

University of Southampton

Automatic Tracking of 3D Vocal Tract Features
During Speech Production Using MRI

by

María Susana Avila García

A thesis submitted for the degree of
Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

December 2006

University of Southampton

ABSTRACT

School of Electronics and Computer Science

Doctor of Philosophy

Automatic Tracking of 3D Vocal Tract Features

During Speech Production Using MRI

by María Susana Avila García

Magnetic resonance imaging has many advantages for visualising the process of speech production, but an important disadvantage is the long scanning acquisition time relative to the characteristic time of articulator motion of tenth of second. Southampton Dynamic Magnetic Resonance Imaging is a technique developed in a previous project to solve this problem. This technique achieves an *apparent* high temporal resolution suitable for dynamic studies. Consequently, a large number of images can be generated describing the evolution of the vocal tract shape. This makes a manual extraction of the vocal tract shape a tedious and time consuming process. The aim of this project firstly is to improve and extend the SDMRI method, and secondly, to determine the outline of the vocal tract automatically. Different feature extraction techniques were analysed and two of them were combined to make a new automatic shape extraction tool, i.e. the active shape models and the Hough transform. Active shape models describe the shape of the articulators while the Hough transform locates it with no initialisation. Initially, the new algorithm was tested analysing isolated magnetic resonance images for extracting tongue shapes; however, although the results were satisfactory the algorithm often fails when multiple solutions are present. A global analysis of the image sequence overcomes these difficulties and the dynamic Hough transform was adapted for our purposes. Experimental results reveals that the algorithm does indeed find the correct shape and position of the tongue and also that it is robust under noisy conditions. The model was extended to other articulators, i.e. the lips. This approach leads to a new algorithm for

automatic extraction of articulatory shape in magnetic resonance image sequences as evident in the results presented in this thesis.

Acknowledgments

I would like to thank the Mexican National Council for Science and Technology (CONACyT) for supporting me with the scholarship number 144969, without which this work would not have been possible.

I would like to thank Dr. John Carter and Prof. Bob Damper for giving me the chance to be part of this project, and for their support throughout the different stages of this project. I would like to acknowledge Dr. Carter for his collaboration with his algorithms for generating random numbers and converting text files into wav files. I would like to specially thank Dr. Carter for his help on difficult situations presented on my personal life. Many thanks for supporting me when I needed it the most! I would like to continue learning and collaborating with them in the future.

I would also like to thank Dr. Ryan C.N. D'Arcy, Dr. Steven Beyea, Dr. Yannick Marchand and James Rioux at the Institute for Biodiagnostics (Atlantic) Neuroimaging Research Laboratory for giving me the opportunity of collecting data for my project and for all their help through this process. I would like to acknowledge Josh Bray for his time and the algorithms developed for the gating pulse sequence acquisition. Dr. Beyea and James for collaborating with me in the analysis of the data and for sharing their knowledge. I would like to thank our 'Subjects' whose name I cannot mention for privacy reasons, many thanks for their collaboration and patience during the scanning sessions.

I am also grateful for the information and noise reduction algorithms provided by Ioannis Andrianakis, PhD candidate at the Institute of Sound and Vibration Research at the University of Southampton. I would like to thank Dr. Pelopidas Lappas for sharing his

knowledge and the dynamic Hough transform algorithm. I also would like to thank my friend Dr. Gabriela Gonzalez for her feedback on my thesis.

Personally, I would like to thank God. Thanks for giving me the opportunity to discover that more than a Father you are my best friend. Thanks for taking care of me when I needed it the most. Thanks for making me feel your presence, your support and your love. I love you.

I would like to dedicate and greatly thank my family. My parents, J. Antonio Avila Gutierrez and Eugenia Garcia Gonzalez who let me be who I am. Because you offered me the most valuable support in my whole life and because in the most difficult moments of our lives you never gave up. I am not going to forget the favourite saying of my Father: *'Dios aprieta pero no ahorca'*. I love you very much Papiringo and Mamiringa. My brother Francisco Antonio Avila Garcia, many thanks for expressing your feelings Carnalito. I always remember that tender hug you gave me the time I was leaving home. I love you my Bodoquito.

I would specially thank Marco Bianchetti, for your support and for being an important part of my present life. Many thanks for giving us the chance of living what we are living now! You are one of the most wonderful gifts God gave me. Te amo mi Guerito!

Thanks go to my friends in Mexico for their support. I would also like to say it has been a pleasure to share this wonderful experience with my new family in Southampton since the first day I arrived until now.

Finally, I just would like to say... Viva Mexico!!!!!!!!!!!!!!

knowledge and the dynamic Hough transform algorithm. I also would like to thank my friend Dr. Gabriela Gonzalez for her feedback on my thesis.

Personally, I would like to thank God. Thanks for giving me the opportunity to discover that more than a Father you are my best friend. Thanks for taking care of me when I needed it the most. Thanks for making me feel your presence, your support and your love. I love you.

I would like to dedicate and greatly thank my family. My parents, J. Antonio Avila Gutierrez and Eugenia Garcia Gonzalez who let me be who I am. Because you offered me the most valuable support in my whole life and because in the most difficult moments of our lives you never gave up. I am not going to forget the favourite saying of my Father: *'Dios aprieta pero no ahorca'*. I love you very much Papiringo and Mamiringa. My brother Francisco Antonio Avila Garcia, many thanks for expressing your feelings Carnalito. I always remember that tender hug you gave me the time I was leaving home. I love you my Bodoquito.

I would specially thank Marco Bianchetti, for your support and for being an important part of my present life. Many thanks for giving us the chance of living what we are living now! You are one of the most wonderful gifts God gave me. Te amo mi Guerito!

Thanks go to my friends in Mexico for their support. I would also like to say it has been a pleasure to share this wonderful experience with my new family in Southampton since the first day I arrived until now.

Finally, I just would like to say... Viva Mexico!!!!!!!!!!!!!!

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction | 1 |
| 1.1 | Motivation | 1 |
| 1.2 | Thesis Outline | 4 |
| 2 | Speech Production Research Using MRI | 6 |
| 2.1 | Introduction | 6 |
| 2.2 | Speech Production | 7 |
| 2.3 | Vocal Tract Modelling | 8 |
| 2.4 | Methods for Acquiring Vocal Tract Data | 10 |
| 2.5 | Speech and MRI | 15 |
| 2.6 | Speech and Retrospective Gating | 16 |
| 2.7 | Dynamic MRI in Speech | 17 |
| 3 | Feature Extraction | 19 |
| 3.1 | Introduction | 19 |
| 3.2 | Arbitrary Shape Feature Extraction Techniques | 20 |
| 3.2.1 | Template Matching | 20 |
| 3.2.2 | Hough Transform | 20 |
| 3.2.3 | Deformable Templates | 21 |
| 3.2.4 | Snakes | 22 |
| 3.2.5 | Active Shape Models | 23 |
| 3.3 | Conclusions | 24 |
| 4 | Active Shape Hough Transform | 25 |
| 4.1 | Introduction | 25 |
| 4.2 | Active Shape Models | 26 |
| 4.3 | Combining Active Shape Models and the Hough Transform | 28 |
| 4.4 | Application to MRI Data | 30 |
| 4.4.1 | Noise Reduction and Edge Detection | 31 |
| 4.4.2 | Generation of the Tongue Model | 36 |
| 4.5 | Conclusions | 41 |
| 5 | Active Shape Dynamic Hough Transform | 43 |
| 5.1 | Introduction | 43 |
| 5.2 | Dynamic Hough Transform | 44 |
| 5.3 | Combining Active Shape Models and the Dynamic Hough Transform | 48 |

| | | |
|----------|---|------------|
| 5.4 | Application to MRI Sequences | 49 |
| 5.5 | Conclusions | 52 |
| 6 | Experimental Evaluation | 54 |
| 6.1 | Introduction | 54 |
| 6.2 | Chamfer Distance Metric | 56 |
| 6.3 | Ground-Truth Data Set | 57 |
| 6.3.1 | Isolated Tongue Shapes Experiments | 57 |
| 6.3.2 | Full Edge Images Experiments | 62 |
| 6.4 | Unseen Data Set | 66 |
| 6.5 | Synthetic Data | 70 |
| 6.6 | Noise Response | 71 |
| 6.7 | Extending the Model to Lips | 75 |
| 6.8 | Summary | 80 |
| 7 | Collecting Data in Halifax | 89 |
| 7.1 | Introduction | 89 |
| 7.2 | Data Acquisition | 90 |
| 7.2.1 | Speech Recording | 90 |
| 7.2.2 | Image Acquisition | 92 |
| 7.3 | Data Analysis | 94 |
| 7.3.1 | Analysis of the Speech Data | 95 |
| 7.3.2 | Analysis of Gating Pulses | 98 |
| 7.4 | Synchronisation | 100 |
| 7.5 | Reconstruction | 102 |
| 7.6 | Summary | 104 |
| 8 | Evaluating Results | 106 |
| 8.1 | Introduction | 106 |
| 8.2 | Results | 107 |
| 8.3 | Experimental Evaluation | 111 |
| 8.4 | Application of the ASHT Algorithm on Halifax Data | 119 |
| 8.5 | Conclusions | 126 |
| 9 | Conclusions and Future Work | 130 |
| 9.1 | Introduction | 130 |
| 9.2 | Discussion | 131 |
| 9.3 | Future Work | 135 |
| 9.3.1 | Improving speed | 135 |
| 9.3.2 | Improving Data Acquisition | 136 |
| A | Magnetic Resonance Imaging | 138 |
| A.1 | Introduction | 138 |
| A.2 | Basic Principles | 139 |
| A.3 | Slice selection | 141 |
| A.4 | Fourier Imaging | 143 |

| | | |
|----------|---|------------|
| A.5 | Imaging Techniques | 145 |
| A.6 | Scanners | 146 |
| A.7 | Artefacts | 147 |
| A.8 | Safety | 147 |
| B | Southampton Dynamic Magnetic Resonance Imaging | 149 |
| B.1 | Introduction | 149 |
| B.2 | Basic Theory | 150 |
| B.2.1 | Data Acquisition | 150 |
| B.2.2 | Synchronisation | 151 |
| B.2.3 | Reconstruction | 153 |
| B.3 | Mohammad Results | 154 |
| | Bibliography | 160 |
| | List of Authors’s Relevant Publications | 168 |

List of Figures

| | | |
|-----|---|----|
| 2.1 | Sagittal view of speech articulators of the vocal tract, (1) lips, (2) teeth, (3) tooth root, (4) tongue tip, (5) tongue blade, (6) tongue dorsum, (7) epiglottis, (8) vocal folds, (9) larynx, (10) pharynx, (11) velum, (12) soft palate, (13) hard palate. Reproduced from O'Shaughnessy (2000). | 8 |
| 2.2 | Two dimensional transducer attached to the upper lip, lower lip and jaw. Reprinted from Muller and Abbs (1979). | 11 |
| 2.3 | Electromagnetic midsagittal articulometer. Reprinted from Perkell et al. (1992). | 12 |
| 4.1 | Landmark and edge points defined to illustrate the voting process shown in Figure 4.2. In (a) the four landmark points considered in the model definition are shown, and the remaining images show the shapes and the edge points considered in the searching process. | 30 |
| 4.2 | Voting process. The landmarks and edge points considered in this example are defined in Figure 4.1. The sequence represents the voting process for the first edge point of the shape presented in Figure 4.1(b). The first edge point in the shape (outlined with a solid line) is defined as the fourth "d" landmark point for all possible shapes in the set, outlined with dash lines; centroids are represented by a cross mark. | 30 |
| 4.3 | Example of voting process. The landmarks and edge points considered in this example are defined in Figure 4.1(a),(b). The target is presented with solid outline, while the dashed outlined shapes represent the matching shape in the process. The first image represents the matching of the centroid of the target and the shape used, placing the point (1) as the (a) landmark point. Then, the same edge point in defined as the (b),(c) and (d) landmark points. In the second row, the edge point (2) is defined as the (a),(b),(c) and (d) landmark points respectively, incrementing the same centroid position when the edge point is placed in the (c) landmark position. Finally, the third row shows the iterative process for the edge point (3), the centroid, as mentioned before, is incremented in one when the (d) landmark point is used. The ideal peak will have a value equal to the number of edge points. | 31 |
| 4.4 | Comparison between Gaussian (a) and anisotropic filtering (b). Gaussian image was obtained using a window size of 4×4 and $\sigma = 4.5$. Anisotropic image was obtained setting $K = 90\%$ and $t = 10$. The anisotropic filtered image shows sharper edges, in particular, the neck, the larynx, the pharynx and the upper lip. | 34 |

| | | |
|------|---|----|
| 4.5 | Images c_N , c_S , c_E and c_W which contain the edge information in the anisotropic filtering process. The images c_N and c_S give the horizontal edge definition whilst vertical edge information is provided by c_E and c_W . Differences between c_N and c_S are almost imperceptible due to the definition of this coefficients in terms of the $\nabla_N I_{i,j}$ and $\nabla_S I_{i,j}$ differences. This applies to c_E and c_W | 34 |
| 4.6 | The complete edge detection image of the first frame. | 35 |
| 4.7 | Comparison between Sobel and Canny edge detectors. Better results are obtained with the anisotropic filtered image due to the noise reduction were spurious details were reduced. Very similar results were obtained using the Sobel and the Canny edge detectors on the anisotropic filtered image. | 35 |
| 4.8 | Hand labelled midsagittal frame of the vocal tract. | 36 |
| 4.9 | Vocal tract shape. (a) Sum of all the vocal tract shapes outlined by Mohammad. Thicker boundaries indicates motion and/or deformation of the articulators. (b) shows the average of the sum and with different grey level where the motion and deformation of the articulators can be observed. | 37 |
| 4.10 | Selection of vocal tract shape points. The centroid of the shape is calculated, then a vertical line is drawn and the angle $\theta(45^\circ)$, on both sides defines the set of points to be discarded. | 37 |
| 4.11 | (a) Original tongue shape, extracted from the outlined data set generated by Mohammad. (b) Reduced tongue shape (c) Interpolated and smoothed tongue shape with 61 points | 38 |
| 4.12 | Superimposed tongue shapes. | 38 |
| 4.13 | Tongue shapes generated using one eigenvalue and 9 different steps. It can be observed how the tongue blade is raised towards the hard palate. | 39 |
| 4.14 | Sequence of training tongue shapes. | 40 |
| 4.15 | The ASHT algorithm was tested using one mode of variation model in 11 steps First row shows three edge images selected from the resulting 39 frames of the SDMRI sequence generated by Mohammad. Second row shows results of the application of the HT combined with the ASM method in an isolated analysis. The resulting tongue shape is superimposed in the edge image. | 41 |
| 5.1 | Process to track arbitrary deforming shapes in a sequence using the new form of the Hough transform and the new adaptation of the DHT. First, binary images are generated; then, the HT method is applied to generate multidimensional voting spaces for each image. Finally, the new adaptation of the DHT is applied using a dynamic programming formulation to generate the optimal path. | 46 |
| 5.2 | Relationship among parameters: peak, velocity and direction, reprinted from Lappas et al. (2001). Two double element state variables are defined, first variable is composed by (u_{t-1}, u_t) and the second one by (u_t, u_{t+1}) . The cost function is then based on a 3 frame basis with two-element state variables. The cost function for these two variables is defined by Equation 5.5. | 47 |

| | | |
|------|---|----|
| 5.3 | Coarse to fine dynamic programming formulation. In (a) the original dynamic programming scheme is presented. In the remaining graphs, a possible progress of a coarse to fine formulation is presented. The definition of super states is represented by rectangles containing different states. First superstates are defined in (b) and the optimal path for these supervariables is determined and shown with a solid bold line. A finer optimisation is performed in (c). Finally, the optimal path is determined in (d). The super states defined in the optimal path are refined iteratively until the optimal path is formed by single peaks. Reprinted from Raphael (2001). | 48 |
| 5.4 | Comparative results between the ASHT and the ASDHT algorithms applied to full edge images. The algorithms were tested using one mode of variation model in 11 steps. First row shows three edge images selected from the resulting 39 frames of the MRI sequence generated by Mohammad. Second row shows results of the application of the HT combined with the ASM method in an isolated analysis. Finally, third row shows the results of extracting the tongue shape during the global analysis of the sequence by applying the ASDHT algorithm. The resulting tongue shape is superimposed in the edge image. | 52 |
| 6.1 | Tongue shape of the frame 39 of the training data sequence and it is corresponding chamfer distance image. The estimated shape (c) is superimposed on the chamfer distance image (d). The error calculated for this match is 1.59; the difference between both shapes is appreciated in (e) with the target shape in grey and the fitted result in black. | 57 |
| 6.2 | Average rms error (chamfer distance) for each frame using different scaling factors for experiment on isolated tongue images using a one eigenvalue model. Minimum error for experiment 1.B was found in frame 27, while the maximum error was found for frame 39. | 59 |
| 6.3 | Superimposed results obtained for the sequence of extracted tongues for the experiment 1.B. The tongue tip is not well fitted by the one eigenvalue model in frames from 15 to 23, while the tongue blade is well fitted in most of the frames. | 60 |
| 6.4 | Average rms error (chamfer distance) for each frame using different scaling factors. The minimum sequence error was calculated for experiment 1.H. | 61 |
| 6.5 | Superimposed results obtained for the sequence of extracted tongues for the experiment 1.H. | 61 |
| 6.6 | Set of edge images used in the full edge experiments. | 63 |
| 6.7 | Average rms error (chamfer distance) for each frame using different scaling factors. | 64 |
| 6.8 | Superimposed results obtained for the sequence of extracted tongues for the experiment 2.B. | 65 |
| 6.9 | Average rms error (chamfer distance) for each frame using different scaling factors. | 66 |
| 6.10 | Superimposed results obtained for the sequence of extracted tongues for the experiment 2.F. The tongue tip is better in this sequence than in the one eigenvalue model. However, it does not match the correct shape on frames 13, and 15 to 19, where the model considers part of the lower lip as part of the tip tongue. | 67 |

| | | |
|------|---|----|
| 6.11 | Average rms error in terms of chamfer distance for tongue extraction using one and two eigenvalues for training data: (a) isolated tongue data: (b) full edge images. | 68 |
| 6.12 | Superimposed results obtained for the sequence of extracted unseen tongue shapes for the experiment 3 using a one eigenvalue model. | 68 |
| 6.13 | Superimposed results obtained for the sequence of extracted unseen tongue shapes for the experiment 3 using a two eigenvalue model. | 69 |
| 6.14 | Sequence of full edge images tongues shapes for the experiment 4. | 70 |
| 6.15 | Superimposed results obtained for the sequence of full edge unseen images using a one-eigenvalue model. | 71 |
| 6.16 | Superimposed results obtained for the sequence of full edge unseen images using a two-eigenvalue model. | 72 |
| 6.17 | Average rms error (chamfer distance) for tongue extraction using one and two eigenvalues for unseen data: (a) isolated tongue data: (b) full edge images. | 72 |
| 6.18 | Average rms error (chamfer distance) between extracted shape and ground-truth using one and two eigenvalues for synthetic data. | 73 |
| 6.19 | Training data set sequence corrupted with 10% of noise. | 74 |
| 6.20 | Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 10% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image. | 75 |
| 6.21 | Training data set sequence corrupted with 20% of noise. | 76 |
| 6.22 | Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 20% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image. | 77 |
| 6.23 | Training data set sequence corrupted with 30% of noise. | 78 |
| 6.24 | Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 30% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image. | 79 |
| 6.25 | Training data set sequence corrupted with 40% of noise. | 80 |
| 6.26 | Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 40% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image. | 81 |
| 6.27 | Results of the algorithm using contaminated images with noise. | 81 |
| 6.28 | Set of training lip shapes, extracted from the sequence of 39 images outlined by Mohammad. | 82 |
| 6.29 | Results obtained (black) using a one eigenvalue model for the lips over isolated lip images (grey) of the training data set. | 82 |
| 6.30 | Results obtained (black) using a two eigenvalue model for the lips over isolated lip images (grey) of the training data set. | 83 |
| 6.31 | Results obtained (black) using a one eigenvalue model for the lips over full edge images (grey) of the training data set. | 84 |
| 6.32 | Lips and tongue training data set formed by 39 shapes extracted from the labelled data set of Mohammad. | 85 |
| 6.33 | Results obtained using a one eigenvalue model for the tongue and lips over the full edge images of the training data set. | 86 |

6.34 Results obtained using a one eigenvalue model for the tongue and lips over the full edge images of an unseen data set. 87

7.1 Audio recording system. A fine tube was attached to the head coil. Speech sounds were conducted to the built in microphone inside the magnet room. The speech was collected for another microphone taped to the intercom built in the console in the control room. A Soundmac audio digital card digitised the audio which was saved using a LabView program. This program, collects the pulse gating sequences generated by the scanner and acquired by an IMAQ acquisition card. 91

7.2 Spectrograms of preliminary /pasi/ recordings example. Clipping effects are present in (a) and eliminated in (b). Points A and B enclose the spectral information of /a/; information of /s/ and /i/ is either not sufficient or non existent for achieving a satisfactory segmentation of the tokens during the synchronisation stage. 92

7.3 Seven sagittal planes collected during no speech (static position) for subject 02. 93

7.4 Acquisition sequence. Each row for each plane is collected at a time. The same row for subsequent planes is collected in a sequential order. NP denotes the number of planes to be collected. 94

7.5 Parts of an original (a) and a filtered speech file (b) generated for the first experiment of the Subject 04. The filter applied was the minimum mean squared error and minimum statistics noise estimation. 96

7.6 Audio file segmentation 1. Points mark the starting and ending points of the /a/ phonemes of the token /pasa/. The first /a/ duration, $d_{a1,1}$, is defined by the interval [a,b], while the second /a/ duration, $d_{a2,1}$, is defined by the interval [c,d]. The /s/ duration, $d_{s,1}$, is defined as the interval [b,c] between the first and second /a/'s. Analogously, the apparent /p/ duration, $d_{p,2}$, should be defined as the interval [d,e] defined between the end second /a/ of the last token, (d), and the beginning of the first /a/ of the next token, (e). 97

7.7 Audio file segmentation 2. The interval of time to be considered in the synchronisation stage, corresponding to the /p/ articulation, could be composed by two intervals of time: [d,e] posterior to the last token, $d_{p1,2}$, and [f,g] anterior to the next token, $d_{p2,2}$ 97

7.8 Audio file segmentation 3. The interval of time to be considered in the synchronisation stage, corresponding to the /p/ articulation, could be composed by the interval of time [e,f] anterior to the next token, $d_{p2,2}$ 98

7.9 Format of the gating pulse sequence. First and second pulse are separated by a time interval of 60 ms. Some files presented second and third pulses (a) as a double pulse (b), due to the sampling frequency used. In general, two prepulses are generated 63ms before the 672 pulses corresponding to each row in the K-space matrix. 99

7.10 Illustration of the pulse delay between the first pulse in the sequence and the point when the scanner start the acquisition. 100

7.11 Graph of the audio pulse delays. It shows the delay between the audio data and the first pulse sent by the scanner in each measurement. The audio files plotted in the graph were selected randomly; the notation SubjectY.X denotes the audio file X of the subject Y. The line model is defined by Equation 7.1. 101

| | | |
|------|--|-----|
| 7.12 | Synchronisation stage. The phase is defined for each row. The phase does not follow a sequential order due to the acquisition process adopted. | 102 |
| 7.13 | Acquisition sequence with a T_R of 120 ms and an acquisition time (at) of 1.7 ms. The collection of the corresponding row in the seven planes is done in T_R intervals. The actual acquisition time is defined by the parameter at , the remaining time between rows is assumed to be used for the associated processing of the row. | 103 |
| 8.1 | First audio segmentation scheme which considers the complete token duration in the reconstruction. | 108 |
| 8.2 | Sequence of 5 frames reconstructed for the first segmentation scheme. Frames labelled as /p/ are those reconstructed for the interval of time between the last /a/ of the token and the first one. | 108 |
| 8.3 | Sequence of 15 frames reconstructed for the first segmentation scheme. Frames labelled as /p/ are those reconstructed for the interval of time between the last /a/ of the token and the first one. | 108 |
| 8.4 | Sequence of 40 frames reconstructed for the first segmentation scheme. Frames labelled as /p/ are those reconstructed for the interval of time between the last /a/ of the token and the first one. | 109 |
| 8.5 | Second audio segmentation scheme, where two intervals of time labelled as /p/ duration are considered in the reconstruction. These intervals are defined before and after the pronunciation of each /asa/ subtoken. | 110 |
| 8.6 | Sequence of 5 frames reconstructed for the second segmentation scheme. Frames labelled as /p/ are those reconstructed for intervals of time of 140 ms after the last /a/ of a token and before the first /a/ of the next one. | 110 |
| 8.7 | Sequence of 15 frames reconstructed for the second segmentation scheme. Frames labelled as /p/ are those reconstructed for intervals of time of 140 ms after the last /a/ of a token and before the first /a/ of the next one. | 111 |
| 8.8 | Sequence of 40 frames reconstructed for the second segmentation scheme. Frames labelled as /p/ are those reconstructed for intervals of time of 140 ms after the last /a/ of a token and before the first /a/ of the next one. | 112 |
| 8.9 | Third audio segmentation scheme, where the interval labelled as /p/ duration is considered previous to the pronunciation of the /asa/ sub token. | 113 |
| 8.10 | Sequence of 5 frames reconstructed for the third segmentation scheme. Frames labelled as /p/ are those reconstructed for an interval of time of 240 ms taken before the first /a/ of the each token. | 113 |
| 8.11 | Sequence of 15 frames reconstructed for the third segmentation scheme. Frames labelled as /p/ are those reconstructed for an interval of time of 240 ms taken before the first /a/ of the each token. | 113 |
| 8.12 | Sequence of 40 frames reconstructed for the third segmentation scheme. Frames labelled as /p/ are those reconstructed for an interval of time of 240 ms taken before the first /a/ of the each token. | 114 |

| | | |
|------|---|-----|
| 8.13 | Static and reconstructed images of the Subject 04 of the Halifax experiments to illustrate noticeable differences in image quality by visual inspection. The reconstructed image is noisy due to motion artefacts, missing information problems during the reconstruction, susceptibility problems and probably for problems with the reconstruction because the token was not repeated exactly enough. | 115 |
| 8.14 | Raw MR images acquired in Southampton for the middle plane of the subject PJ , while the token /pasi/ is repeated. | 117 |
| 8.15 | Raw MR images acquired in Halifax for the middle plane of the subject 04 , while the token /pasa/ is repeated. | 117 |
| 8.16 | Sustained images captured for the Subject PJ in the Southampton experiments. | 119 |
| 8.17 | Sustained images captured for the Subject 04 in the Halifax experiments. . . | 119 |
| 8.18 | Results obtained for the sustained image subject 04. The algorithm fails detecting the tongue shape. | 120 |
| 8.19 | Non-speech and sustained images used to extend the tongue model. | 120 |
| 8.20 | Set of 15 tongue shapes extracted from non-speech and sustained images collected in Halifax. | 121 |
| 8.21 | Extended training data set with the tongue shapes aligned. This set includes the subset of 15 non-speech and sustained tongue shapes collected in Halifax. | 122 |
| 8.22 | Superimposed tongue shapes. (a) Sum of all the 54 original tongue shapes. (b) Sum of all the 54 aligned tongue shapes | 123 |
| 8.23 | Results obtained of applying the algorithm ASDHT over the original sequence of 39 tongue shapes labelled by Mohammad, using the extended model of the tongue with one eigenvalue. | 123 |
| 8.24 | Results obtained of applying the algorithm ASDHT over the original sequence of 39 tongue shapes labelled by Mohammad, using the extended model of the tongue with two eigenvalues. | 124 |
| 8.25 | Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with one eigenvalue. | 125 |
| 8.26 | Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with two eigenvalues. | 125 |
| 8.27 | Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with three eigenvalues. . . . | 125 |
| 8.28 | Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with four eigenvalues. | 126 |
| 8.29 | Results obtained of applying the algorithm ASHT over the set of full edge images for subject 01 (first row) subject 02 (second row), subject 03 (third row) and subject 04 (fourth row). | 127 |
| A.1 | Proton properties. (a) Single proton spins with a magnetic moment μ , (b) Precessing effect around the magnetic field B_0 , with a Larmor frequency ω_0 . . | 140 |

A.2 Net magnetisation M_0 . (a) Protons align either parallel or antiparallel to the magnetic field B_0 , constituting two energy states, the higher state, E_1 , with all the protons aligned parallel to B_0 , and the lower state, E_2 , with all the protons aligned anti-parallel to B_0 . (b) The protons are precessing out of phase, therefore the net magnetisation M_0 lies in the longitudinal axis. . . . 141

A.3 Slice selection. The use of a gradient magnetic field G_z determines the bandwidth of radio frequencies exciting the desired protons. G_z varies from 0.3 to 0.7 T defining the desired slice, the strength of the magnetic field for such a slice (0.59 to 0.61 T) the radio frequency bandwidth, (25.13 – 25.98 MHz), which is defined by equation A.1. Reproduced from Hashemi and Bradley (1997) 142

A.4 Procedure for slice selection. 143

A.5 Image reconstruction of the sampled K -space matrix. 144

A.6 Pulse sequence diagram for the spin-echo sequence for Fourier imaging. . . . 145

B.1 Data acquisition. Example of the image acquisition, while the subject is repeating the word /pasi/. The scanner starts approximately at 17% of the pronunciation of the third token, and finishes at 65% of the fifth one. 152

B.2 Data synchronisation. Raw images are composed by rows with different phases. Matrices with the same phase or range of phases must be defined. 152

B.3 Generation of matrices with the same phase. A matrix for the phase of 15% is generated. Copying rows from the raw matrices with the phase value of 15%. However, this process will generate matrices with oversampled rows (Row 0), and missing rows (Row $N - 1$). 153

B.4 Acquisition sequence used. Each plane is collected at a time. Rows for a specific plane are collected in a sequential order on intervals of T_R 155

B.5 Digitised audio data. In (a) the subject starts to pronounce /pasi/; then, the scanner starts to collect: (b) the left (c) the mid and (d) the left sagittal planes; the image acquisition finishes in (e). 156

B.6 Range of phases defined for the token /pasi/. Each range is defined as 25% and the number of phoneme is defined by N_ph 157

B.7 Example of definition of phase for the first row. When the time the first row is determined, it can be deduced the phoneme that was pronounced at that time. The phase value is calculated using Equation B.2. 157

B.8 Sequence of 39 midsagittal frames reconstructed by Mohammad for the subject PJ in the experiment 4. 158

List of Tables

| | | |
|-----|---|-----|
| 4.1 | Eigenvalues of the covariance matrix derived from the aligned tongue shapes. | 39 |
| 6.1 | Average rms error (chamfer distance) for different scaling factors used in the first set of experiments using a one eigenvalue model. | 59 |
| 6.2 | Average rms error (chamfer distance) for different scaling factors used in the first set of experiments 1 for the two eigenvalue model. | 62 |
| 6.3 | Average rms error (chamfer distance) for different scaling factors used in the second set of experiments for one eigenvalue model. | 62 |
| 6.4 | Average rms error (chamfer distance) for different scaling factors used in the second set of experiments for one eigenvalue model. | 64 |
| 6.5 | Average rms error (chamfer distance) for the data sets used in the experiments using two eigenvalue models. | 74 |
| 6.6 | Eigenvalues of the covariance matrix derived from the aligned lip shapes. . . | 76 |
| 6.7 | Eigenvalues of the covariance matrix derived from the aligned tongue and lip shapes. | 79 |
| 6.8 | Average rms error (chamfer distance) for the data sets used in the three sets of data for one and two eigenvalue models. | 87 |
| 7.1 | Number of scans collected for the experiments conducted. | 94 |
| 7.2 | Number of wav files collected for each subject with the number of scan on them. | 95 |
| 8.1 | Eigenvalues of the covariance matrix derived from the aligned tongue shapes. | 121 |
| B.1 | Acquisition parameters used by Mohammad in the fourth experiment. | 155 |
| B.2 | Results generated by Mohammad for the fourth experiment. | 159 |

Chapter 1

Introduction

1.1 Motivation

An important challenge in speech research is the measurement of the vocal tract shape during real-time speech production. The study of the dynamic behaviour of the speech articulators, (e.g. the tongue) is important for the full understanding of speech production. It is difficult to obtain data from the articulators since they are not easily accessible. The vocal tract shape is complex and time-varying and there is not a model to describe this shape.

Some of the vocal tract articulators are inaccessible for natural speech studies and medical imaging techniques, such as magnetic resonance imaging (MRI), are the most useful methods to obtain information of them. MRI has been used by many speech researchers due to its advantages such as good contrast of soft tissue, of which most of the vocal tract articulators are formed; and the possibility to scan any orientation. MRI data can be collected for static and dynamic studies. Static studies generally consists of acquiring data when the subject is sustaining a sound, a single phone. Images obtained for these studies are usually good, but these do not provide any information about the movements of the articulators when they go from the pronunciation of one phoneme to another. This transition is referred as coarticulation.

Dynamic studies have been used for studying the dynamic behaviour of the vocal tract articulators. These studies involved the acquisition of images while the subject is speaking using different pulse sequences with relatively short acquisition times. However, some speech articulators move relatively fast, such as the tongue which has occasional fast movements (50-100Hz) (Perkell et al., 1992; Kiritani, 1986; Horiguchi and Bell-Berti, 1987). Recent advances in MRI technology have led to the development of real-time imaging with rates of between 5 (Demolin et al., 2002) and 9 (Narayanan et al., 2004) images per second. However, while these need expensive machines as well as requiring sophisticated imaging techniques, there are alternatives.

Southampton dynamic magnetic resonance imaging (SDMRI), described by Mohammad (1999), is one of the approaches for obtaining information of images with an apparent high temporal resolution. This technique consists of acquiring, simultaneously, MR images and the speech data of a subject repeating a nonsense word, i.e. a word that has no meaning but is chosen to demonstrate some phonetic effect. These images and audio data are synchronised to reconstruct multiplanar images of the vocal tract with an apparent sampling rate of 63 Hz, which is an increase by a factor of 136 over the rate of the scanner used in the experiments conducted. This sampling rate combined with the spatial resolution of the images allows movement velocities of up to 12 cm/sec to be clearly distinguished, which is suitable for capturing most articulatory movements. However, occasional fast movements of the tongue exceed this resolution; in such cases, the tongue will appear blurred. Even though the images were reconstructed from incomplete data, this method was shown to give very good information about the vocal tract dynamics. The tendency is to achieve a better temporal resolution in the image acquisition so that the generation of a larger number of images to describe the vocal tract dynamics as well as possible. The manual labelling and extraction of the vocal tract shape becomes a more tedious and time consuming process, which can be overcome with an automatic vocal tract shape extraction.

The aim of this project was firstly, to understand, analyse and reimplement this method and then to use the resulting images to automatically extract the vocal tract shape to visualise the vocal tract dynamics in 3D.

For vocal tract shape extraction, it is necessary to consider that the images obtained from the SDMRI method are noisy and with it some articulators blurred because their fast movements exceed the resolution of such a method. The missing data problem related to the SDMRI technique also contributes.

The automatic extraction of the vocal tract shape requires two conditions to be satisfied. First, as automatic extraction is desired, the method should need no initialisation; and second, as the vocal tract shape is a complex non-parametric shape, this is there is no simple model to describe it, then it is necessary either to generate a model for the vocal tract shape or to develop a feature extraction method where no model is needed.

Different feature extraction techniques were analysed, but none satisfied both conditions at the same time. However, the characteristics of the problem suggested that we combine two computer vision techniques in a new implementation of the Hough transform, called active shape Hough transform (ASHT). A model of the vocal tract shape is generated using active shape models (ASMs),(Cootes et al., 1995), which are based on the generation of a model using principal component analysis (PCA). The ASMs will deform according with the main modes of variation found in the training data set. The Hough transform (Hough, 1962; Illingworth and Kittler, 1988) is an evidence-gathering technique based on a voting process where a model of the shape is required as well. No initialisation is required which makes it suitable for automatic extraction.

Although this method can achieve a good automatic shape extraction, it fails when multiple solutions are present. A global analysis of the sequence overcomes this problem. Therefore, the dynamic Hough transform (DHT) was adopted to our purposes. Different experiments were performed to test the algorithm over specific scenarios. The extension of the model, as information of additional shapes may be included, is relatively straightforward.

Although the data set used initially consists of sequences of MR images in three sagittal planes, it was considered not enough for fully describing the vocal tract dynamics in 3D, because these planes did not cover the full body of the articulators. Therefore, data was collected at the Institute for Biodiagnostics (Atlantic) Neuroimaging Research Laboratory in

Halifax, Canada, using a 4T scanner. More sagittal planes were collected from 4 different subjects.

1.2 Thesis Outline

This thesis continues with an introduction of speech production research using MRI in Chapter 2. A brief introduction to vocal tract anatomy is presented, followed by a review of different approaches to vocal tract modelling. The most common methods used by speech researchers are described and the MRI method is presented in more detail, focussing on the dynamic MRI methods.

In Chapter 3, a review of different feature extraction techniques is presented. This is followed in Chapter 4 by a description of a new combination of two feature extraction techniques, active shape models and the Hough transform. This new form of the Hough transform, called the active shape Hough transform (ASHT), is tested for the extraction of the tongue shape from MRI sequences. The problem of choosing the optimal solution when multiple candidates are present in the accumulator generated during the evidence gathering process is presented in Chapter 5.

Experimental results of the tongue shape extraction using ground-truth, unseen and synthetic data sets are presented and discussed in Chapter 6. An evaluation of the algorithm performance with controlled contaminated images is also presented. The extension of the model is evaluated as information for the lip is added to the initial tongue shape model. Again the extended algorithm is tested using ground-truth and unseen tongue shape sequences.

Then, the collection of data performed in Halifax, Canada is described in Chapter 7. The experimental work and the corresponding analysis and reconstruction of the image sequences is presented. In Chapter 8, the experimental results of applying the tongue model over the static and sustained images is presented. The extension of the tongue model is performed as a set of 15 tongue shapes is included in the set of training images. The evaluation of the results is presented.

Finally, the conclusions drawn from this research are discussed in Chapter 9. Possible directions for future work on this project are presented.

Chapter 1 Introduction

The purpose of this research is to investigate the effects of various factors on the production of a specific product. The study is designed to provide a comprehensive understanding of the production process and to identify the key factors that influence the outcome. The research is organized into several chapters, each focusing on a different aspect of the production process. Chapter 2 provides a detailed overview of the production process, including the raw materials, the manufacturing process, and the final product. Chapter 3 discusses the various factors that can affect the production process, such as the quality of the raw materials, the skill of the workers, and the efficiency of the equipment. Chapter 4 presents the results of the research, showing the relationship between the various factors and the production outcome. Chapter 5 discusses the implications of the research findings for the production process, and Chapter 6 provides a summary of the research and a list of recommendations for future work. The research is conducted using a combination of qualitative and quantitative methods, including interviews, surveys, and experiments. The data collected is analyzed using statistical techniques to identify the key factors that influence the production outcome. The research findings are presented in a clear and concise manner, making it easy for readers to understand the results and their implications. The research is a valuable contribution to the field of production management, providing a detailed understanding of the production process and the factors that influence it. The findings of the research can be used to improve the production process and to increase the efficiency of the manufacturing process. The research is a model for future research in the field of production management, providing a clear and concise overview of the production process and the factors that influence it.

Chapter 2

Speech Production Research Using MRI

2.1 Introduction

An important aspect of understanding speech production is to know how the vocal tract articulators deform and interact to produce speech sounds. Shape configurations adopted by the vocal tract articulators can be modelled for different purposes such as speech synthesis, study and evaluation of patients with deficit of speech functions and animation applications. Different methods have been used to collect data from such articulators; the most commonly used are medical-imaging based, since they overcome articulator inaccessibility. However, other methods, such as electropalatography and electromagnetic midsagittal articulometer systems, are also quite popular among speech researchers.

