. 691–707.

Bing, P., Hui-min, X., Tao, H. & Asundi, A. (2009), 'Measurement of coefficient of thermal expansion of films using digital image correlation method', *Polymer Testing*, vol. 28, pp. 75–83.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning.*, Information Science and Statistics, 1st edn, Springer Science+Business Media.

Bogdanov, A. V. (2008), 'Neuroinspired architecture for robust classifier fusion of multisensor imagery', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1467–1487.

Carreira-Perpinan, M. A. & Renals, S. (1998), 'Dimensionality reduction of electropalatographic data using latent variable models', *Speech Communication*, vol. 26, pp. 259–282.

Chen, C.-S., Yeh, C.-W. & Yin, P.-Y. (2009), 'A novel fourier descriptor based image alignment algorithm for automatic optical inspection', *Journal of Visual Communication and Image Representation*, vol. 20, pp. 178–189.

Cho, T. & Keating, P. (2009), 'Effects of initial position versus prominence in english', *Journal of Phonetics*, vol. 37, pp. 466–485.

Christensen, J. M., Fletcher, S. G. & McCutcheon, M. J. (1992), 'Esophageal speaker articulation of /s,z/: A dynamic palatometric assesment', *Journal of Communication Disorders*, vol. 25, pp. 65–76.

CompleteSpeech (2008), Online:. `http://www.completespeech.com` Last Accessed: 01-10.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine Learning*, vol. 20, pp. 273–297.

da Silva Sousa, J. R. F., Silva, A. C., de Paiva, A. C. & Nunes, R. A. (2009), 'Methodology for automatic detection of lung nodules in computerized tomography images', *Computer methods and programs in biomedicine*.

Dagenais, P. A. (1995), 'Electropalatography in the treatment of articulation/phonological disorders', *Journal of Communication Disorders*, vol. 28, pp. 303–329.

de Albuquerque, V. H. C., de Alexandria, A. R., Cortez, P. C. & Tavares, J. M. R. S. (2009), 'Evaluation of multilayer perceptron and self-organizing map neural network topologies applied on microstructure segmentation from metallographic images', *NDT and E International*, vol. 42, pp. 644–651.

Deshpande, M. S., Kakade, A. C., Chaukar, D. A., Gore, V. T., Pai, P. S., Chaturvedi, P. & D'Cruz, A. K. (2009), 'Validation and assessment of voice-related quality of life in indian patients undergoing total laryngectomy and primary tracheoesophageal puncture', *Head and Neck*, pp. 37–44.

Diaz, G., Gonzalez, F. A. & Romero, E. (2009), 'A semi-automatic method for quantification and classification of erythrocytes infected with malaria parasites in microscopic images', *Journal of Biomedical Informatics*, vol. 42, pp. 296307.

Dromey, C. & Sanders, M. (2009), 'Intra-speaker variability in palatometric measures of consonant articulation', *Journal of Communication Disorders*, vol. 42, pp. 397–407.

Duxans, H. & Bonafonte, A. (2003), 'Estimation of gmm in voice conversion including unaligned data', *Proceedings of EuroSpeech - Geneva*, pp. 861–864.

Duxans, H., Erro, D., Perez, J., Diego, F., Bonafonte, A. & Moreno, A. (2006), 'Voice conversion of non-aligned data using unit selection', *TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, Spain*, pp. 237–242.

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.

Erro, D. (2008), 'Research and work at upc', Online:. `http://gps-tsc.upc.es/veu/personal/derro/` Last Accessed: 08-09.

Erro, D. & Moreno, A. (2007), 'Frame alignment method for cross-lingual voice conversion', *InterSpeech*.

Eslava, D. E. (2008), Intra-Lingual And Cross-Lingual Voice Conversion Using Harmonic Plus Stochastic Models, PhD thesis, Universitat Politecnica de Catalunya.

Fagan, M. J., Ell, S. R., Gilbert, J. M., Sarrazin, E. & Chapmanc, P. M. (2008), 'Development of a (silent) speech recognition system for patients following laryngectomy', *Medical Engineering and Physics*, vol. 30, pp. 419425.

Felps, D., Bortfeld, H. & Gutierrez-Osuna, R. (2009), 'Foreign accent conversion in computer assisted pronunciation training', *Speech Communication*, vol. 51, pp. 920–932.

Francis, W. N. & Kucera, H. (1982), *Frequency Analysis of English Usage*, Houghton Mifflin Company.

Garcia, M., Sanchez, C. I., Lopez, M. I., Abasolo, D. & Hornero, R. (2009), 'Neural network based detection of hard exudates in retinal images', *Computer Methods and Programs in Biomedicine*, vol. 93, pp. 919.

Gershikov, E., Lavi-Burlak, E. & Porat, M. (2007), 'Correlation-based approach to color image compression', *Signal Processing: Image Communication*, vol. 22, pp. 719733.

Gonzalez, R. C. & Woods, R. E. (2002), *Digital Image Processing*, 2nd edn, Prentice-Hall.

Hardcastle, W., Jones, W. & Knight, C. (1989), 'New developments in electropalatography: A state-of-the-art report', *Clinical Linguistics and Phonetics*, vol. 3, no. 1, pp. 1–38.

Helmuth, J. A., Burckhardt, C. J., Greber, U. F. & Sbalzarini, I. F. (2009), 'Shape reconstruction of subcellular structures from live cell fluorescence microscopy images', *Journal of Structural Biology*, vol. 167, pp. 1–10.

Hotta, K. (2008), 'Robust face recognition under partial occlusion based on support vector machine with local gaussian summation kernel', *Image and Vision Computing*, vol. 26, pp. 1490–1498.

Huckvale, M. (2007), 'Prorec - prompt and record version 1.2', Online:. `www.phon.ucl.ac.uk` Last Accessed: 06-09.

Huckvale, M. (2008), 'Sfs - speech filing system', Online:. `http://www.phon.ucl.ac.uk/resource/sfs/` Last Accessed: 03-09.

Kazi, R., Kanagalingan, 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.

Kim, H.-C., Kim, D. & Bang, S. Y. (2002), 'Face recognition using the mixture-of-eigenfaces method', *Pattern Recognition Letters*, vol. 23, pp. 1549–1558.

Krishnan, M. M. R., Pal, M., Bomminayuni, S. K., Chakraborty, C., Paul, R. R., Chatterjee, J. & Ray, A. K. (2009), 'Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis - an svm based approach', *Computers in Biology and Medicine*.

Kubert, H. L., Stepp, C. E., Zeitels, S. M., Gooey, J. E., Walsh, M. J., Prakash, S. R., Hillman, R. E. & Heaton, J. T. (2009), 'Electromyographic control of a hands-free elctrolarynx using neck strap muscles', *Journal of Communication Disorders*, vol. 42, pp. 211–225.

Kunttu, I., Lepisto, L., Rauhamaa, J. & Visa, A. (2006), 'Multiscale fourier descriptors for defect image retrieval', *Pattern Recognition Letters*, vol. 27, pp. 123–132.

Lauder, J. (2007), 'Electrolarynx', Online:. `http://www.electrolarynx.com/` Last Accessed: 03-07.

Lee, M., Roan, M. & Smith, B. (2009), 'An application of principal component analysis for lower body kinematics between loaded and unloaded walking', *Journal of Biomechanics*, vol. 42, pp. 2226–2230.

Lestrel, P., Kanazawa, E. & Wolfe, C. (2009), 'Sexual dimorphism in the craniofacial complex: Differences in shape using fourier descriptors', *HOMO Journal of Comparative Human Biology*, vol. 60, pp. 239–290.

Li, B. & Meng, M. Q.-H. (2009), 'Computer-aided detection of bleeding regions for capsule endoscopy images', *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1032–1039.

Liu, H. & Ng, M. L. (2007), 'Electrolarynx in voice rehabilitation', *Auris Nasus Larynx*, vol. 34, pp. 327–332.

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.

Maglogiannis, I. G. & Zafiropoulos, E. P. (2004), 'Characterization of digital medical images utilizing support vector machines', *BMC Medical Informatics and Decision Making*.

Malagon-Borja, L. & Fuentes, O. (2009), 'Object detection using image reconstruction with pca', *Image and Vision Computing*, vol. 27, pp. 2–9.

Mathworks (2008), *MATLAB: The language of technical computing. R2008a Help Documentation*, The Mathworks.

May, A. (2010), 'On the tip of the tongue', Online. `http://www.therapytimes.com/content=0402J84C487EB084404040441` Last Accessed: 09-10.

McGrath, M. & Summerfield, Q. (1985), 'Intermodal timing relations and audio-visual speech recognition by normal-hearing adults', *Journal of the Acoustic Society of America*, vol. 77, no. 2, pp. 678–685.

Menesatti, P., Aguzzi, J., Costa, C., Garcia, J. A. & Sarda, F. (2009), 'A new morphometric implemented video-image analysis protocol for the study of social modulation in activity rhythms of marine organisms', *Journal of Neuroscience Methods*, vol. 184, pp. 161–168.