Magnetic resonance imaging (MRI) has been extensively used and different techniques have been developed to collect information for either static, sustained phonemes, or dynamic speech. Different techniques are available to collect information for dynamic speech including dynamic magnetic resonance imaging; real time magnetic resonance imaging; and cine magnetic resonance imaging; all with a common aim to collect as many images as possible to describe the movements and deformation of the articulators during the production of a speech

sequence of interest. Similar techniques, such as retrospective gating, are used to describe the dynamics of different organs within the human body, such as the heart. Southampton dynamic magnetic resonance imaging (SDMRI) is a method reported by Mohammad (1999) as being capable of generating sequences with a rate of 63 Hz; this method has been adopted with slight modifications and analysed in the present work.

This chapter is intended to describe briefly basic concepts involved in speech production. Next, different approaches to vocal tract articulator modelling are presented. Then, the most commonly-used methods for obtaining data about such articulators are introduced; special emphasis is placed on MRI-based methods, since this work aims to extract automatically the vocal tract shape from MRI sequences. A brief introduction to this method is presented.

2.2 Speech Production

Speech is generated as a necessity to communicate ideas and the realisation of such ideas into speech sounds is the subject of the study of speech production. An idea is generated within the brain; this sends the corresponding information to each body part involved within the process. First, the lungs provide the airflow and pressure, that is the source of speech; then, the airflow is modulated by the vocal folds and finally, the vocal tract articulators move and deform to produce a specific sound (O'Shaughnessy, 2000).

The vocal tract is a duct composed of muscular and bony parts. Figure 2.1 shows a cross sectional view of the vocal tract. One of the articulators which deforms the most during speech production is the tongue; its different parts are: (4) the tip, (5) the blade, (6) the dorsum. It consists of 12 interactive pairs of muscles and some passive tissues. Its position and deformation is relevant in speech sound production. It is beyond the scope of this work to discuss in detail the structure of the tongue. For a more detailed reference see for example, O'Shaughnessy (2000); Fletcher (1992); Hardcastle (1976).

The vocal tract configuration varies according with the sound or phoneme to be pronounced. The manner of articulation is concerned with the way the airflow is impeded by vocal tract constrictions, (O'Shaughnessy, 2000). Most of the vowels, for example, are produced with

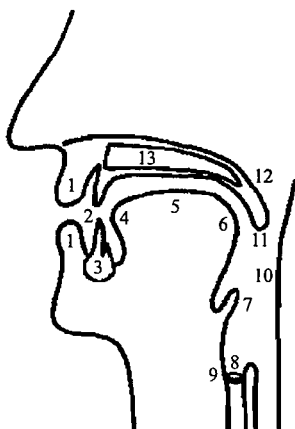


FIGURE 2.1: Sagittal view of speech articulators of the vocal tract, (1) lips, (2) teeth, (3) tooth root, (4) tongue tip, (5) tongue blade, (6) tongue dorsum, (7) epiglottis, (8) vocal folds, (9) larynx, (10) pharynx, (11) velum, (12) soft palate, (13) hard palate. Reproduced from O'Shaughnessy (2000).

a relatively open vocal tract while consonants require a more obstructed one (Stone and Lundberg, 1996). For example, stop or plosive consonants, such as /p/, involve a complete closure and a subsequent released of a vocal tract obstruction (O'Shaughnessy, 2000).

2.3 Vocal Tract Modelling

The vocal tract shape can be modelled using different approaches, for example, in some speech synthesis applications, the vocal tract has been modelled as a group of tubes interconnected to generate speech sounds, where the vocal tract is defined as a filter and the vocal folds as a modulator. Models of the vocal tract articulators have been developed using different methods such as: principal component analysis (PCA), which is a statistical-based method; and, area functions, which are sets of cross-sectional areas A_m (for $m = 1, \dots, N$) for N short sections constituting the vocal tract.

Beautemps et al. (1996) reported a model of the vocal tract based on synchronised cineradiographic pictures, video and audio sequences. They followed a guided principal component analysis (PCA) procedure, to define tongue parameters. The analysis of the

complete tongue contour was rejected due to the evidence reported by Gabioud (1994), where the poor ability to describe the tongue tip from such an approach was shown. Therefore, the tongue tip shape was isolated for fully describing its shape deformation. Globally, the tongue shape was predicted with a variance of 88% followed by a conversion from midsagittal shapes to area functions.

A catalogue of area functions for 12 vowels, 3 nasals and 3 plosives was presented by Story et al. (1996) based on MRI data. Using the same imaging method and electropalatography data, Narayanan et al. (1997) reported the analysis of the vocal tract geometry. Using ultrasound data collected for 11 English vowels in two consonant context (Stone et al., 1997) determined the principal components of cross sections of tongue shapes. The first two components covered 93% of the variance of the data. Using the same guided PCA procedure followed by Beautemps et al. (1996), Badin et al. (1998) reported a 3D statistical articulatory model of the vocal tract based on MRI data. DeLucia and Kochman (2000) described a non-iterative algorithm for generating area functions.

In a later work, Beautemps et al. (2001) achieved a tongue variation of 96%. The articulatory model was controlled by nine parameters: jaw height JH, tongue blade TB, tongue dorsum TD, tongue tip TT, tongue advance TA, larynx height LY, lip height LH, lip protrusion LP and lip vertical position LV. A three-dimensional linear articulatory modelling of the tongue, lips and face was reported by Badin et al. (2002). They derived this model from sets of MRI data and profile video sequences, using a 1 T MRI scanner to collect the midsagittal MR images and separate sessions for the video acquisition. Their results revealed that 3D features can be predicted from midsagittal images.

A different approach based on biomechanical theory was presented by Napadow et al. (2002). It compares the tongue to a common engineering device, using an analog model. This model defines what muscular elements are involved in tongue deformation, focussed on sagittal bending movements of the tongue reaching the hard palate.

Engwall (2003) reported a 3D model of the tongue based on MRI, EMA and EPG data. The 3D MRI data set collected consisted of 54 images of 18 parallel axial, oblique and coronal image subsets. The model was defined by six linear parameters: jaw height, tongue tip,

tongue blade, tongue dorsum, tongue advance and tongue width. A total of 78% of the variance was covered. Models of the palate and teeth were introduced as well.

For animation purposes, models for lips, jaw, cheeks and tongue have been developed. The most important aspect in animated speech is to make it look very realistic, so that expressions and face movements in general should be well synchronised with the actual appearance of the character as well as with the word or sentence to be pronounced. In this case the tongue is modelled in a very rough way, because details about its deformation are not that important because it is not very visible, but the interaction between tongue and lips is quite relevant. Ma and Cole (2004) include in their research a 3D model for the tongue. Another approach for modelling is the use of deformable template models. Deformable templates for lips, tongue and full mouth were developed by King and Parent (2005) for creating more realistic and intelligible animated speech, covering coarticulation effects. However, this model was based on muscle contractions, so that effects such as tongue/lips interaction are not fully described. The tongue model includes information of the tip, dorsum and what they called left and right wings; however, it does not include information of the root and dorsum, nor interactions with teeth and lips.

Recently, Takemoto and Honda (2006) calculated vocal tract area functions from sagittal images collected using a 3D cine-MRI technique. First, a vocal tract midline is calculated, then, images are sectioned at 2.5 mm intervals along the midline; finally, each section is measured to calculate the corresponding area function. This 3D cine-MRI technique is based on a synchronised sampling method presented by Masaki et al. (1999), where a sequence of trigger pulses synchronise the token repetition and the data acquisition. A total of 20 sagittal planes and sets of 2D cine-MRI with 56 frames per slice were collected while 640 utterances were repeated.

2.4 Methods for Acquiring Vocal Tract Data

Different methods for acquiring information of the vocal tract movements have been used. These methods can be either invasive or non-invasive. Invasiveness refers to the physical

interaction of the equipment used to collect data with the vocal tract articulators. Non-invasive methods offer best results for studies for natural speech.

Some examples of invasive methods are those based on transducing, which is a technique used for registering the articulatory movements. In an early paper, Muller and Abbs (1979) describes the refinements made in a prototype system for transducing lip and jaw movements in the midsagittal plane. This prototype is shown in Figure 2.2.

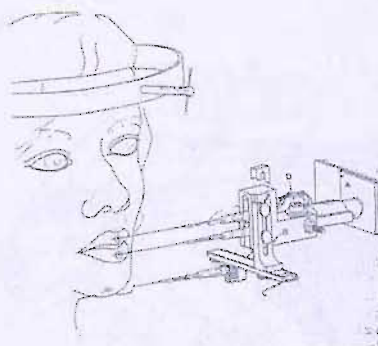


FIGURE 2.2: Two dimensional transducer attached to the upper lip, lower lip and jaw.
Reprinted from Muller and Abbs (1979).

Perkell et al. (1992) developed two electromagnetic midsagittal articulometer (EMMA) systems for transducing articulatory movements in a midsagittal plane on supraglottal structures, during speech production. The principle of operation is based on the alternating magnetic field movement transducer system. Magnetic fields pass through transducer coils, fixed on the midline of the vocal tract articulators, inducing an alternating signal. This signal is proportional to the distance between the transducer and the transmitter. Figure 2.3(a) shows a schematic midsagittal view of a subject with nine possible measurement points (indicated by filled circles), corresponding to the interesting areas, in this case the effective measurement area covers a radius of 75 mm. Figure 2.3(b) shows a three transmitter system. Transmitters are labelled with 'T', side plates with 'P', head mount with 'H' and circular collar with 'C'. Some disadvantages of this method are: the preparing time and care needed to attach the transducers; the misalignment errors produced; and its limitation to collect information along the midsagittal plane. However, when the experiment is running the system is reasonable efficient, providing large amounts of accurate information of the articulators along the midsagittal plane (Perkell et al., 1992).

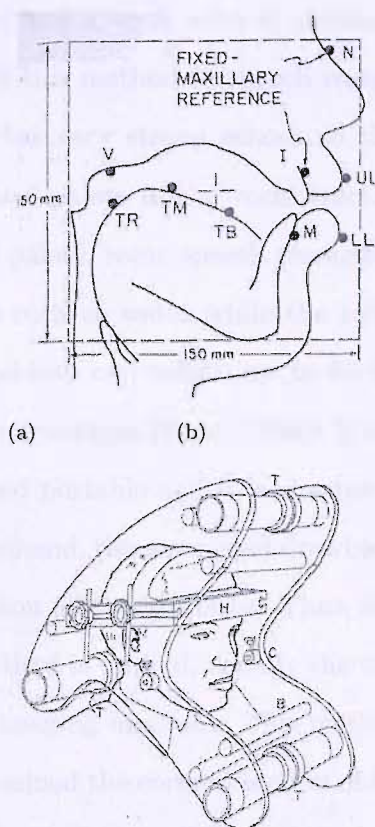


FIGURE 2.3: Electromagnetic midsagittal articulometer. Reprinted from Perkell et al. (1992).

Electropalatography is another invasive method that can give information about the tongue and palate interaction. This method consists of an acrylic palate that covers the hard palate and the inner and outer surfaces of the teeth (Stone and Lundberg, 1996). A specific number of electrodes are embedded along the surface of this palate and these electrodes give the information of the contact position of the tongue at different times during speech sound production.

Most of the examples of non-invasive methods are imaging methods. The most commonly used are: ultrasound, X-Ray, computer tomography (CT) and magnetic resonance imaging (MRI). Ultrasound is a medical imaging method based on the application of high frequency sound waves, typically between 1 and 15 MHz. This method works under the principle of reflection. When an ultrasound pulse is sent into the body this pulse can either reflect or penetrate depending on the boundary met. If it reflects then a strong echo is recorded;

otherwise the pulse penetrates and a weak echo is obtained. An important characteristic to consider when working with this method in speech research is the reflection coefficients. The air-soft tissue boundary has very strong echoes so that imaging beyond air space is impossible, for example the hard palate in the vocal tract. For imaging other parts of the vocal tract including the hard palate, some speech researchers can use contrast agents with different reflection coefficients such as water while the subject is swallowing (Epstein and Stone, 2005). Ultrasound machines can collect up to 30 frames per second (Stone et al., 1998). This method has some advantages (Pope, 1999): it does not have any known harmful effect, the equipment is safe and portable and it is cheaper compared with another medical imaging methods. On the other hand, the associated drawbacks are: it can not penetrate bone and it has almost 100% reflection at air interfaces. Thus, the number of speech articulators that can be imaged by this method is limited, usually the tongue. Clarity of image is poorer compared with other medical imaging methods. This method has been to study mainly the tongue, Stone et al. (1998) examined the coronal section of the mid-tongue using ultrasound. Recently Stone (2005) describes how to collect and analyse tongue ultrasound images in coronal and sagittal planes to reconstruct 3D tongue surface motion.

X-ray imaging is based on the application of high-frequency electromagnetic radiation to the human body. These X-rays can either pass through or be absorbed by different body structures. These rays are exposed to a photographic film (Pope, 1999). This method can image bones with an excellent resolution, contrary to soft tissues which have poor resolution. A disadvantage of this method is that these radiations can be hazardous. Another method, X-ray computed tomography (CT) is based on X-ray application but the attenuated X rays are collected by detectors and converted to digital information processed by a computing device for image reconstruction (Cho et al., 1993). This method has the same disadvantage of the X-ray imaging technique because of the X-ray application.

Magnetic resonance imaging is a technique commonly used in medicine for imaging the human body. This technique has been demonstrated to be a safe imaging method with no known side effects. MRI is based mainly on the application of magnetic fields and radio frequency (RF) pulses to image a desired slice of the human body. Nowadays, different pulse sequences

and acquisition techniques have been developed. A more detailed explanation of MRI is presented in Appendix A. Advantages of using MRI over other imaging methods in speech research are, first, superior image quality and contrast of soft tissues, which includes most of the vocal tract articulators. In medicine, this advantage means that artificial contrast agents need not to be used. Although in some special cases, gadolinium can be used as a contrast agent. Second, the ability to scan any plane of interest at any angle and finally, its safety to the subject, although it is necessary to follow some controls and precautions, because of potential hazards related to foreign magnetic objects within the body. A review of safety of strong, static magnetic fields is presented by Schenck (2000). Contrarily, this method has some associated drawbacks in speech research such as:

1. The inability to show bone and teeth, which have a low hydrogen content, so that interactions between with the vocal tract articulators can not be accurately studied. Consequently, air cannot be distinguished from bone or teeth on MRI images. However, some contrast agents can be used for imaging such parts; for example, the teeth can be impregnated with gadolinium to obtain MRI signals from them.
2. It has long scanning times; there is a trade-off between spatial and temporal resolutions, for either static or dynamic studies.
3. The high acoustical noise associated with some scanners; this may confuse the subject during the token pronunciation; however, different manufacturers, such as Phillips has reduced the associated noise by 85% (Koninklijke Philips Electronics NV, 2003).
4. The vocal tract articulators are surrounded by air, which makes the MR signal sensitive to susceptibility problems; these refer to the differences between magnetic properties of the air-tissue boundaries; although these differences are small, they may adversely distort the image if the magnetic strength of the main magnet is high. Usually scanners with a magnetic strength of 0.5 to 1.5 T are used in speech research to reduce these susceptibility problems.

2.5 Speech and MRI

MRI has been widely used in speech research due to its ability for imaging soft tissue at any orientation plane. It has been used for studying the vocal tract shape during the speech production by means of static and dynamic studies. Static studies attempt to characterise the position of the vocal tract shape while a phone is sustained. Baer et al. (1991) presented a study of the vocal tract shape and its dimensions for vowels using magnetic resonance imaging. This analysis was carried out obtaining vocal tract shapes associated with sustainable speech sounds. Dynamic studies acquire data while the subject is pronouncing in a natural way a token or sentence of tokens defined to study the vocal tract configuration in specific conditions.

Static and dynamic studies have a trade-off between spatial and temporal resolution. Superior image quality is achieved in static studies; generally, in dynamic studies blurred images with possible missing information can be generated. Long scanning times for speech purposes are involved in both studies; sustaining a speech sound for a relatively long time could not guarantee the fixed position of the articulators; while in dynamic speech, even with low acquisition times and fast scanning techniques, it can not acquire enough frames to clearly describe the articulator motion. Information of the deformation of the vocal tract articulators can not be extracted from static studies. Although the subject is commonly well trained for dynamic experiments, the exact reproduction of the token is not warranted. Engwall (2000) presents a comparative study of static and dynamic speech. The differences analysed between the two studies, static and dynamic, were focused on jaw position, lip protrusion, linguopalatal distance and tongue contours.

The term ‘image quality’ is referred in this work as the definition of boundaries on the image, as well as the contrast between different structures on it, appreciated by visual inspection. In our case boundaries between air, bone, and tissue should be well defined.

2.6 Speech and Retrospective Gating

One of the most complex human organs studied in medical imaging research is the heart. This organ has very specific motion conditions that make it one of the closest cases to speech research on vocal tract motion.

The retrospective gating imaging technique consists of acquiring data from an imaging scanner and an electrocardiogram (ECG) at the same time in an interrupted sequence. A synchronisation stage between the data acquired is carried out. A sequence of images of different phases of the heartbeat cycle is the result obtained. Some cardiac imaging approaches are based on this technique (Albers et al., 2003; Roerdink and Zwaan, 1993). Some interesting characteristics of the heart motion research are:

1. Its movement is periodic; a heartbeat is repetitive, approximately constant between heartbeats. This characteristic can not be completely assured due to different conditions during the acquisition and mainly of the healthy conditions of the subject used.
2. Its fast movement; this means, the method used for imaging the heart must be fast enough for reconstructing the complete heart cycle; reconstruction techniques are required for overcoming technology constraints.
3. The heart is highly deforming, its muscles contract and expand deforming its shape.
4. Short scanning times are required for a satisfactory reconstruction of heart movements.
5. Artefacts caused for breathing movements must be addressed.

There are many similarities between the heart motion and dynamic speech research. In this work, a sequence of images for a subject pronouncing a specific nonsense word is desired. For achieving this, images are acquired while the subject is repeating a selected word. Then, the token repetition is considered as a periodic movement of the vocal tract articulators similar to the heart cycle, as the word is the same and the subject is previously trained for achieving the most constant possible repetition. Speech articulators move relatively fast. Most of the

speech movements of the tongue have a bandwidth of less than 15 Hz, but occasional fast movements of the tongue have a bandwidth range of 50-100 Hz (Perkell et al., 1992; Kiritani, 1986; Horiguchi and Bell-Berti, 1987). A satisfactory description of the tongue motion requires a repetition time (T_R) of 10 ms.

2.7 Dynamic MRI in Speech

Different approaches for studying the dynamics of the vocal tract articulators have been carried out. A dynamic study was reported by Demolin et al. (1997). They adapted an ultra fast implementation of the Turbo Spin Echo (TSE) sequence to achieve a dynamic continuous monitoring of the vocal tract with a resolution of 4 images per second, using a 1.5 T Philips Gyroscan with T_R of 250 ms and a partial Fourier acquisition of 60%, acquiring an image matrix of 32×128 .

A successful technique for obtaining information of the dynamic behaviour of the vocal tract shape was described by Mohammad (1999): the Southampton dynamic MRI method (SDMRI). This method consists of acquiring, simultaneously, MR images and the speech data of a subject repeating a word. Images and audio data are synchronised off-line to reconstruct multi-planar images of the vocal tract with an apparent sampling rate of 63 Hz, which is an increase by a factor of 136 over the rate of the scanner used in the experiments conducted. This sampling rate combined with the spatial resolution of the images allows movement velocities of up to 12 cm/sec to be clearly distinguished, which is suitable for capturing most articulatory movements. In addition, the maximum velocity of velar motion is 14.1 cm/sec, but moderately fast tongue tip motion velocity is 80 cm/sec (Perkell et al., 1992; Engelke et al., 1996). The subject does not need to be phonetically trained and is not involved in the measurement procedure of the audio data, providing a more natural speech environment.

The use of tagged cine-magnetic resonance imaging (tMRI), which has tag lines used to view muscle motion, has been used to create 3D models of the tongue motion during dynamic speech (Stone et al., 2000; Dick et al., 2000; Stone et al., 2001). A mechanical model, that represents local, homogeneous, internal tongue deformation during speech was developed by

Stone et al. (2001). It was used a TR of 14ms, a field of view (FOV) of 24 cm and a slice thickness of 7 mm.

Real time MRI has been used to study the dynamics of the vocal tract deformation by Demolin et al. (2002). They used the Turbo Spin Echo Zoom sequence with an actual time resolution of 4 to 6 images per second, with image matrices of 32×128 , to show that this technique could be used to study the relative movements of the main articulators involved in speech production.

Narayanan et al. (2004) used spiral k -space acquisitions with a low flip-angle gradient echo pulse sequence, achieving an acquisition rate of 8-9 images per second and reconstruction rates of 20-24 images per second using a sliding window. They used a total TR of 5.5 ms with a slice thickness of 5 mm. Complete images were acquired every 110 ms. They proposed a possible automatic segmentation and tracking of the real-time data using Kalman snakes and optical flow.

Recently, Takemoto and Honda (2006) reported a 3D cine-MRI technique based on a synchronised sampling method presented by Masaki et al. (1999), where a sequence of trigger pulses synchronise the token repetition and the data acquisition. This 3D cine-MRI technique or multiplanar 2D cine-MRI was performed on a Shimadzu-Marconi ECLIPSE 1.5 T PowerDrive, scanner capable of collecting 2D cine-MRI, with a multiplanar acquisition rate of 30 frames per second. A total of 20 sagittal planes and sets of 2D cine-MRI with 56 frames per slice were collected while 640 utterance were repeated. They calculated vocal tract area functions from such sagittal images.

In this work, to reconstruct the vocal tract shape a large set of data must be acquired. Multiplanar acquisitions are required to achieve a 3D reconstruction of the vocal tract shape. Subsequently, a large number of images are expected to be reconstructed. Manual labelling of those images can be a very time consuming and tedious process. Thus, an automatic extraction of those features of interest is desired. Hence, it is important to analyse different available techniques for feature extraction and noise reduction. This will be introduced in the next chapter, where feature extraction techniques are presented.

Chapter 3

Feature Extraction

3.1 Introduction

One of the most challenging task in computer vision is the image analysis. Analysing an image involves detecting, extracting and understanding features in it according to a specific application such as remote sensing, security, medical and sea imaging. In medical imaging, manual labelling of features is a good approach when the number of features to be extracted and the set of images are short. Otherwise this task could be really tedious and time consuming. The tendency in dynamic speech studies is to collect a large number of images to describe the dynamics of the vocal tract shape when a specific phoneme or sequence of phonemes is produced. Hence, the importance of automating this task.

In this project a large number of images can be produced by using the SDMRI method. These images tend to be noisy and blurred, due to the dynamic behaviour of the vocal tract articulators. The aim of this project is the automatic outlining of the vocal tract shape in midsagittal SDMRI sequences. As an automatic extraction is desired there are two aspects that need to be covered for the feature extraction technique to be used: It must provide a model of the deforming vocal tract shape and no initialisation is required. This chapter introduces the most common feature extraction techniques.

3.2 Arbitrary Shape Feature Extraction Techniques

There is an extensive variety of feature extraction techniques. Generally, they need edge or binary images to achieve their task. Arbitrary shape extraction involves either the generation of a parametric model describing such a shape or the deformation of a free contour to achieve the desired shape. Techniques such as template matching, Hough transform, snakes, deformable templates and active shape models are introduced in this section.

3.2.1 Template Matching

Template matching is a method of defining the position and pose of a template in an image. This technique is based on finding the best match for a shape. However, this technique is not suitable for deforming arbitrary shapes. Subsequently, a template is often stored as a discrete set of points where rotation and scaling may not be optimally implemented. The Mellin transform scales a template but its best performance is achieved with continuous functions, while the Fourier Mellin transform scales and rotates a template but it has many problems with discrete implementations (Nixon and Aguado, 2002). A problem using this technique for the entire vocal tract shape is that this is a deforming shape.

3.2.2 Hough Transform

The Hough transform which was proposed by Hough (1962) is another technique that locates shapes in images. This technique, originally used for extracting lines, circles and ellipses, is based on the evidence gathering approach, where a transformation from the image space into the parameter space (Hough space) is performed. The Hough transform does not require any initialisation and is robust to noisy and occluded conditions. Although it has high computational requirements, such requirements are much less than those needed by template matching.

An extension of the Hough transform designed to find arbitrary shapes is the generalised Hough transform (GHT), which was reported by Ballard (1981). This technique is suitable for

extracting those shapes that cannot be defined by a parametric model. The shape description used during the searching process is detailed in a table, called the *R*-table, which contains distance and the angle values calculated for each point in the shape to a previously defined reference point.

Other methods are based on object tracking. The vocal tract shape and even each of its components can be stated as objects to be tracked along a sequence of images. Different approaches derived from the Hough transform have been proposed for detecting moving shapes in sequences of images. Nash et al. (1997) proposed a new technique for dynamic feature extraction, the velocity Hough transform (VHT). This technique includes motion in the evidence gathering process to detect moving parametric shapes. A disadvantage is that it assumes constant linear velocity, which limits its application.

Grant et al. (2002) fused two evidence-gathering techniques, the VHT and the Fourier-descriptor template representation. This variant of the VHT was referred as the continuous-template variant of the VHT or the CVHT. They used Fourier descriptors instead of a discrete representation of the arbitrary shape, avoiding the effects of discretisation such as distortions or missing points when the template is scaled or rotated. This technique requires no initialisation or training, with a good tolerance to noise and occlusion.

The dynamic velocity Hough transform reported by Lappas et al. (2001), is a method for tracking parametric objects, extending the velocity Hough transform (VHT) so that arbitrary velocity is allowed. The problem is reformulated as tracking an object which changes its velocity, scaling or rotating smoothly frame to frame. The problem is reduced to finding an optimal path of an object within the parameter space. The optimal path is defined by means of a dynamic programming method.

3.2.3 Deformable Templates

Deformable templates are a combination of shapes that are allowed to change in size and orientation while retaining their spatial relationship. An important disadvantage of this technique is that a model of the vocal tract cannot be provided. A possible analysis of the

implementation of this technique could be performed by taking each component of the vocal tract separately; however, a model of such articulators must be defined.

Deformable templates for lips, tongue and full mouth have been developed by King and Parent (2005) for animation purposes. More realistic and intelligible animated speech was intended, covering coarticulation effects. However, this model was based on muscle contractions, so that effects such as tongue lips interaction are not fully described.

3.2.4 Snakes

An active contour or snake is a commonly used method for extracting flexible shapes. A snake is an energy-minimizing spline guided by external constraint forces and influenced by image forces that pull it toward features such as lines and edges (Kass et al., 1988). A snake is defined as a set of points which aims to enclose a target feature, the feature to be extracted, and is expressed as an energy minimization process (Nixon and Aguado, 2002). The functional energy E_{snake} is defined as follows:

$$E_{snake} = \int_{s=0}^1 E_{int}(v(s)) + E_{image}(v(s)) + E_{con}(v(s)) ds, \quad (3.1)$$

where $s \in [0, 1]$ is the normalised length around the snake, $v(s)$ is the set of points, E_{int} is the internal energy, which controls the behaviour of the snake, E_{image} is image energy, which attracts the snake to the selected low-level features such edges points and E_{con} is the constraint energy that allows higher level information to control the snake's evolution.

Initial approaches refer to closed and shrinking contours. The greedy algorithm (Williams and Shah, 1992) for snakes gives a simple implementation of a snake that shrinks. The initialisation must contain the shape to be extracted. The greedy method iterates around the snake finding the local minimum energy at snake points. A problem with this method is that the local minimum found may not always be the best minimum.

Snakes have two main disadvantages: the initial contour needs to be close enough to the desired contour and they have poor convergence to concave boundaries. Some extensions of

the snake algorithm evolve open and expanding contours. Cohen (1991) proposed a model that makes the curve behave like a balloon, which is inflated by an additional force. These expanding techniques could be useful for finding the vocal tract shape, considering that the snake should expand through the lips and down to the larynx. There are many snake techniques reported for extracting shapes. Geodesic active contours, reported by Caselles et al. (1997), are based on active contours evolving in time according to intrinsic geometric measures of the image. The proposed approach is based on the relation between active contours and the computation of geodesic or minimal distances. Yezzi (1997) described a geometric snake model for segmentation of magnetic resonance imaging (MRI), computer tomography (CT) and ultrasound medical imagery. This method is based on the defining feature-based metrics on a given image. An interesting method is that presented by Cohen and Kimmel (1997). This is based on the interpretation of the snake as a path of minimal length, or as a path of minimal cost, between two end points for opened contours and one point for a closed contour. This approach needs two end points as initialisation. If a separate analysis and detection of the components of the vocal tract is done, this method could be used for joining end points of such components. Xu and Prince (1998) developed a new external force for active contours, called gradient vector flow (GVF), that is computed as a diffusion of the gradient vectors of a grey-level or binary edge map derived from the image. This approach largely solves the problem of poor convergence to concave boundaries as reported by its authors. Snake methods have been applied to MR images, as reported by Makowski et al. (2002), presenting an active contour segmentation method for 2D structures in MR images. This method consists of two phases. The first phase, called the *balloon phase*, is defined to find an approximate placement of contour vertexes. The second phase, *snake phase*, fine tunes the model obtained in the balloon phase. Anti-tangling features are introduced to improve segmentation of complex shapes.

3.2.5 Active Shape Models

Cootes et al. (1995) described active shape models (ASMs) as a method for finding shapes in an image from a model that can vary with global shape constraints making this the principal

difference of this method with the active contour models (snakes). The technique relies in the representation of the desired shape by point distribution models (PDMs). A PDM represents an object as a set of labelled points. Applying limits to the parameters of the model enforces global shape constraints ensuring that any new examples generated are similar to those in the training set. ASM has been widely implemented in medical imaging segmentation and shape extraction. In speech research, this method has been implemented by Matthews et al. (2002) for parameterising lip image sequences. A more detailed explanation of this method is described in the next Chapter.

3.3 Conclusions

In this Chapter, the most common feature extraction techniques have been introduced. Our problem is focused on the automatic extraction of an arbitrary and deforming shape. However, none of the techniques presented fully cover the requirements of our problem. The combination of two feature extraction techniques is proposed in the following Chapter: active shape models, which are used in the generation of a shape model, and the Hough transform (HT) which uses the model to find the appropriate parameters and position of the shape to be extracted.

Chapter 4

Active Shape Hough Transform

4.1 Introduction

Different techniques can be used to extract arbitrary parametric shapes; most of them require either some form of initialisation or a parametric model to describe the shape as presented in the previous chapter. However, they do not satisfy the requirements for an automatic and arbitrary shape extraction.

The combination of two feature extraction techniques is proposed in this thesis to achieve such an extraction automatically: the active shape models (ASM's) and the Hough transform (HT). The ASM method provides a model of the target, which deforms accordingly to the training data set, and the HT requires no initialisation whatsoever to locate a model.

This chapter describes the implementation of the ASM method as a shape description in a new form of the Hough transform. First, a description of the generation of active shape models is presented, followed by the implementation of this model using a new form of the HT algorithm. Then, the application of this new feature extraction technique is applied to magnetic resonance images for tongue shape extraction. Finally, conclusions drawn from this technique are presented.

4.2 Active Shape Models

The process of generating an active shape model (ASM) is defined as stated by Cootes et al. (1995). First, a consistent set of landmark points is defined for each shape in the training set. The vector Ψ_i denotes the n points of the i th shape in the set as follows:

$$\Psi_i = (x_{i0}, y_{i0}, x_{i1}, y_{i1}, \dots, x_{in-1}, y_{in-1})^T,$$

where the super index T denotes the transpose of the vector. Second, all the shapes must be aligned. The alignment process between two shapes, consists of rotating, scaling and translating one shape to align it with another selected shape; this is achieved by minimising the weighted sum, given by:

$$E_j(\Psi_j, \Psi_i) = (\Psi_i - M(s_j, \theta_j)[\Psi_j] - \mathbf{t}_j)^T \mathbf{W} (\Psi_i - M(s_j, \theta_j)[\Psi_j] - \mathbf{t}_j),$$

where M denotes the transformation matrix associated with a θ rotation, s and \mathbf{t} represent the scaling factor and the translating vector respectively. This matrix is defined as follows:

$$M(s, \theta)[x_{jk}, y_{jk}]^T = ((s \cos \theta)x_{jk} - (s \sin \theta)y_{jk}, (s \sin \theta)x_{jk} + (s \cos \theta)y_{jk})^T,$$

$$\mathbf{t}_i = (t_{xj}, t_{yj}, \dots, t_{xj}, t_{yj}),$$

and \mathbf{W} represents a diagonal matrix of weights, w_k , for each point. These weights are defined as follows:

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

where R_{kl} denotes the distance between the points k and l in a shape and $V_{R_{kl}}$ represents the variance in the distance of the shapes.

The alignment of all the shapes consists of an iterative process, as defined in Algorithm 1.

The mean shape $\bar{\Psi}$ of a set of N shapes is defined as:

$$\bar{\Psi} = \frac{1}{N} \sum_{i=1}^N \Psi_i, \quad (4.1)$$

Algorithm 1 Algorithm for aligning the shapes.

Define a reference shape (Ψ_i)

FOR (Shape (Ψ_i))

 Calculate the weighted sum $E_j(\Psi_j, \Psi_i)$

DO

 Calculate the mean $\bar{\Psi}$ shape as in Equation 4.1.

 Mean shape parameters are normalised

 FOR (Shape(Ψ_i))

 Calculate the weighted sum $E_j(\Psi_j, \bar{\Psi})$

WHILE (Process does not converge)

Thirdly, a principal component analysis is performed over the aligned set of shapes. The covariance matrix \mathbf{S} is calculated to capture the variation in the landmark points around the mean shape as follows:

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N d\Psi_i d\Psi_i^T,$$

where $d\Psi_i$ is the deviation from the mean of each shape in the set defined by:

$$d\Psi_i = \Psi_i - \bar{\Psi},$$

Then, the decomposition of the \mathbf{S} matrix, $\mathbf{S}\mathbf{p}_k = \lambda_k \mathbf{p}_k$, defines its principal axes and modes of variation denoted by the eigenvectors \mathbf{p}_k and the eigenvalues λ_k respectively. The number of eigenvalues and eigenvectors considered in the model will determine the variation covered. The total variation is calculated as the sum of the resulting eigenvalues as:

$$\lambda_T = \sum_{i=1}^{2n} \lambda_i,$$

and the variation of the k eigenvalue λ_k is defined by:

$$V_{\lambda_k} = \frac{\lambda_k}{\lambda_T}.$$

The variation desired to be covered by the model to generate new shapes defines the number of the first N_e eigenvectors and eigenvalues to be used. Hence, a matrix \mathbf{P} composed by the first N_e eigenvectors and a vector \mathbf{b} with N_e elements (eigenvalues) weighting the collaboration of each eigenvector are generated. The variation of these weights is limited, as recommended by Cootes et al. (1995), to the range:

$$-3\sqrt{\lambda_k} \leq b_k \leq 3\sqrt{\lambda_k}.$$

Finally, the active shape model is defined as:

$$\Psi = \bar{\Psi} + \mathbf{P} \cdot \mathbf{b}. \quad (4.2)$$

In the original formulation, the shape was fitted by an iterative refinement. An initial estimate, pose and shape parameters are defined. These parameters are updated according to an adjustment criteria, such as those derived from the image gradient information. In this new implementation this model is used as a shape description in a new form of HT.

4.3 Combining Active Shape Models and the Hough Transform

In this section an algorithm to find arbitrary shapes and deforming shapes is proposed. This algorithm uses an active shape model to find a target shape using a similar voting

scheme, which characterises the Hough transform. A voting space is defined to accumulate the centroid and model parameter information involved in the searching process. At the end, the maximum peak, will represent the solution. This solution must provide information not only of the position of the shape, but also the shape parameters of the model including scaling and rotating factors if required.

Search consists of performing a voting process over each edge image as described in Algorithm 2. For each edge pixel, all possible centroids of all possible shapes are added into a multi-dimensional accumulator space whose dimensions are chosen to represent potential centroid positions, (x, y) , plus the involved number of eigenvalues and possibly scaling and rotating factors as well. Once all edge pixels have been processed, the largest peak in the accumulator is considered as the solution. The Algorithm 2 addresses the feature extraction of an active shape model using one eigenvalue, however, the extension to more eigenvalues is straightforward. Clearly, the size of the accumulator depends on the number of steps into which the various dimensions are discretised. An odd number of steps is suggested, so that the mean shape is included in the searching process.

Algorithm 2 Pseudo code implementation of the ASM model implementation in a new form of generalised Hough transform using one eigenvalue.

```

FOR (Edge Pixels ( $e_x, e_y$ ))
  FOR (Landmark point ( $l_x, l_y$ ))
    FOR (nStepsEig1 ( $b1$ ))
      Generate the model
      Calculate centroid ( $c_x, c_y$ )
      Accumulator[ $c_x$ ][ $c_y$ ][ $b1$ ] ++
Search in Accumulator for Maximum Peak

```

The voting process is illustrated with the example presented in Figure 4.1. A model with four landmark points is considered in Figure 4.1(a); the four remaining tongue shapes shown in Figure 4.1 are defined as the set of all the possible shapes to be considered in the searching process. In Figure 4.2 the third edge pixel, marked in Figure 4.1(b), is considered as the (a) landmark point of all the possible shapes, and the centroids of these shapes are added up in the accumulator space. Here, the centroid is marked with a cross.

An ideal scenario, where the resulting shape in the voting process is examined is presented in Figure 4.3. The centroid of the target shape is incremented until reach the value of 3, which

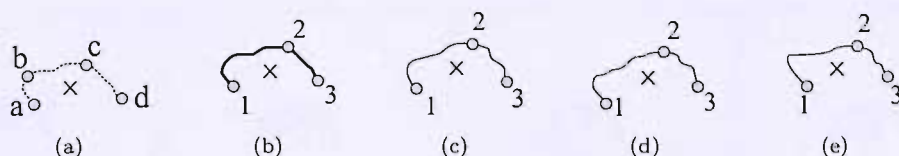


FIGURE 4.1: Landmark and edge points defined to illustrate the voting process shown in Figure 4.2. In (a) the four landmark points considered in the model definition are shown, and the remaining images show the shapes and the edge points considered in the searching process.

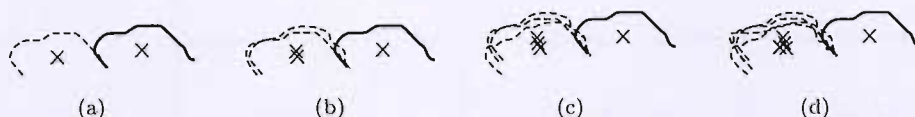


FIGURE 4.2: Voting process. The landmarks and edge points considered in this example are defined in Figure 4.1. The sequence represents the voting process for the first edge point of the shape presented in Figure 4.1(b). The first edge point in the shape (outlined with a solid line) is defined as the fourth "d" landmark point for all possible shapes in the set, outlined with dash lines; centroids are represented by a cross mark.

is the same number of the searching points.

The maximum peak is considered as the solution for this voting process. Such a peak will define centroid position, scaling and rotating factors and the values of the weights for each eigenvector considered in the model.

4.4 Application to MRI Data

The former algorithm was applied to the SDMRI sequences obtained by Mohammad (1999). Images from such sequences are very noisy, hence, a noise reduction filter was used before the algorithm was applied to obtain better results. Then, a model is generated for the tongue, which is the most deforming articulator; this is used as a shape description for the new ASHT algorithm. This algorithm was tested and the results obtained are presented.

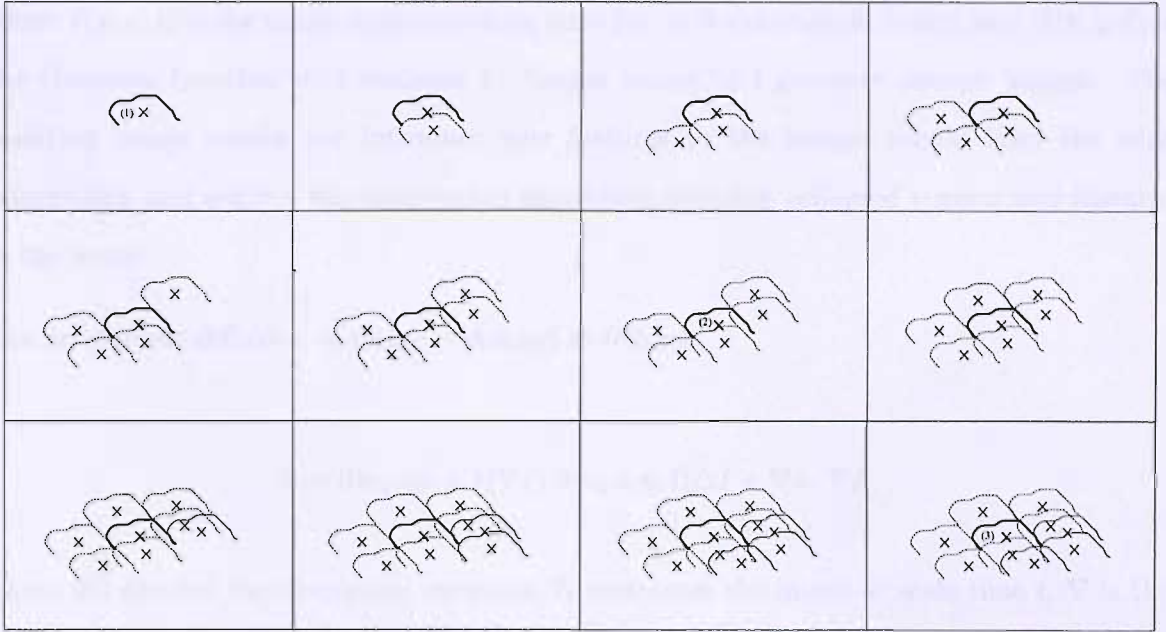


FIGURE 4.3: Example of voting process. The landmarks and edge points considered in this example are defined in Figure 4.1(a),(b). The target is presented with solid outline, while the dashed outlined shapes represent the matching shape in the process. The first image represents the matching of the centroid of the target and the shape used, placing the point (1) as the (a) landmark point. Then, the same edge point is defined as the (b),(c) and (d) landmark points. In the second row, the edge point (2) is defined as the (a),(b),(c) and (d) landmark points respectively, incrementing the same centroid position when the edge point is placed in the (c) landmark position. Finally, the third row shows the iterative process for the edge point (3), the centroid, as mentioned before, is incremented in one when the (d) landmark point is used. The ideal peak will have a value equal to the number of edge points.

4.4.1 Noise Reduction and Edge Detection

A common issue with some noise reduction techniques, such as Gaussian filtering, is the loss of edge information. The anisotropic filtering technique, introduced by Perona and Malik (1990), achieves different criteria: causality, immediate localisation and piecewise smoothing. Consequently, since the preservation of the edge information is paramount, this filter has been applied to SDMRI images presented in this work as reported by Perona and Malik (1990).

This technique is based on the scale-space technique, where coarser images are generated by convolving the original image with a Gaussian kernel of variance t :

$$I(x, y, t) = I_0(x, y) * G(x, y; t),$$

where $I(x, y, t)$ is the image sequence along time (t), I_0 is the original image, and $G(x, y; t)$ is the Gaussian function with variance t . Larger values of t generate coarser images. The resulting image should not introduce new features to the image, should keep the edge information and achieve the inter-region smoothing avoiding collapsed regions and features in the image.

The anisotropic diffusion equation is defined as follows:

$$I_t = \text{div}(c(x, y, t)\nabla I) = c(x, y, t)\Delta I + \nabla c \cdot \nabla I,$$

where div denotes the divergence operator, I_t represents the image at scale time t , ∇ is the gradient operator, Δ is the Laplacian operator and $c(x, y, t)$ is the diffusion coefficient. This diffusion coefficient may satisfy the second and third criteria if a suitable choice is done.

The authors have shown that diffusion in which the conduction coefficient is chosen locally as:

$$c(x, y, t) = g(\|\nabla I(x, y, t)\|), \quad (4.3)$$

will not only preserve, but also sharpen the edges of the image. They showed that Equation 4.3 can be solved iteratively on discrete images leading to:

$$I_{i,j}^{t+1} = I_{i,j}^t + \lambda[c_N \cdot \nabla_N I + c_S \cdot \nabla_S I + c_E \cdot \nabla_E I + c_W \cdot \nabla_W I]_{i,j}^t \quad (4.4)$$

where I^t is the derived image, and I^0 is the original image, λ is chosen with an interval where the stability of the numerical scheme is guaranteed as $0 \leq \lambda \leq 0.25$. N, S, E and W are the mnemonic subscripts for North, South, East, West. ∇ (not to be confused with ∇ the gradient operator) is defined as follows:

$$\nabla_N I_{i,j} \equiv I_{i-1,j} - I_{i,j},$$

$$\nabla_S I_{i,j} \equiv I_{i+1,j} - I_{i,j},$$

$$\nabla_E I_{i,j} \equiv I_{i,j+1} - I_{i,j} ,$$

$$\nabla_W I_{i,j} \equiv I_{i,j-1} - I_{i,j} ,$$

and diffusion coefficients are given by:

$$c_{N_{i,j}}^t = g(| \nabla_N I_{i,j}^t |) ,$$

$$c_{S_{i,j}}^t = g(| \nabla_S I_{i,j}^t |) ,$$

$$c_{E_{i,j}}^t = g(| \nabla_E I_{i,j}^t |) ,$$

$$c_{W_{i,j}}^t = g(| \nabla_W I_{i,j}^t |) .$$

Different functions g can be used, as Perona and Malik (1990) suggested, the function used in this project was:

$$g(\nabla I) = \exp(-(\|\nabla I\|/K)^2) , \quad (4.5)$$

where K , is constant. Perona and Malik (1990) proposed to calculate a histogram of the absolute value of the gradient throughout the image and set K to the 90% value of its integral at every iteration.

Comparative results, with Gaussian filtering, are shown in Figure 4.4. By visual inspection, the noise reduction seems to be the same in both cases; however, edges in the Gaussian filtering look more blurred than in the anisotropic one. Sharper edges are observed in the neck, the larynx, the pharynx and the upper lip of the anisotropic filtered image.

The preservation of the edges is achieved by the $c_{N_{i,j}}^t, c_{S_{i,j}}^t, c_{E_{i,j}}^t, c_{W_{i,j}}^t$ images. These images provide the *gradient* or edge information in the four different directions, north, south, east and west; c_N and c_S define the horizontal edges whereas the vertical edges are defined by c_E and c_W as shown in Figure 4.5. Differences between c_N and c_S are almost imperceptible due to the definition of this coefficients in terms of the $\nabla_N I_{i,j}$ and $\nabla_S I_{i,j}$ differences. This applies to c_E and c_W . The complete edge image with both, horizontal and vertical edges could be defined as:

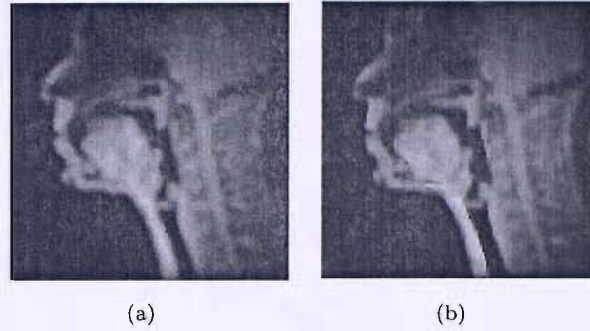


FIGURE 4.4: Comparison between Gaussian (a) and anisotropic filtering (b). Gaussian image was obtained using a window size of 4×4 and $\sigma = 4.5$. Anisotropic image was obtained setting $K = 90\%$ and $t = 10$. The anisotropic filtered image shows sharper edges, in particular, the neck, the larynx, the pharynx and the upper lip.

$$edge = \sqrt{\left(\frac{c_N + c_S}{2}\right)^2 + \left(\frac{c_E + c_W}{2}\right)^2}$$

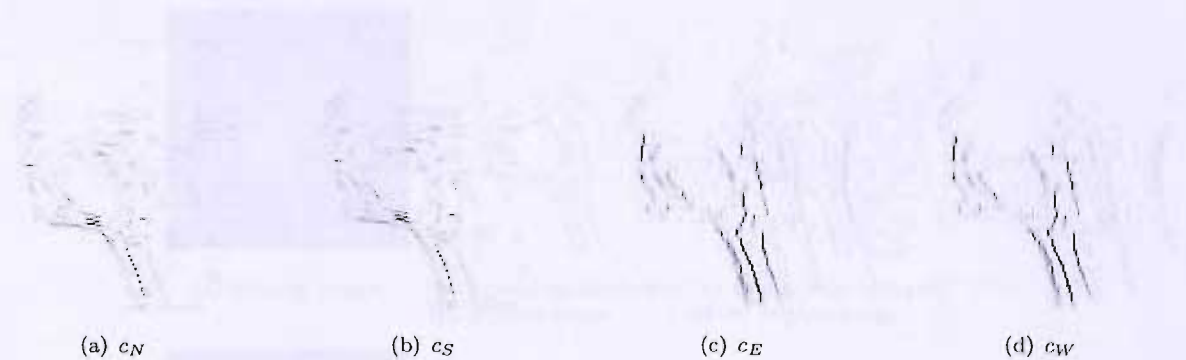


FIGURE 4.5: Images c_N , c_S , c_E and c_W which contain the edge information in the anisotropic filtering process. The images c_N and c_S give the horizontal edge definition whilst vertical edge information is provided by c_E and c_W . Differences between c_N and c_S are almost imperceptible due to the definition of this coefficients in terms of the $\nabla_N I_{i,j}$ and $\nabla_S I_{i,j}$ differences. This applies to c_E and c_W .

This edge information is shown in Figure 4.6. As it can be observed, the edges for the hard and soft palate are not defined; while information for the tongue tip and the tongue blade is poor as well.

For comparison, results obtained for the application of the Sobel and Canny edge detectors over the anisotropic filtered images are presented in Figure 4.7. The edges derived from the filtered images have less spurious details due to the use of the noise reduction filter.

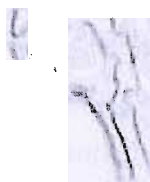


FIGURE 4.6: The complete edge detection image of the first frame.

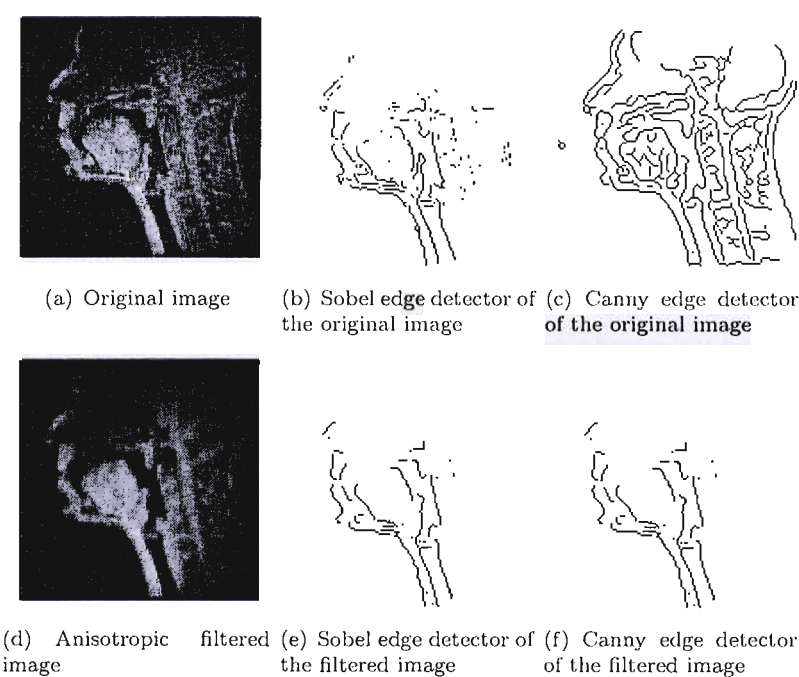


FIGURE 4.7: Comparison between Sobel and Canny edge detectors. Better results are obtained with the anisotropic filtered image due to the noise reduction were spurious details were reduced. Very similar results were obtained using the Sobel and the Canny edge detectors on the anisotropic filtered image.

4.4.2 Generation of the Tongue Model

The vocal tract shape is a complex and deforming shape. First, an analysis of the dynamic behaviour of the vocal tract shape was performed. The vocal tract shape was extracted from a set of 39 images, previously hand labelled by Mohammad (1999). From these images, the labelled articulators were: upper lip, hard palate, velum, pharynx, epiglottis, tongue, lower lip and the tooth root, which is not properly an articulator, but it was labelled in the midsagittal images as can be observed in Figure 4.8.

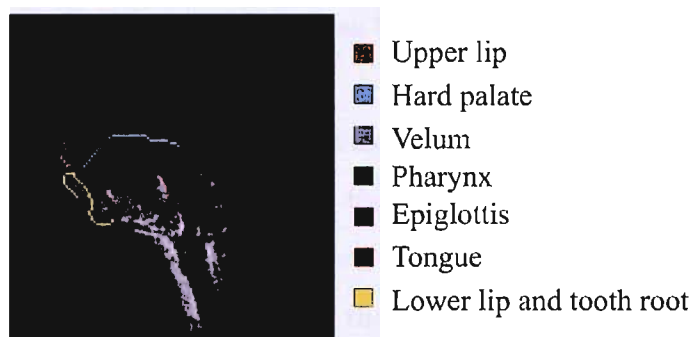


FIGURE 4.8: Hand labelled midsagittal frame of the vocal tract.

A superimposition of the vocal tract shapes was performed in order to analyse the motion of the articulators, in a visual manner as shown in Figure 4.9. For this sequence of images (generated while a subject was pronouncing the word /pasi/) the upper lip shows less movement than the lower lip, the hard palate, velum and pharynx appear almost static, this can be inferred by the thin boundaries. By contrast, the epiglottis, tongue and lower lip are the most deforming articulators. In fact, the tongue appears to be the most deforming one.

The tongue, recognised in this sequence as the most deforming articulator, was chosen to be the first articulator to be modelled. The tongue is formed by several muscles and four main parts can be identified: the root, the tip, the blade and the dorsum. The set of landmark points for the tongue shape is defined as the complete set of points describing it. However, such set of landmark points must be consistent, so that, first, all the points from an specific region of the shape are selected. These regions are defined by calculating the centroid of the shape, drawing a vertical line, and defining an aperture angle on both sides of this line; the

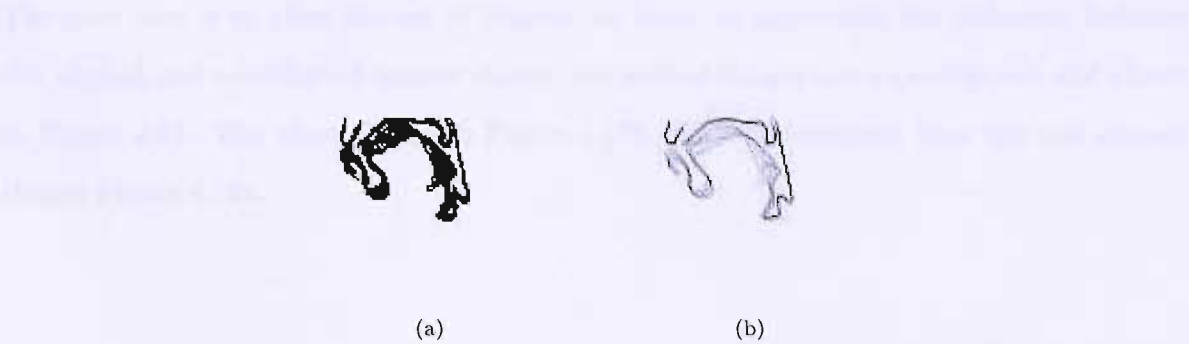


FIGURE 4.9: Vocal tract shape. (a) Sum of all the vocal tract shapes outlined by Mohammad. Thicker boundaries indicates motion and/or deformation of the articulators. (b) shows the average of the sum and with different grey level where the motion and deformation of the articulators can be observed.

set of points contained within the area delimited by this angle is discarded; thus, the set of points excluded from the region delimited by these angles are the tongue shape points to be considered, as illustrated in Figure 4.10. Then, all the shapes are redefined with the same number of points. This number is defined as the maximum number of points for a shape in the training set. Interpolation and smoothing operations are applied to each shape in the set.

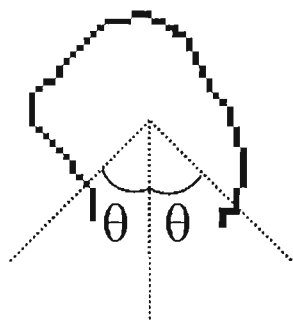


FIGURE 4.10: Selection of vocal tract shape points. The centroid of the shape is calculated, then a vertical line is drawn and the angle $\theta(45^\circ)$, on both sides defines the set of points to be discarded.

Each tongue shape is defined, by 61 points. An example of a tongue shape generated by the reduction step is shown in Figure 4.11b and the result obtained from a posteriori interpolation and smoothing is presented in Figure 4.11c.

The next step is to align the set of shapes. In order to appreciate the difference between the aligned and non-aligned tongue shapes, the sets of images are superimposed and shown in Figure 4.12. The aligned shapes, Figure 4.12b, are more compact than the non aligned shapes Figure 4.12a.

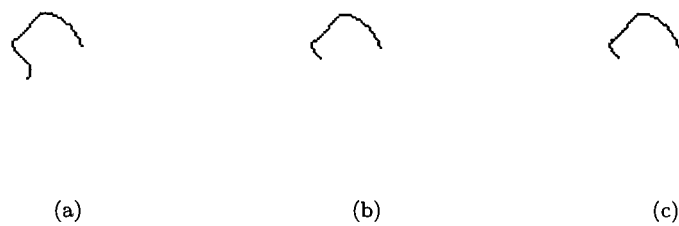


FIGURE 4.11: (a) Original tongue shape, extracted from the outlined data set generated by Mohammad. (b) Reduced tongue shape (c) Interpolated and smoothed tongue shape with 61 points

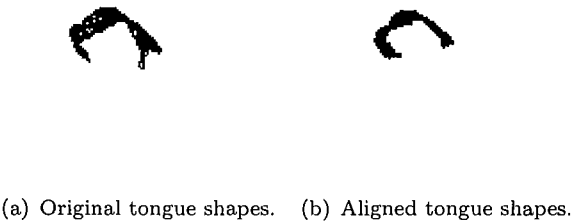


FIGURE 4.12: Superimposed tongue shapes.

The principal component analysis is performed on this set of aligned shapes and the covariance matrix is calculated. The principal axes of the set are described by the eigenvectors and eigenvalues of this matrix; these describe the most significant modes of variation. The total number of eigenvalues and eigenvectors calculated were 122, summing a total of 120.89. According to Table 4.1, the first eigenvalue gives the higher variation with 72% and 90% of the total variation is covered by the first two eigenvalues.

When varying the contribution of the eigenvectors in the shape generation, it can be appreciated how each eigenvalue modifies certain characteristics of the shape. For example,

| λ | Eigenvalue | Percentage | Accumulated percentage |
|-------------|------------|------------|------------------------|
| λ_1 | 87.399 | 72.29 | 72.29 |
| λ_2 | 21.854 | 18.07 | 90.37 |
| λ_3 | 2.8542 | 2.36 | 92.73 |
| λ_4 | 1.5748 | 1.30 | 94.03 |
| λ_5 | 1.3519 | 1.11 | 95.15 |
| λ_6 | 1.1174 | 0.92 | 96.07 |

TABLE 4.1: Eigenvalues of the covariance matrix derived from the aligned tongue shapes.

if tongue shapes are generated using just one eigenvalue the tongue blade is raised toward the hard palate as shown in the first row in Figure 4.13.

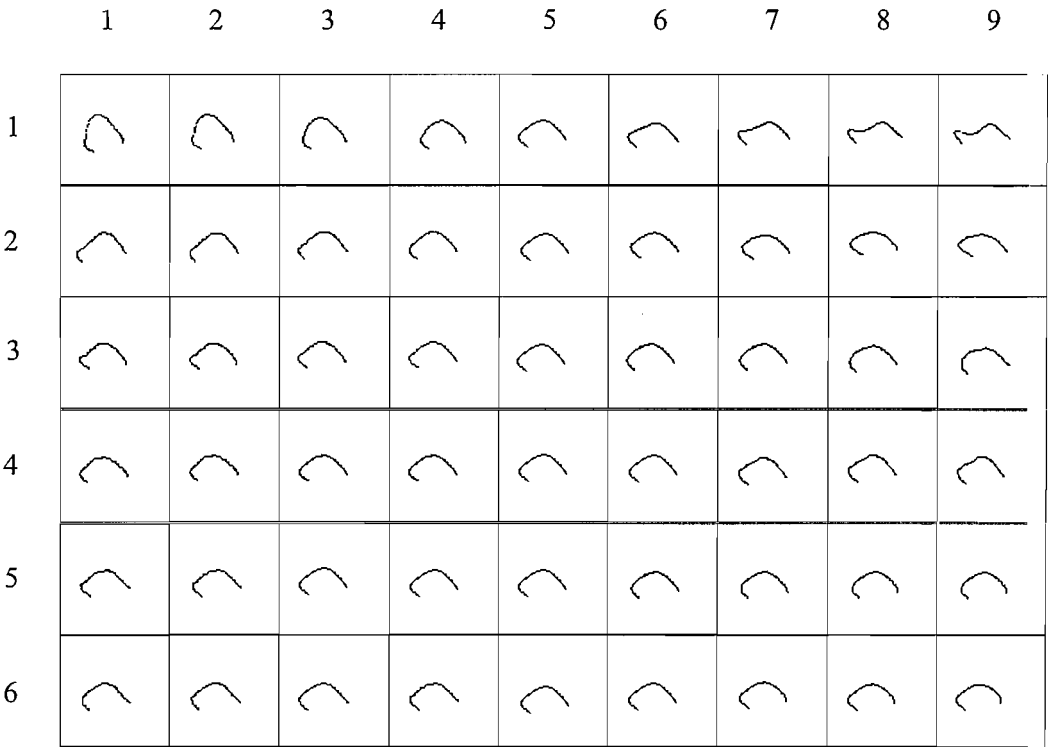


FIGURE 4.13: Tongue shapes generated using one eigenvalue and 9 different steps. It can be observed how the tongue blade is raised towards the hard palate.

Figure 4.14 shows the training set of images. According to these results, this algorithm provides good results since the set of images generated when varying different eigenvalues fits satisfactorily the tongue shapes of the training set.

This algorithm was tested using the set of images generated by Mohammad using one mode of variation with 11 steps. Some results are presented in Figure 4.15. As it can be observed,

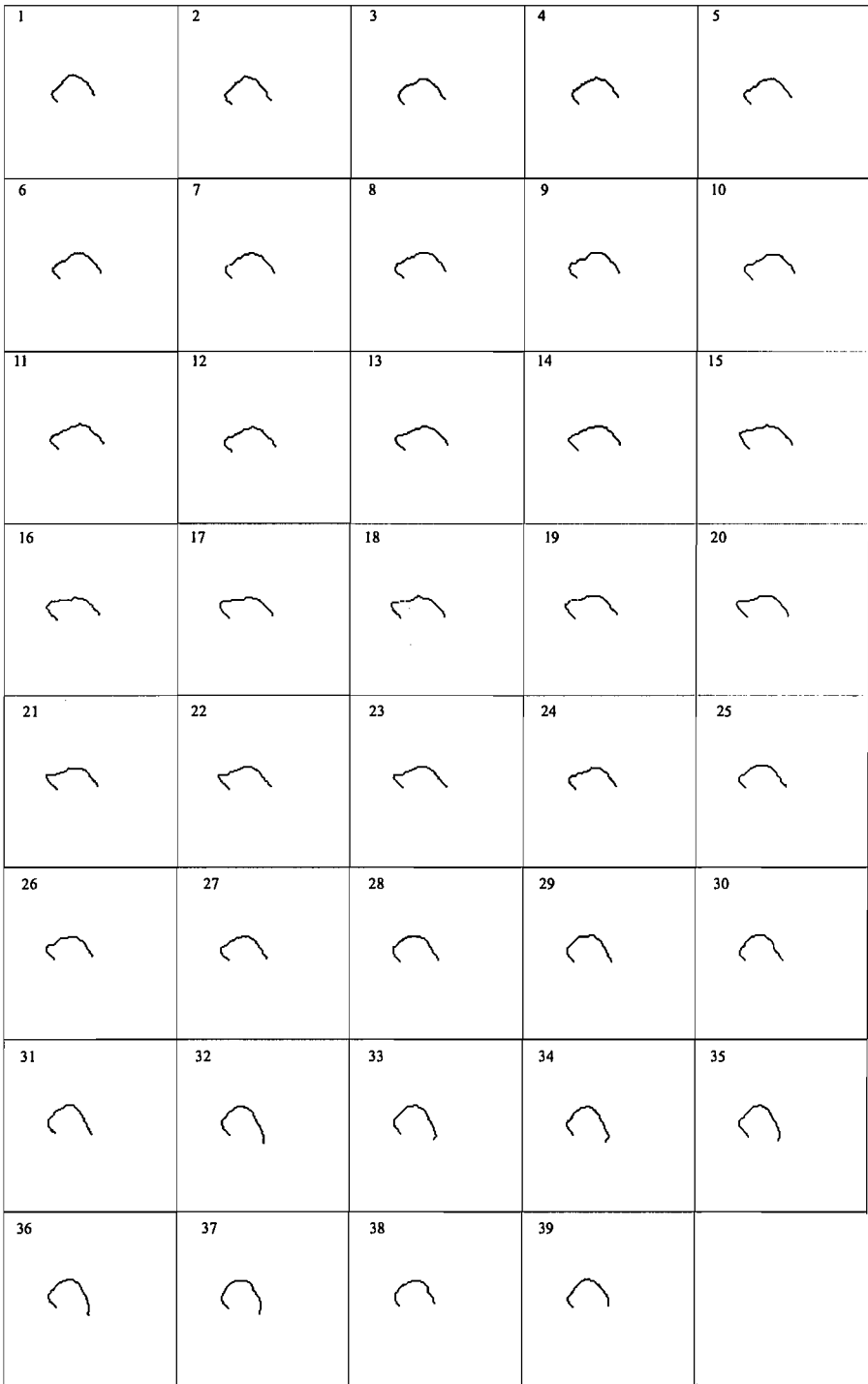


FIGURE 4.14: Sequence of training tongue shapes.

tongue shape for frame 05 was well fitted, however, results for frame 09, as multiple candidates where present, a wrong one was chosen.

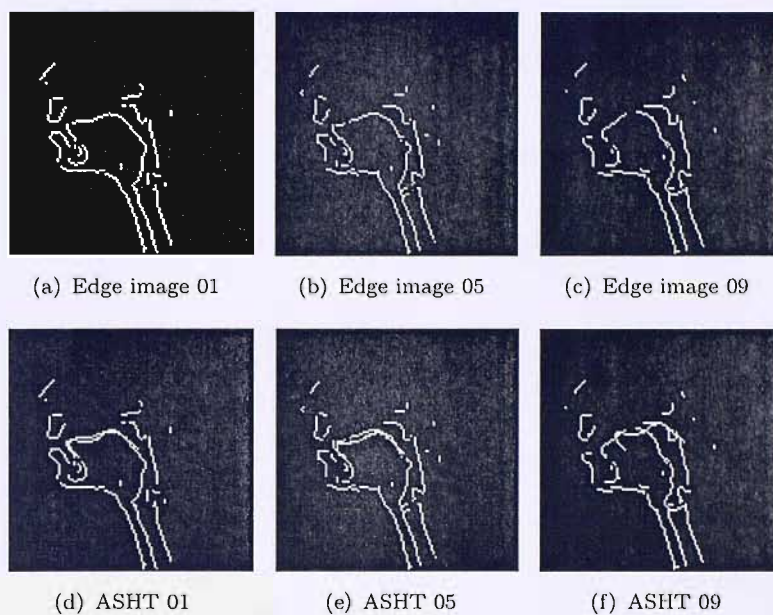


FIGURE 4.15: The ASHT algorithm was tested using one mode of variation model in 11 steps First row shows three edge images selected from the resulting 39 frames of the SDMRI sequence generated by Mohammad. Second row shows results of the application of the HT combined with the ASM method in an isolated analysis. The resulting tongue shape is superimposed in the edge image.

4.5 Conclusions

Although this new form of the HT gives good results with some images, it can go terribly wrong, when selecting incorrect positions and tongue shapes due to missing information, noise or the existence of multiple candidates. In this case, any peak can be chosen, and very different solutions can be provided, since there is no criteria to chose a specific peak. This problem arises from the fact that each image in the sequence is considered individually, or ‘locally’, with no regard whatsoever for the globally best solution. Hence, it is necessary to restrict the search in some appropriate manner. One possibility is to assume that all the parameters vary smoothly over the sequence. One particular extension of the Hough transform, the dynamic Hough transform (Lappas et al., 2002) does exactly this, and can be

used to solve the problem. The adaptation of this technique for our purposes is covered in the next chapter.

Chapter 5

Active Shape Dynamic Hough Transform

5.1 Introduction

The active shape Hough transform (ASHT) is a new form of the Hough transform described in the previous chapter. It uses an active shape model as a shape description to find an object shape within an image by using a voting process. An accumulator is generated during this process and the maximum peak generated is selected as the optimal solution. However, this fails to give the correct solution when multiple peaks are generated since no peak discarding scheme is defined.

The global analysis of a sequence of images exploits both spatial and temporal information to achieve a satisfactory shape extraction in the presence of multiple candidates in the voting space. This analysis is performed by some techniques such as velocity Hough transform (VHT) and the dynamic Hough transform (DHT). These techniques provide a good response in noisy environments and good detection of distorted and occluded objects based on an evidence gathering process.

The VHT technique described by Nash et al. (1997) is an algorithm for tracking parametric shapes with linear or constant velocity. Although the VHT can give good results in some

scenarios, usually the object shapes to be tracked cannot be described with a parametric model or the velocity cannot be modelled.

The DHT technique reported by Lappas et al. (2001), extends the VHT to track objects with arbitrary velocity and motion. The problem is reformulated as tracking an object which changes its velocity, scaling or rotating smoothly frame to frame. The problem is reduced to finding an optimal path of an object within the parameter space. The optimal path is defined by means of a dynamic programming method. Although this technique reduces the computational cost comparing to the VHT, this is still high.

However, an object may vary or deform its shape along a sequence. In this chapter an extension of the DHT is presented. This algorithm is capable of tracking arbitrary and deforming shapes on a sequence with no initialisation required.

This chapter presents first, a brief introduction to DHT, followed by the adaptation of this method for arbitrary shape extraction on image sequences. Finally, the application of the new algorithm on magnetic resonance images for extracting tongue shapes using one and two eigenvalue models is described.

5.2 Dynamic Hough Transform

The analysis of image sequences for object boundary tracking has been approached with different Hough transform extensions. The velocity Hough transform (VHT), reported by Nash et al. (1997), is a method for tracking parametric shapes with constant or linear velocity in a sequence. However, this method is limited not only to parametric shapes with constant velocity but also for its inherent high computational cost. The dynamic Hough transform, reported by Lappas et al. (2001), is an extension of the VHT with no constant velocity constraints. The DHT method is focussed in the global analysis of the image sequence providing tolerance to noise and detection of disconnected segments of its boundary.

The general process of applying the DHT in a sequence is illustrated in Figure 5.1. The method works on binary images so that, first, an edge extraction must be performed over

the image sequence. Then, new Hough transform using an active shape model as shape description is applied and a set of accumulator spaces is generated for each image. The dimensionality of these spaces will depend on the number of shape features to consider in the evidence gathering process, such as centroid position (x,y), rotating and scaling factors and shape parameters. Then, the DHT is applied considering the appropriate energy cost function and constraints to control the movement of the centroid and the deformation of the shape. The optimal path is calculated by means a dynamic programming formulation.

Lappas et al. (2001) defined the cost function for this method based on the path coherence function reported by Sethi and Jain (1987). This function is based on smooth changes in velocity and direction and considering as candidates for the trajectory motion those points with the maximum peaks. The first constraint adds up the peak values of the accumulator space as follows:

$$E_1 = \sum_{t=1}^N Peak_t, \quad (5.1)$$

where $Peak_t$ represents the peak value on the t frame. The second constraint restricts abrupt changes in direction by:

$$E_2 = \sum_{t=2}^{N-1} |\theta_{t-1} - \theta_t|, \quad (5.2)$$

where θ_t represents the direction between t and $t + 1$ frames. Finally, the third constraint limits big changes in velocity by:

$$E_3 = \sum_{t=2}^{N-1} |V_{t-1} - V_t|, \quad (5.3)$$

where V_t represents the velocity between t and $t + 1$ frames. These constraints define the cost function as follows:

$$E = w_1 E_1 - w_2 E_2 - w_3 E_3, \quad (5.4)$$

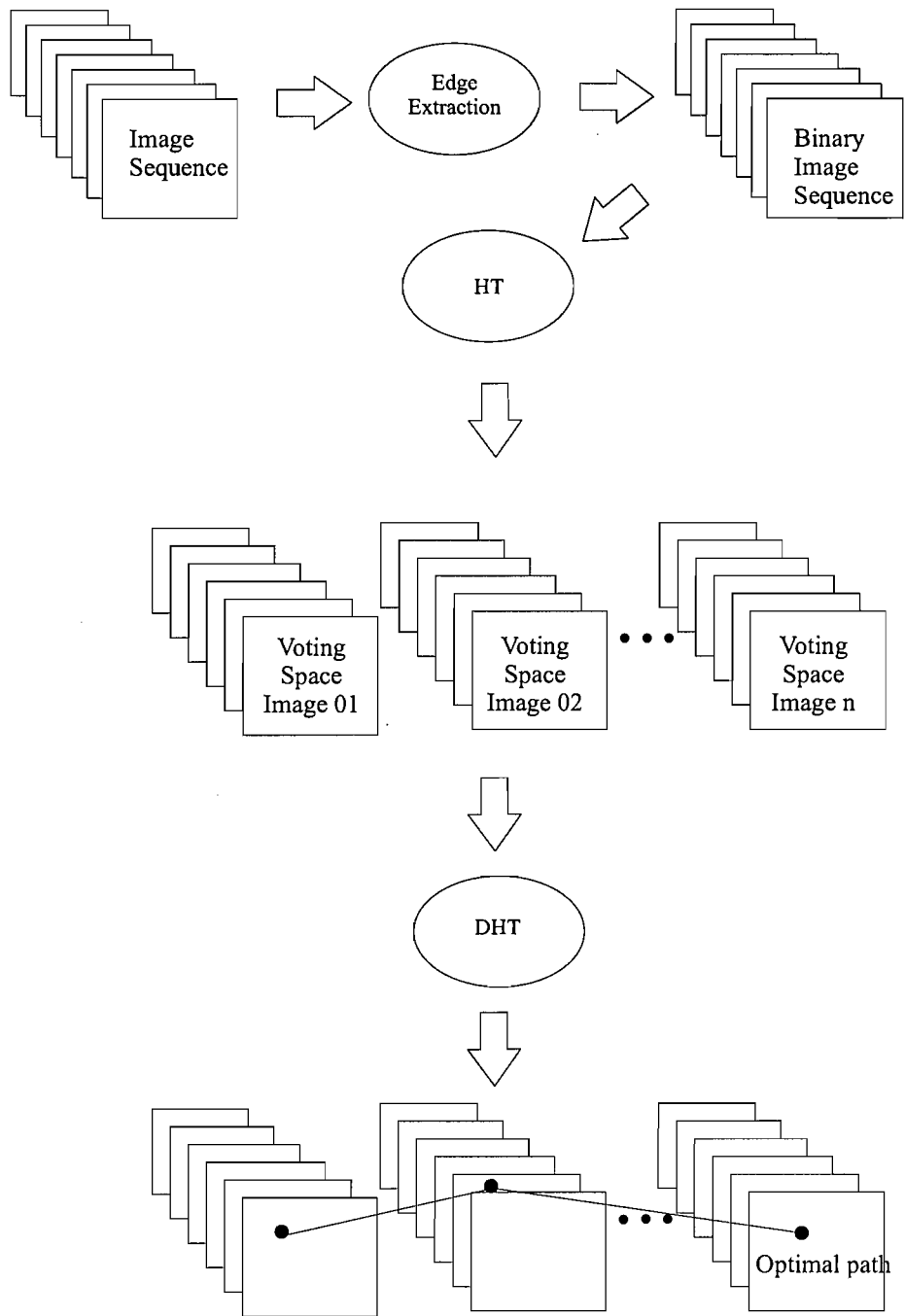


FIGURE 5.1: Process to track arbitrary deforming shapes in a sequence using the new form of the Hough transform and the new adaptation of the DHT. First, binary images are generated; then, the HT method is applied to generate multidimensional voting spaces for each image. Finally, the new adaptation of the DHT is applied using a dynamic programming formulation to generate the optimal path.

which has to be maximised. The w_1, w_2 and w_3 values are the corresponding weights defining the contribution of each energy term. These weights are defined to support changes in velocity or direction.

The problem is focussed on solving the optimal motion trajectory, which maximise Equation 5.4 using a time delayed dynamic programming algorithm. This algorithm is applied due to the characteristics of smooth changes between frames, where double state variables are defined to accomplish the second constraint defined by Equation 5.2, and the third one by Equation 5.3. The relationship of such parameters and constraints is illustrated in Figure 5.2. Here, two-element state vectors are defined and the energy function E_{t-1} is given by:

$$E_{t-1}(u_{t-1}, u_t, u_{t+1}) = w_1 \text{Peak}_t - w_2 |\theta_{t-1} - \theta_t| - w_3 |v_{t-1} - v_t|. \quad (5.5)$$

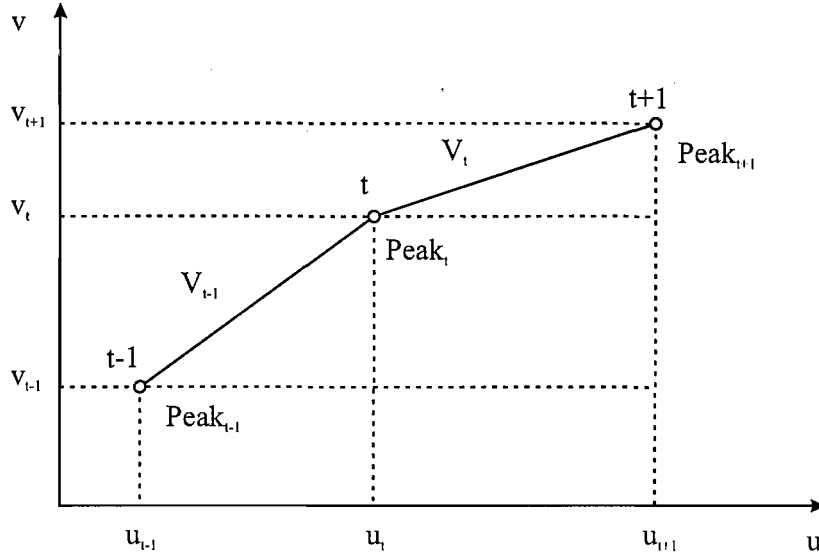


FIGURE 5.2: Relationship among parameters: peak, velocity and direction, reprinted from Lappas et al. (2001). Two double element state variables are defined, first variable is composed by (u_{t-1}, u_t) and the second one by (u_t, u_{t+1}) . The cost function is then based on a 3 frame basis with two-element state variables. The cost function for these two variables is defined by Equation 5.5.

The dynamic programming formulation used by Lappas et al. (2001) is that reported in Raphael (2001). This introduces the concept of super states as the aggregations of states. An initial coarser optimisation is performed using these superstates and refining the optimisation by an iterative process until the optimal path is defined for single states. This

process is illustrated in Figure 5.3. In Figure 5.3(a) the traditional formulation of dynamic programming is presented, where all the possible paths are considered. In Figure 5.3(b) the superstates are defined as aggregations of three states. The optimal path for these superstates is defined and represented by a bold line. Then, the superstates are refined into smaller aggregations as illustrated in Figure 5.3(c); the advantage is that in the refining process all the remaining superstates are not considered anymore. This refining process is iterated until the global optimal solution is determined as shown in Figure 5.3(d).