Moerman, K. M., Holt, C. A., Evans, S. L. & Simms, C. K. (2009), 'Digital image correlation and finite element modeling as a method to determine mechanical properties of human soft tissue in vivo', *Journal of Biomechanics*, vol. 42, pp. 1150–1153.

Moller, M. F. (1993), 'A scaled conjugate gradient algorithm for fast supervised learning', *Neural Networks*, vol. 6, pp. 525–533.

Monasterio, V., Laguna, P. & Martinez, J. P. (2009), 'Multilead analysis of t-wave alternans in the ecg using principal component analysis', *IEEE Transaction on Biomedical Engineering*, vol. 56, no. 7, pp. 1880–1890.

Morgan, N. & Bourlard, H. (1995), 'Continuous speech recognition', *IEEE Signal Processing Magazine*, pp. 25–42.

Most, T., Tobin, Y. & Mimran, R. C. (2000), 'Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production', *Journal of Communication Disorders*, vol. 33, pp. 165–181.

Muhlbaier, M. D., Topalis, A. & Polikar, R. (2009), 'Learn++nc: Combining ensemble of classifiers with dynamically weighted consult-and-vote for efficient incremental learning of new classes', *IEEE Transactions on Neural Networks*.

Murty, K. S. R. & Yegnanarayana, B. (2008), 'Epoch extraction from speech signals', *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613.

Nabney, I. T. (2002), *NETLAB Algorithms for Pattern Recognition*, Advances in Pattern Recognition, 4th edn, Springer-Verlag.

Nabney, I. T. (2003), 'Netlab', Online:. `http://www.ncrg.aston.ac.uk/netlab/index.php` Last Accessed: 06-09.

Nanni, L., Lumini, A., Lin, Y.-S., Hsu, C.-N. & Lin, C.-C. (2010), 'Fusion of systems for automated cell phenotype image classification', *Expert Systems with Applications*, vol. 37, pp. 1556–1562.

Naqa, I. E., Grigsby, P. W., Apte, A., Kidd, E., Donnelly, E., Khullar, D., S.Chaudhari, Yang, D., Schmitt, M., Laforest, R., Thorstad, W. L. & Deasy, J. O. (2009), 'Exploring feature-based approaches in pet images for predicting cancer treatment outcomes', *Pattern Recognition*, vol. 42, pp. 1162–1171.

Navarra, J., Hartcher-OBrien, J., Piazza, E. & Spence, C. (2009), 'Adaptation to audiovisual asynchrony modulates the speeded detection of sound', *PNAS*, vol. 106, no. 23, pp. 91699173.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W. & Spence, C. (2005), 'Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration', *Cognitive Brain Research*, vol. 25, pp. 499 507.

Ng, M. L., Kwok, C.-L. I. & Chow, S.-F. W. (1997), 'Speech performance of adult cantonese-speaking laryngectomees using different types of alaryngeal phonation', *Journal of Voice*, vol. 11, no. 3, pp. 338–344.

Ogden, C. K. (1937), *Basic English and Grammatical Reform*, Cambridge: The Orthological Institute.

Ohbuchi, R., Nakazawa, M. & Takei, T. (2003), 'Retrieving 3d shapes based on their appearance', *Proc. 5th ACM SIGMM Workshop on Multimedia Information Retrieval*.

Orozco, J. & Garcia, C. A. R. (2003), 'Detecting pathologies from infant cry applying scaled conjugate gradient neural networks', *Proceedings of the European Symposium on Artificial Neural Networks, Belgium*, pp. 349–354.

O'Shaughnessy, D. (2003), 'Interacting with computers by voice: Automatic speech recognition and synthesis', *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1272–1305.

Pawar, P. V., Sayed, S. I., Kazi, R. & Jagade, M. V. (2008), 'Current status and future prospects in prosthetic voice rehabilitation following laryngectomy', *Journal of Cancer Research Therapy*, vol. 4, no. 4, pp. 186–191.

Rabiner, L. R. (1989), 'A tutorial on hidden markov models and selected applications in speech recognition', *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286.

Rao, K. S. (2009), 'Voice conversion by mapping the speaker-specific features using pitch synchronous approach', *Computer Speech and Language*.

Reddy, N. P. & Buch, O. A. (2003), 'Speaker verification using committee neural networks', *Computer Methods and Programs in Biomedicine*, vol. 72, pp. 109–115.

Roach, P. (2002), *A little encyclopedia of phonetics*, Online:. `http://www.personal.reading.ac.uk/~llsroach/peter/` Last Accessed: 06-08.

Russell, M. J., Rubin, D. M., Marwala, T. & Wigdorowitz, B. (2009*a*), 'Pattern recognition and feature selection for the development of a new artificial larynx', *WC09 IFMBE Proceedings*, vol. 25/IV, pp. 736–739.

Russell, M. J., Rubin, D. M., Marwala, T. & Wigdorowitz, B. (2009*b*), 'A voting and predictive neural network system for use in a new artificial larynx', *IEEE Proceedings of the 2nd International Conference in Biomedical and Pharmaceutical Engineering*, pp. 1–4.

Russell, M. J., Rubin, D. M., Wigdorowitz, B. & Marwala, T. (2008), 'The artificial larynx: A review of current technology and a proposal for future development', *NBC 2008 Proceedings*, vol. 20, pp. 160–163.

Sataloff, R. T. (2007), 'Esophageal speech', Plural Publishing Online. `www.origin8.nl/medical/esophgea.htm` Last Accessed: 03-07.

Saykol, E., Gudukbay, U. & Ulusoy, O. (2005), 'A histogram-based approach for object-based query-by-shape-and-color in image and video databases', *Image and Vision Computing*, vol. 23, pp. 1170–1180.

Schapire, R. E. (2002), 'The boosting approach to machine learning an overview', *MSRI Workshop on Nonlinear Estimation and Classification*.

Selver, M. A., Kocaoglu, A., Demir, G. K., Dogan, H., Dicle, O. & Guzeli, C. (2008), 'Patient oriented and robust automatic liver segmentation for pre-evaluation of liver transplantation', *Computers in Biology and Medicine*, vol. 38, pp. 765 784.

Shoureshi, R. A., Chaghajerdi, A., Aasted, C. & Meyers, A. (2003), 'Neural-based prosthesis for enhanced voice intelligibility in laryngectomees', *Proc. IEEE EMBS Conference on Neural Engineering*, pp. 173–176.

Smith, L. I. (2006), 'A tutorial on principal component analysis', Online:. `http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf` Last Accessed: 04-08.

Soquet, A., Saerens, M. & Lecuit, V. (1999), 'Complimentary cues for speech recognition', *Proceedings of the International Conference of Phonetic Science, San Francisco*, pp. 1645–1648.

Tack, J. W., Qiu, Q., Schutte, H. K., Kooijman, P. G. C., a Meeuwis, C., van der Houwen, E. B., Mahieu, H. F. & Verkerke, G. J. (2008), 'Clinical evaluation of a membrane-based voice producing element for laryngectomized women', *Head and Neck*, pp. 1156–1166.

Takahashi, H., Nakao, M., Kikuchi, Y. & Kaga, K. (2005), 'Alaryngeal speech aid using an intra-oral electrolarynx and a miniature fingertip switch', *Auris Nasus Larynx*, vol. 32, pp. 157–162.

Tian, S., Mua, S. & Yina, C. (2007), 'Sequence-similarity kernels for svms to detect anomalies in system calls', *Neurocomputing*, vol. 70, no. 4-6, pp. 859–866.

Torrejano, G. & Guimaraes, I. (2009), 'Voice quality after supracricoid laryngectomy and total laryngectomy with insertion of voice prosthesis', *Journal of Voice*.

Toutios, A. & Margaritis, K. (2006), 'Learning electropalatograms from acoustics', *IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Toutios, A. & Margaritis, K. (2008), 'Estimating electropalatographic patterns from the speech signal', *Computer Speech and Language*, vol. 22, pp. 346–359.

Trenn, S. (2008), 'Multilayer perceptrons: Approximation order and necessary number of hidden units', *IEEE Transactions on Neural Networks*, vol. 19, no. 5, pp. 836–844.

Tresp, V., Hu, Y. H. & Hwang, J.-N. (2001), *Handbook for Neural Network Signal Processing*, CRC Press.

Tsantis, S., Dimitropoulos, N., Cavouras, D. & Nikiforidis, G. (2009), 'Morphological and wavelet features towards sonographic thyroid nodules evaluation', *Computerized Medical Imaging and Graphics*, vol. 33, pp. 9199.

UPenn (2008), 'Timit sentences by the linguistic data consortium, university of pennsylvania', Online. `http://www.ldc.upenn.edu/Catalog/docs/LDC2008S09/manual/html/node16.html` Last Accessed: 03-09.

van de Graaff, K. M. (2002), *Human Anatomy*, 6th edn, McGraw-Hill Higher Education.

van Wassenhove, V., Grant, K. W. & Poeppel, D. (2007), 'Temporal window of integration in auditory-visual speech perception', *Neuropsychologia*, vol. 45, pp. 598607.