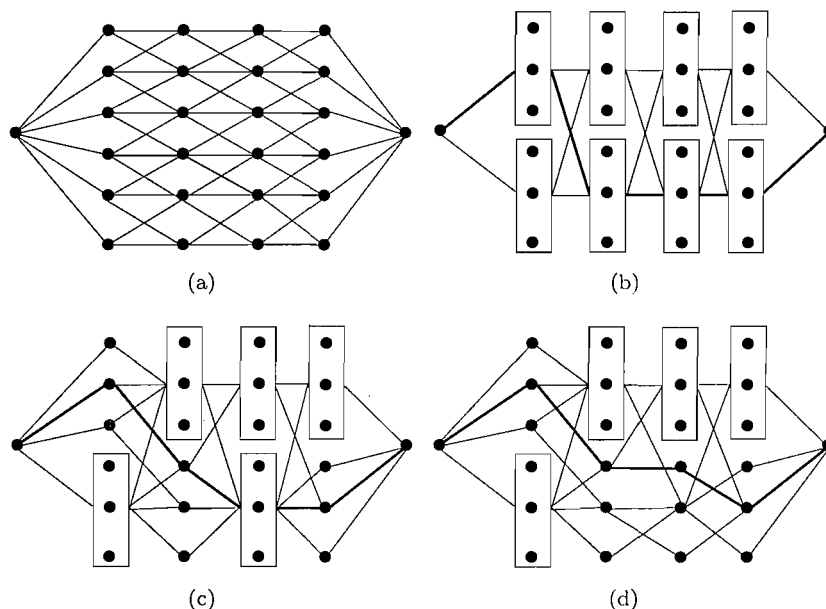


FIGURE 5.3: Coarse to fine dynamic programming formulation. In (a) the original dynamic programming scheme is presented. In the remaining graphs, a possible progress of a coarse to fine formulation is presented. The definition of super states is represented by rectangles containing different states. First superstates are defined in (b) and the optimal path for these supervariables is determined and shown with a solid bold line. A finer optimisation is performed in (c). Finally, the optimal path is determined in (d). The super states defined in the optimal path are refined iteratively until the optimal path is formed by single peaks.

Reprinted from Raphael (2001).

5.3 Combining Active Shape Models and the Dynamic Hough Transform

Some Hough transforms are based on tracking geometrical shapes, so that a parametric model can be used. However, the extension for tracking arbitrary shapes has been addressed by

some researchers. The generalised Hough transform reported by Ballard (1981), is a method of tracking arbitrary shapes where the shape description is contained in an R -table. This table, contains the shape description in terms of the distance and the angle formed between boundary shape points and a reference point.

Zamora et al. (2003) reported a combination of the GHT with active shape models for detecting the cervical and lumbar vertebrae using a two step approach: first, the application of the generalised Hough transform to find an estimate for the position of the desired shape; and second, the application of the ASM method to refine the extraction, using the estimate provided by the GHT as initialisation. They used a customised generalised Hough transform, as reported by Tezmol et al. (2002), where the characteristic R -table of the shape is generated using a mean of 50 templates obtained from a manual land marking process. The R -table contained information for different (r, α) values, where r and α are the distance and the angle, respectively, between edge points and a reference point. This approach applied the generalised Hough transform with a rigid template and consequently a second stage was needed to deform such a template to fit it as possible to the target shape. In a later work, Howe et al. (2004) reported a three stage process. The first two remain as previously reported in Zamora et al. (2003), and the third stage included the isolated application of the active shape model for each single vertebrae.

In this thesis, a new combination of the dynamic Hough transform and the new form of the Hough transform using the active shape models as a shape description is presented. This technique tracks arbitrary and deforming shapes in image sequences. The complexity of this method will depend not only on the number of modes of variations considered in the searching process but also on the scaling, translating and rotating degrees of freedom.

5.4 Application to MRI Sequences

Initially, the searching process for the tongue on MRI sequences is tested. First, the ASHT algorithm is applied to each edge image in the sequence to generate its corresponding accumulator space. As it was mentioned in Chapter 4, the first two modes of variation

comprise the 90% of the variation of the tongue shape. In the formulation presented in this section two modes of variation are considered in the searching process.

Therefore, the voting space for this sequence is then defined as a multidimensional space which includes the centroid position (x, y) , the first two modes of variation (b_1, b_2) , and the scaling factor S . An additional variable required is the time dimension.

The DHT is based on the assumption that changes in the Hough voting parameters such as position, shape descriptors and scaling factor vary smoothly from frame to frame, so that constraints for each parameter can be defined. Analysing the MRI sequence generated by Mohammad, it can be observed that even though the tongue is not suffering a translational motion, it is deforming. This deformation causes the centroid of the tongue of two consecutive frames to change its position, since the tongue is deforming smoothly from frame to frame the centroid position and the shape parameters change smoothly as well.

The cost function is based on smooth changes of the centroid position, the scaling factor and the first mode of variation for the set of points representing the maximum peaks. The first constraint considers the maximum peaks in the voting space, as defined in Equation 5.1, as follows:

$$E_1 = \sum_{t=1}^N Peak_t, \quad (5.6)$$

where $Peak_t$ represents the peak value on the t frame. The second constraint restricts abrupt changes in the first mode of variation by:

$$E_2 = \sum_{t=2}^{N-1} |b1_{t-1} - b1_t|, \quad (5.7)$$

where $b1_t$ represents the mode of variation between the t and $t+1$ frames. The third constraint limits big changes in centroid position or centroid velocity, as defined in 5.3 as:

$$E_3 = \sum_{t=2}^{N-1} |V_{t-1} - V_t|, \quad (5.8)$$

where V_t represents the velocity of the centroid between the t and $t + 1$ frames. Finally, the fourth constraint limits big changes in the scaling factor by,

$$E_4 = \sum_{t=2}^{N-1} |S_{t-1} - S_t|, \quad (5.9)$$

where S_t represents the scaling factor between the t and $t + 1$ frames. These constraints redefine the cost function as follows,

$$E = w_1 E_1 - w_2 E_2 - w_3 E_3 - w_4 E_4, \quad (5.10)$$

which has to be maximised. The w_1, w_2, w_3 and w_4 values are the corresponding weights defining the contribution of each energy term. These weights are defined to support larger or smaller changes in the corresponding features.

This method was applied on the MRI sequences generated by Mohammad using one mode of variation in 11 steps. Initially, the standard dynamic programming formulation was performed to find the optimal path. Comparative results between the isolated and the global analysis are presented in Figure 5.4. As it can be observed, results obtained from the global analysis improved those obtained by the isolated analysis. This is more evident in results for frame 09, where in the presence of multiple candidates a wrong one was chosen, while in the global analysis the correct candidate was selected.

Although results obtained from the ASDHT algorithm were better, the computational cost due to practical implementations was high. The simulations were performed using a PC with processor Pentium 4 at 4 GHz and 1.5 GB in RAM. The initial slow version of this algorithm was implemented in Matlab using the traditional formulation of dynamic programming with execution times in the order of days; for example, finding the tongue shape in edge images of 128×128 pixels with 11 steps for the first eigenvalue was achieved with an execution time of 28 hours. The reimplementations of this algorithm in the C language and using the formulation of the coarse to fine dynamic programming decrease the computational time of the same simulation mentioned in just 37 seconds, making the method more practical. Detecting the

tongue shape in sequences of 39 frames of 56×56 pixels using a model of two eigenvalues and 9 steps for each eigenvalue variable and 9 step for scaling steps was achieved with an execution time of 2.7 hours.



FIGURE 5.4: Comparative results between the ASHT and the ASDHT algorithms applied to full edge images. The algorithms were tested using one mode of variation **model** in 11 steps. First row shows three edge images selected from the resulting 39 frames of the MRI sequence generated by Mohammad. Second row shows results of the application of the HT combined with the ASM method in an isolated analysis. Finally, third row shows the results of extracting the tongue shape during the global analysis of the **sequence** by applying the ASDHT algorithm. The resulting tongue shape is superimposed in the **edge image**.

5.5 Conclusions

In this chapter an algorithm to extract arbitrary and deforming shapes in image sequences with no initialisation required was described. This algorithm performs a global analysis and

it is focused on the definition of an optimal path. The optimal solution was obtained by applying a cost function which considers the maximum peaks in the voting space and smooth in changes in feature parameters such as centroid position, modes of variation of the shape and scaling factors. The computational cost was greatly reduced when the coarse to fine dynamic programming formulation was adopted for finding the optimal path. Preliminary results were presented to show how this method overcome the difficulties to chose the most appropriate candidate in the voting space. A more extensive evaluation is presented in the next chapter.

Chapter 6

Experimental Evaluation

6.1 Introduction

In previous chapters the active shape Hough transform (ASHT), which is a new form of the Hough transform (HT), has been described. This method uses an active shape model (ASM) as a shape description in an evidence gathering process which generates a multidimensional voting space. This process is performed over an individual edge image where the maximum peak in the generated accumulator space represents the optimal solution; however, when multiple solutions are present, any candidate can be chosen.

These difficulties can be overcome by performing a global analysis of the image sequence using an adaptation of the dynamic Hough transform (DHT). This new adaptation provides an optimal solution by defining a cost function to track object shapes with smooth changes in shape, position, and scaling factor.

This algorithm, as many Hough transform extensions, suffers from high computational cost. In this project, some implementation aspects contributed to increase this cost; for example, the programming language used to develop the application, the multidimensional voting space defined and the dynamic programming formulation used. In this project the algorithms were initially implemented in Matlab, followed by a reimplementations in the C language to improve the performance. The voting space was large since each analysis was made over sequences of

more than 20 frames each. Furthermore, it was necessary to consider a representative number of steps for scaling and eigenvalues which controlled the modes of variation of the model.

This algorithm was tested using midsagittal MRI sequences of the vocal tract. In these sequences the subjects were pronouncing the token /pasi/, which requires most of the vocal tract articulators to move and deform. An analysis of the movements made by the different articulators has shown that the tongue is the most challenging articulator to be tracked, therefore, the analysis was initially focused on the tongue. In advance, it can be said that results demonstrated the ASDHT method does find the tongue shape on MRI sequences and the model is shown to be easily extended as information of the lips was included in the tongue model. The algorithm was tested to find these shapes.

The performance of the ASDHT algorithm was tested on different sets of data: MRI sequences of seen or training data, unseen data sets and synthetic tongue image sequences. The tests performed on the training images measure the performance of the method on the ground-truth data set as it is referred in this document. Unseen images, which are images of a different subject, were employed to evaluate the performance of the algorithm on unseen data sets. Synthetic images were generated using the tongue model described in previous chapters to measure the effectiveness of the algorithm to find the correct shape over a known set of eigenvalues.

In this chapter the automatic extraction of the deforming tongue shape in midsagittal MRI sequences is evaluated. First, a brief introduction to the chamfer distance metric is given, this was adopted as a metric to measure the effectiveness of this method. Then a description of the experiments is detailed, followed by the results for each experiment also including an evaluation of the algorithm under noisy conditions. Subsequently, the results for the extension of the model which includes information of the lips are detailed. To conclude, a summary of the evaluation of the experiments is presented.

6.2 Chamfer Distance Metric

In terms of evaluation, simply comparing the resulting eigenvalues is not effective, because the difference between eigenvalues does not give an estimation of the variation in the final shape. The comparison needs to be made in the image plane, comparing how close the estimated shape is to the tongue shape. To measure the effectiveness of the algorithms, a chamfer distance metric was used (Borgefors, 1988). This metric has been demonstrated, as Thiel and Montanvert (1992), to be a good approximation of the Euclidean distance between neighbouring pixels with a ‘low computational cost’. This is achieved by placing an estimated shape over a distance image of the target shape. First, a distance image of the edge image was generated. In a distance image each position for a non-edge pixel is assigned with a distance value to the nearest edge pixel and all the edge pixel positions are assigned with a zero value. For example, Figure 6.1(a) shows an edge image and Figure 6.1(b) its corresponding distance image, where the edge pixel positions were assigned with zero (in black) and the non-edge pixel with a distance value (in grey). The closer to the edge pixels the smaller (darker) the value.

The distance value was defined sequentially with 3×3 forward and backward masks, as described by Borgefors (1988). The mask was centred on each pixel of the image and its value will be the minimum of the set of sums of the corresponding pixel in the image and the mask. Once a distance image has been calculated, the estimated shape is placed over it. The difference (or deviation) of the estimated shape from the target shape, that will be referred as the error value, is calculated as the root mean square (rms) average as follows:

$$Error = \frac{1}{c_1} \sqrt{\frac{1}{N} \sum_{i=1}^N d_i^2},$$

where N is the number of points in the resulting tongue shape and d_i is the corresponding distance value in the distance image for the i th pixel. The final division by c_1 is to compensate the minimum distance value in the mask. In Figure 6.1, a target shape is shown in (a), its corresponding distance image in (b) and the estimated shape in (c). The estimated shape

is placed over the distance image as shown in (d) and the error value is calculated. In this case the error is 1.59, and the result is superimposed on the target image in (e), to visually appreciate the difference between both shapes, the target shape is outlined in grey and the fitted result in black.

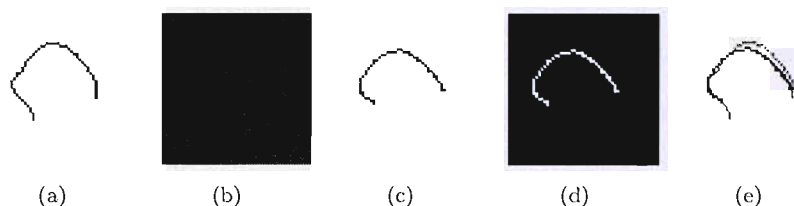


FIGURE 6.1: Tongue shape of the frame 39 of the training data sequence and it is corresponding chamfer distance image. The estimated shape (c) is superimposed on the chamfer distance image (d). The error calculated for this match is 1.59; the difference between both shapes is appreciated in (e) with the target shape in grey and the fitted result in black.

6.3 Ground-Truth Data Set

The evaluation of the algorithm on the ground-truth data set was designed to perform experiments on isolated tongue shapes and full edge images. This set comprises a sequence of 39 images from subject PJ previously labelled by Mohammad (1999). The size of the original images was 128×128 pixels. The first experiment tests the extraction on isolated training tongue shapes; the second experiment used the full edge images. In general, different scaling factors were tested using models with one and two eigenvalues covering 72 and 90% of the tongue variation respectively.

6.3.1 Isolated Tongue Shapes Experiments

In the isolated tongue shape experiments, the tongue shapes were extracted from the isolated hand-labelled tongue data used for training. The term ‘isolated’ is used to define only the edge points defining the tongue shapes. Therefore, these experiments will evaluate the fitting of the tongue model to the target, since the image data set is clean with no noise on it (each non-tongue edge pixel is considered as noise in these experiments). However, it is important

to note that each original tongue shape was rotated, scaled and translated to align all the tongue shapes with the corresponding mean shape, which generated small variations from the original set of labelled tongue shapes.

In these experiments, the multidimensional voting space was defined as a four dimensional space with $N_{rows} \times N_{cols} \times N_{sc} \times Nb_1$ where N_{rows} and N_{cols} denote the size of the image, N_{sc} is the number of discrete scaling steps and Nb_1 denotes the number of discrete steps for the first eigenvalue.

Instead of employing the whole image, windows of 56×56 pixels containing the tongue shapes were extracted from the full size images to reduce the size of the accumulator space. The number of discretisation steps for each parameter was chosen as non prime numbers to allow the definition of clusters during the dynamic programming stage. The scaling factor was used with 9 steps over different ranges. Different step values were tested and 0.02 found to be large enough to vary the tongue shape smoothly. The shape parameter, denoted by b_1 , was allowed to vary in the range $-3\sqrt{\lambda_k} \leq b_1 \leq 3\sqrt{\lambda_k}$, as suggested by Cootes et al. (1995), as suitable limits to generate new shapes. The number of discrete steps for the first eigenvalue was defined in 9 steps as well, which is an odd number to include the mean shape during each analysis, and with the same number of steps in both ranges (low and high).

Different ranges for the scaling factor were tested as shown in Table 6.1. In experiment 1.A no scaling factor was used. As can be observed, the minimum average of error was obtained in experiment 1.B. The corresponding results for each experiment in a frame by frame basis can be appreciated in Figure 6.7 and the sequence of results for the experiment 1.B is presented in Figure 6.3. The target tongue shapes are outlined in grey and the resulting fitted models are presented in black. As mentioned in Chapter 4, the first mode of variation of the tongue shape model controls the blade part of the tongue. Therefore, frames from 27 to 38 where the main variation of the tongue is in the blade part of the tongue are well fitted. In contrast, the tongue tip in frames 15 to 23 cannot be entirely fitted. The minimum error of 0.69 for this experiment was produced for frame 27 and the maximum error of 1.60 was obtained for frame 39. Frame 16 is the second worst match obtained with an error of 1.51 and visually it can be appreciated how the tip of the tongue cannot be well fitted. It is important to mention

| Experiment | Scaling factor | Error |
|------------|------------------------------------|-------|
| 1.A | No scaling | 1.16 |
| 1.B | From 0.92 to 1.08 in steps of 0.02 | 1.05 |
| 1.C | From 0.96 to 1.12 in steps of 0.02 | 1.12 |
| 1.D | 1 - 1.16 in steps of 0.02 | 1.13 |

TABLE 6.1: Average rms error (chamfer distance) for different scaling factors used in the first set of experiments using a one eigenvalue model.

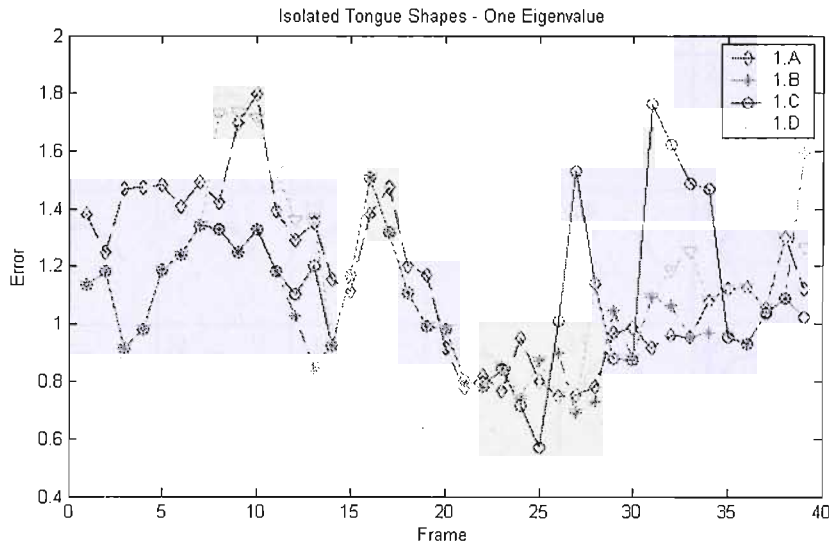


FIGURE 6.2: Average rms error (chamfer distance) for each frame using different scaling factors for experiment on isolated tongue images using a one eigenvalue model. Minimum error for experiment 1.B was found in frame 27, while the maximum error was found for frame 39.

that the tongue shapes were reduced during the model generation to generate a consistent set of points to define the tongue shape. For this reason the ends are not well matched.

In experiments for the two eigenvalue model a different testing scheme was adopted. Based on the evidence that with a one eigenvalue model a very good approximation of the position and shape of the tongue was obtained, the value of the first eigenvalue was restricted according with that results as follows:

$$-3\sqrt{\lambda_k} + stepB_1(B_1 - 1) \leq b_1 \leq -3\sqrt{\lambda_k} + stepB_1(B_1 + 1),$$

where B_1 is resulting step counter of the first eigenvalue in the fitted model and $stepB_1$ is

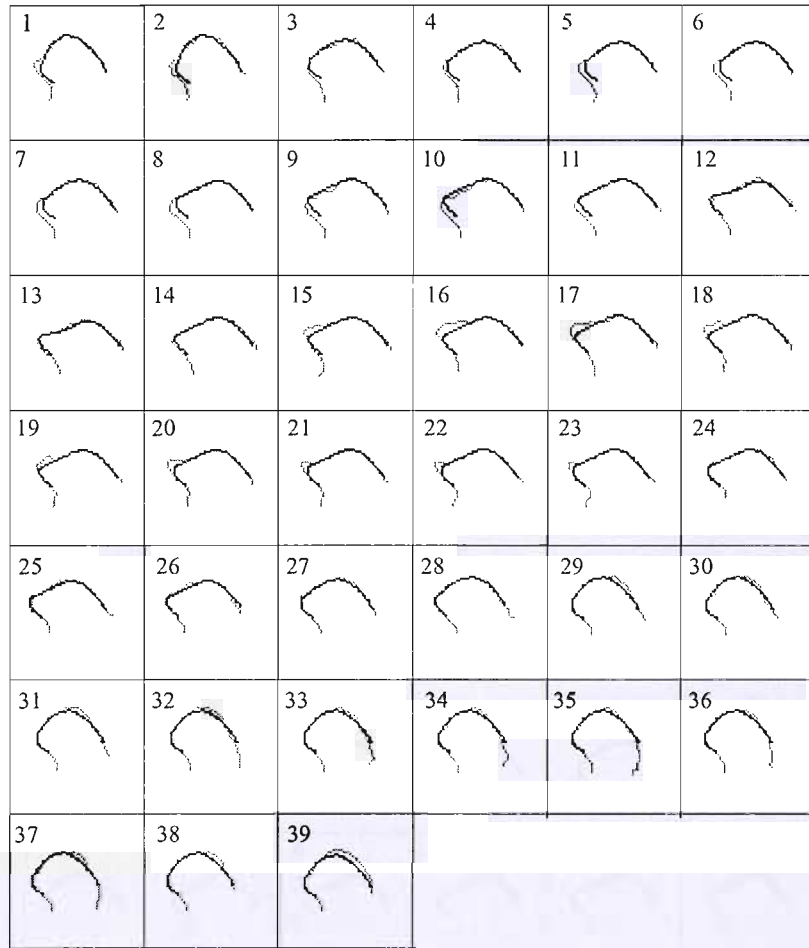


FIGURE 6.3: Superimposed results obtained for the sequence of extracted tongues for the experiment 1.B. The tongue tip is not well fitted by the one eigenvalue model in frames from 15 to 23, while the tongue blade is well fitted in most of the frames.

the step value used in such experiments. The second eigenvalue, b_2 , is allowed to vary in the range $-3\sqrt{\lambda_2} \leq b_2 \leq 3\sqrt{\lambda_2}$.

A summary of the results obtained for the two eigenvalue model are presented in Table 6.2 and the results in a frame by frame basis for such experiments are shown in Figure 6.4. The sequence of results for each tongue (grey) with the fitted model (black) superimposed are presented in Figure 6.5. As can be observed frames from 1 to 3 and from 15 to 20 are fitted better than in the one eigenvalue experiments shown in Figure 6.3. However, frames from 29 to 38 were fitted better in the one eigenvalue experiments. Although the algorithm produces better results for some frames using the one eigenvalue model, the error generated for the sequence 1.H is less than the one obtained for the 1.B.

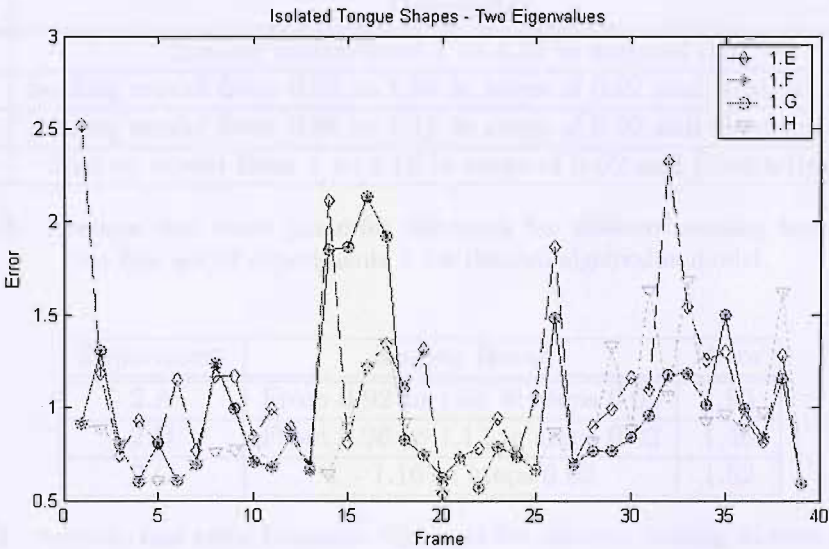


FIGURE 6.4: Average rms error (chamfer distance) for each frame using different scaling factors. The minimum sequence error was calculated for experiment 1.H.

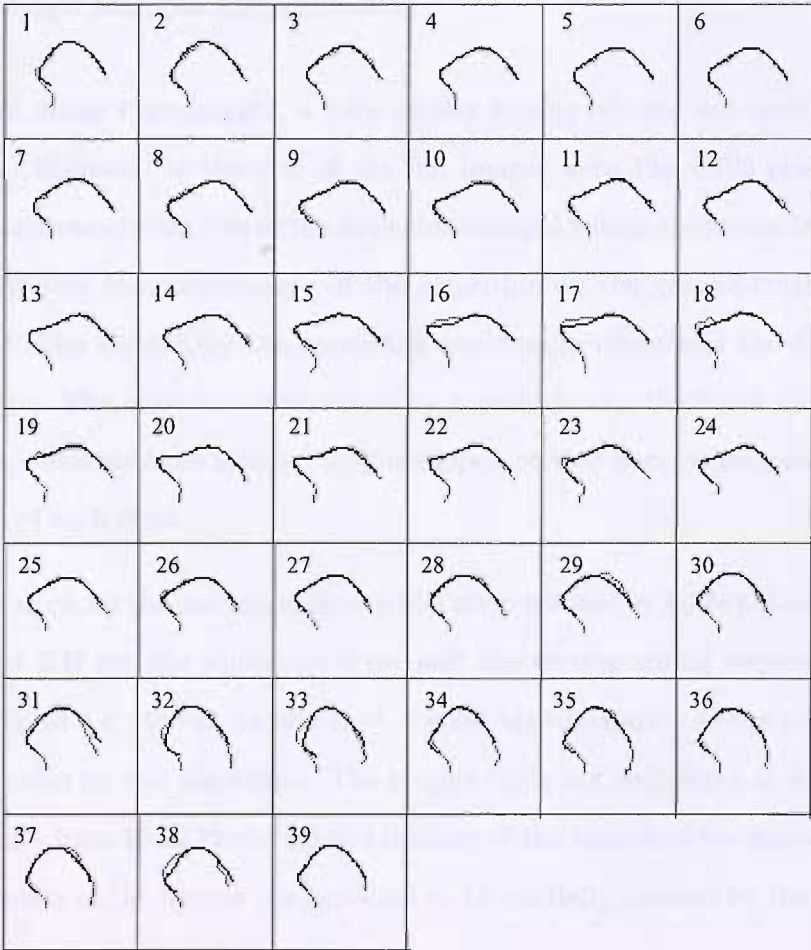


FIGURE 6.5: Superimposed results obtained for the sequence of extracted tongues for the experiment 1.H.

| Experiment | Description | Error |
|------------|---|-------|
| 1.E | Scaling model from 1 to 1.16 in steps of 0.02 | 1.11 |
| 1.F | Scaling model from 0.92 to 1.08 in steps of 0.02 and Restricting b1 | 1.01 |
| 1.G | Scaling model from 0.96 to 1.12 in steps of 0.02 and Restricting b1 | 0.98 |
| 1.H | Scaling model from 1 to 1.16 in steps of 0.02 and Restricting b1 | 0.88 |

TABLE 6.2: Average rms error (chamfer distance) for different scaling factors used in the first set of experiments 1 for the two eigenvalue model.

| Experiment | Scaling factor | Error |
|------------|---------------------------------|-------|
| 2.A | From 0.92 to 1.08 in steps 0.02 | 1.93 |
| 2.B | From 0.96 to 1.12 in steps 0.02 | 1.46 |
| 2.C | 1 - 1.16 in steps 0.02 | 1.52 |

TABLE 6.3: Average rms error (chamfer distance) for different scaling factors used in the second set of experiments for one eigenvalue model.

6.3.2 Full Edge Images Experiments

In the full edge image experiments, a very similar testing scheme was used using different scaling factors. However, as the size of the full images were 128×128 pixels as shown in Figure 6.6, the corresponding size of the multidimensional voting space was increased. These experiments will test the performance of the algorithm on the ground-truth data set with some degree of noise defined by the remaining non-tongue edges and the missing pixels in the tongue shape. The error was determined by superimposing the fitted model on chamfer distance images obtained from isolated tongue shapes, so that non-tongue pixels do not affect the calculation of such error.

The results obtained for the one eigenvalue model are presented in Table 6.3 and in Figure 6.7. The experiment 2.B got the minimum error and the corresponding sequence of frames is presented in Figure 6.8. As can be observed, a good approximation of the tongue shape and position is provided by this algorithm. The tongue tip is not well fitted in frames from 1 to 9, while in frames from 10 to 19 the tip and dorsum of the tongue of the model do not fit the target. These parts of the tongue are expected to be partially covered by the two eigenvalue model.

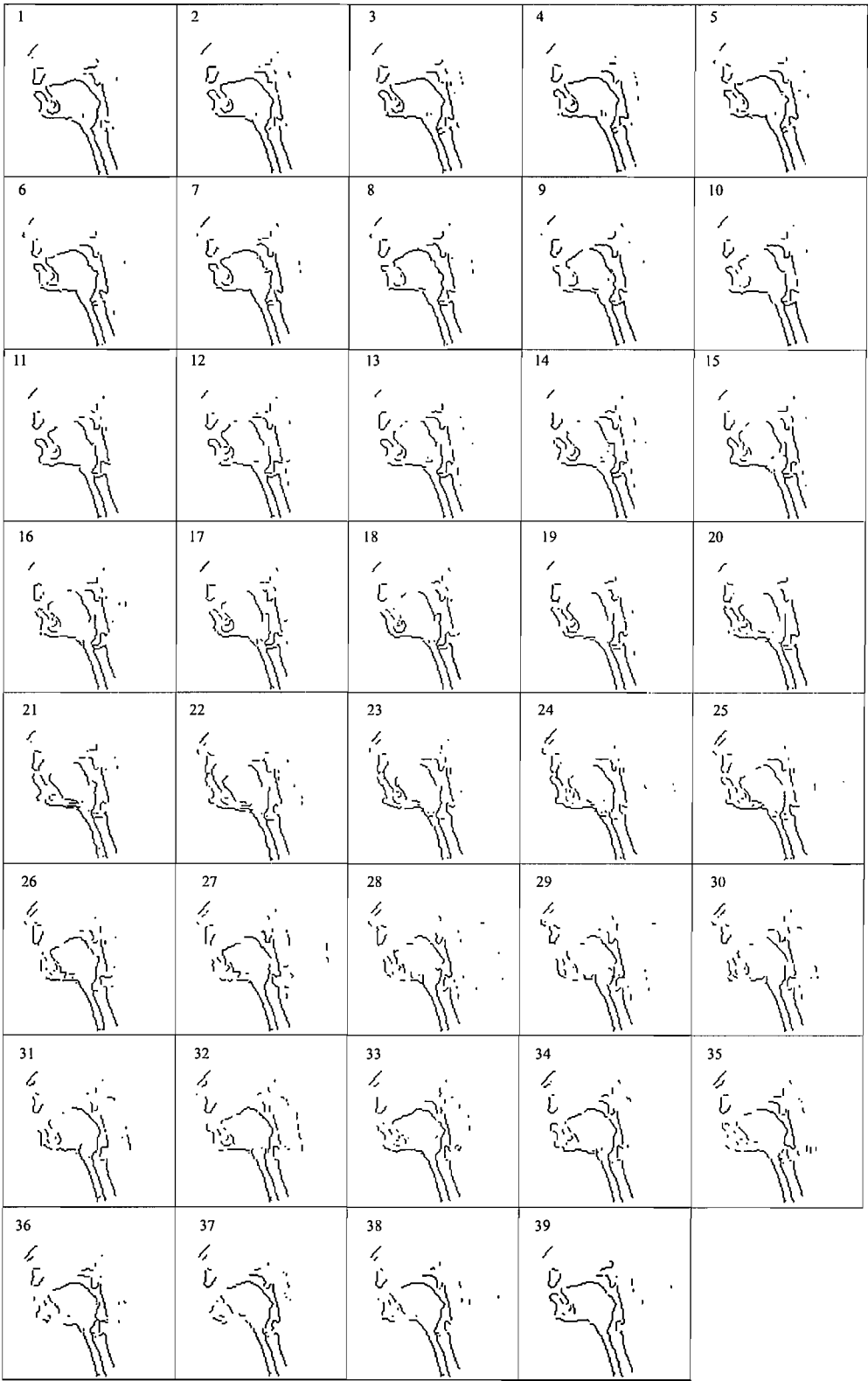


FIGURE 6.6: Set of edge images used in the full edge experiments.

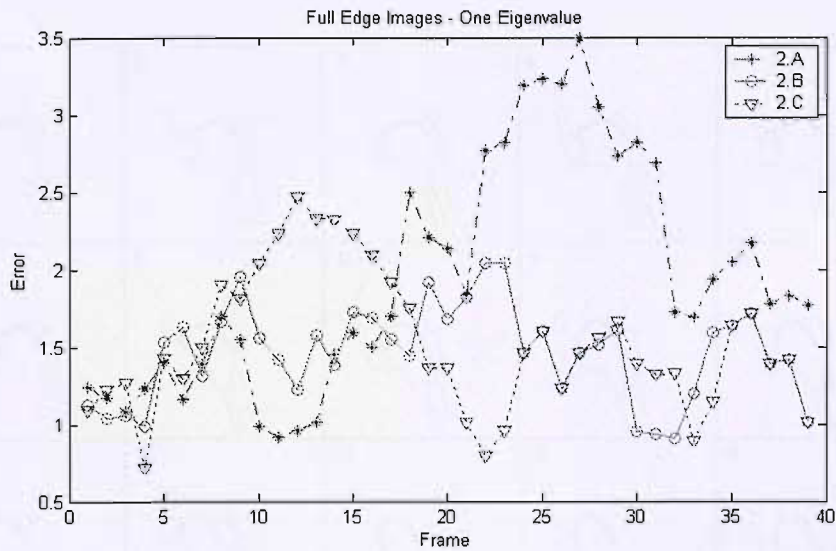


FIGURE 6.7: Average rms error (chamfer distance) for each frame using different scaling factors.

| Experiment | Scaling factor | Error |
|------------|---------------------------------|-------|
| 2.D | From 0.92 to 1.08 in steps 0.02 | 1.68 |
| 2.E | From 0.96 to 1.12 in steps 0.02 | 1.57 |
| 2.F | 1 - 1.16 in steps 0.02 | 1.54 |

TABLE 6.4: Average rms error (chamfer distance) for different scaling factors used in the second set of experiments for one eigenvalue model.

The corresponding experiments on full edge images were performed on the two eigenvalue model. As can be observed in Table 6.4 and in Figure 6.9 the minimum error was obtained in experiment 2.F, however, this error is larger than the minimum error obtained in experiment 2.B for the one eigenvalue experiments. The sequence of results are presented in Figure 6.10. Although the first twelve frames show a better fit than the one eigenvalue model, frames 13 and 15 to 19 seem to be attracted by the edges corresponding to the lower lip. This might be occasioned by the lack of information of the tongue in such frames. Frames 32 to 34 are better fitted in experiment 2.F than in the 2.B.

In Figure 6.11 a comparison of the better results obtained using one and two eigenvalues in isolated images (a) and full edge images are presented.

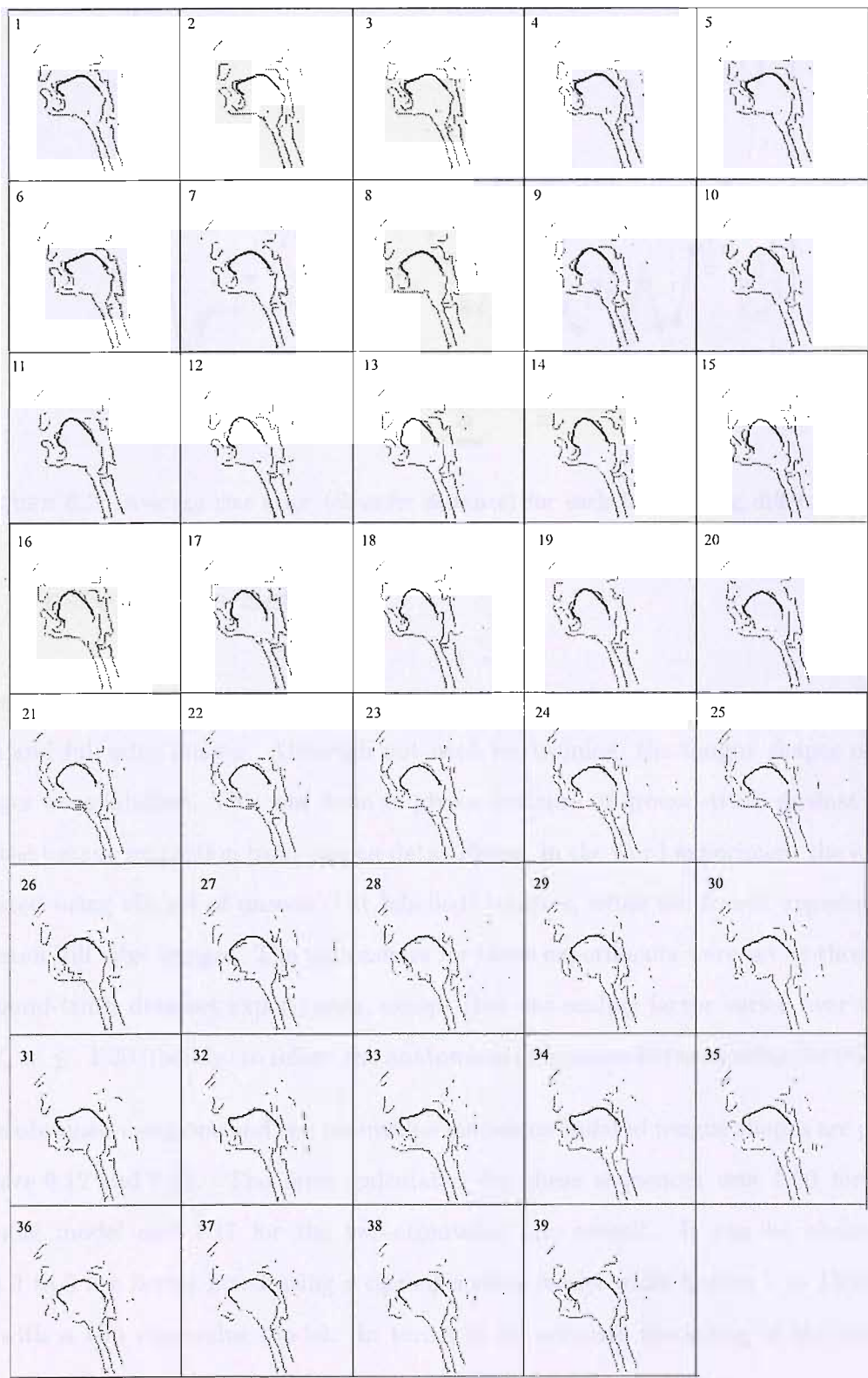


FIGURE 6.8: Superimposed results obtained for the sequence of extracted tongues for the experiment 2.B.

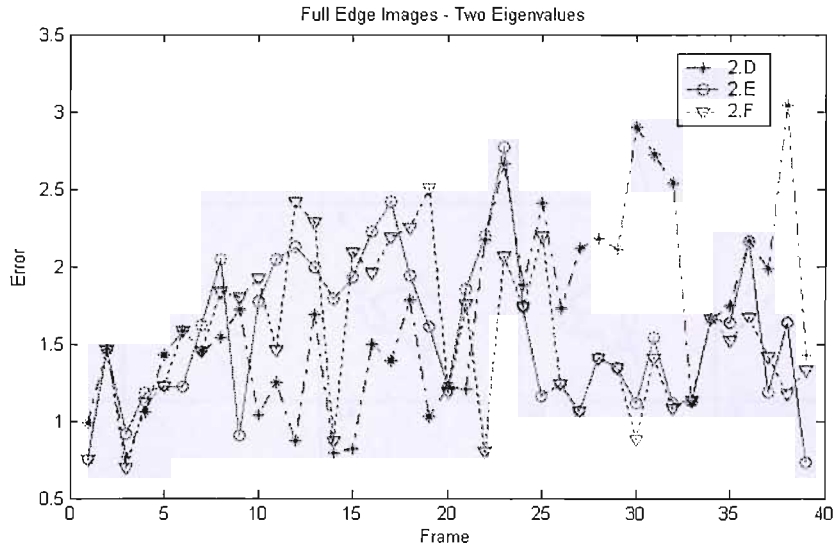


FIGURE 6.9: Average rms error (chamfer distance) for each frame using different scaling factors.

6.4 Unseen Data Set

The tests performed on the unseen data set comprises experiments on isolated tongue shape images and full edge images. Although not used for training, the tongue shapes of this set of images were labelled. This was done to give a measure of ground-truth against which to assess the tongue extraction from unseen data. Hence, in the third experiment the extraction was tested using the set of unseen (but labelled) tongues, while the fourth experiment used the unseen full edge images. The parameters for these experiments were set as those used in the ground-truth data set experiments, except that the scaling factor varied over the range $1.14 \leq s \leq 1.30$ (mostly) to reflect the anatomical differences between subjects SG and PJ.

Results obtained using one and two eigenvalue models on isolated tongue shapes are presented in Figure 6.12 and 6.13. The error calculated for these sequences was 1.60 for the one eigenvalue model and 1.17 for the two-eigenvalue one overall. It can be observed that frames 3 to 5 are better fitted using a one eigenvalue model while frames 7 to 15 are better fitted with a two eigenvalue model. In terms of an accurate modelling of the vocal tract configuration results obtained with the two eigenvalue model are better.

The next experiment was performed on full edge unseen images. This set is presented in

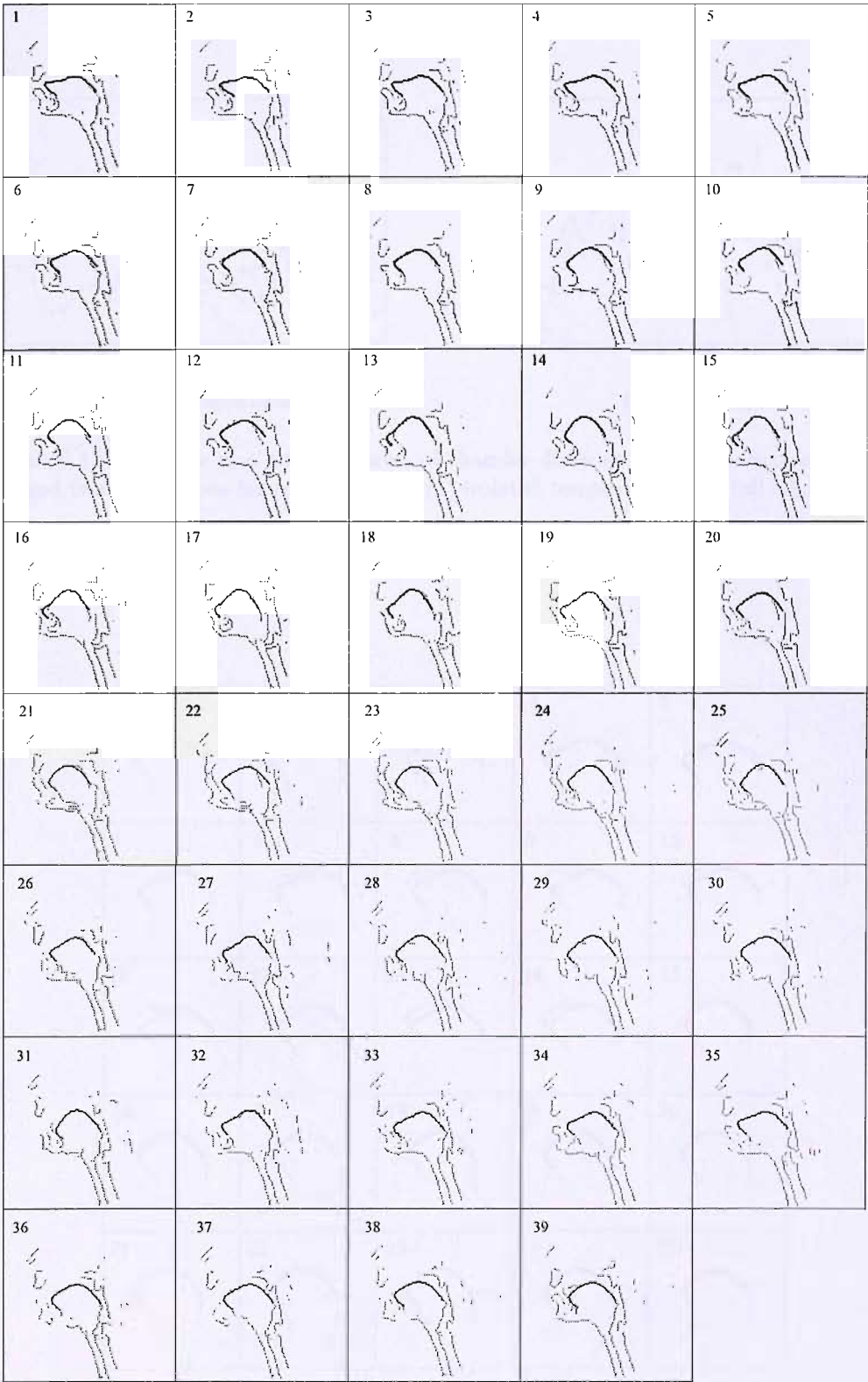


FIGURE 6.10: Superimposed results obtained for the sequence of extracted tongues for the experiment 2.F. The tongue tip is better in this sequence than in the one eigenvalue model. However, it does not match the correct shape on frames 13, and 15 to 19, where the model considers part of the lower lip as part of the tip tongue.

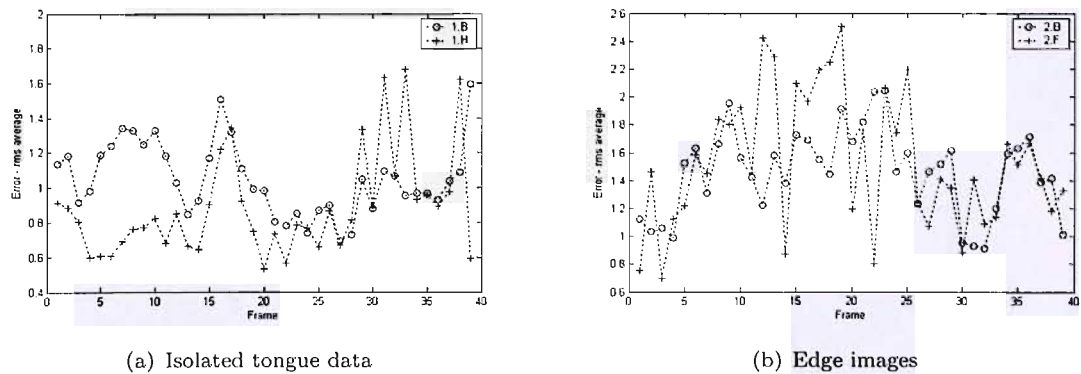


FIGURE 6.11: Average rms error in terms of chamfer distance for tongue extraction using one and two eigenvalues for training data: (a) isolated tongue data: (b) full edge images.

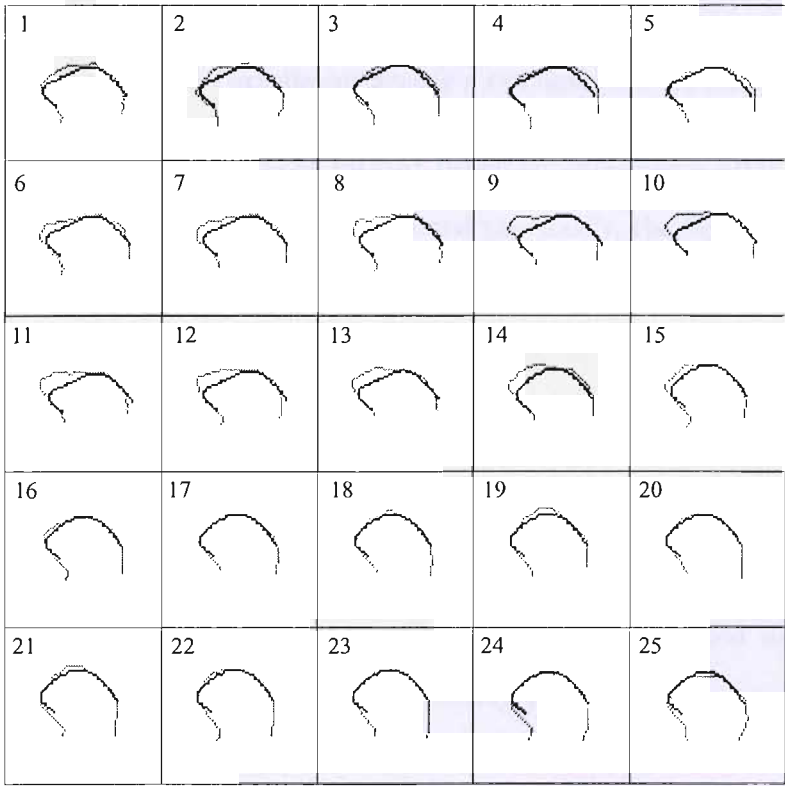


FIGURE 6.12: Superimposed results obtained for the sequence of extracted unseen tongue shapes for the experiment 3 using a one eigenvalue model.

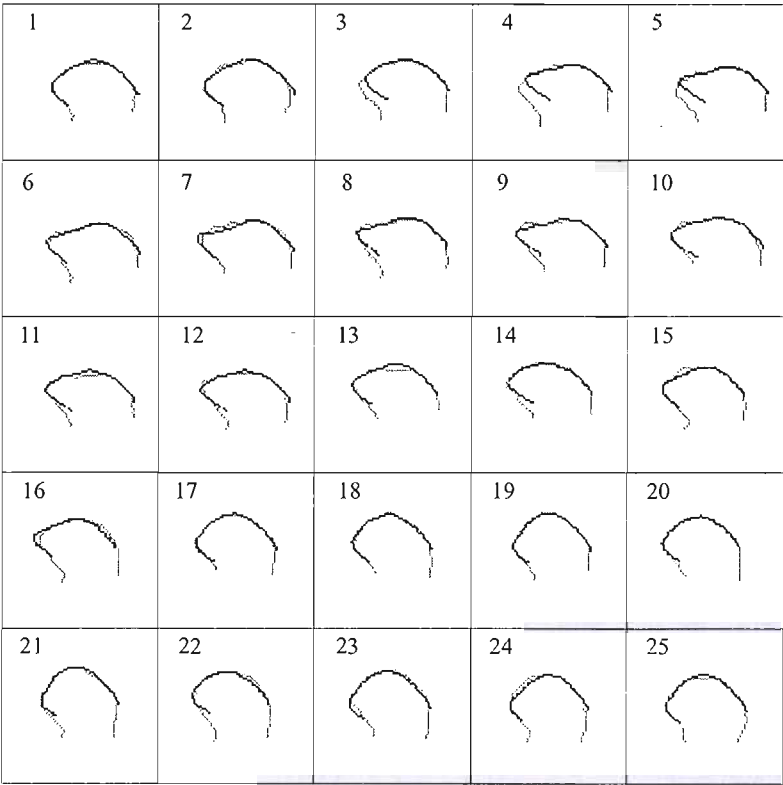


FIGURE 6.13: Superimposed results obtained for the sequence of extracted unseen tongue shapes for the experiment 3 using a two eigenvalue model.

Figure 6.14. The original images were filtered using an anisotropic filter followed by the application of a Sobel edge detector. As mentioned previously, the Sobel edge detector was chosen by simplicity since the results offered by more sophisticated edge detectors were very similar. Results obtained using one and two eigenvalue models on these sets of edge images are presented in Figure 6.15 and 6.16, the errors for this sequence were 1.36 and 1.25 for the one and two eigenvalue models respectively. It can be observed that frames from 6 to 14 provides a smaller error in the two eigenvalue results than in the corresponding one eigenvalue. This comparison frame by frame for both models can be appreciated in Figure 6.17. As can be seen in Figure 6.17(a), the tongue shape is well detected in the isolated tongue data. The average rms error for two eigenvalues is not noticeably poorer than for the seen data. This is very encouraging for practical application.

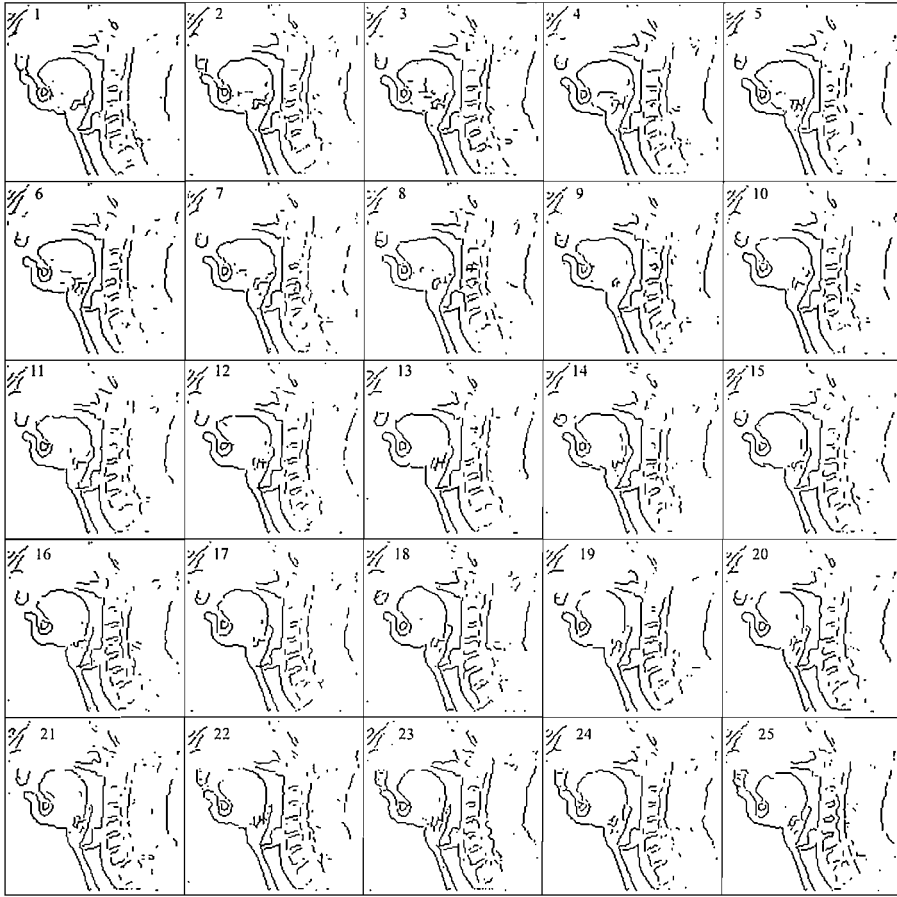


FIGURE 6.14: Sequence of full edge images tongues shapes for the experiment 4.

6.5 Synthetic Data

The third set of experiments were conducted over 50 synthetic sequences generated to measure the effectiveness of the model to find the correct shape over a know set of eigenvalues, data for which the ground-truth was known, without the subjectivity of manual labelling.

Experiments were conducted for 50 synthetic sequences of 20 frames each. A three-eigenvalue model with no scaling factor was used to generate these sequences. The model used for generating these sequences had three eigenvalues and no scaling factor was used. The first and second eigenvalues were selected randomly over a defined range. Then the third eigenvalue was defined from these two eigenvalues to generate a valid or possible tongue shape. These sequences were analysed with one and two eigenvalues. Results are presented in Figure 6.18. The average error obtained for such sequences was of 1.58 using a one eigenvalue model and

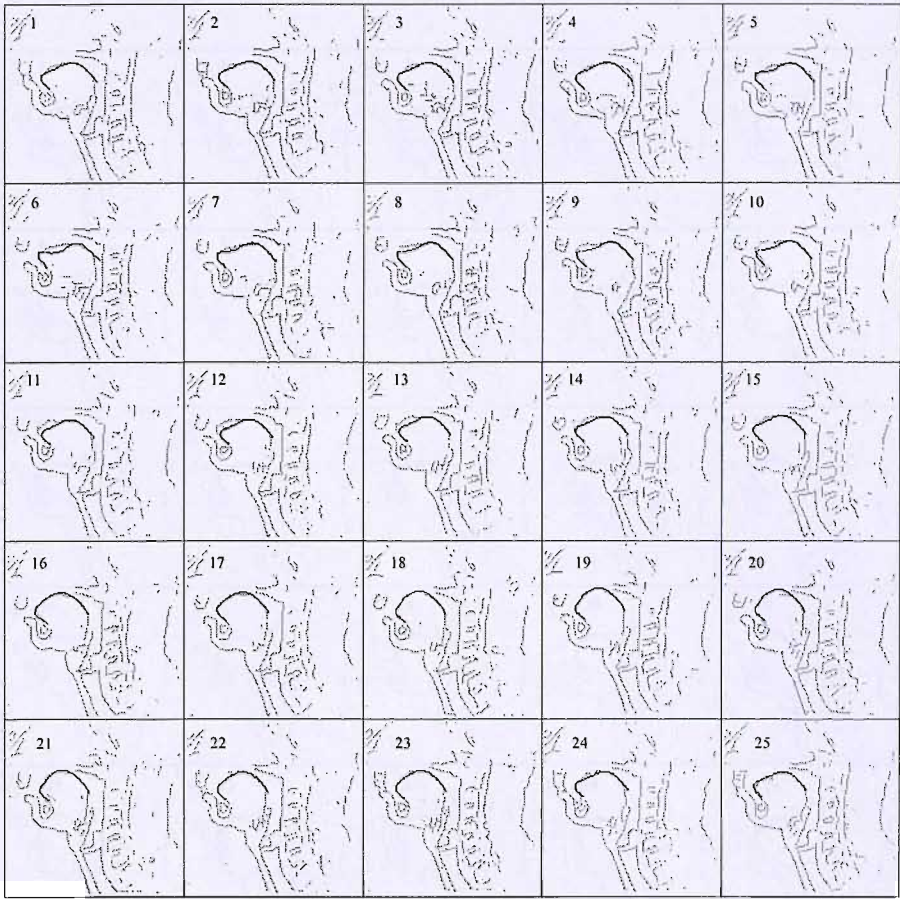


FIGURE 6.15: Superimposed results obtained for the sequence of full **edge unseen images** using a one-eigenvalue model.

1.16 using a two eigenvalue model, which confirms that in general better results are **obtained** when more eigenvalues are considered in the shape generation.

6.6 Noise Response

The ASDHT algorithm was tested on noisy conditions using a two eigenvalue model. The ground-truth data set, which consists on the set of training isolated tongue images, was corrupted with noise using a program to generate random numbers developed in the C language. The pixels were corrupted by inverting their value, so that pixels in black are set into white and vice versa. These sequences were corrupted using 11 different combinations of seed values in the generation of random numbers with 10, 20, 30 and 40% of level of noise. In Figure 6.19 a sequence of tongue shapes with 10% of noise is **presented** and **Figure 6.20** shows

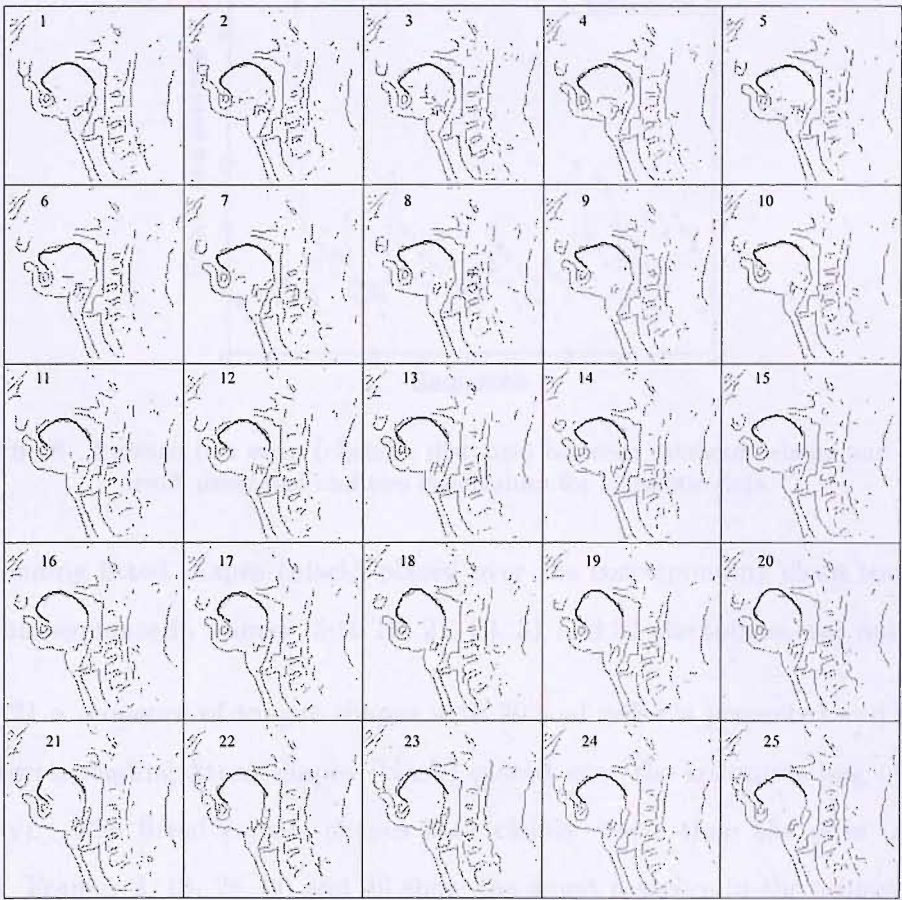


FIGURE 6.16: Superimposed results obtained for the sequence of full edge unseen images using a two-eigenvalue model.

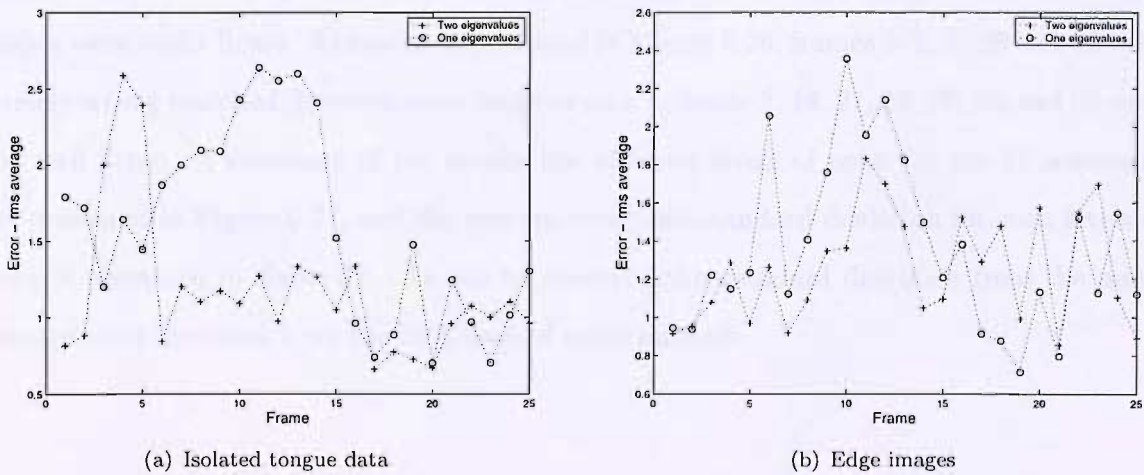


FIGURE 6.17: Average rms error (chamfer distance) for tongue extraction using one and two eigenvalues for unseen data: (a) isolated tongue data: (b) full edge images.

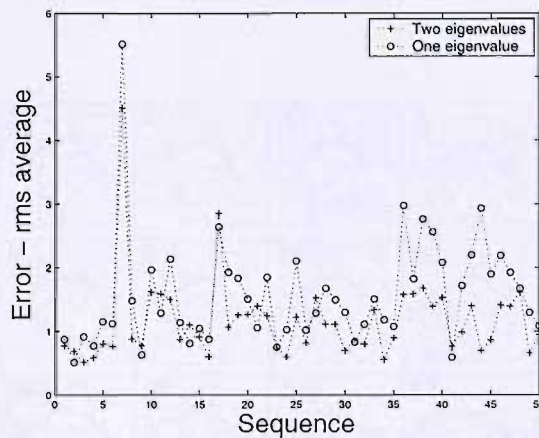


FIGURE 6.18: Average rms error (chamfer distance) between extracted shape and ground-truth using one and two eigenvalues for synthetic data.

the corresponding fitted shapes (black) placed over the corresponding clean tongue images (grey). In this sequence in frames 15 to 17, 21, 23, 31 and 33 the tongue was not well fitted.

In Figure 6.21 a sequence of tongue shapes with 20% of noise is presented and Figure 6.22 shows the corresponding fitted shapes (black) placed over the corresponding clean tongue images (grey). The fitted tongue shapes look clearly worst than the ones presented in Figure 6.20. Frames 2, 13, 28, 30 and 36 show the worst matches in the sequence. Similar results are presented for the 30% level of noise in Figures 6.23 and 6.24. Although results for frames 19, 28, 30, 34 and 38 were bad, frames 10 to 12, 21, 27, 29, 31 and 39 were well fitted.

With a 40% of noise augmented to the image sequence, as shown in Figure 6.25, some tongue shapes were badly fitted. As can be appreciated in Figure 6.26, frames 4, 5, 6, 29 and 32 were terribly wrong matched, however some tongues such as frame 2, 14, 21, 24, 27, 25, and 26 were still well fitted. A summary of the results the different levels of noise for the 11 sequences are presented in Figure 6.27, and the average error and standard deviation for such levels of noise is presented in Table 6.5. As can be observed the standard deviation from the mean average error increases from the 30% level of noise onwards.

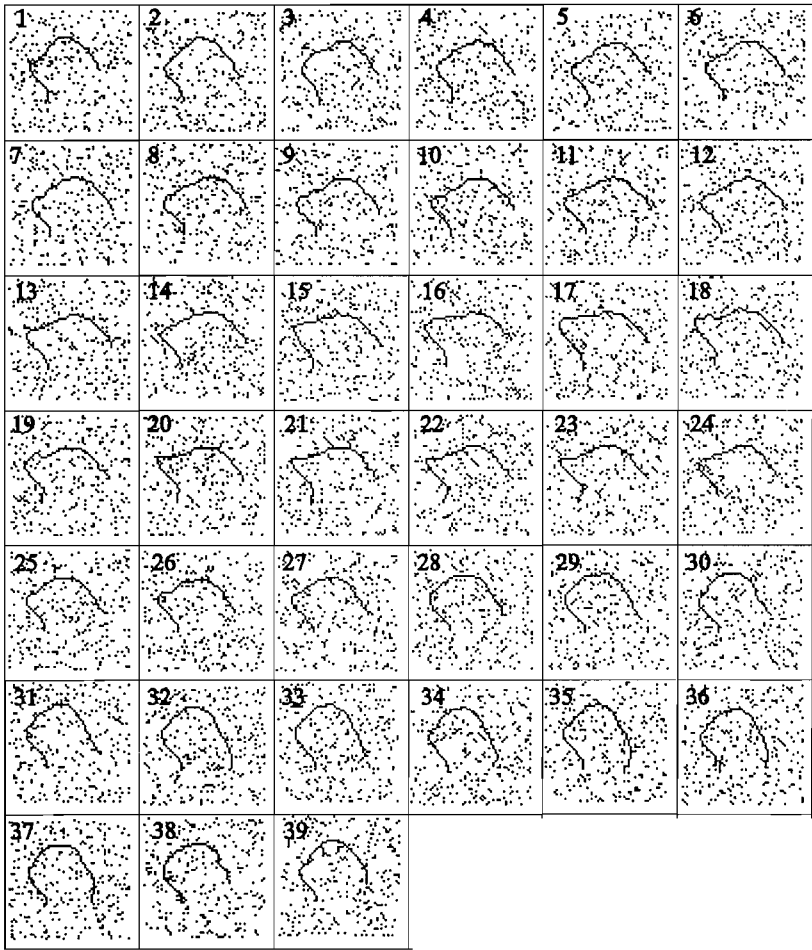


FIGURE 6.19: Training data set sequence corrupted with 10% of noise.

| Level of Noise | Average Error | Average Standard Deviation |
|----------------|---------------|----------------------------|
| 10 | 1.04 | 0.40 |
| 20 | 1.19 | 0.48 |
| 30 | 1.58 | 0.80 |
| 40 | 1.85 | 0.90 |

TABLE 6.5: Average rms error (chamfer distance) for the data sets used in the experiments using two eigenvalue models.

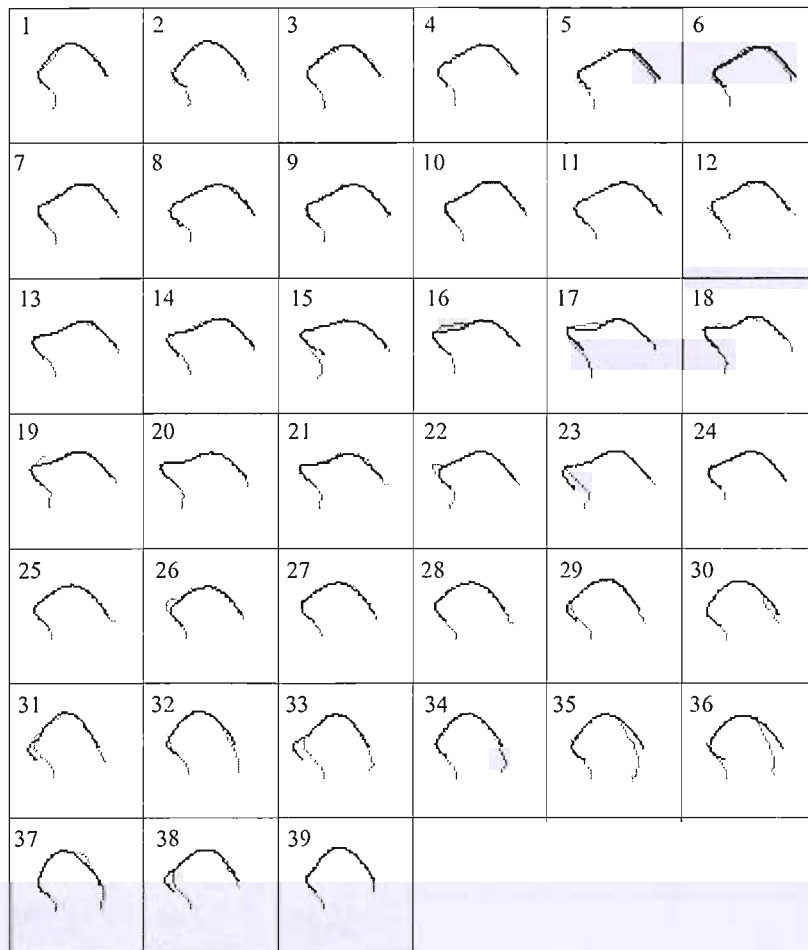


FIGURE 6.20: Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 10% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image.

6.7 Extending the Model to Lips

A model for the lips was generated from the original training data set of 39 frames as described in Chapter 4. The set of training lips, shown in Figure 6.28, was defined by a consistent set of 67 landmark points for each shape. The most significant modes of variation for this model are defined in Table 6.6. As can be observed, the first mode of variation for this model is not as significant as the first mode of variation of the tongue model, as the first one covers only the 47.77% of the variation. This means that seven eigenvalues should be used in the model to cover the same variation covered for the tongue model in previous experiments.

Tests were performed over the ground-truth data set for the lips using one and two eigenvalue

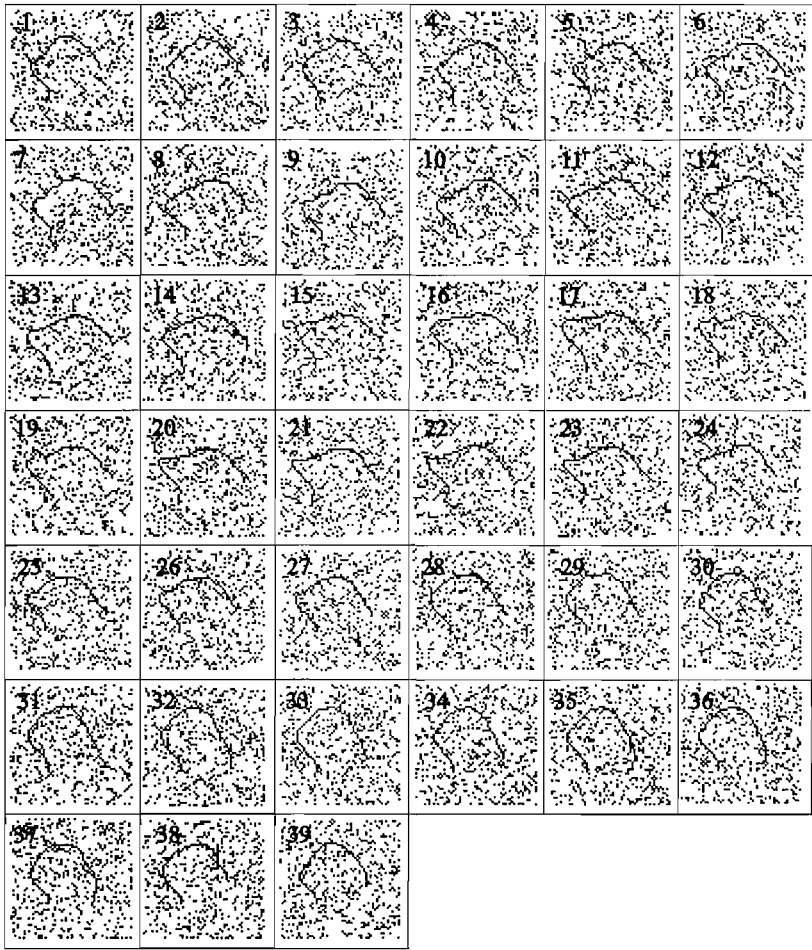


FIGURE 6.21: Training data set sequence corrupted with 20% of noise.

| λ | Eigenvalue | Percentage | Accumulated percentage |
|-------------|------------|------------|------------------------|
| λ_1 | 40.8872 | 47.77 | 47.77 |
| λ_2 | 21.1497 | 24.71 | 72.48 |
| λ_3 | 3.0984 | 9.33 | 81.81 |
| λ_4 | 1.6225 | 3.62 | 85.43 |
| λ_5 | 1.3619 | 1.90 | 87.33 |
| λ_6 | 1.1794 | 1.59 | 88.92 |
| λ_7 | 0.9625 | 1.38 | 90.30 |

TABLE 6.6: Eigenvalues of the covariance matrix derived from the aligned lip shapes.

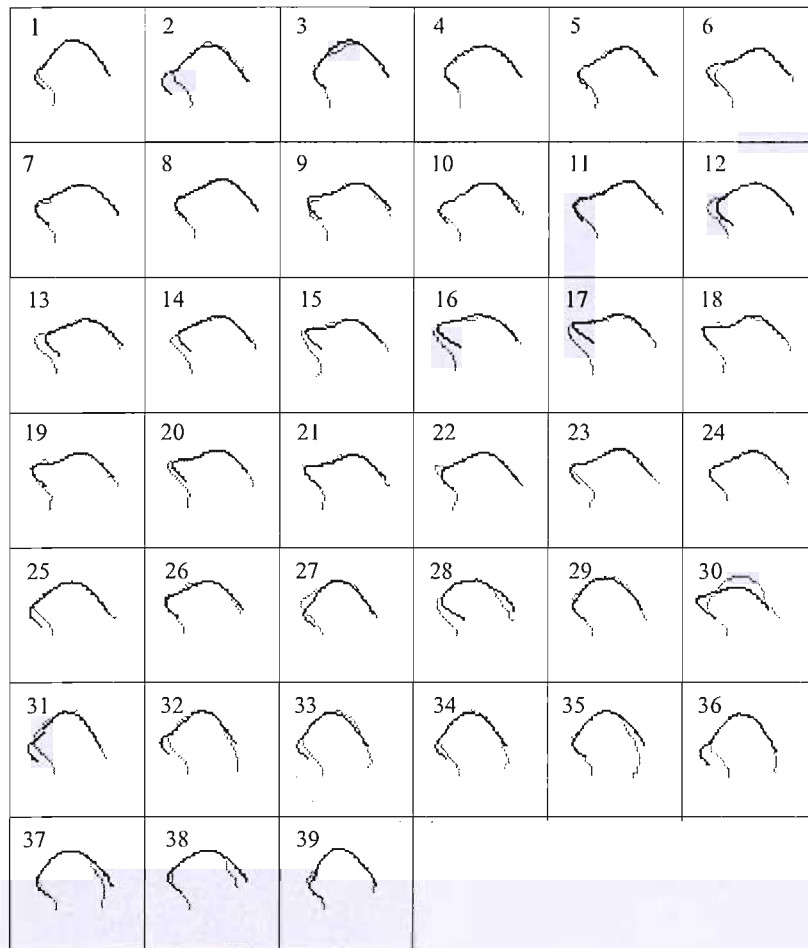


FIGURE 6.22: Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 20% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image.

models, the results are presented in Figures 6.29 and 6.30. From visual inspection it can be appreciated that the two eigenvalue lip model fits better the lips than using one eigenvalue. In terms of accurate modelling of the vocal tract configuration the results for the two eigenvalue model are better; as can be observed in frames 1 to 3 and 37 to 39 the lips are configured to pronounce the phoneme /p/, such a configuration cannot be appreciated in the corresponding frames of the results for the one eigenvalue model. The same tests were performed over full edge images of the ground-truth data set. Results are presented for one eigenvalue model in Figure 6.31.