Wagner, J. L. (2005), Exploration of lip shape measures and their association with tongue contact patterns, Master's dissertation, Brigham Young University.

Wang, J., Plataniotis, K. & Venetsanopoulos, A. (2005), 'Selecting discriminant eigenfaces for face recognition', *Pattern Recognition Letters*, vol. 26, pp. 1470–1482.

WorldEnglish (2009), 'The 500 most commonly used words in the english language', Online:. `http://www.world-english.org/english500.htm` Last Accessed: 06-08.

Yadav, R. B., Nishchal, N. K., Gupta, A. K. & Rastogi, V. K. (2008), 'Retrieval and classification of objects using generic fourier, legendre moment, and wavelet zernike moment descriptors and recognition using joint transform correlator', *Optics and Laser Technology*, vol. 40, pp. 517–527.

Yu, Z. Q., Chow, P. S. & Tan, R. B. (2007), 'Quantification of particle morphology by boundary fourier transform and generic fourier transform', *Chemical Engineering Science*, vol. 62, pp. 3777–3786.

Zhang, D. & Lu, G. (2003), 'General image retrieval using shape and texture features', *In Proc. of the 7th International Conference on Internet and Multimedia Systems and Applications*, pp. 585–589.

Zhang, D. & Lu, G. (2004), 'Review of shape representation and description techniques', *Pattern Recognition*, vol. 37, pp. 1–19.

Zhang, D. & Lu, G. (2005), 'Study and evaluation of different fourier methods for image retrieval', *Image and Vision Computing*, vol. 23, pp. 33–49.

Zhang, L., Lin, F. & Zhang, B. (2001), 'Support vector machine learning for image retrieval', *Proc. of IEEE Int. Conf. on Image Processing*, pp. 721–724.

Appendix A

Ethics Approval

Ethics approval was given for human testing of the artificial larynx on the principal investigators. A copy of the approval is shown in Figure A.1. The ethics approval was granted on 14 December 2007 and the reference number is: R14/49 Russell

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Russell

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER M071124**
PROJECT	Research Toward the Development of an Artificial Larynx
INVESTIGATORS	Ms M Russell
DEPARTMENT	Electrical & Info Engineering
DATE CONSIDERED	07.11.30
DECISION OF THE COMMITTEE*	APPROVED UNCONDITIONALLY

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 07.12.14 **CHAIRPERSON** .....
(Professors PE Cleaton-Jones, A Dhai, M Vorster, C Feldman, A Woodiwiss)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof D Rubin

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10005, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

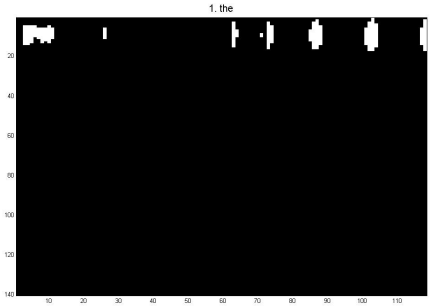
PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Figure A.1: The Ethics Approval Document

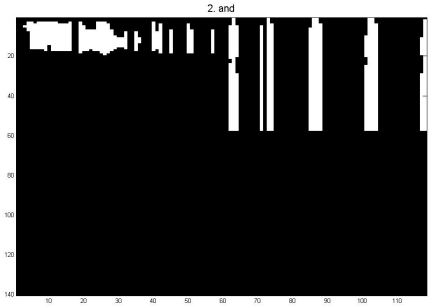
Appendix B

Space-Time Plots of the 50 Words

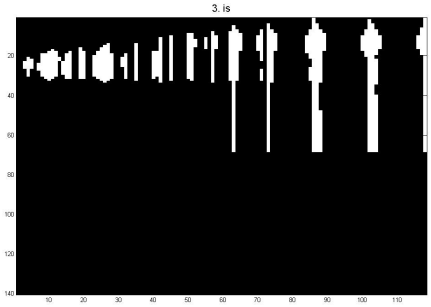
Table B.1: Space-Time plots of words 1-4



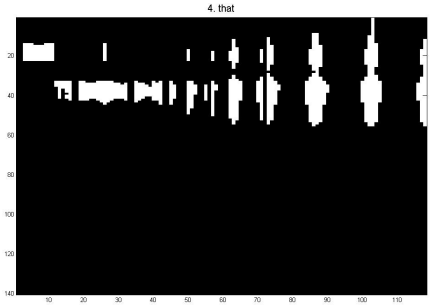
Word 1: the



Word 2: and

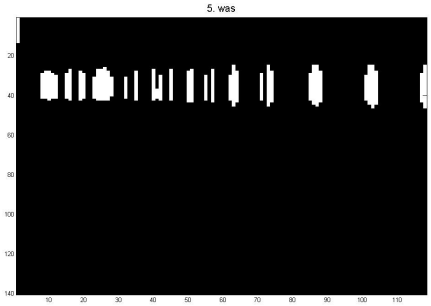


Word 3: is

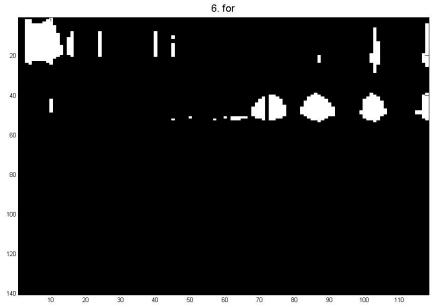


Word 4: that

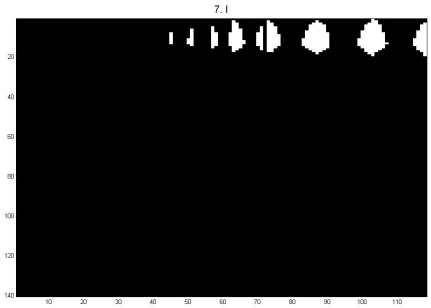
Table B.2: Space-Time plots of words 5-12



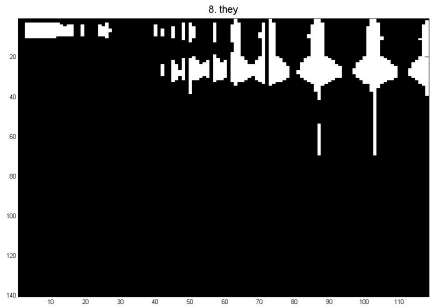
Word 5: was



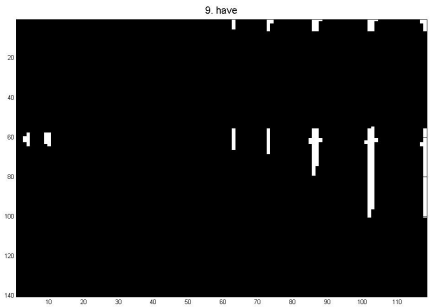
Word 6: for



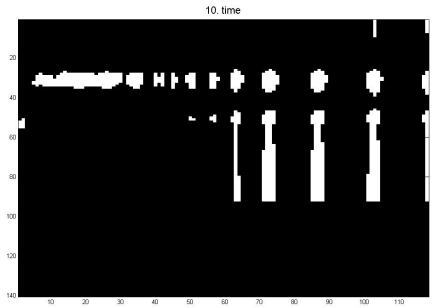
Word 7: I



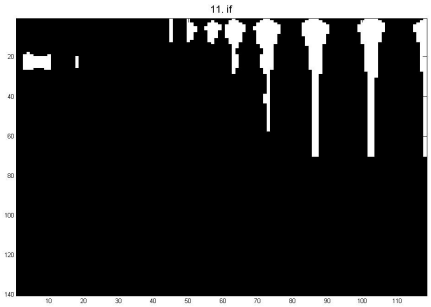
Word 8: they



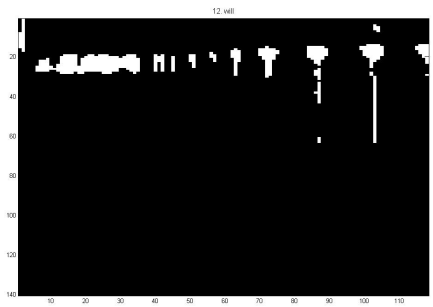
Word 9: have



Word 10: time

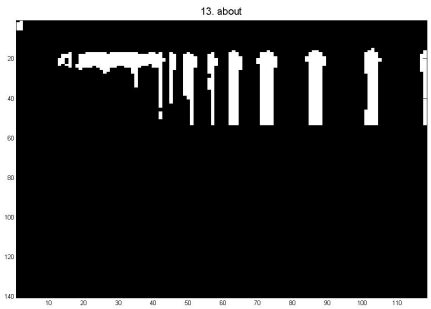


Word 11: if

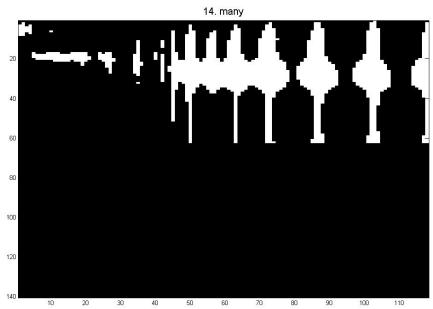


Word 12: will

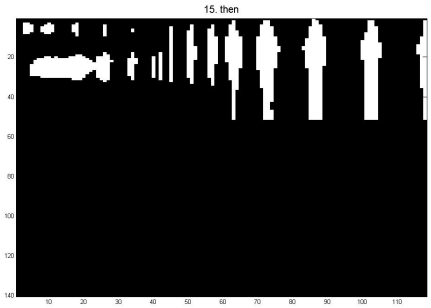
Table B.3: Space-Time plots of words 13-20