The tongue model was extended in order to evaluate how the algorithm works with a model of more than one articulator. Information for the lips from the original set of 39 frames was

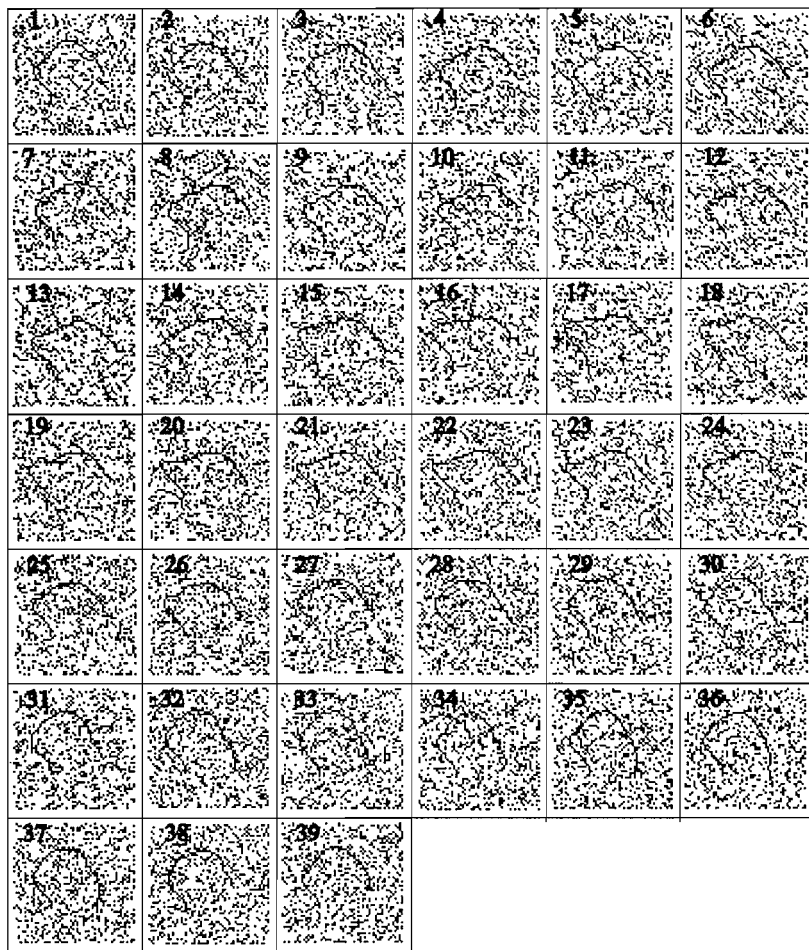


FIGURE 6.23: Training data set sequence corrupted with 30% of noise.

included and the set of training tongue and lip shapes is shown in Figure 6.32. The model was defined as described previously and the corresponding set of modes of variation for this model are shown in Table 6.7. As can be observed the variation of the shape increases, so that in order to cover the 90% of the variation at least 6 eigenvalues must be used. Basically, this increases the size of the accumulator space in the ASDHT algorithm. However, experiments were conducted to test the effectiveness of the algorithm for such articulators using a one eigenvalue model. Results of testing the lip and tongue model with one eigenvalue over ground-truth data is presented in Figure 6.33 with an error calculated of 2.48. In general the shape was very well located and the tongue is fitted better than the lips and the lower lip achieved better matches than the upper one. The same test were performed over unseen data and the error calculated for this sequences is 2.25. The corresponding results are presented in Figure 6.34. Surprisingly, the error is less than the one for the ground-truth full edge

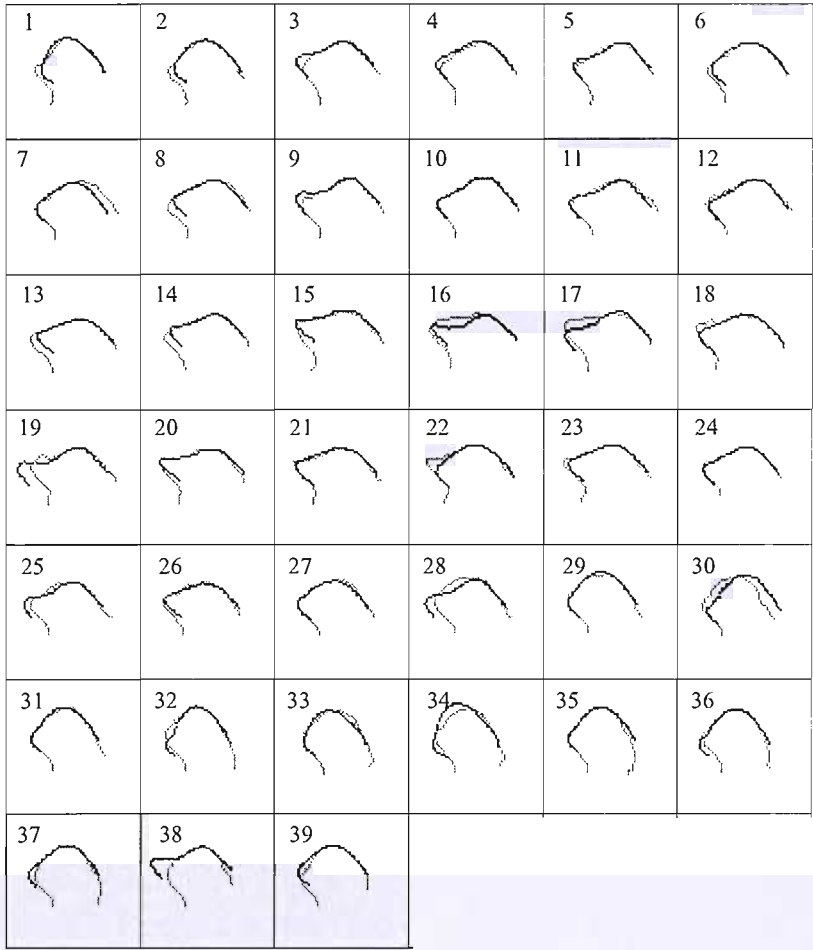


FIGURE 6.24: Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 30% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image.

images. In these results can be observed that the tongue shape is not very well fitted in frames 1 to 6, and 9 to 15. The lips appeared smaller and better fits were obtained for the lower lip than the upper one.

| λ | Eigenvalue | Percentage | Accumulated percentage |
|-------------|------------|------------|------------------------|
| λ_1 | 209.5848 | 53.13 | 53.13 |
| λ_2 | 59.3508 | 15.05 | 68.18 |
| λ_3 | 42.5836 | 10.79 | 78.97 |
| λ_4 | 23.4840 | 5.95 | 84.92 |
| λ_5 | 11.5390 | 2.93 | 87.85 |
| λ_6 | 8.0744 | 2.05 | 89.9 |

TABLE 6.7: Eigenvalues of the covariance matrix derived from the aligned tongue and lip shapes.

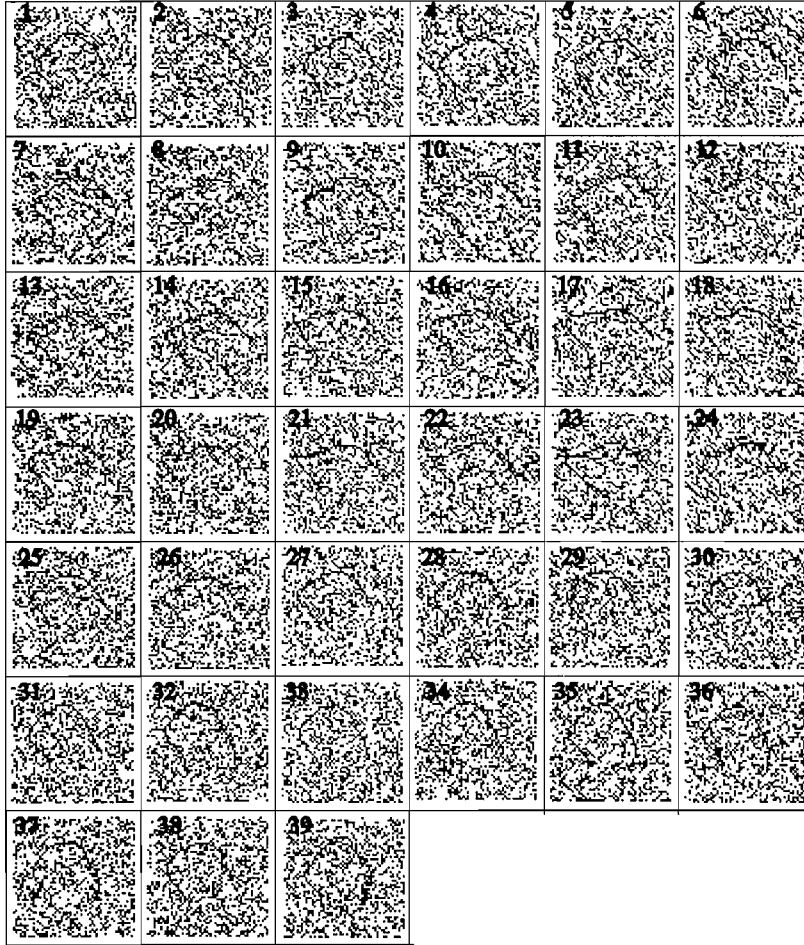


FIGURE 6.25: Training data set sequence corrupted with 40% of noise.

6.8 Summary

In order to measure how successful the ASDHT algorithm is, a chamfer distance metric was used. In this way, the comparison should be made in the image plane, comparing how close the estimated shape is to the tongue shape. To measure the effectiveness of the algorithms, a chamfer distance metric was used (Borgefors, 1988).

Initial exploratory work with the new algorithm indicated that using just one eigenvalue was sufficient to locate an excellent approximation of the tongue shape and position. Obviously, the computation with just one eigenvalue is much faster than the application of the two-eigenvalue model. To speed computation, the searching region for the two-eigenvalue model was restricted to (64×64) windows and the range of variation of the first eigenvalue was

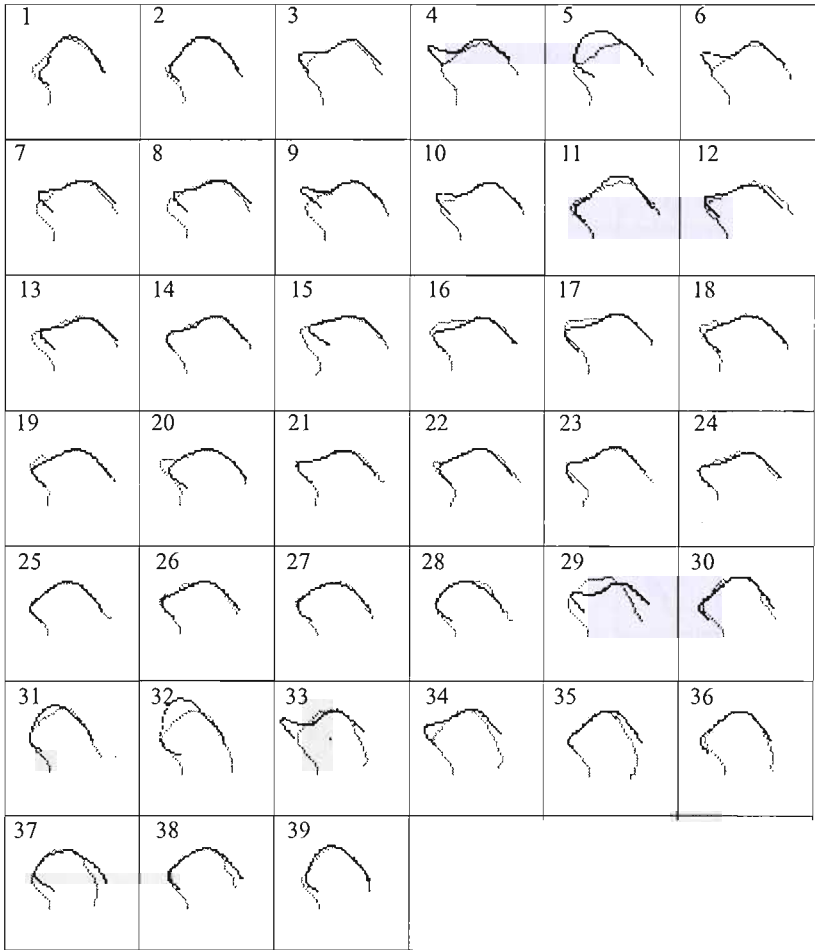


FIGURE 6.26: Results obtained using a two eigenvalue tongue model in a data set sequence corrupted with 40% of noise. Fitted tongue shapes (black) is placed over the corresponding clean shape image.

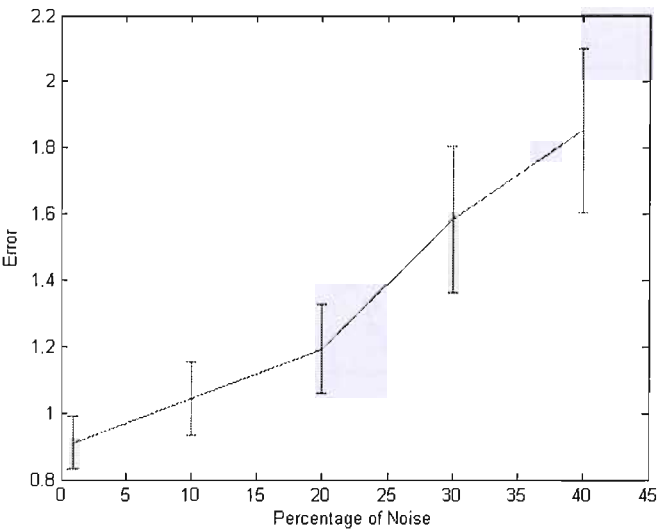


FIGURE 6.27: Results of the algorithm using contaminated images with noise.

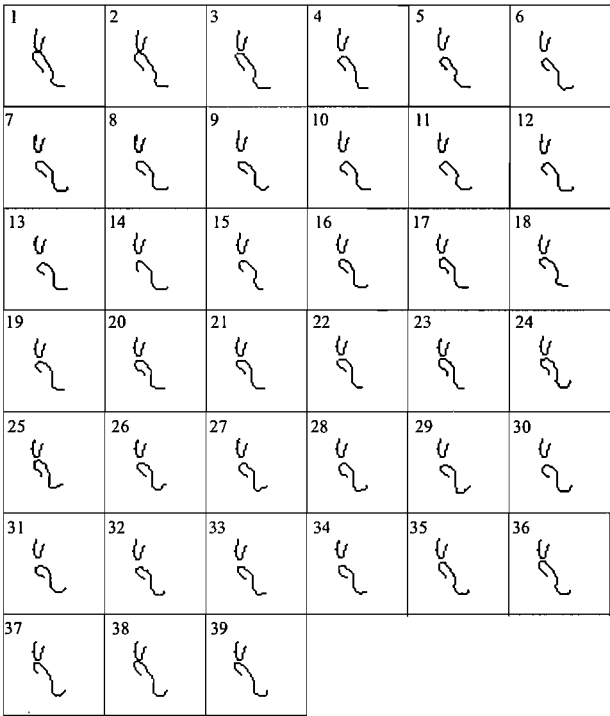


FIGURE 6.28: Set of training lip shapes, extracted from the sequence of 39 images outlined by Mohammad.

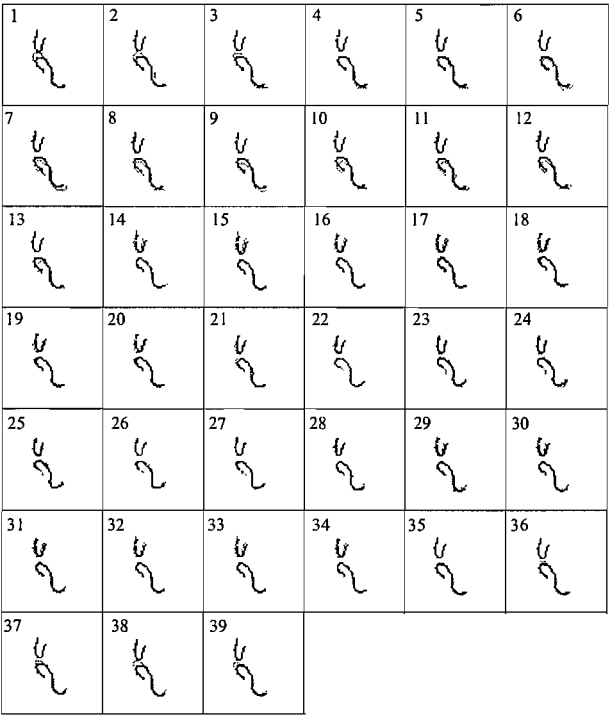


FIGURE 6.29: Results obtained (black) using a one eigenvalue model for the lips over isolated lip images (grey) of the training data set.

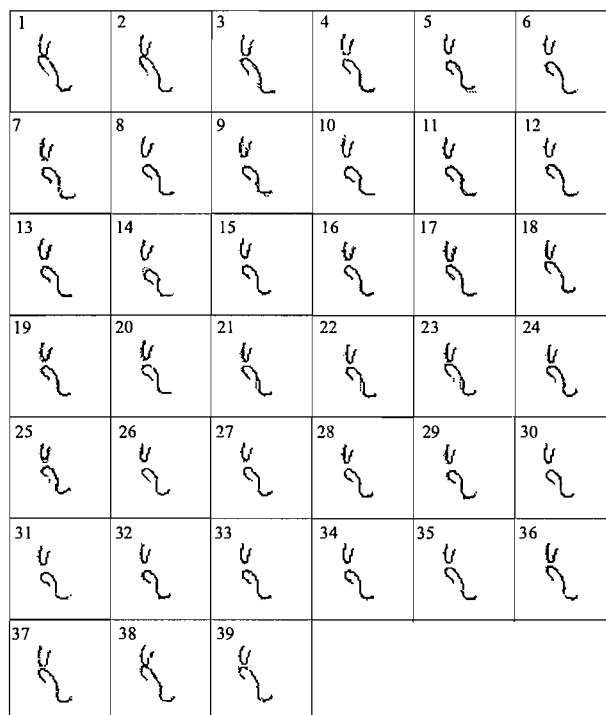


FIGURE 6.30: Results obtained (black) using a two eigenvalue model for the lips over isolated lip images (grey) of the training data set.

restricted as well to one step before and after the corresponding value estimated in the first search.

Experiments conducted using the ground-truth data set shown that the average rms error is typically less than 1 pixel per edge point. Results are generally better when two eigenvalues are used to construct the model. The scaling factor did not vary too much from 1.

On the other hand, experimental results achieved on unseen data are encouraged by the fact that there is only minor deterioration in performance relative to the seen data. This opens up the possibility of training the ASM on limited data from just one speaker and then applying the model successfully to a range of other speakers.

In summary, the average error for each data set is given in Table 6.8. In general, the use of two eigenvalues is better than using just one and testing on the training data gives the lowest error but this is not the situation in practice. However, the performance on unseen data (from a different subject) is not dramatically worse. Using synthetic data gives us a

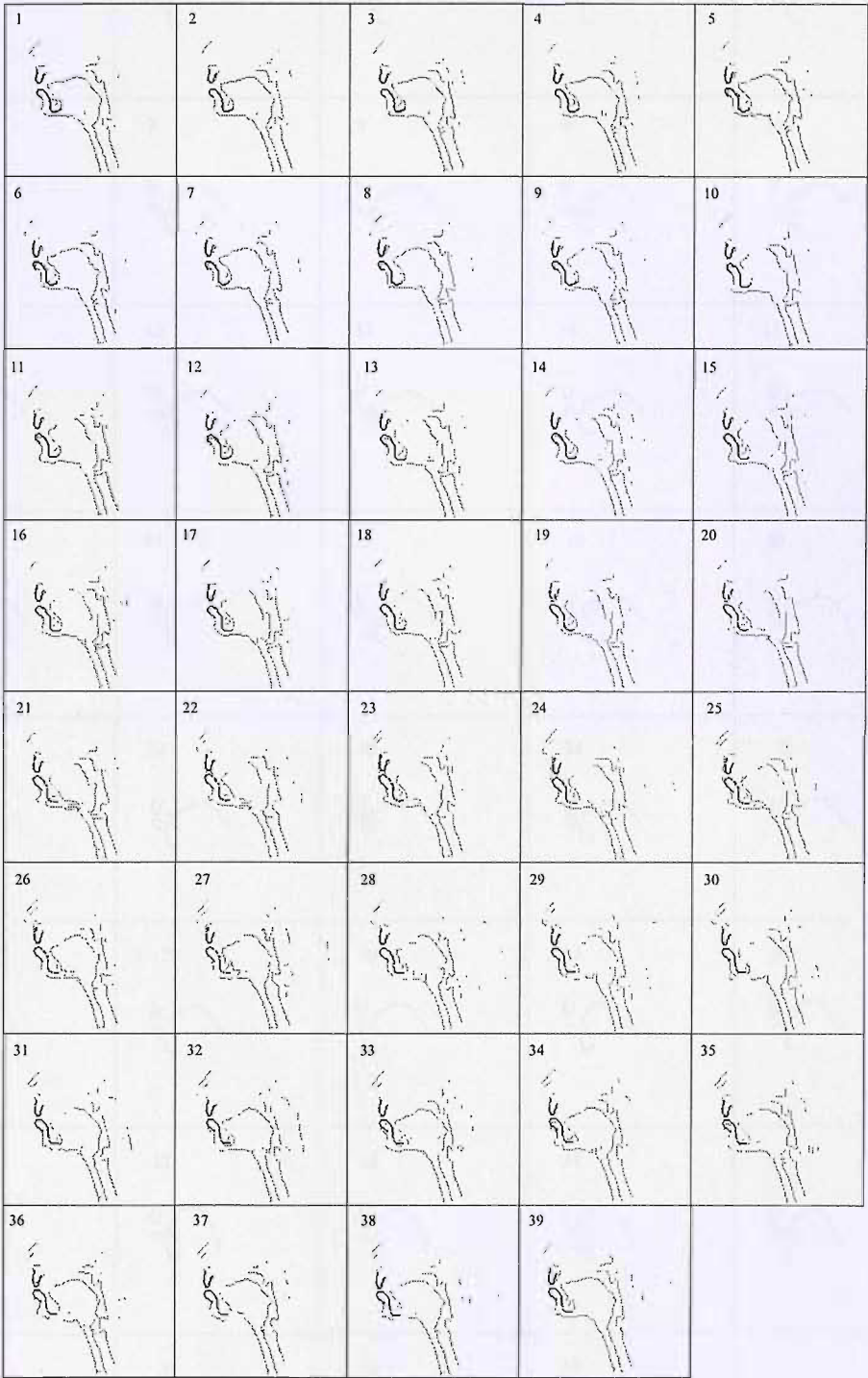


FIGURE 6.31: Results obtained (black) using a one eigenvalue model for the lips over full edge images (grey) of the training data set.

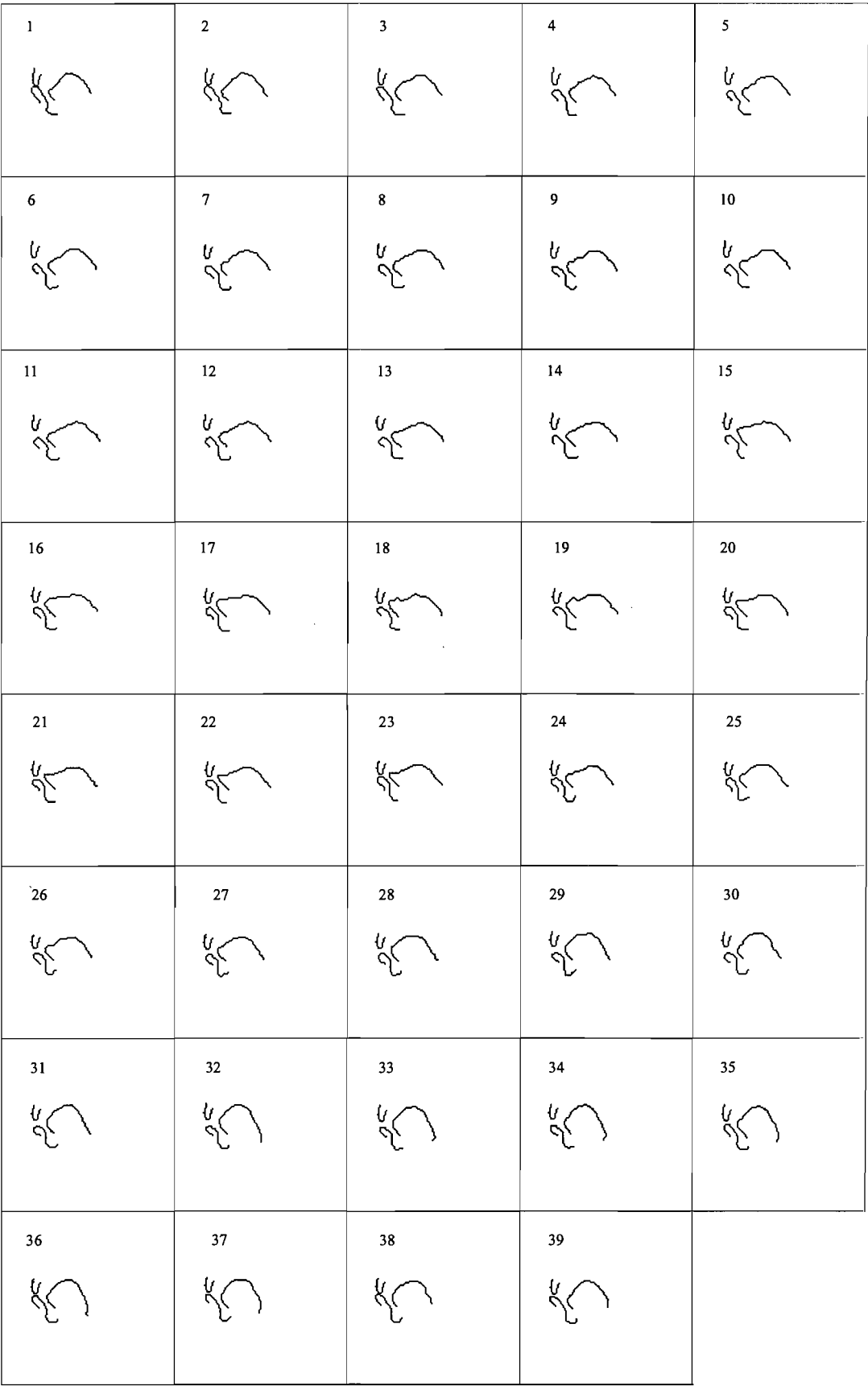


FIGURE 6.32: Lips and tongue training data set formed by 39 shapes extracted from the labelled data set of Mohammad.

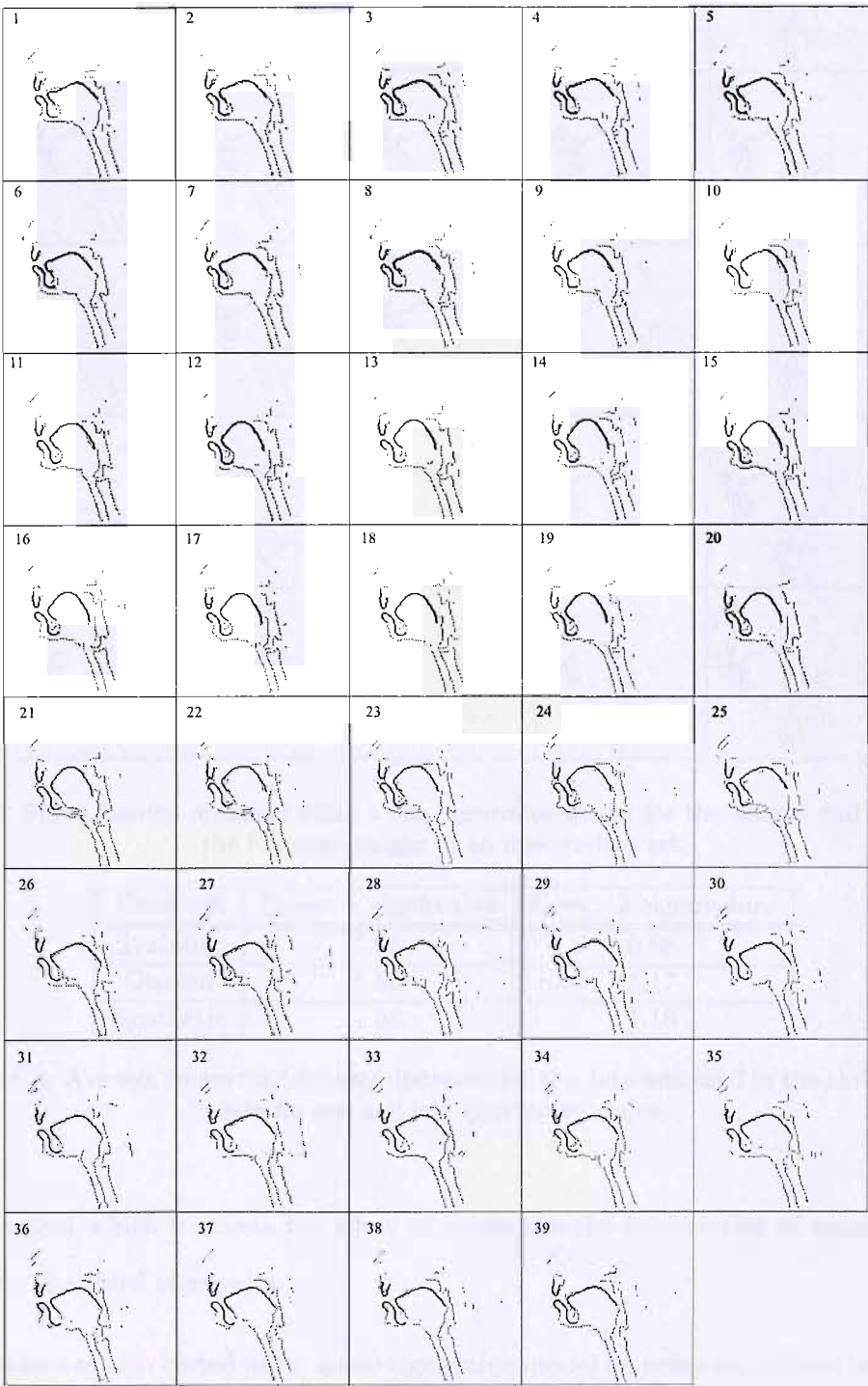


FIGURE 6.33: Results obtained using a one eigenvalue model for the tongue and lips over the full edge images of the training data set.

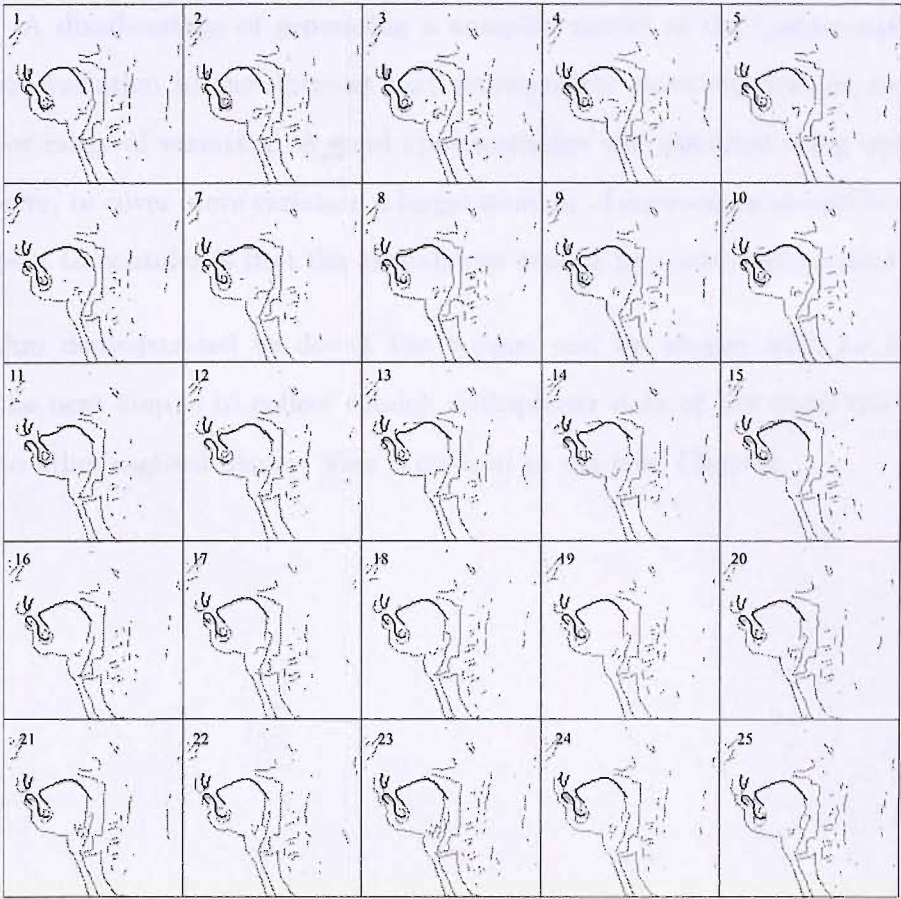


FIGURE 6.34: Results obtained using a one eigenvalue model for the tongue and lips over the full edge images of an unseen data set.

| Data set | Error: 1 eigenvalue | Error: 2 eigenvalues |
|-----------|---------------------|----------------------|
| Training | 1.05 | 0.88 |
| Unseen | 1.60 | 1.17 |
| Synthetic | 1.58 | 1.16 |

TABLE 6.8: Average rms error (chamfer distance) for the data sets used in the three sets of data for one and two eigenvalue models.

reference against which to assess the effect of issues like the subjectivity of manual labelling and ignoring the third eigenvalue.

The algorithm was also tested using a two eigenvalue model on noisy conditions by corrupting 10, 20, 30 and 40% of the images. Although results were presenting major deterioration for some frames with a 30% of the noise, some were still well fitted.

The model was extended to test the performance of the algorithm over more that one

articulator. A disadvantage of generating a complex model of the tongue and lip shapes involves more variation in the data set and consequently more eigenvalues are needed to cover a major range of variation. A good approximation was obtained using one eigenvalue model, however, to cover more variation a larger number of eigenvalues should be considered. Another aspect to consider is that the articulators cannot be scaled independently.

The algorithm demonstrated to detect the tongue and lip shapes with no initialisation required. The next step is to collect enough multiplanar data of the vocal tract to extend this model to other sagittal planes. This is covered in the next Chapter.

Chapter 7

Collecting Data in Halifax

7.1 Introduction

The application of the ASDHT algorithm has been applied to midsagittal MRI sequences to extract the tongue shape. The extension of the model to more articulators such as the lips was shown to be straightforward. The application of this algorithm to more sagittal planes was desired to describe the 3D vocal tract shape dynamics, however, the set of data inherited from Mohammad include the sequences for three midsagittal planes, which were not enough. The opportunity of collecting multiplanar images using a 4T scanner was offered by the Institute for Biodiagnostics (Atlantic) Neuroimaging Research Laboratory in Halifax Canada. Using that facility, data was collected from four subjects pronouncing the nonsense word /pasa/ in seven sagittal planes. These seven sagittal planes were considered as sufficient information to achieve an appropriate 3D representation of the vocal tract dynamics.

This chapter details this experimental work and the corresponding analysis performed over the collected data. First, the acquisition stage and the parameters used during the scanning sessions are detailed. Then, the analysis of the data is presented followed by their synchronisation. Finally, the image reconstruction is described.

7.2 Data Acquisition

During this stage, three different sets of data were collected: speech recordings, MR images and gating pulses generated by the scanner when each row of the raw matrices is collected. Four subjects were used: The first and second were native female speakers of Canadian English (S01) and Mexican Spanish (S02), while the third and fourth were native male speakers of British English (S03) and French (S04). Images were collected using a Oxford Instruments 4.0 T scanner and the audio was collected from the intercom using a Soundmac card while the gating pulse sequences were collected using an IMAQ PCI-1407 acquisition card using LabView. The data collection is detailed in the following sections.

7.2.1 Speech Recording

The speech recording system was set up as illustrated in Figure 7.1. Inside the magnet room, a fine tube with a diameter of 5 mm was attached to the head coil; it was placed as close as possible to the mouth position. This tube conducted the speech sounds from the mouth to the built in microphone inside the magnet room. Then, the audio data was collected from the intercom of the console using a microphone taped to it. The audio was collected using a Soundmac digital audio card with a sampling frequency of 11025 Hz, 16 bits and mono.

Initially, the token to be repeated during the scanning sessions was /pasi/, in order to be consistent with the set of data generated by Mohammad. However, the range of frequencies permitted by the intercom was interfering with frequency information of the /s/ and /i/ phonemes; speech frequencies above 2.5 KHZ were cut off by the intercom as can be appreciated in Figure 7.2(b). Consequently, relevant information, defined above this limit, for phonemes /s/ and /i/ was not present in the spectrograms of the preparatory recordings.

In the other hand, the audio gain was carefully adjusted to avoid clipping effects as illustrated in Figure 7.2; as can be observed a clipping effect is present in (a) and eliminated in (b) after some adjustments made on the recording system. Frequency information contained in the range marked between the points A and B corresponds to the /a/ phoneme; however, information about formant frequencies for /s/ and /i/ phonemes is not available. Hence, the

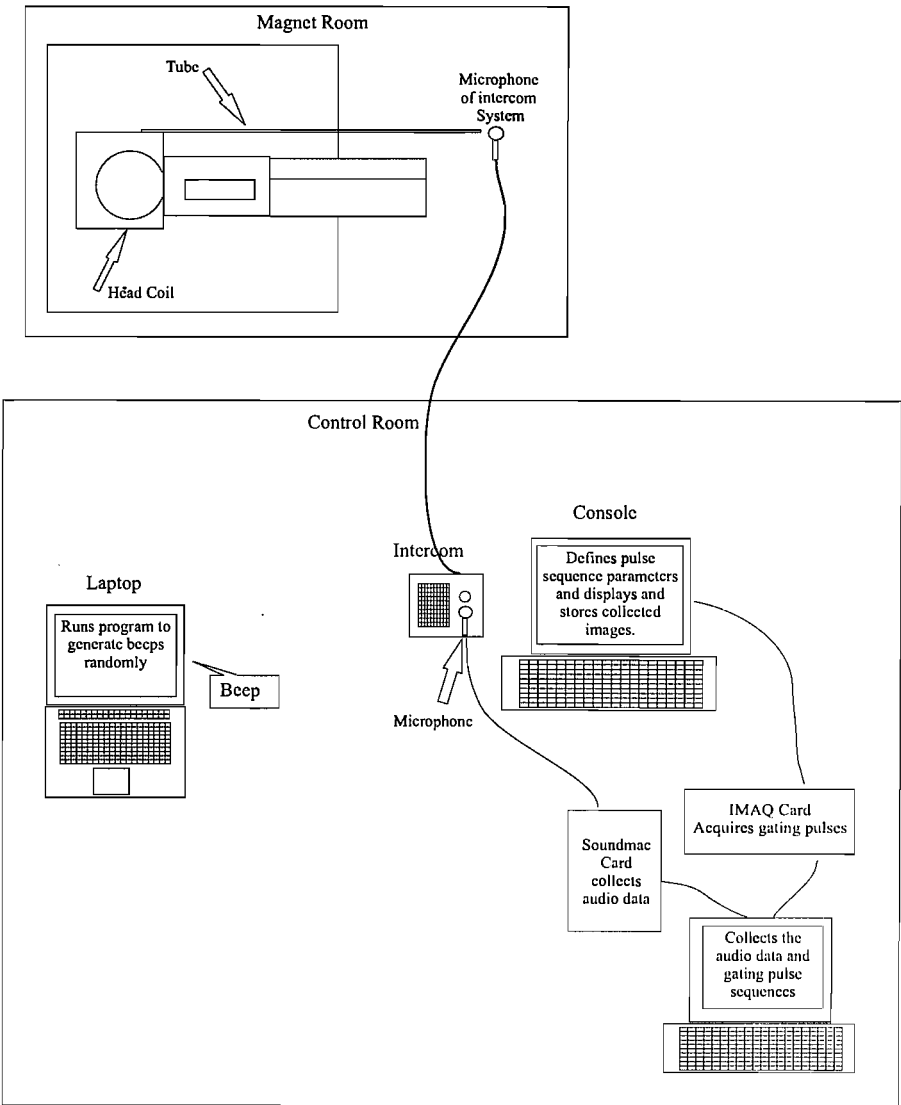
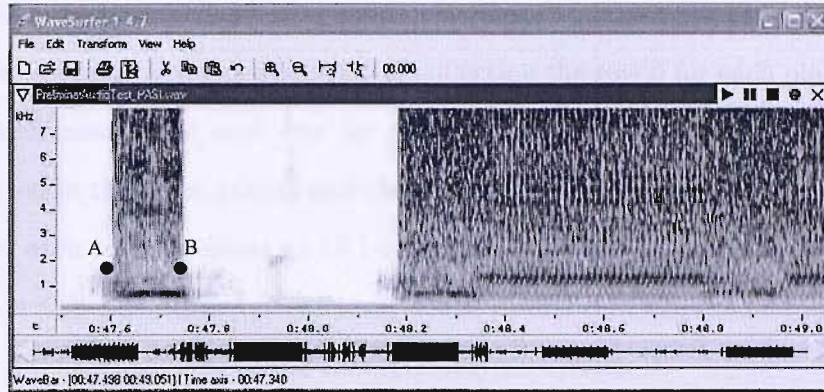
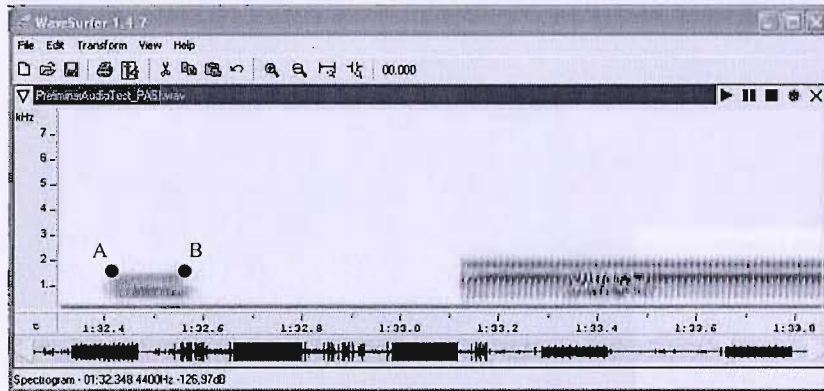


FIGURE 7.1: Audio recording system. A fine tube was attached to the head coil. Speech sounds were conducted to the built in microphone inside the magnet room. The speech was collected for another microphone taped to the intercom built in the console in the control room. A Soundmac audio digital card digitised the audio which was saved using a LabView program. This program, collects the pulse gating sequences generated by the scanner and acquired by an IMAQ acquisition card.

token selected to study in these experiments was /pasa/, since the information of the /a/ phoneme was present and easily detected in the spectrograms. During the experiment, the subject was instructed to repeat the token as regular and constant as possible and to take a breath and continue whenever the air was running out.



(a)



(b)

FIGURE 7.2: Spectrograms of preliminary /pasi/ recordings example. Clipping effects are present in (a) and eliminated in (b). Points A and B enclose the spectral information of /a/; information of /s/ and /i/ is either not sufficient or non-existent for achieving a satisfactory segmentation of the tokens during the synchronisation stage.

7.2.2 Image Acquisition

Preparatory scans were conducted in order to define the acquisition parameters and achieve the desired spatial and temporal resolutions. MR images were acquired using an Oxford Instruments 4.0 T scanner and a gradient echo multi-slice pulse sequence with T_R of 120 ms, TE of 4.58 ms and an acquisition time (at) of 1.712 ms. A coil around the head was used

with a FOV defined as $340 \times 240 \text{ mm}$. Thus, all the vocal tract articulators were covered. In Figure 7.3 the set of seven sagittal planes acquired from the subject S02 during a static position is presented. Seven sagittal planes were collected in approximately 11.5 s, with a thickness of 5 mm and a separation of 5 mm. The image acquisition is illustrated in Figure 7.4. As can be appreciated, the scanner may start collecting data at any stage of the pronunciation of a token. The scanning sequence consists of collecting the row 0 for each plane in a T_R time of 120 ms, which means that each row for each plane was collected in 17.14 ms. Then, the row 1 is collected in the seven planes and the row 2 and so on. Although the time within the T_R interval for each row is defined as 17.14 ms, the actual acquisition time is defined by the parameter Δt , which was defined as 1.71 ms, which means that the scanner spends this time acquiring the data and the remaining time possibly processing it.

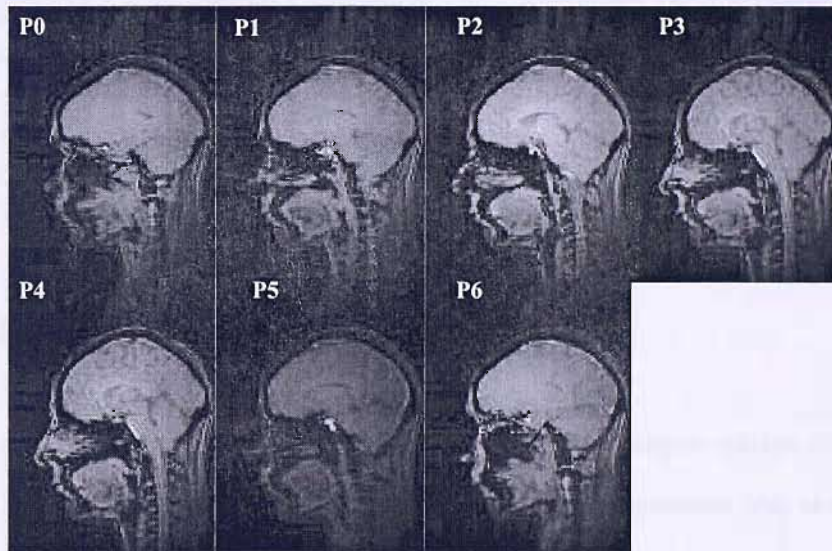


FIGURE 7.3: Seven sagittal planes collected during no speech (static position) for subject 02.

The set of images collected for each subject consisted of non speech images, a sustained /a/, a sustained /s/ and dynamic images. Non speech and sustained images were collected for reference for the reconstructed dynamic images. Table 7.1 indicates the number of scans recorded for each subject.

The scanner generates a timing signal when every row is collected. These gating pulses were recorded using a IMAQ PCI-1407 acquisition card at a sampling rate of 2KHz and saved into text files using LabView; these were used in the synchronisation stage.

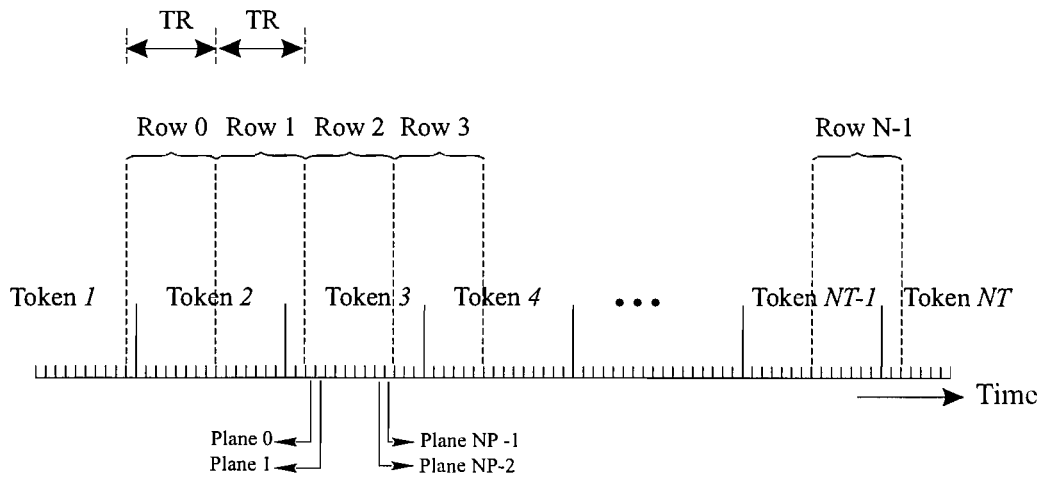


FIGURE 7.4: Acquisition sequence. Each row for each plane is collected at a time. The same row for subsequent planes is collected in a sequential order. NP denotes the number of planes to be collected.

| Subject | Non Speech | Sustained /a/ | Sustained /s/ | Dynamic |
|---------|------------|---------------|---------------|---------|
| S01 | 1 | 1 | 1 | 30 |
| S02 | 2 | 1 | 1 | 40 |
| S03 | 2 | 1 | 1 | 39 |
| S04 | 1 | 2 | 1 | 40 |

TABLE 7.1: Number of scans collected for the experiments conducted.

7.3 Data Analysis

During this stage, the noise in the audio data is reduced and the segmentation of the audio files in individual acquisitions, and subsequently, in tokens and phonemes was analysed. There were two main factors that affected the quality of the audio data, and made complicated the segmentation of it. First, the noise of the scanner, which was reduced using a noise reduction filter; and second, the volume the subject was speaking at, since for some subjects the volume was quite low and the phoneme information was poorer. Therefore, in the analysis of the audio data a special emphasis was made on the data corresponding to the subject 04 since this set was sufficiently clear to achieve a satisfactory phoneme and token segmentation. The text files with the information of the timing pulses were analysed. Although, there was not a written reference to the structure of such sequence, the correspondence with the audio and the image data was explored.

7.3.1 Analysis of the Speech Data

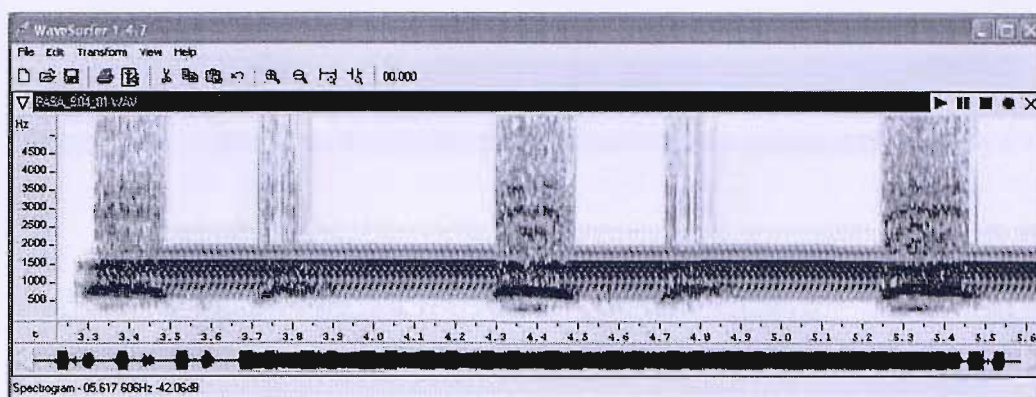
In general, the audio data (recorded in wav format) included more than one MRI acquisition, since the time consumed during the saving process of such files was relatively long within each acquisition. A posterior segmentation was performed to generate an audio file per acquisition. This segmentation was performed using WaveSurfer and PRAAT software to visualise the spectral information of the phonemes and to hear the audio files. The number of wav files recorded per subject is specified in Table 7.2.

| Subject | Number of Files (wav) | Number of Acquisitions |
|---------|-----------------------|------------------------|
| S01 | 6 | 27 |
| S02 | 2 | 40 |
| S03 | 2 | 34 |
| S04 | 2 | 40 |

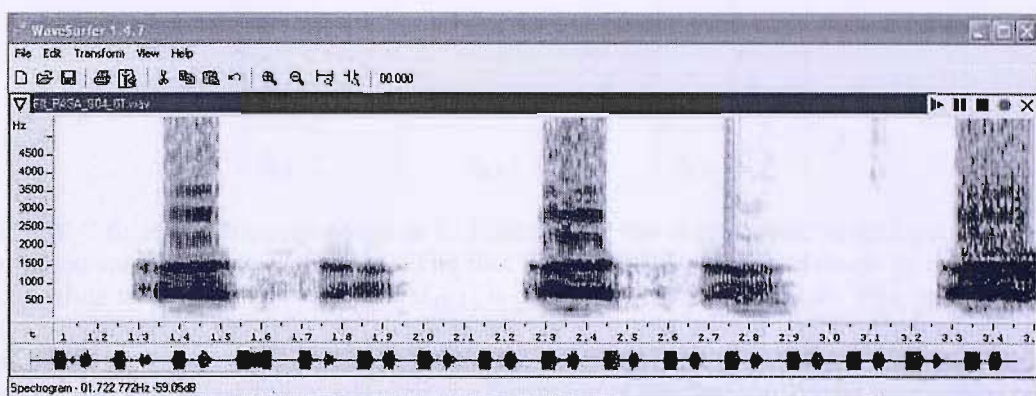
TABLE 7.2: Number of wav files collected for each subject with the number of scan on them.

The audio files were filtered using a speech enhancement algorithm with minimum mean squared error and minimum statistics noise estimation (Ephraim and Malah, 1984; Martin, 2001). Since this algorithm works analysing the data statistically the part of the audio file which contains the initial burst of noise from the scanner was not cleaned as the remaining parts of the audio. Therefore, the filter was applied forward and backwards. Results, as shown in Figure 7.5, were acceptable to measure phone and token durations, since the spectral information for the /a/ phonemes was clearly presented.

Ideally, phone durations should be measured as shown in Figure 7.6; a preliminary audio file of the subject S04 is used to illustrate the process. This file does not contain noise and the spectral information of the phonemes can be clearly appreciated. Since information for /p/ and /s/ phonemes is not available in the collected audio data, the duration of /s/, denoted by $d_{s,1}$, was inferred as the interval [b,c] defined between the first /a/ with duration $d_{a1,1}$, and the second /a/, $d_{a2,1}$. Similarly, the apparent duration of /p/, $d_{p,2}$, for the second token, should be inferred as the interval [d,e] defined between the final /a/ of the first token with duration $d_{a2,1}$, and the first /a/ of the next one with duration $d_{a1,2}$. The term *apparent* is used because real duration of /p/ is in fact shorter. However, $d_{p,2}$ is relatively long; the



(a) Filtered audio file



(b)

FIGURE 7.5: Parts of an original (a) and a filtered speech file (b) generated for the first experiment of the Subject 04. The filter applied was the minimum mean squared error and minimum statistics noise estimation.

subject is making a small pause before pronouncing the next token. The activity of the articulators is unpredictable during these pauses, since the subject could be: preparing for the pronunciation of the next token, holding the position of the ultimate phone or breathing. Then, two additional schemes for segmenting the audio information were considered in order to define the activity of the articulators during this pause and which segmentation should be used in this project. In the second segmentation the time to consider in the synchronisation could be determined by taking an arbitrary time, $d_{p1.2}$, posterior to the pronunciation of the ultimate token and a time, $d_{p2.2}$, previous to the next token as illustrated in Figure 7.7. The third approach to segment the audio file is presented in Figure 7.8 where a period of time previous the first /a/ pronunciation in each token is considered as the duration of the /p/.

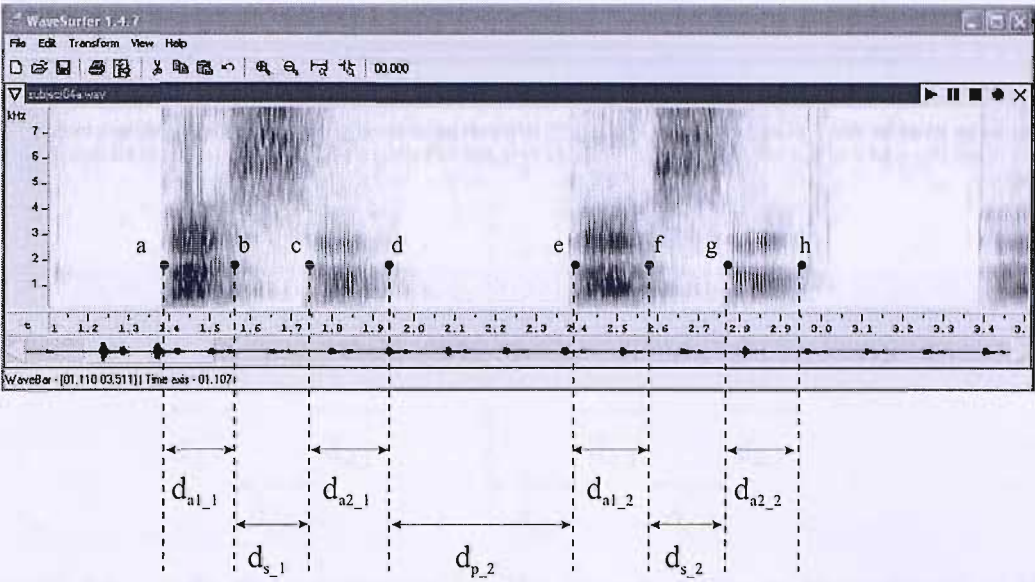


FIGURE 7.6: Audio file segmentation 1. Points mark the starting and ending points of the /a/ phonemes of the token /pasa/. The first /a/ duration, d_{a1_1} , is defined by the interval [a,b], while the second /a/ duration, d_{a2_1} , is defined by the interval [c,d]. The /s/ duration, d_{s_1} , is defined as the interval [b,c] between the first and second /a/’s. Analogously, the apparent /p/ duration, d_{p_2} , should be defined as the interval [d,e] defined between the end second /a/ of the last token, (d), and the beginning of the first /a/ of the next token, (e).

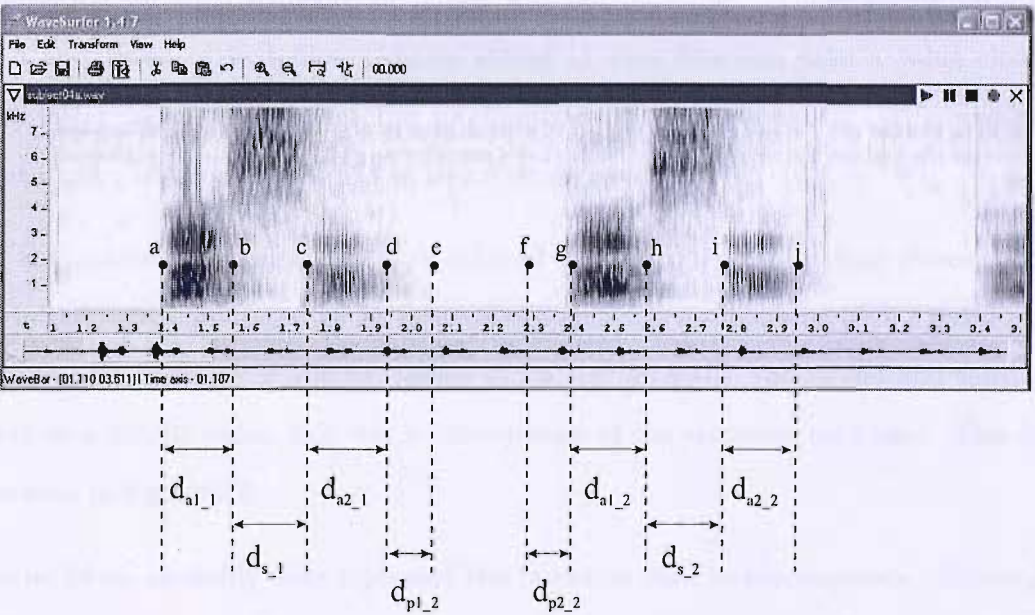


FIGURE 7.7: Audio file segmentation 2. The interval of time to be considered in the synchronisation stage, corresponding to the /p/ articulation, could be composed by two intervals of time: [d,e] posterior to the last token, d_{p1_2} , and [f,g] anterior to the next token, d_{p2_2} .

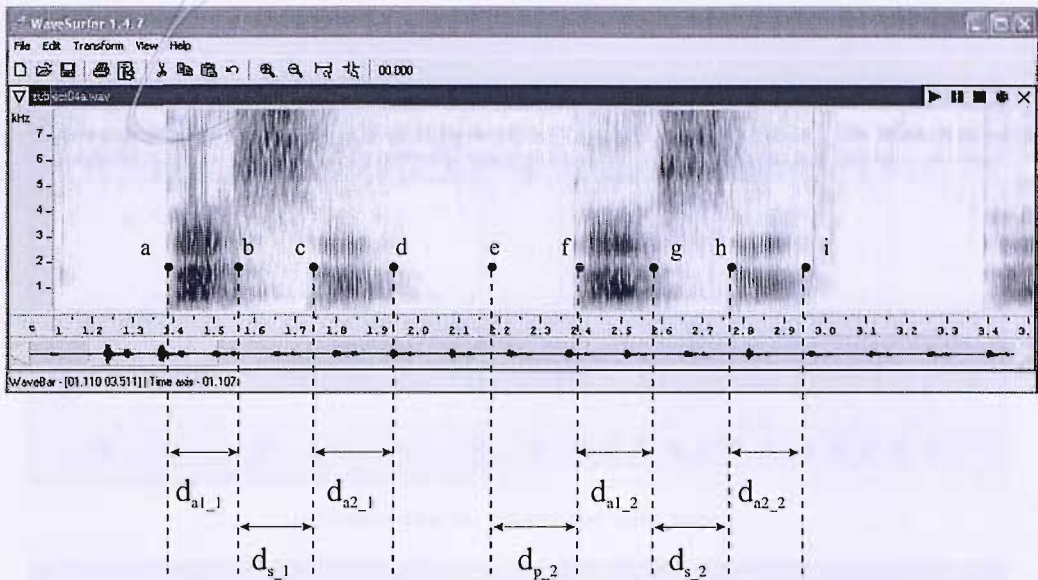


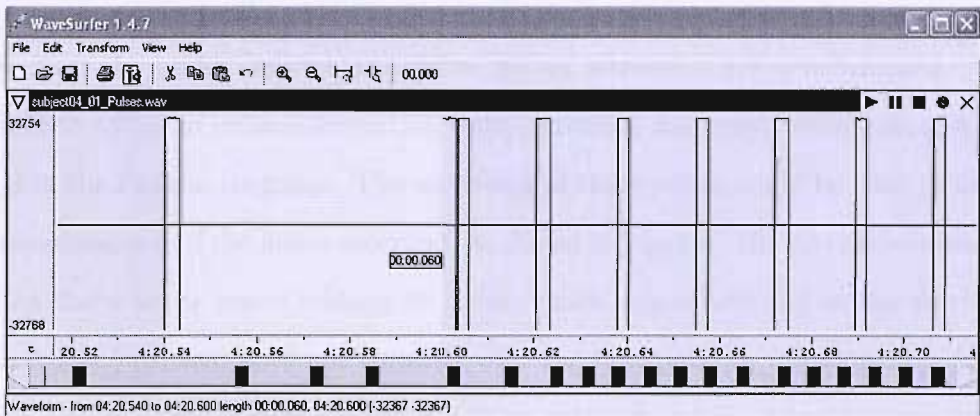
FIGURE 7.8: Audio file segmentation 3. The interval of time to be considered in the synchronisation stage, corresponding to the /p/ articulation, could be composed by the interval of time [e,f] anterior to the next token, $d_{p2.2}$.

7.3.2 Analysis of Gating Pulses

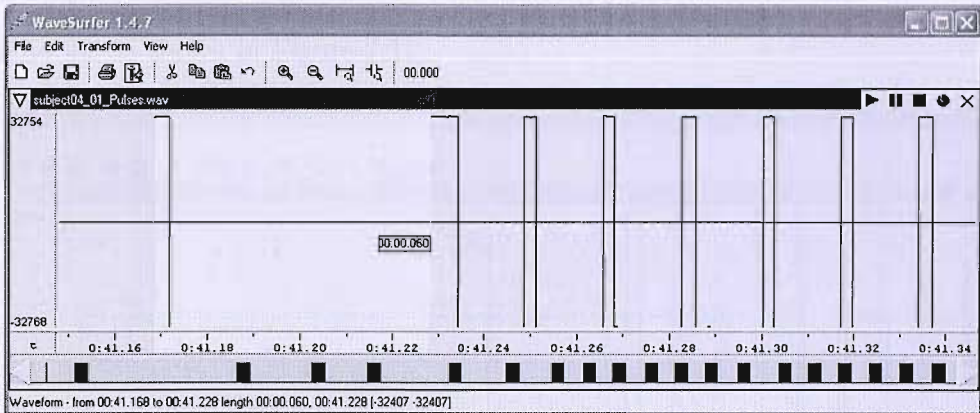
The scanner generates a gating pulse when a single row is collected; however, proper documentation for such sequence of pulses was not available. These gating pulses were acquired and saved in text files. Values stored in such files can take a value close to 0 or 5; the latter describing a pulse. The total number of pulses should be 672, considering 96 rows per image; however, either 673 or 674 were registered.

During the analysis of these files it was inferred that a sequence of gating pulses is defined by two initial pulses separated by an interval of 60 ms and a succession of 672 pulses equally distributed with intervals of approximately 17 ms. Occasionally, the second and third pulses appeared as a double pulse; this was a consequence of the sampling rate used. This format is illustrated in Figure 7.9.

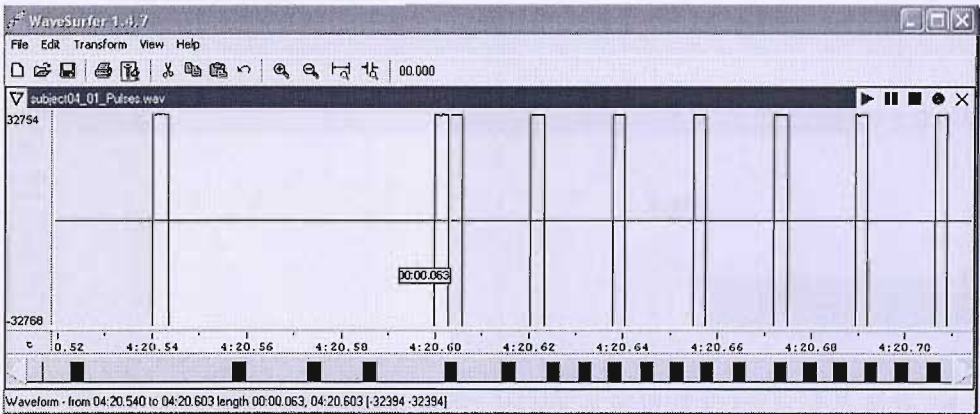
The initial 60 ms probably does represent the inversion time in the sequence. Although the audio files were segmented for each acquisition, the corresponding synchronisation with the gating pulse sequences was performed over the original audio files.



(a) Initial interval. Second and third pulses.



(b) Initial interval. Second and third pulses as a double pulse.



(c) Initial interval. Interval of 63 s before the 672 gating pulses.

FIGURE 7.9: Format of the gating pulse sequence. First and second pulse are separated by a time interval of 60 ms. Some files presented second and third pulses (a) as a double pulse (b), due to the sampling frequency used. In general, two prepulses are generated 63 ms before the 672 pulses corresponding to each row in the K-space matrix.

7.4 Synchronisation

In this section the audio, images and pulse gating sequences are synchronised. The text files generated with the pulse information were converted into wav format, using a program developed in the Python language. The waveform of these pulses could be used to compare it with the spectrogram of the audio recorded, as shown in Figure 7.10. As can be observed, the point when the scanner starts making its noise, which was considered as the starting point for the image acquisition, does not coincide with the first pulse in the sequence; an interval of 77 ms is observed. This interval, which will be referred as time delay, increases along the audio file within each scan. These delays were measured and plotted against their occurrence in the audio file as shown in Figure 7.11.

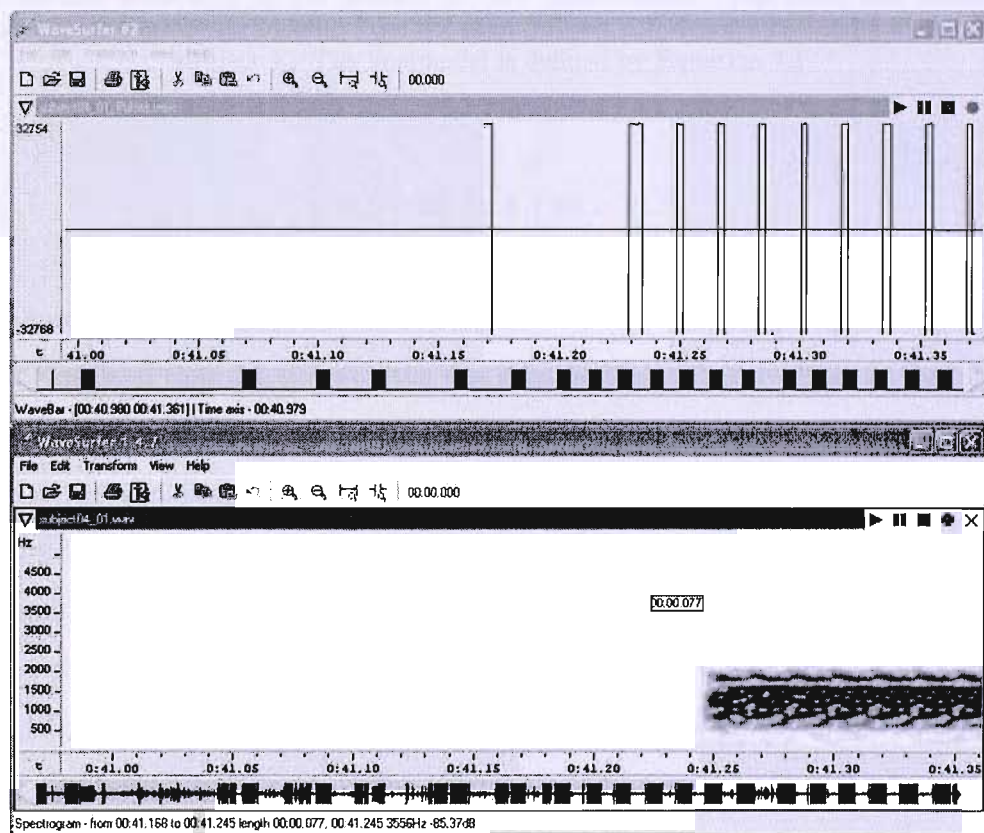


FIGURE 7.10: Illustration of the pulse delay between the first pulse in the sequence and the point when the scanner start the acquisition.

Applying a linear regression model to these data the model is represented as a straight line determined by

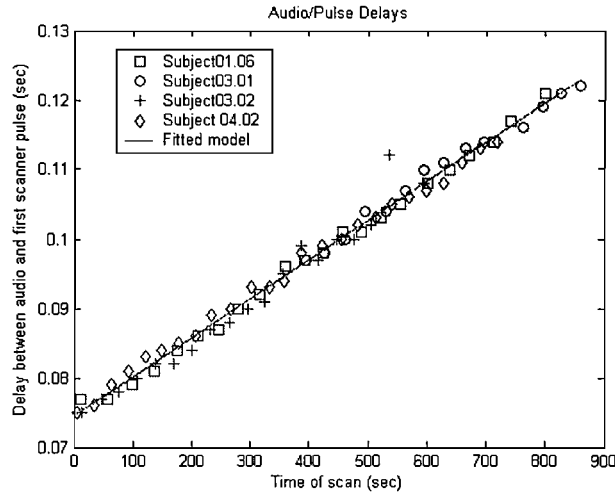


FIGURE 7.11: Graph of the audio pulse delays. It shows the delay between the audio data and the first pulse sent by the scanner in each measurement. The audio files plotted in the graph were selected randomly; the notation SubjectY.X denotes the audio file X of the subject Y. The line model is defined by Equation 7.1.

$$y = 1 \times 10^{-4}x + 7.45 \times 10^{-2} \quad (7.1)$$

where x represents the time (sec) of the sample and y the resulting delay (sec) on the audio file. It was assumed that the slope on the line is caused by a slightly different sampling rate. The initial offset of 74.5 ms could be caused by one or more of the following aspects: a delay between the row acquisition and the pulse generation by the scanner; the inversion recovery time involved in the acquisition pulse sequence, which was defined as 60 ms; and, a delay due to the velocity of sound. If the inversion recovery time is considered within the initial offset, and considering that the first row is collected 63 ms after the initial pulse, then a remaining delay of 11.5 ms still need to be explained. It can be caused by a delay in the transmission of the sound. Different sound transmission intervals must be considered: first the one between the mouth and the microphone inside the magnet room; then the electronic transmission from inside the magnet to the intercom and finally, the transmission between the microphone taped in the intercom to the audio acquisition card. If the speed of the sound is considered to be 343.371 m/s and a distance between the speakers mouth and the microphone in the magnet room is approximately 3 m then the time delay is 8.73 ms.

The next step is to assign a phase value to each row in the raw matrices; this value will represent the moment the row was collected during the pronunciation of the token. Since the T_R was of 120 ms the phase calculation could not be made as reported by Mohammad, where the phase of the first row was calculated and the following row phases were defined as a step increment. Now, the phase calculation must be performed for each row.

The phase assignment is illustrated in Figure 7.12, where the phase is considered to take values between 0 and 100% and steps of 25% are used for each phoneme. The phase of a row will take a value in the range of the phoneme pronounced at the moment it was collected. For example row 0 of the plane 0 was collected at the end of the first token during the pronunciation of the /a/. Then, a value between 75 and 100% will be assigned to it. Subsequently, new matrices with rows with the same phase are generated. As mentioned before matrices will have oversampled and missing rows.

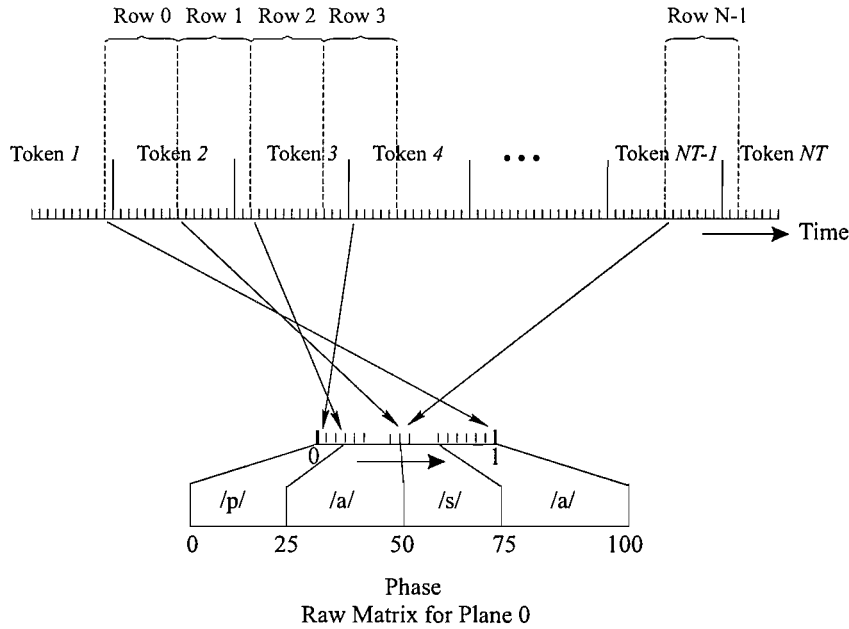


FIGURE 7.12: Synchronisation stage. The phase is defined for each row. The phase does not follow a sequential order due to the acquisition process adopted.

7.5 Reconstruction

The number of frames to be reconstructed, as defined by Mohammad, was defined by the total average duration of the token and the T_R . Actually, in the acquisition sequence used

in his experiments the T_R was the inter-row time, which is the interval of time between the acquisition of two consecutive rows in the same plane. However, the acquisition sequence used in the present work is different. As illustrated in Figure 7.13, the T_R time comprises the collection of a specific row in all the planes collected. Therefore, the time for collecting a single row in a single plane is defined as $T_R/\#Planes$ in this case 17 ms. However, the time that actually the scanner spend collecting the information is defined by the acquisition time (at) parameter, which in this case was defined as 1.712 ms. The number of expected frames N_{EF} was defined by Mohammad as follows:

$$N_{EF} = \overline{D}_{tok}/T_R, \quad (7.2)$$

where \overline{D}_{tok} is the token duration. However, the number of reconstructed frames could depend on: the T_R , the inter-row time or the at parameters. Considering that the average of the complete duration of the token of 1.14s, which includes the time when the subject is making a small pause, then if the T_R parameter is considered in this calculation, the number of expected frames will be approximately 9. However, a sequence with such a number of frames will not provide a good tracking of the vocal tract dynamics since the resulting temporal resolution will be smaller than the velocity of most of the articulators, especially the tongue.

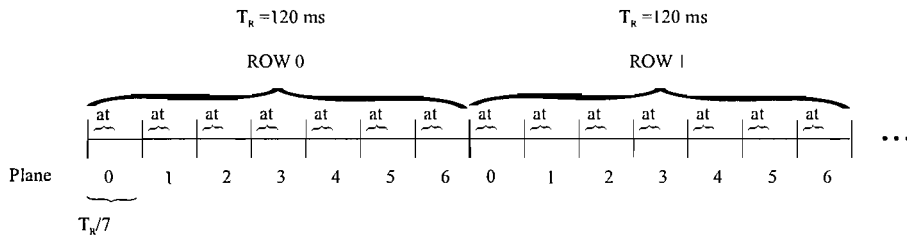


FIGURE 7.13: Acquisition sequence with a T_R of 120 ms and an acquisition time (at) of 1.7 ms. The collection of the corresponding row in the seven planes is done in T_R intervals. The actual acquisition time is defined by the parameter at , the remaining time between rows is assumed to be used for the associated processing of the row.

In the contrary, if the at parameter is considered in this calculation the number of frames becomes much bigger, around 670. Although this number leads to a temporal resolution of 1.71 ms, which should be suitable for these purposes, the number of raw matrices needed to achieve such results must be bigger than 40. The concept of inter-row acquisition time,

which is in this case is 17.14ms, could be then introduced; this leads to a total of expected frames of 67 frames.

The final matrices generated with rows with the same range of phase will have missing and oversampled rows. This problem is overcome by using the same criterion reported by Mohammad (1999) and presented in Appendix B. These criterion involves the average for the oversampled rows and a borrowing one for the missing rows.

7.6 Summary

In this Chapter the details of the data acquisition performed at the Institute for Biodiagnostics (Atlantic) Neuroimaging Research Laboratory in Halifax Canada were presented. Different issues were addressed with the audio recording mechanism and two main problems were mentioned: the noise generated by the scanner and the range of frequencies transmitted by the intercom. The noise generated by the scanner was reduced by applying of the minimum mean squared error and minimum statistics noise estimation filter. In the other hand the limitations imposed by the intercom were sorted out by using a different token for the subject to repeat, so that instead of collecting data for the pronunciation of the token /pasi/, the experiments were performed using the token /pasa/.

In comparison with the experiments conducted by Mohammad, a new set of data was collected. This set corresponds to the timing or gating pulses generated by the scanner when each row of the raw matrices is collected. This pulses were synchronised with the audio data files. However, a delay was found in such synchronisation. Although a proper documentation of this gating pulse sequence was not provided, the delay was assumed to be caused by a slightly different sampling rate of such pulse sequences.

In the other hand, the images were acquired using a pulse sequence which collects the same row for each plane in a T_R interval. Since the acquisition sequence for the experiments used by Mohammad was different, a new scheme to defined the phase for each row is defined. The number of reconstructed images was also defined in a different manner, because the T_R has

a different interpretation. A detailed comparison of the results obtained for the new set of data collected is presented in the following Chapter.

ating Results

Chapter 8

Evaluating Results

8.1 Introduction

The acquisition of data to generate MRI sequences for extracting the vocal tract dynamics in seven sagittal planes was detailed in the previous Chapter. Three sets of data were collected: MR images, audio recordings and gating pulses. The audio data was analysed, segmented and the phoneme and token durations were measured. The gating pulses were analysed and synchronised with the audio data. The synchronisation of this data with the MR images was performed followed by the corresponding image reconstruction.

In advanced it can be said that the results of the dynamic reconstruction obtained from the experiments performed in Halifax were not as good as expected. In spite of this fact, the algorithm ASHT was applied to isolated non-speech and sustained images using the original tongue model developed in Chapter 4. Considerable anatomical differences led the model to fail in finding a correct match for this set of unseen images. Therefore, the model was retrained by including into the training data set the tongue shapes extracted from the non-speech and sustained images obtained from the experiments conducted in Halifax.

In this chapter, first, the results obtained from the reconstruction are presented, followed by the experimental differences compared to the experiments performed by Mohammad (1999) and described in Appendix B. Then, the results obtained from the application of the

ASHT algorithm are presented and the subsequent incorporation of the new non-speech and sustained images in the training data set is detailed followed by the results of the application of the ASHT algorithm to the new set of unseen tongue shapes. Finally, the conclusions are presented.

8.2 Results

The reconstruction of midsagittal sequences for the subject 04 was performed considering three different segmentation schemes. The differences among these schemes rely on the interval considered for the pronunciation of /p/ , since the interval between the last /a/ of a token and the first one of the next token was relatively long. Therefore, the behaviour of the vocal tract articulators might not be consistent since the subject may be preparing for the pronunciation of the next token, breathing or holding the vocal tract articulator position of the last /a/ phoneme of the token. Although just part of this interval of time corresponds to the /p/ pronunciation, for purposes of this work this will be denoted as the duration of /p/ and the corresponding reconstructed frames will be labelled as /p/.

The first audio segmentation scheme, which is illustrated in Figure 8.1, considers the total time of the token pronunciation, since the interval between last /a/ of one token and the first /a/ of the next one is labelled as the /p/ duration. Sequences of 5, 15 and 40 frames were reconstructed and the results are presented in Figures 8.2, 8.3, 8.4 respectively. As can be observed in all the sequences the top part of the images showing the brain and the skull, appears very clear in the images, while the bottom part, comprising the vocal tract articulators, appears blurred. In the sequence of 40 frames, the movement of the vocal tract articulators must appear with smooth transition between the last frame and the first one, as the pronunciation of the token was cyclic.