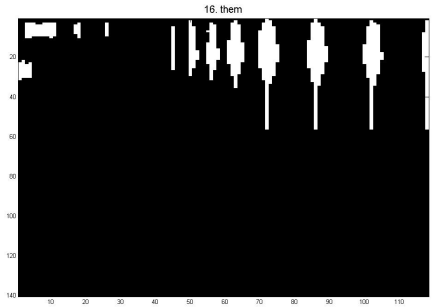
Word 13: about



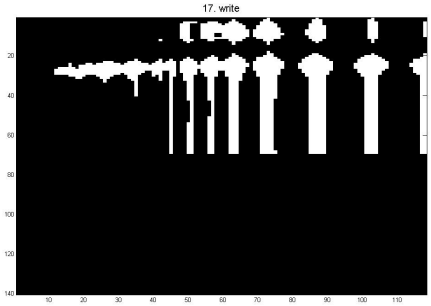
Word 14: many



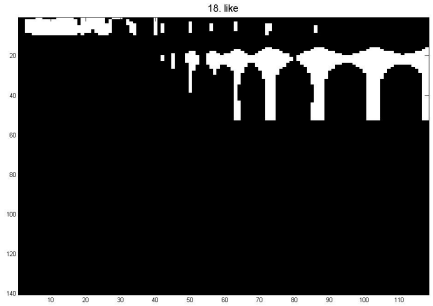
Word 15: then



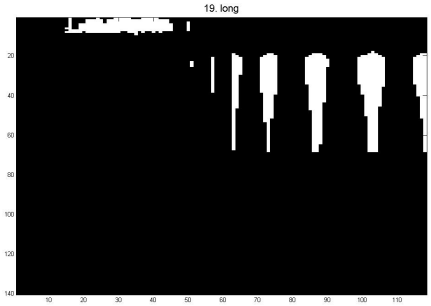
Word 16: them



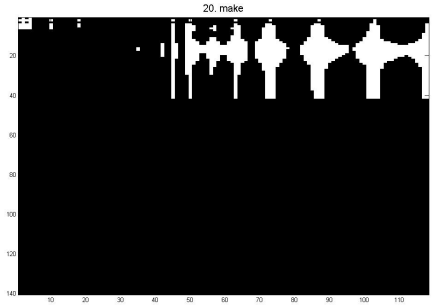
Word 17: write



Word 18: like

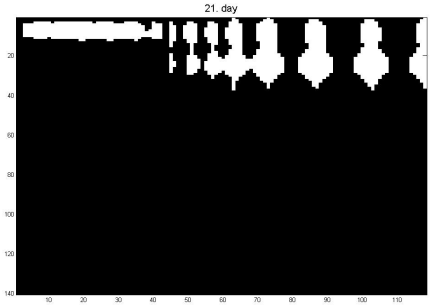


Word 19: long

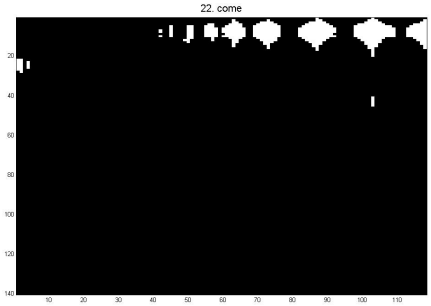


Word 20: make

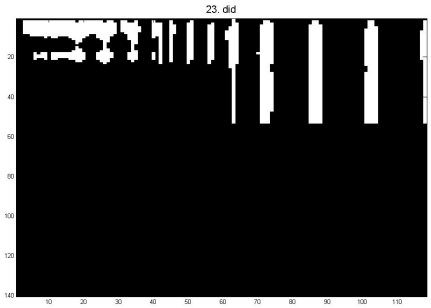
Table B.4: Space-Time plots of words 21-28



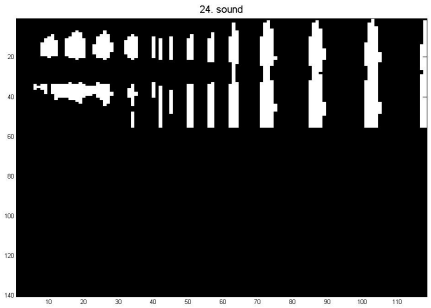
Word 21: day



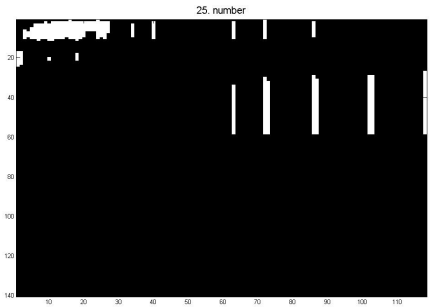
Word 22: come



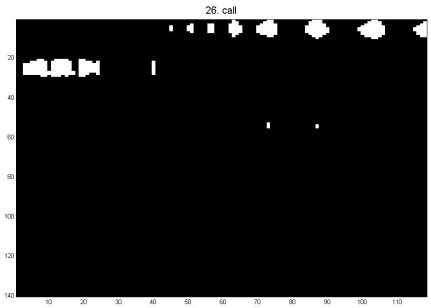
Word 23: did



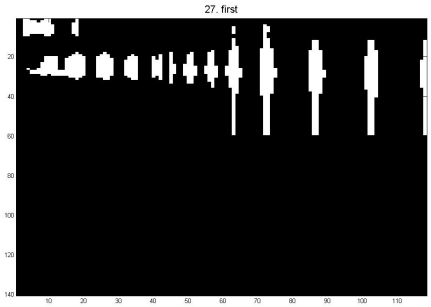
Word 24: sound



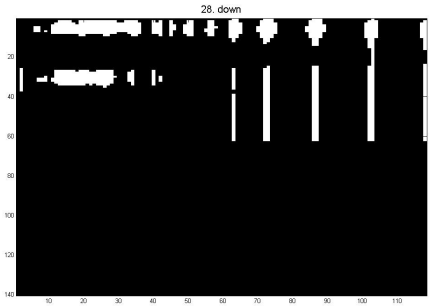
Word 25: number



Word 26: call

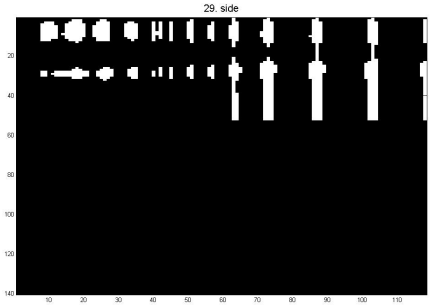


Word 27: first

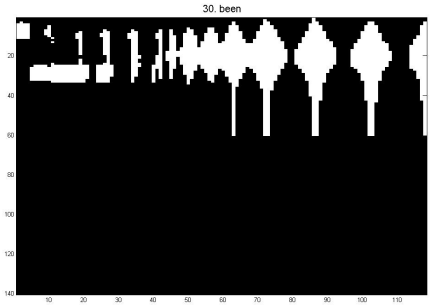


Word 28: down

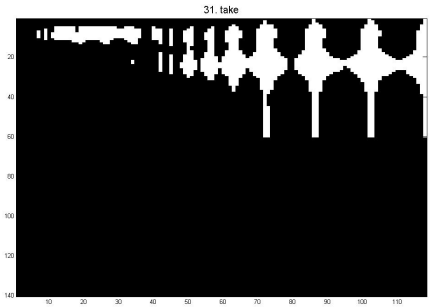
Table B.5: Space-Time plots of words 29-36



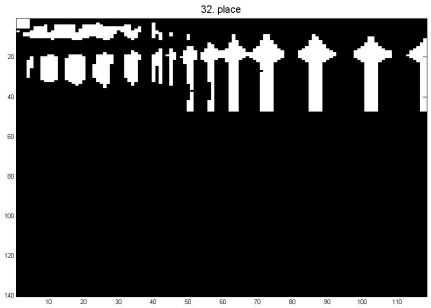
Word 29: side



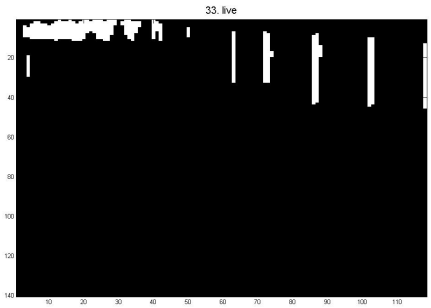
Word 30: been



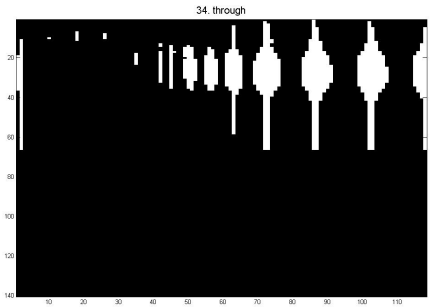
Word 31: take



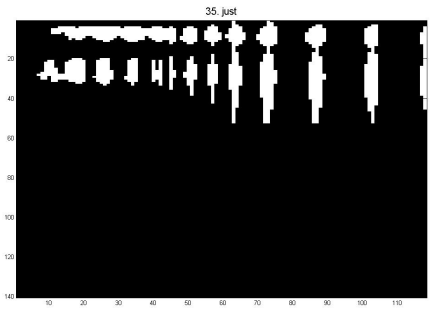
Word 32: place



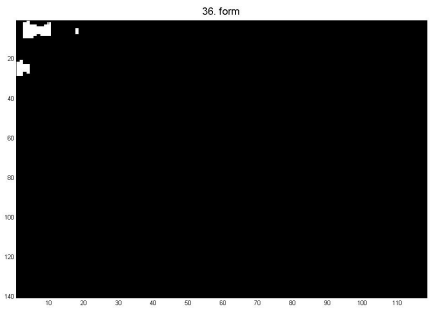
Word 33: live



Word 34: through

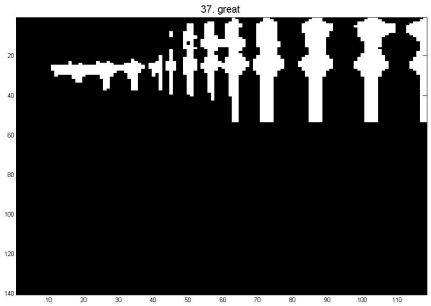


Word 35: just

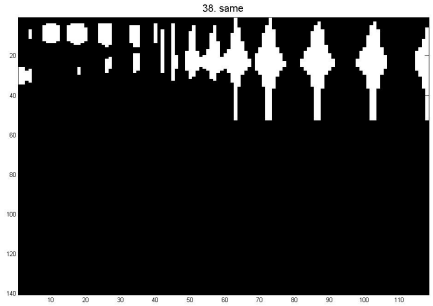


Word 36: form

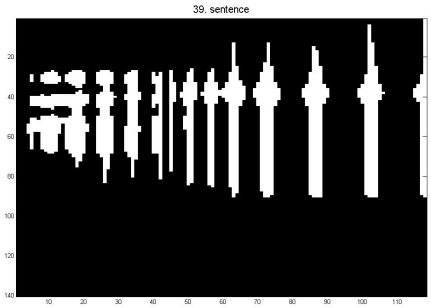
Table B.6: Space-Time plots of words 37-44



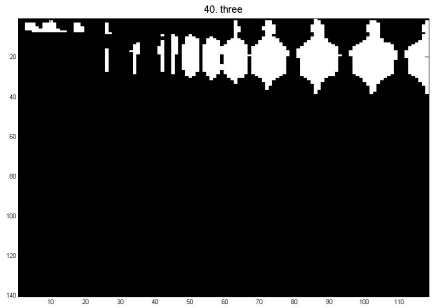
Word 37: great



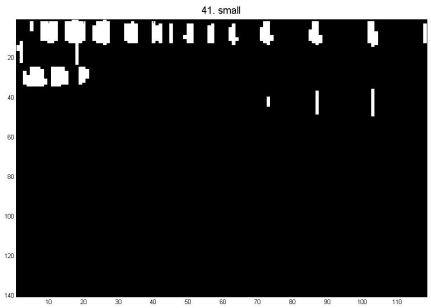
Word 38: same



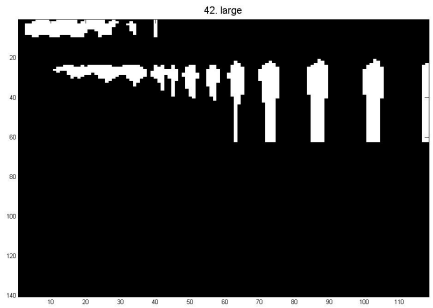
Word 39: sentence



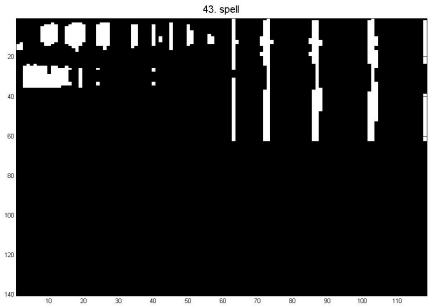
Word 40: three



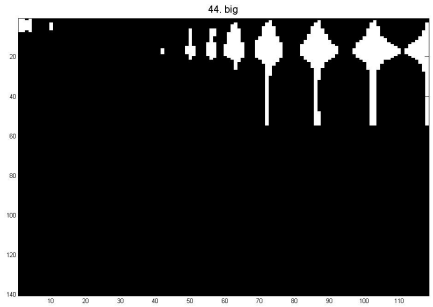
Word 41: small



Word 42: large

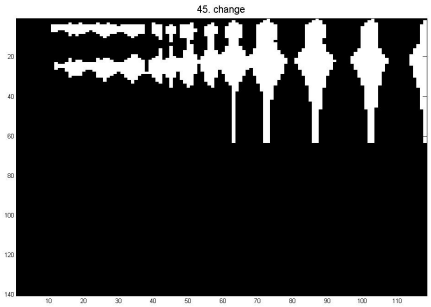


Word 43: spell

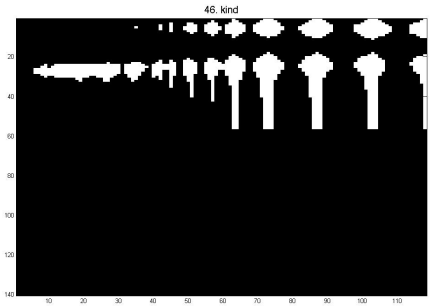


Word 44: big

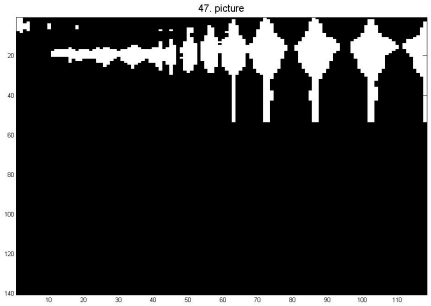
Table B.7: Space-Time plots of words 45-50



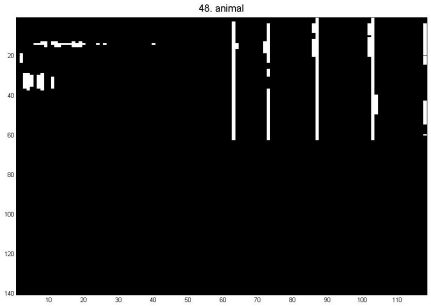
Word 45: change



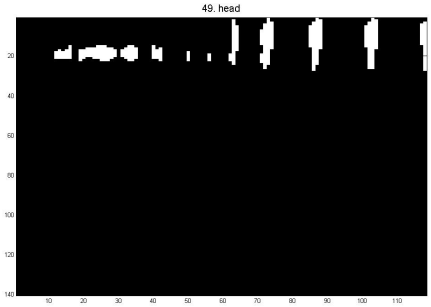
Word 46: kind



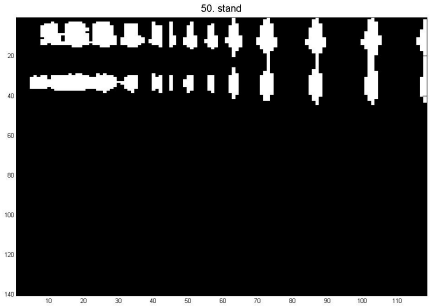
Word 47: picture



Word 48: animal



Word 49: head



Word 50: stand

Appendix C

Grammar

This is the basic grammar used for creating Voting System 2.

Grammar Used in Voting System 2

Word 1 can be followed by:
10, 12, 14, 19, 21, 24, 25, 26, 27, 29, 32, 36, 37, 38, 39, 40, 41, 42, 44, 45, 46, 47, 48, 49, 50

Word 2 can be followed by:
1, 4, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 30, 21, 22, 23, 24, 25, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

Word 3 can be followed by:
1, 2, 4, 6, 10, 13, 14, 18, 19, 21, 24, 25, 27, 38, 34, 35, 37, 40, 41, 42, 44, 46

Word 4 can be followed by:
1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 18, 29, 20, 21, 22, 23, 24, 25, 26, 27, 29, 32, 35, 36, 38, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

Word 5 can be followed by:
1, 4, 6, 7, 13, 19, 21, 24, 27, 28, 34, 35, 37, 41, 42, 44, 46

Word 6 can be followed by:
1, 4, 7, 8, 10, 11, 13, 14, 15, 16, 19, 24, 35, 36, 37, 40, 41, 42, 44, 45, 46

Word 7 can be followed by:
1, 5, 9, 12, 17, 18, 19, 20, 22, 23, 24, 25, 26, 31, 32, 33, 35, 36, 43, 45, 47, 50

Word 8 can be followed by:
4, 9, 12, 15, 17, 18, 19, 20, 22, 23, 24, 25, 26,, 31, 32, 33, 35, 36, 43, 45, 47, 50

Word 9 can be followed by:
1, 4, 7, 8, 10, 13, 14, 16, 19, 22, 24, 27, 30, 35, 40, 41, 42, 44, 45, 46, 47, 48, 49

Word 10 can be followed by:
1, 2, 3, 4, 5, 6, 7, 8, 11, 12, 14, 15, 16, 18

Word 11 can be followed by:
1, 4, 7, 8, 10, 13, 14, 19, 21, 24, 27, 35, 40, 41, 42, 44, 45, 46, 47, 48, 49

Word 12 can be followed by:
1, 4, 7, 8, 10, 13, 14, 17, 18, 19, 20, 21, 22, 24, 26, 31, 32, 33, 35, 36, 37, 40, 41, 42, 44, 45, 46, 50

Word 13 can be followed by:
1, 4, 10, 14, 16, 19, 24, 36, 40, 45, 48

Word 14 can be followed by:
6, 9, 12, 17, 18, 19, 20, 22, 26, 31, 33, 36, 41, 42, 44, 46

Word 15 can be followed by:
1, 4, 7, 8, 10, 11, 12, 13, 14, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

Word 16 can be followed by:
2, 27

Word 17 can be followed by:
1, 2, 4, 6, 11, 13, 14, 19, 41, 42, 44

Word 18 can be followed by:
1, 4, 6, 7, 8, 14, 15, 16

Word 19 can be followed by:
2, 4, 6, 10, 21, 24, 25, 26, 39, 47, 48, 49

Word 20 can be followed by:
1, 2, 4, 6, 10, 13, 14, 16, 24, 37, 40, 41, 42, 44, 45

Word 21 can be followed by:
1, 2, 4, 5, 7, 8, 10, 19

Word 22 can be followed by:
2, 6, 9, 10, 11, 13, 15, 17, 20, 21, 26, 27, 28, 31, 32, 33, 34, 45, 50

Word 23 can be followed by:
1, 4, 7, 8, 10, 14, 21, 22, 26, 31, 37, 39, 41, 42, 44, 45, 48

Word 24 can be followed by:
1, 2, 3, 4, 5, 6, 37

Word 25 can be followed by:
1, 2, 4, 6

Word 26 can be followed by:
1, 4, 5, 6, 10, 11, 12, 13, 14, 15, 16, 27, 28

Word 27 can be followed by:
1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 18, 20, 22, 24, 25, 26, 28, 31, 32, 33, 36, 40, 41, 42, 43, 44, 45, 49, 50

Word 28 can be followed by:
1, 2, 3, 4, 10, 40

Word 29 can be followed by:
3, 4, 5, 6, 28

Word 30 can be followed by:
1, 4, 19, 27, 28, 34, 37, 41, 42, 44, 46

Word 31 can be followed by:
1, 4, 6, 10, 13, 14, 15, 19, 27, 28, 34, 41, 42, 44, 45

Word 32 can be followed by:
1, 4, 5, 6, 13, 14, 16, 26, 40, 41, 42, 44

Word 33 can be followed by:
1, 2, 4, 6, 11, 13, 14, 18, 19, 27, 34, 41, 42, 44

Word 34 can be followed by:
1, 2, 4, 10, 14, 16, 41, 42, 44

Word 35 can be followed by:
1, 4, 6, 7, 8, 9, 11, 13, 15, 17, 20, 22, 23, 26, 30, 31, 32, 33, 37, 40, 41, 42, 44, 45, 47, 49, 50

Word 36 can be followed by:
1, 2, 3, 4, 5, 13, 14, 19, 40, 41, 42, 44

Word 37 can be followed by:
2, 4, 6, 10, 14, 21, 24, 25, 26, 32, 36, 39, 44, 45, 46, 47, 48, 49

Word 38 can be followed by:
3, 4, 6, 10, 15, 20, 21, 24, 25, 26, 29, 32, 36, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50

Word 39 can be followed by:
1, 2, 3, 4, 5, 6, 13, 14, 16

Word 40 can be followed by:
2, 3, 4, 5, 6, 9, 12, 18, 23, 27, 37, 41, 42, 44

Word 41 can be followed by:
2, 3, 5, 24, 25, 32, 39, 43, 45, 47, 48, 49

Word 42 can be followed by:
2, 3, 5, 25, 29, 32, 47, 48, 49

Word 43 can be followed by:
1, 4, 6, 13, 14, 18

Word 44 can be followed by:
2, 3, 5, 21, 24, 25, 32, 39, 45, 47, 48, 49

Word 45 can be followed by:
1, 2, 3, 4, 5, 6, 11, 12, 14, 27, 32

Word 46 can be followed by:
2, 4, 13

Word 47 can be followed by:
1, 4, 5, 6, 13, 14, 16, 40, 41, 42, 44, 45

Word 48 can be followed by:
4, 5, 23, 47, 49

Word 49 can be followed by:
1, 2, 3, 4, 5, 6, 15, 27, 28, 34, 50

Word 50 can be followed by:
1, 4, 5, 6, 11, 14, 15, 28

Appendix D

MLP Hidden Nodes

The results of running the MLP with different numbers of hidden nodes are shown.

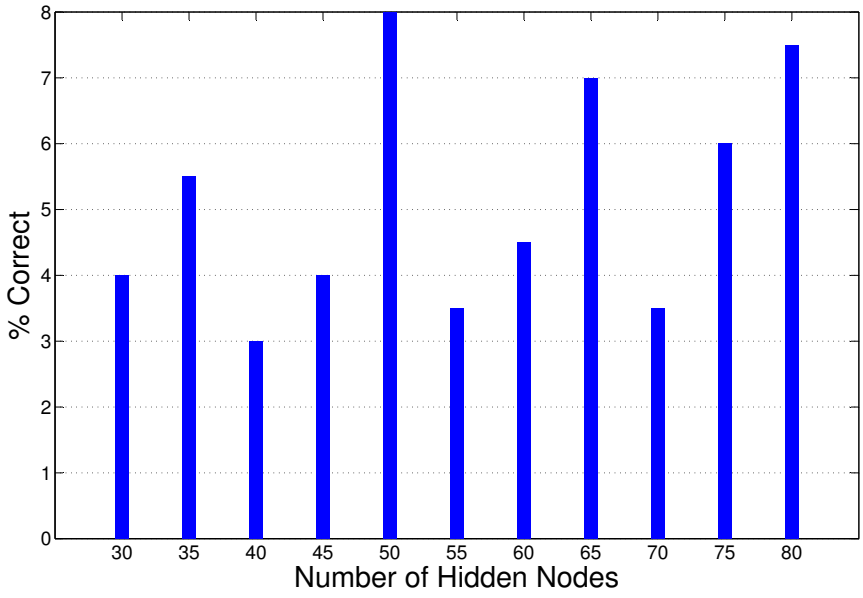


Figure D.1: MLP classification success rate using Principal Components and different numbers of hidden nodes

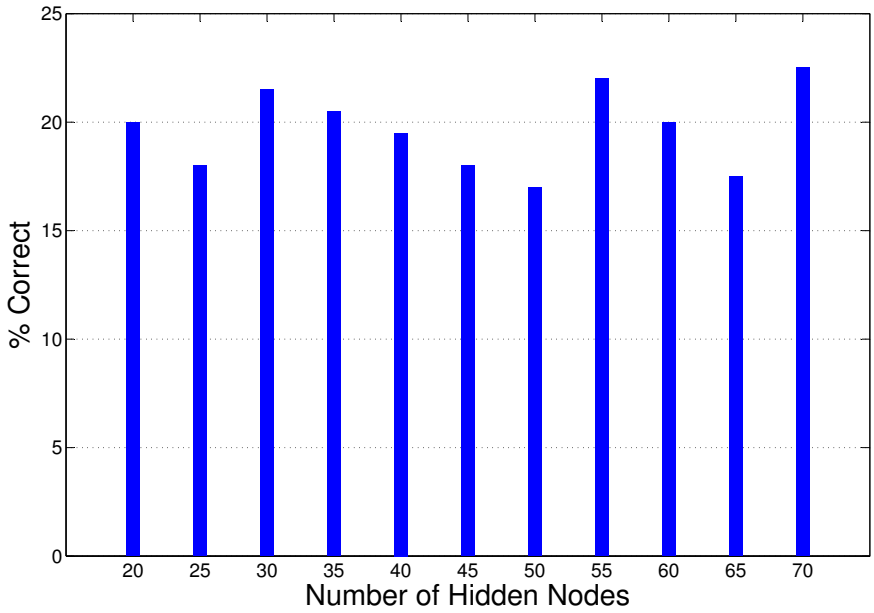


Figure D.2: MLP classification success rate using Fourier descriptors and different numbers of hidden nodes

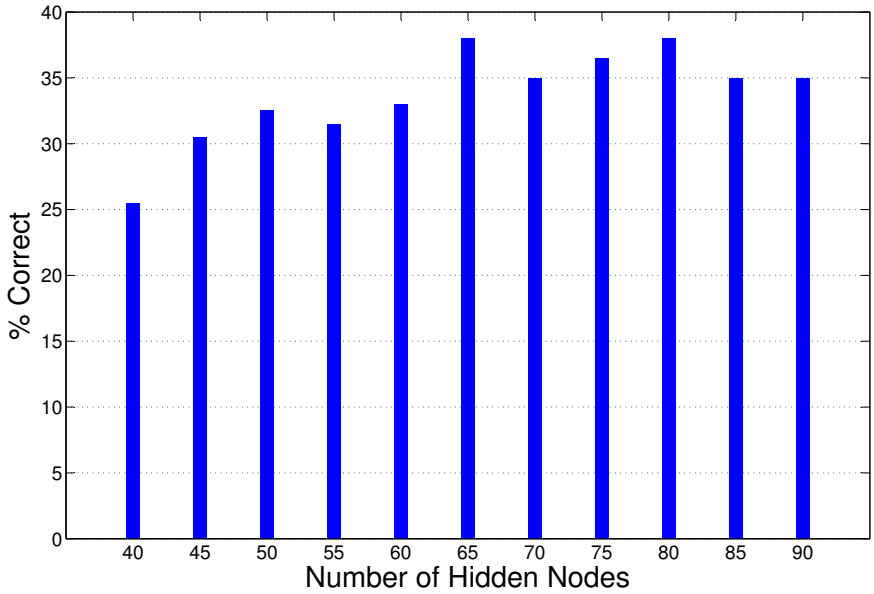


Figure D.3: MLP classification success rate using generic Fourier descriptors and different numbers of hidden nodes

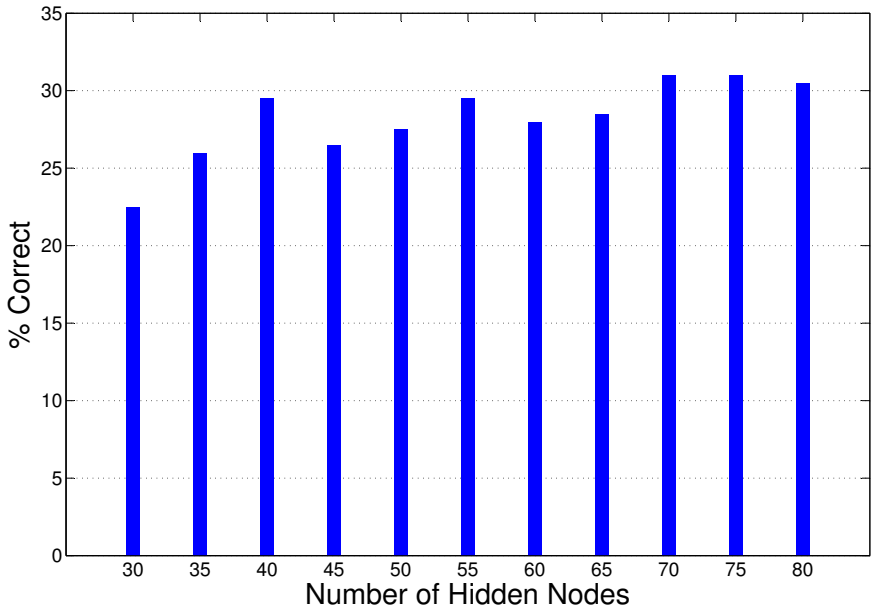


Figure D.4: MLP classification success rate using four image properties and different numbers of hidden nodes

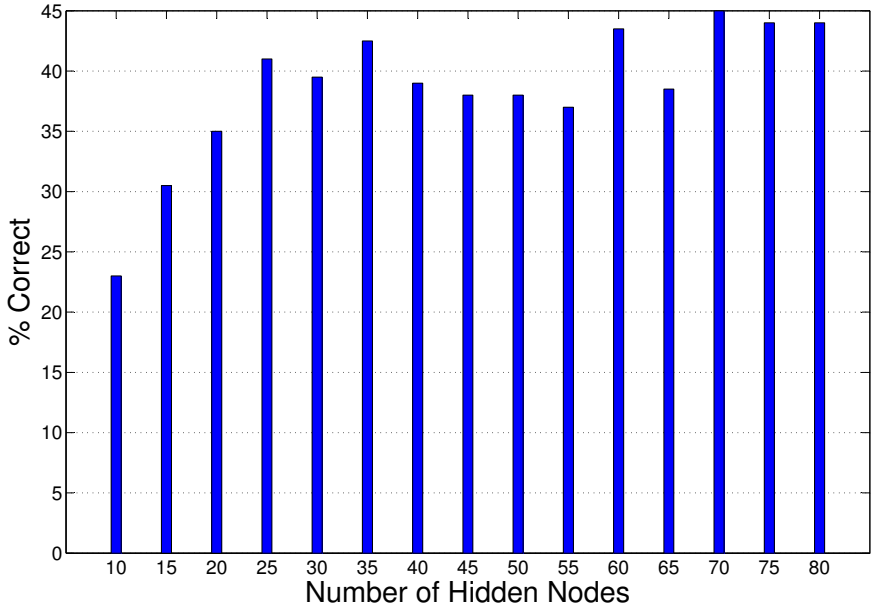


Figure D.5: MLP classification success rate using 13 image properties and different numbers of hidden nodes

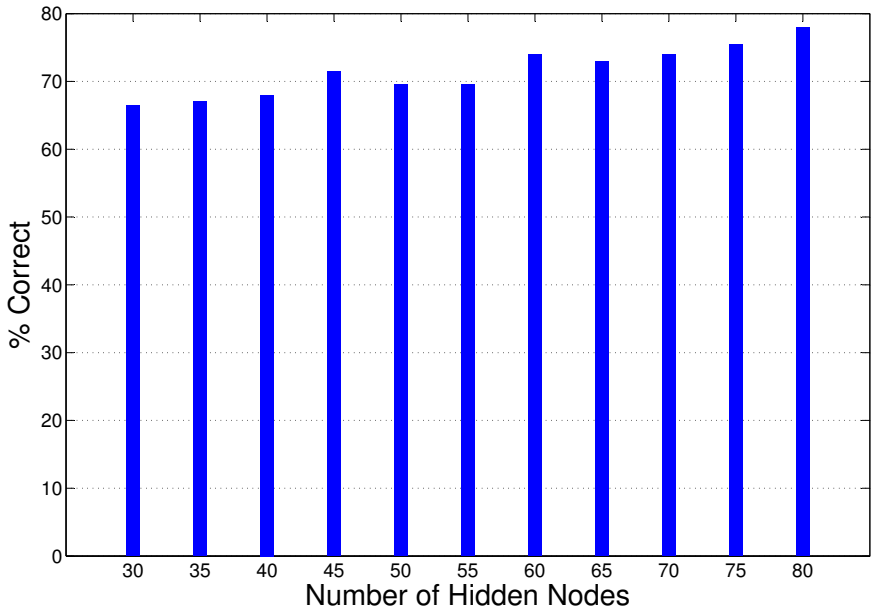


Figure D.6: MLP classification success rate using Correlation Coefficients and different numbers of hidden nodes

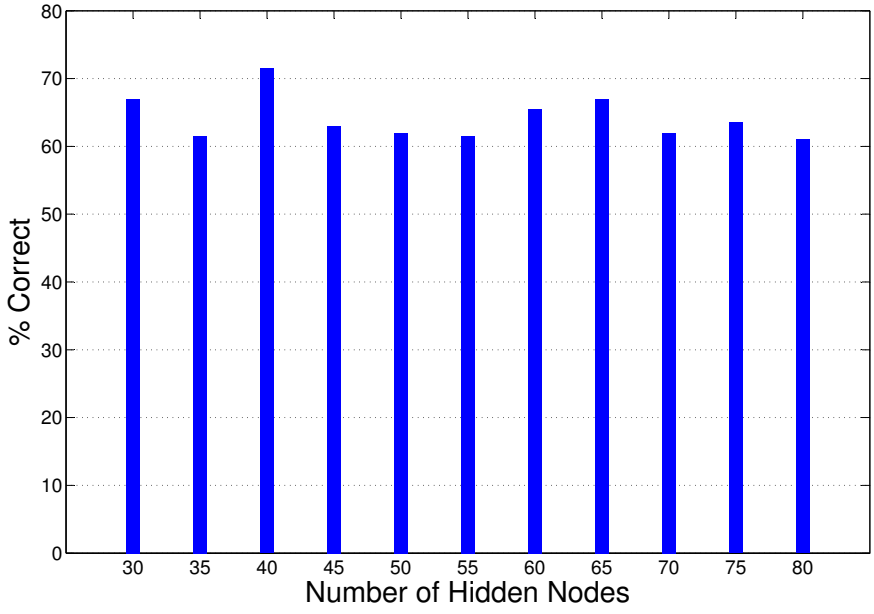


Figure D.7: MLP classification success rate using four image properties and Correlation Number with different numbers of hidden nodes

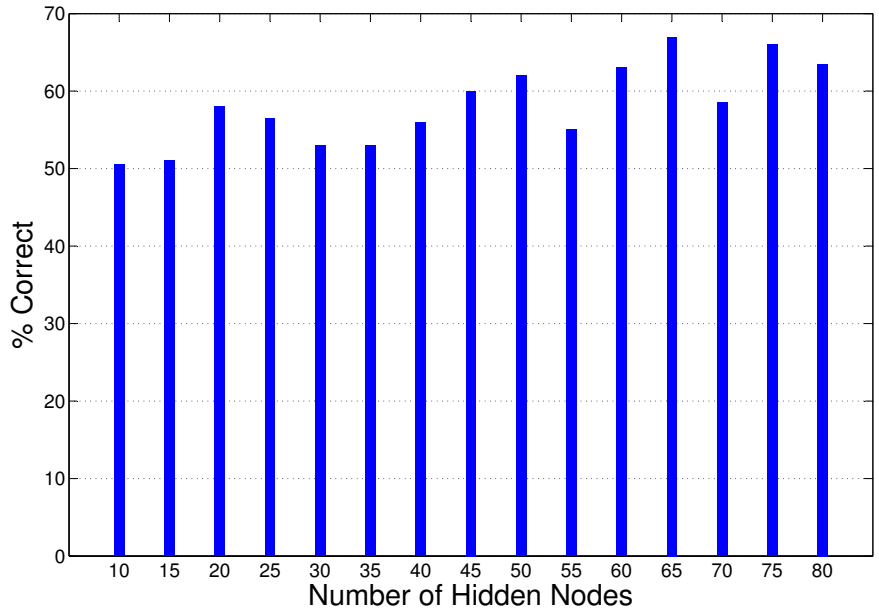


Figure D.8: MLP classification success rate using 13 image properties and Correlation Number with different numbers of hidden nodes

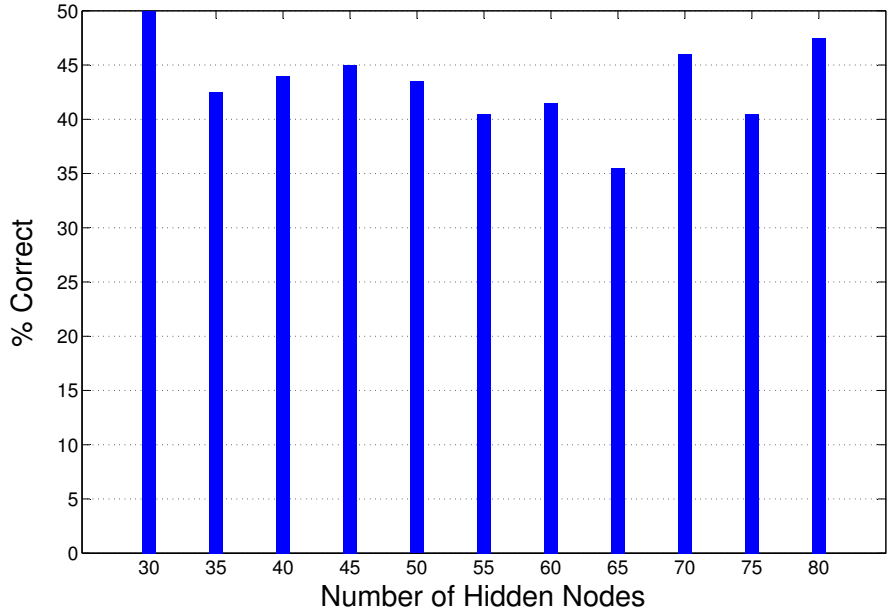


Figure D.9: MLP classification success rate using Fourier descriptors and Correlation Number with different numbers of hidden nodes

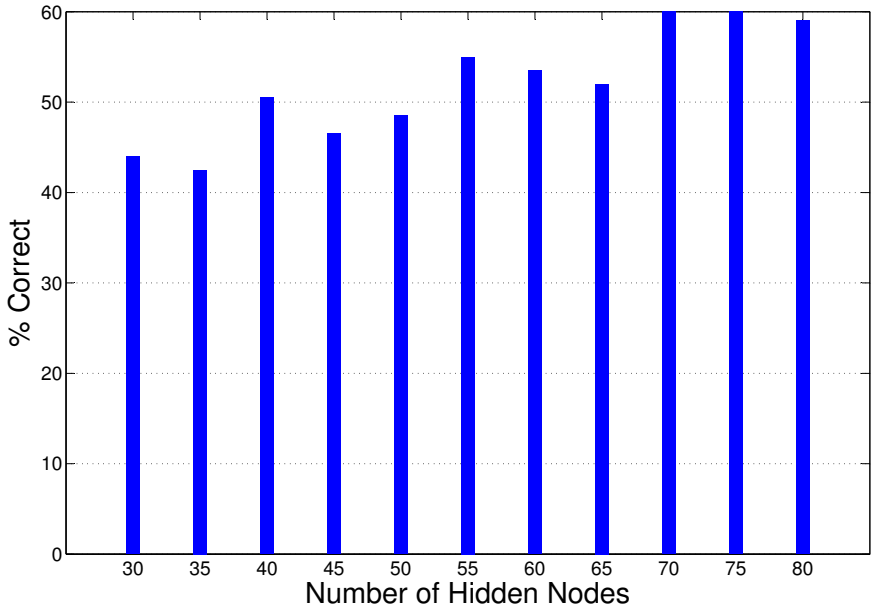


Figure D.10: MLP classification success rate using Fourier descriptors, four image properties and Correlation Number with different numbers of hidden nodes

Appendix E

TIMIT Sentences

The 24 phonetically rich TIMIT sentences are (UPenn, 2008):

1. She had your dark suit in greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. A boring novel is a superb sleeping pill.
4. Call an ambulance for medical assistance.
5. We saw eight tiny icicles below our roof.
6. Each untimely income loss coincided with the breakdown of a heating system part.
7. Jeff thought you argued in favor of a centrifuge purchase.
8. The sermon emphasized the need for affirmative action.
9. Kindergarten children decorate their classrooms for all holidays.
10. Cory and Trish played tag with beach balls for hours.
11. The frightened child was gently subdued by his big brother.
12. The tooth fairy forgot to come when Roger's tooth fell out.
13. Alice's ability to work without supervision is noteworthy.
14. Special task forces rescue hostages from kidnappers.
15. If Carol comes tomorrow, have her arrange for a meeting at two.
16. Military personnel are expected to obey government orders.

17. Laugh, dance, and sing if fortune smiles upon you.
18. The fish began to leap frantically on the surface of the small lake.
19. The easygoing zoologist relaxed throughout the voyage.
20. Brush fires are common in the dry underbrush of Nevada.
21. How much will it cost to do any necessary modernizing and redecorating?
22. Was she just naturally sloppy about everything but her physical appearance?
23. Is a relaxed home atmosphere enough to help her outgrow these traits?
24. The same shelter could be built into an embankment or below ground level.

Appendix F

Published Papers, Articles and Patents

The following published papers resulted from this work:

- M J Russell, D M Rubin, B Wigdorowitz and T Marwala, "The Artificial Larynx: A Review of Current Technology and a Proposal for Future Development", *NBC 2008 Proceedings*, Vol. 20, pp. 160-163, June 2008
- Megan J. Russell, David M. Rubin, Tshilidzi Marwala, Brian Wigdorowitz, "Pattern Recognition and Feature Selection for the Development of a New Artificial Larynx", O. Dossel and W C. Schlegel. (Eds.): *WC 2009 IFMBE Proceedings*, 25/IV, pp. 736-739, 2009
- M. J. Russell, D. M. Rubin, T. Marwala and B. Wigdorowitz, "A Voting and Predictive Neural Network System for use in a New Artificial Larynx", *IEEE Proceedings of the 2nd International Conference in Biomedical and Pharmaceutical Engineering*, in press, 2009

An article about this work was written by Rachel Kremen for MIT Technology Review (www.technologyreview.com). The article can be accessed online at: `http://www.technologyreview.com/biomedicine/24051`. A printout of this article is given at the end of this appendix.

The following patent resulted from this work:

- M.J.Russell; D.M.Rubin; B.Wigdorowitz; T.Marwala; PCT Patent Application PCT/IB2009/006125 An Artificial Larynx (Priority South African Provisional Patent Application 2008/05078, filed 11.06.2008).