The second audio segmentation considers as /p/ duration the interval of time before and after the /asa/ pronunciation, as illustrated in Figure 8.5. This consideration is taken in order to cover possible coarticulation effects between the pronunciation of the last /a/ and the following /p/, in case the subject is spending this period of time preparing the articulators for

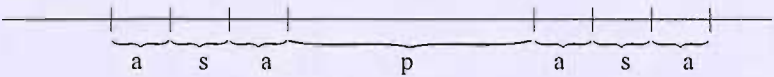


FIGURE 8.1: First audio segmentation scheme which considers the complete token duration in the reconstruction.

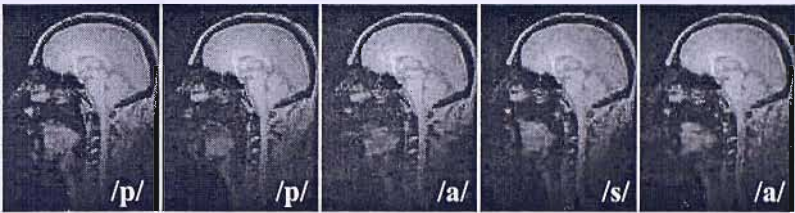


FIGURE 8.2: Sequence of 5 frames reconstructed for the first segmentation scheme. Frames labelled as /p/ are those reconstructed for the interval of time between the last /a/ of the token and the first one.

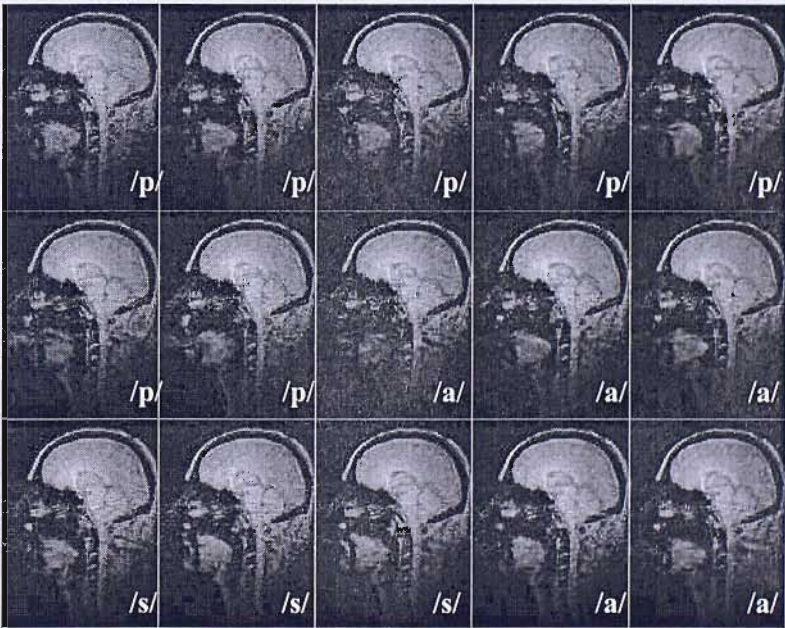


FIGURE 8.3: Sequence of 15 frames reconstructed for the first segmentation scheme. Frames labelled as /p/ are those reconstructed for the interval of time between the last /a/ of the token and the first one.



FIGURE 8.4: Sequence of 40 frames reconstructed for the first segmentation scheme. Frames labelled as /p/ are those reconstructed for the interval of time between the last /a/ of the token and the first one.

the pronunciation of the /p/. Results for intervals of time of 140 ms before and after the /asa/ were calculated for sequences of 5, 15 and 40 frames and presented in Figure 8.6, 8.7, 8.8 respectively. As can be observed, two sets of frames are labelled as /p/ in the beginning and ending of the sequence, which correspond to the 140 ms considered before and after of the /asa/ pronunciation.

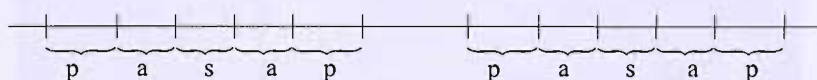


FIGURE 8.5: Second audio segmentation scheme, where two intervals of time labelled as /p/ duration are considered in the reconstruction. These intervals are defined before and after the pronunciation of each /asa/ subtoken.

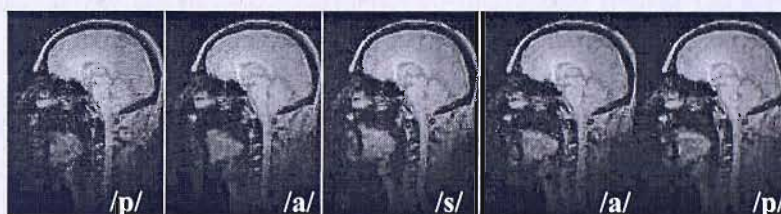


FIGURE 8.6: Sequence of 5 frames reconstructed for the second segmentation scheme. Frames labelled as /p/ are those reconstructed for intervals of time of 140 ms after the last /a/ of a token and before the first /a/ of the next one.

Finally, the third segmentation involves the reconstruction of an interval before the /asa/ pronunciation, as shown in Figure 8.9. Results for the reconstruction of sequences of 5, 15 and 40 frames considering an interval of 240 ms previous to the /asa/ pronunciation are presented in Figures 8.10, 8.11, 8.12. Since an interval of time is discarded during the reconstruction some noticeable changes in the shape of the articulators may be appreciated between the first and last frames of the sequences. The evaluation and analysis of these results are presented in the next section. In all these sequences it can be appreciated that those frames of transition between the /p/ and the /a/ phonemes were the most blurred and noisy within the sequences.

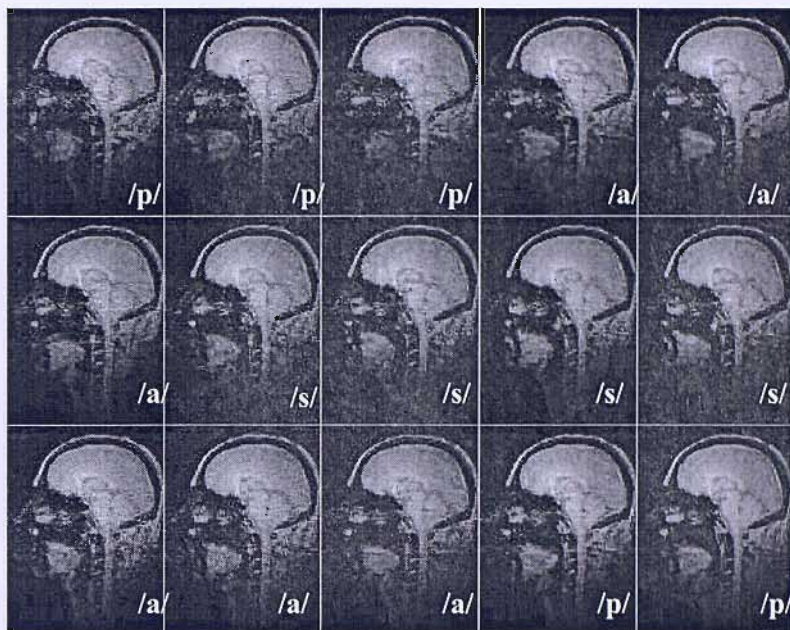


FIGURE 8.7: Sequence of 15 frames reconstructed for the second segmentation scheme. Frames labelled as /p/ are those reconstructed for intervals of time of 140 ms after the last /a/ of a token and before the first /a/ of the next one.

8.3 Experimental Evaluation

The analysis and evaluation of the image quality is based on visual inspection. Although this is a subjective method to evaluate the image quality, differences between frames appear to be sufficiently evident to validate this evaluation criteria. The concept of image quality is used to refer if the shape and the boundary of the vocal tract articulators are sharp, blurred or noisy. An example of this evaluation is presented in Figure 8.13. Here, the static image (Figure 8.13(a)) has the best image quality, since the boundary of the lips and tongue are clearly defined. The image quality for the reconstructed frame (Figure 8.13(b)) is worse than the static one, because the image is noisier, the nose and the tongue boundaries are not very clear and no information of the lower lip is presented. The reconstructed image is noisy due to motion artefacts, missing information problems during the reconstruction, susceptibility problems and probably for problems with the reconstruction because the token was not repeated exactly enough.

If a general visual inspection of the images in the reconstructed sequences is performed, it can be noticed that most of them have a clear boundary and shape definition of the top part



FIGURE 8.8: Sequence of 40 frames reconstructed for the second segmentation scheme. Frames labelled as /p/ are those reconstructed for intervals of time of 140 ms after the last /a/ of a token and before the first /a/ of the next one.

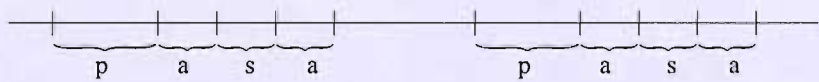


FIGURE 8.9: Third audio segmentation scheme, where the interval labelled as /p/ duration is considered previous to the pronunciation of the /asa/ sub token.

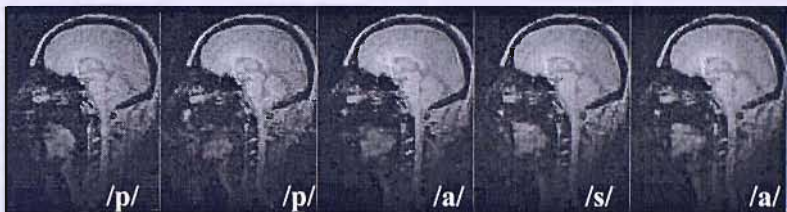


FIGURE 8.10: Sequence of 5 frames reconstructed for the third segmentation scheme. Frames labelled as /p/ are those reconstructed for an interval of time of 240 ms taken before the first /a/ of the each token.

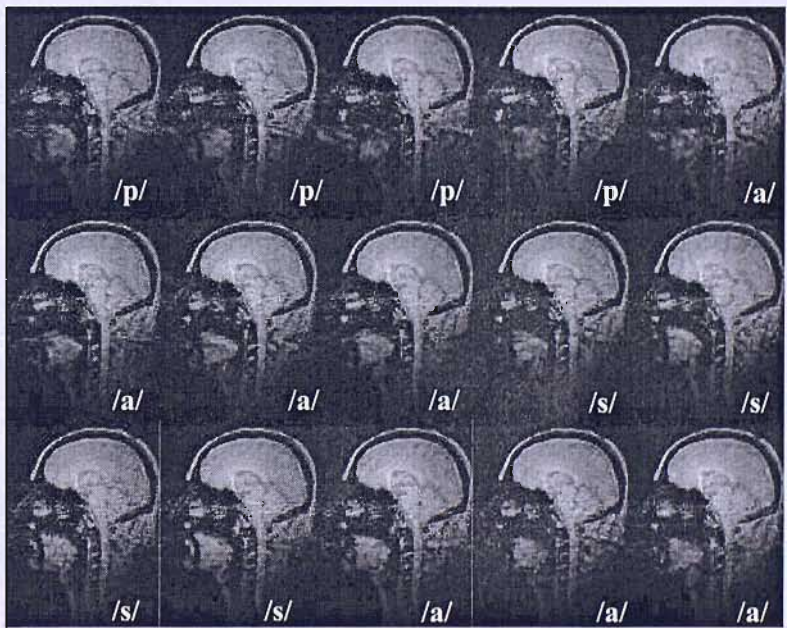


FIGURE 8.11: Sequence of 15 frames reconstructed for the third segmentation scheme. Frames labelled as /p/ are those reconstructed for an interval of time of 240 ms taken before the first /a/ of the each token.

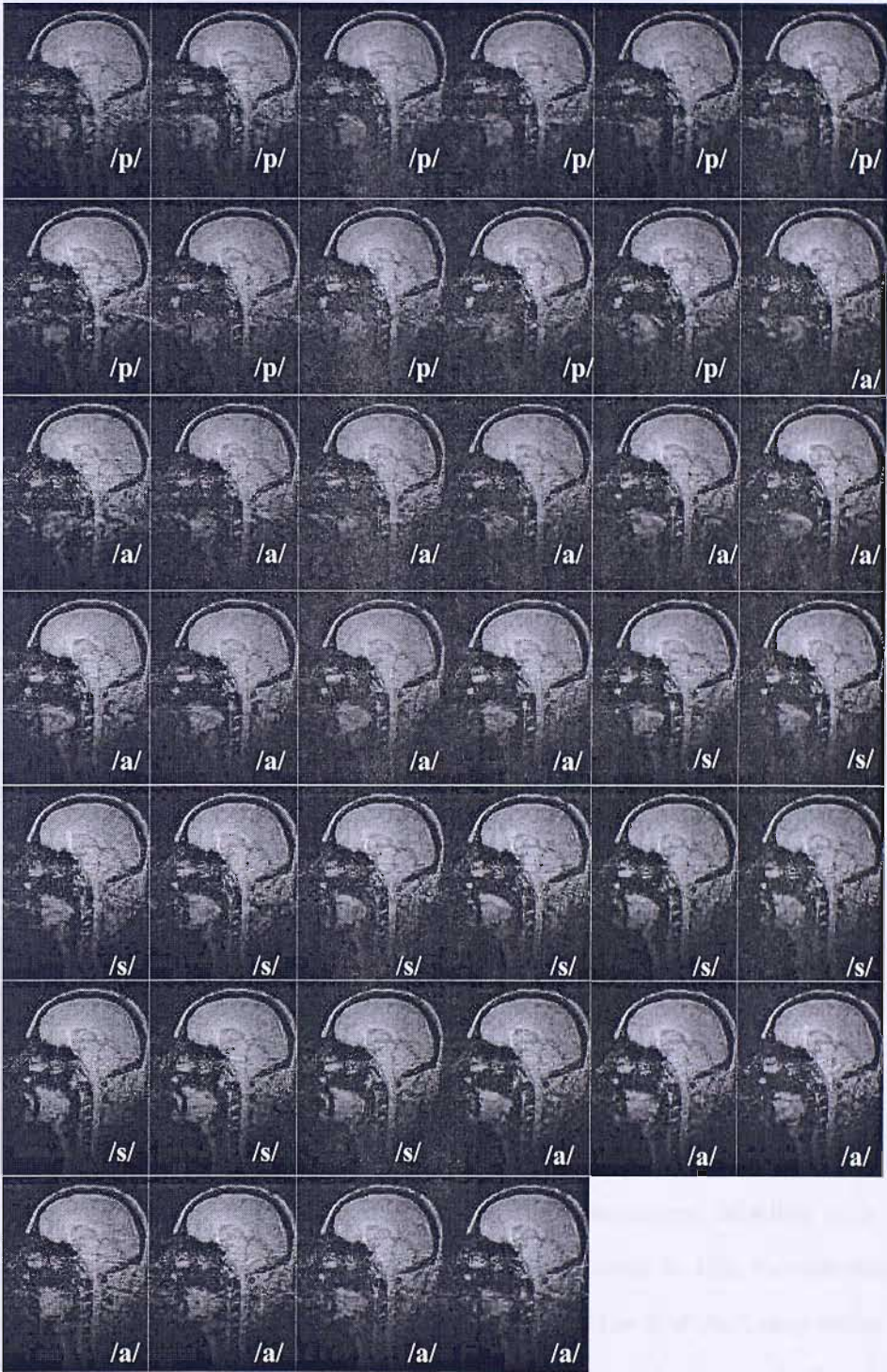


FIGURE 8.12: Sequence of 40 frames reconstructed for the third segmentation scheme. Frames labelled as /p/ are those reconstructed for an interval of time of 240 ms taken before the first /a/ of the each token.

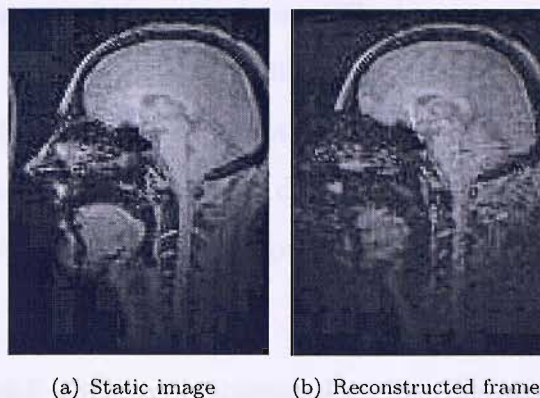


FIGURE 8.13: Static and reconstructed images of the Subject 04 of the Halifax experiments to illustrate noticeable differences in image quality by visual inspection. The reconstructed image is noisy due to motion artefacts, missing information problems during the reconstruction, susceptibility problems and probably for problems with the reconstruction because the token was not repeated exactly enough.

of the head (brain and skull). This is a consequence of the stationary position of this part of the head, which implies that the synchronisation and reconstruction of the images were done correctly. However, in most of the images the bottom part, which comprises the nose, the lips, the tongue, the velum and the jaw appeared blurred and very noisy. This is caused by the movement of the vocal tract articulators during the pronunciation of the token.

In order to compare images that correspond to the same phoneme, it is necessary to describe what the images represent in each scheme. In scheme 1, which comprises the complete token duration, the first set of images marked as /p/ represent the transition from the last /a/ to the next /p/, since the token continuously repeated. The scheme 2 considers two intervals: one before and one after of the first and last /a/ pronunciation of a token. The initial and final frames of these sequences are labelled as /p/, representing the interval before and after the /a/ pronunciation. The last and first images of these sequences, labelled with /p/, may not change smoothly, since there is an interval of time ignored in this reconstruction. The last scheme, which considers an interval of time previous to the first /a/, may suffer from the same transition effect than those of the second scheme due to the interval of time does not considered in the reconstruction.

Images of the sequences for the third scheme appeared to be the most blurred. This is probably caused by the release of the plosive /p/, the airflow caused by this release and

by the consequent movements of the articulators. The burst may generate distortions, such as blurriness and noise, in the images. Additionally, when the missing and oversampled row problem is solved, this noise is drag to other frames in the sequence. This seems not to happen in other sequences, where the set of frames that suffer from such distortions is smaller. However, these distortions always appear when the transition between /p/ and /a/ is happening.

If the middle frame for each phoneme is compared to the corresponding sustained phonemic image presented in Figure 8.17 the tongue shape looks very similar. However, the differences appreciated may be caused by the coarticulation and dynamic effects.

Essential differences can be noticed comparing the reconstructed image obtained in this work with the sequence for the token /pasi/ generated by Mohammad. In the /pasa/ sequence: images are noisier, articulators appeared more blurred, and the deformation of the articulators, specially the tongue, appeared to be minimum.

Factors which affect the image quality are: the dynamic nature of the experiments, the magnetic strength of the scanner, the field of view, space resolution and the acquisition sequence. The movement of the articulators causes a band of noise in the phase encoding (rows) direction. As it can be observed, the band of noise is present only in the region where the moving articulators are located. Static parts, such as the brain, appear satisfactorily defined.

Blurring effects may be occasioned by distortions and inhomogeneities of the magnetic field strength. Stronger magnetic fields achieve better spatial resolutions, however, the associated distortions due to the air-tissue boundaries increase proportionally as well. The air has no susceptibility to magnetic fields, hence, differences in the magnetic fields at the vocal tract articulators make the signal decay faster especially when they are moving. This difference appears to be considerable when raw images obtained from 0.5 T and 4 T scanners are compared. Raw images obtained with the 0.5 T system are presented in Figure 8.14. Although the images are blurred and a band of noise is presented around the moving parts in the image, the articulators can be distinguished. A similar band of noise is present in raw

images obtained with the 4T system as shown in Figure 8.15, however, the articulators are highly distorted.

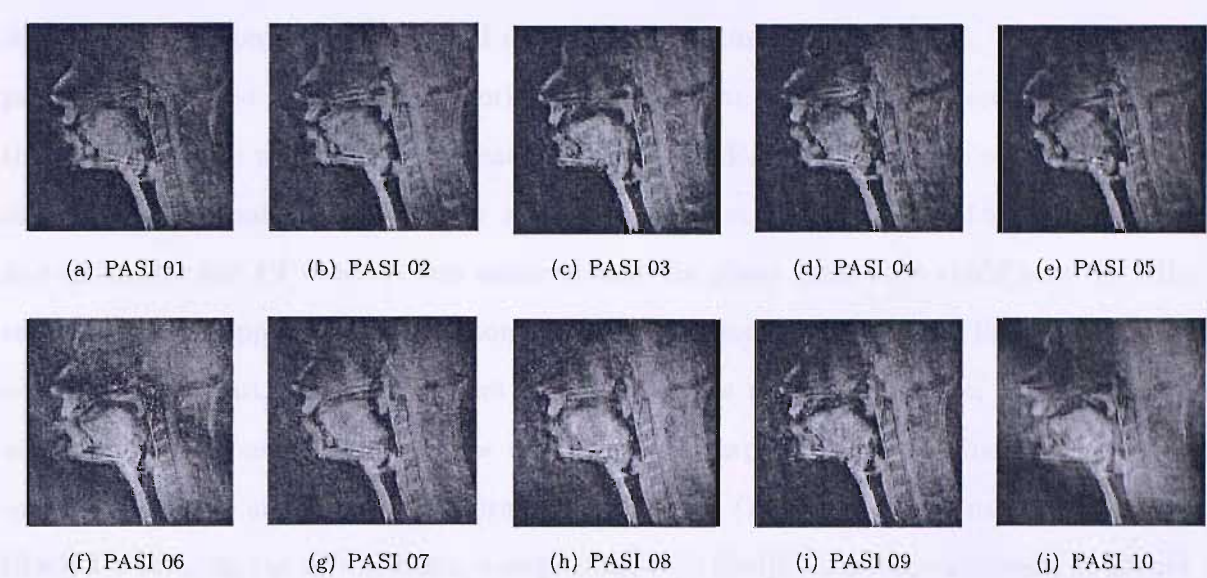


FIGURE 8.14: Raw MR images acquired in Southampton for the middle plane of the subject PJ , while the token /pasi/ is repeated.

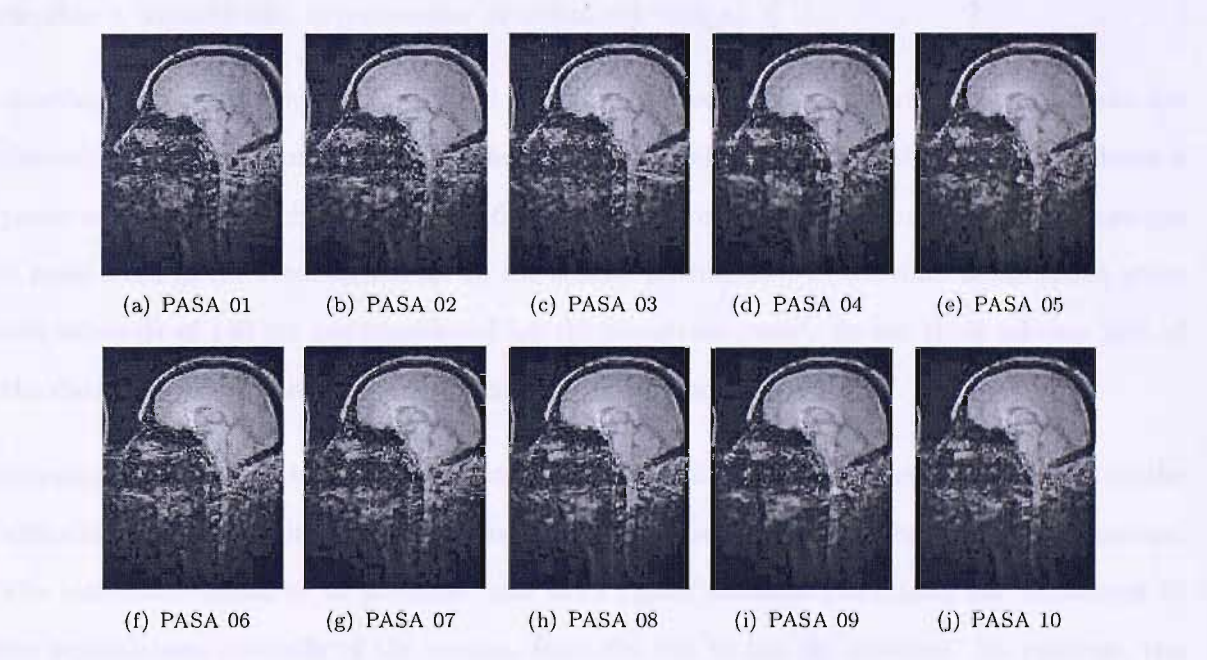


FIGURE 8.15: Raw MR images acquired in Halifax for the middle plane of the subject 04 , while the token /pasa/ is repeated.

The FOV and spatial resolution defined by Mohammad permits the appreciation of the vocal tract articulators in more detail. Consequently, the dynamics of the articulators can be

traced closely. In order to compare these aspects, images for sustained sounds are presented in Figures 8.16 and 8.17.

Additionally, the pulse sequence used may be contributing to such effects. The acquisition pulse sequence used in the present work was the gradient echo multi slice sequence, GEMS; the pulse sequence used by Mohammad was the fast RF-Spoiled gradient echo. The main difference is determined by the inter row sampling rate: 16 ms for the 0.5 T experiments and 120 ms for the 4 T ones. It was assumed that the phase generation could beat the 8 Hz rate, however, it appears that the reconstruction is limited by the Nyquist limit of 8 Hz and consequently no articulator movement is detected. As mentioned before, a controversial aspect in the reconstruction stage is the number of expected frames, since it is defined considering the T_R and the average duration of the token (T_R is defined as inter-row sampling time). Considering the T_R of 120 ms, a small number of final frames (approximately 7) should be reconstructed. In the contrary, considering the acquisition time (at) of 1.712 ms the number of final frames increases around 490. However, the reconstruction of this number of frames requires a considerable large number of initial raw images.

Another factor contributing to blurred effects is the reduction of the raw data set. Data are discarded due to the uncertainty of the articulator behaviour when the subject produces a pause at the end of each token. In the first segmentation scheme the complete token duration is considered in the reconstruction. In the second scheme 22% of the data is discarded when two intervals of 140 ms are considered for the transition /asa/. In the third scheme 26% of the data is discarded with 240 ms considered in the reconstruction.

Sequences correspond to the pronunciation of different tokens. The small differences in the articulator motion during the /pasa/ sequence is a consequence of the repeated /a/ phoneme. The movement required to produced the word /pasi/ involves essentially the movement of the articulators, specially of the tongue, from the /a/ to the /i/ position. By contrast, the tongue does not require extreme deformations during the /pasa/ sequence, since the first and second vowels are the same.

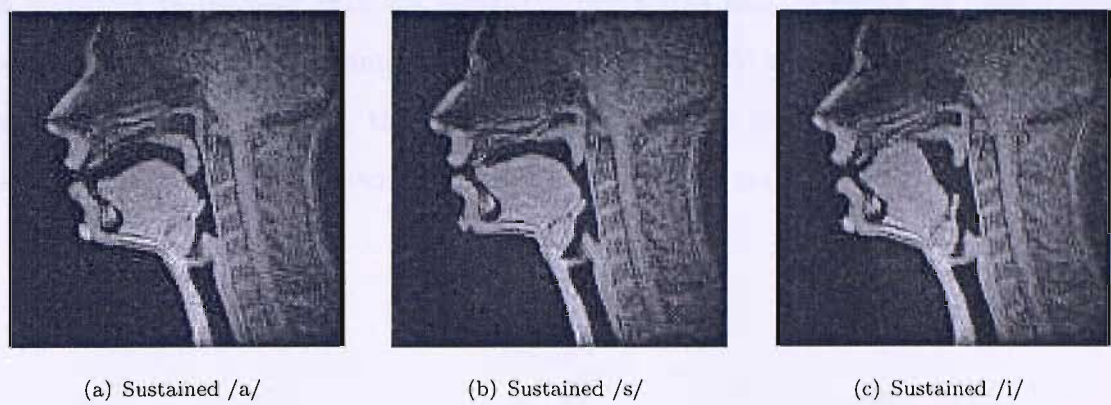


FIGURE 8.16: Sustained images captured for the Subject PJ in the Southampton experiments.

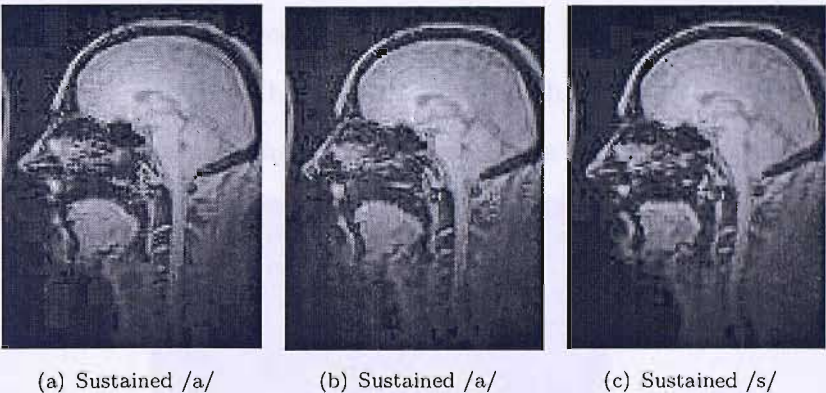


FIGURE 8.17: Sustained images captured for the Subject 04 in the Halifax experiments.

8.4 Application of the ASHT Algorithm on Halifax Data

The vocal tract articulators were very blurred within the sequences of images generated for the data collected in Halifax. The extraction of the tongue shapes using the ASHT method was then performed over the non-speech and sustained images obtained from the four subjects, since these were clear enough to perform the extraction. These images were analysed individually due to anatomical differences from subject to subject. The images could be treated as a sequence of 3 frames to analyse the whole sequence; however, the ASDHT was not used since changes from frame to frame does not vary smoothly.

The ASHT algorithm was applied and the tongue was not fitted well as shown in Figure 8.18.

It is necessary to consider that the algorithm was tested using a model on unseen images generated with a different scanner, spatial resolution, FOV and collected from male and female subjects. In addition, these images were collected for the token /pasa/ which is different from the one used when the images of the original training data set were collected.



FIGURE 8.18: Results obtained for the sustained image subject 04. The algorithm fails detecting the tongue shape.

Therefore, the tongue model was decided to be retrained as the 15 new tongue shapes were included in the training data set. The set of 15 non-speech and sustained images is presented in Figure 8.19 and the corresponding set of tongue shapes are presented in Figure 8.20.

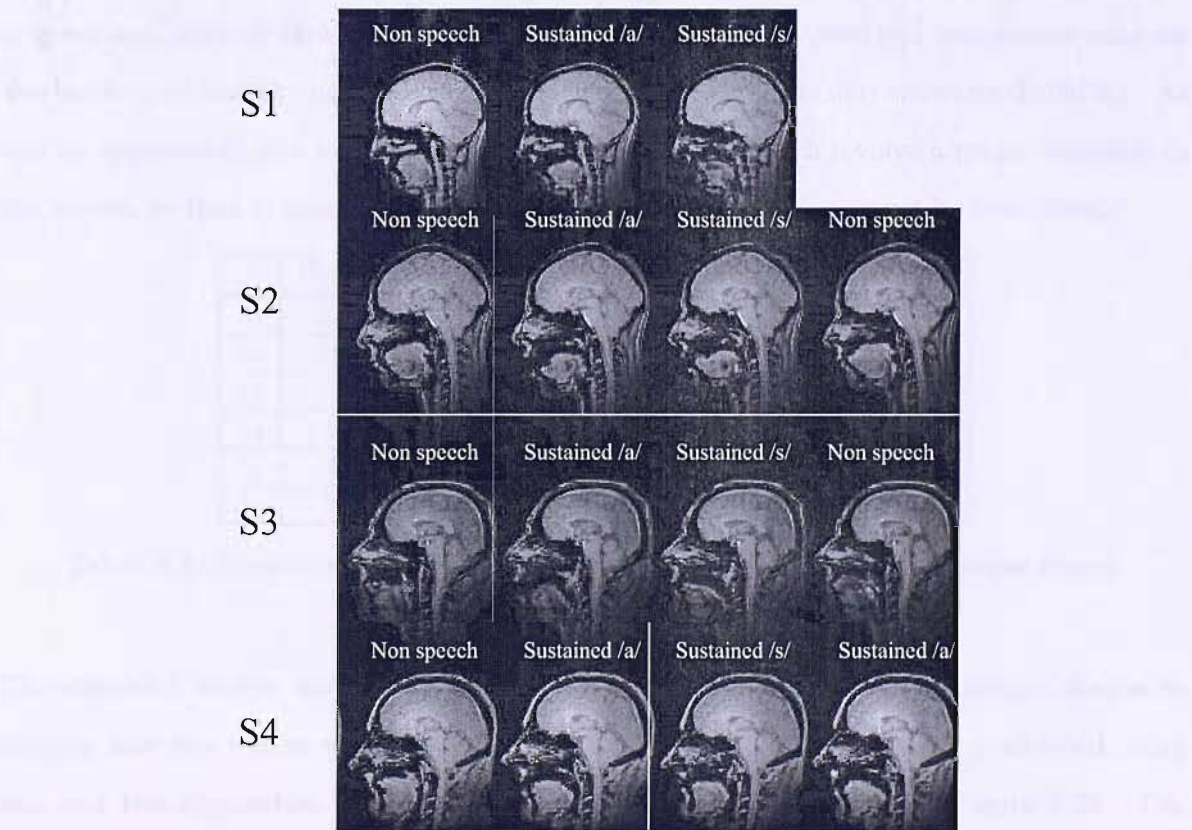


FIGURE 8.19: Non-speech and sustained images used to extend the tongue model.

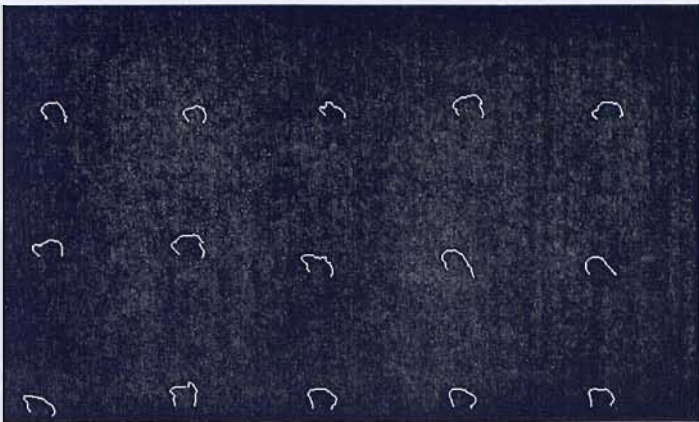


FIGURE 8.20: Set of 15 tongue shapes extracted from non-speech and sustained images collected in Halifax.

The process of generating the new tongue model, which will be referenced as extended tongue model, was the same as previously described. The set of 15 tongue shapes were interpolated to define a consistent set of landmark points, then the complete set of 54 tongue shapes was aligned. The set of 54 aligned shapes is shown in Figure 8.21 and The sum of the original and aligned data set are shown in Figure 8.22. The principal component analysis was performed and the new modes of variation were calculated and shown in Table 8.1. As can be appreciated, the tongue shapes are very irregular, which involve a major variation in the model, so that to cover the 90% of the variation 6 eigenvalues must be considered.

| λ | Eigenvalue | Percentage | Accumulated percentage |
|-------------|------------|------------|------------------------|
| λ_1 | 93.2400 | 53.32 | 53.32 |
| λ_2 | 39.8804 | 22.80 | 76.12 |
| λ_3 | 11.8115 | 6.75 | 82.87 |
| λ_4 | 5.9045 | 3.37 | 86.24 |
| λ_5 | 4.2196 | 2.41 | 88.65 |
| λ_6 | 2.5474 | 1.45 | 90.1 |

TABLE 8.1: Eigenvalues of the covariance matrix derived from the aligned tongue shapes.

The extended tongue model was applied to the original sequence of 39 tongue shapes to analyse how the results were affected. The tongue shape extraction was performed using one and two eigenvalues. The one eigenvalue results are presented in Figure 8.23. The error calculated for this sequence is 1.25. The corresponding results using a two eigenvalue model is presented in Figure 8.24 and the error calculated for this sequence is 1.20. It can

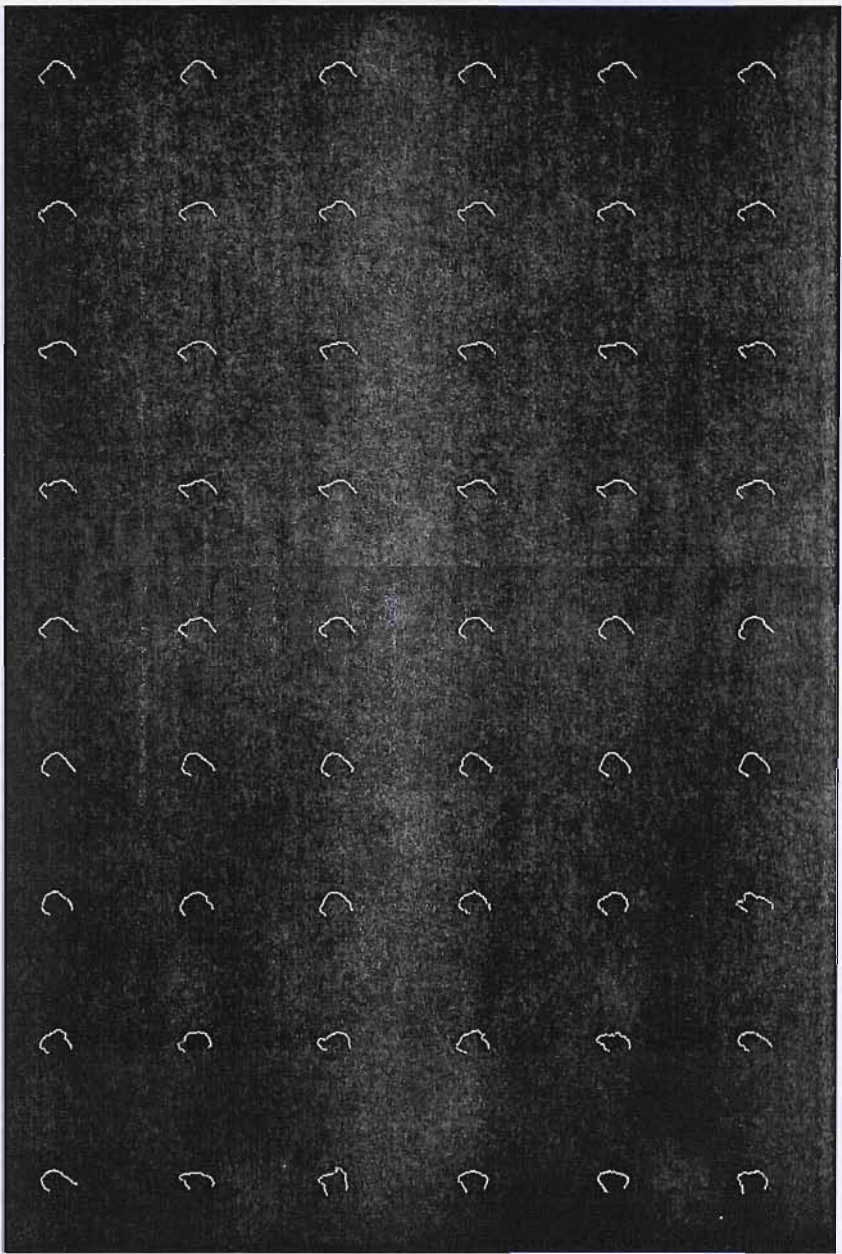


FIGURE 8.21: Extended training data set with the tongue shapes aligned. This set includes the subset of 15 non-speech and sustained tongue shapes collected in Halifax.

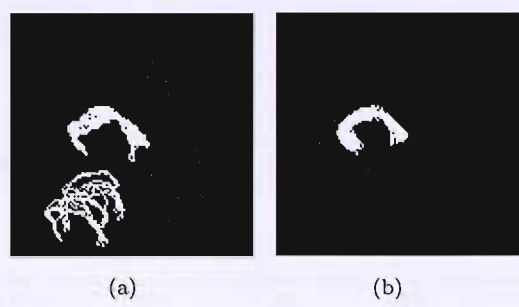


FIGURE 8.22: Superimposed tongue shapes. (a) Sum of all the 54 original tongue shapes. (b) Sum of all the 54 aligned tongue shapes

be appreciated that the first frames in the sequence are better fitted in the two eigenvalue results, however the remaining last frames are better in the one eigenvalue sequence results. The errors were greater than those obtained with the original tongue model.

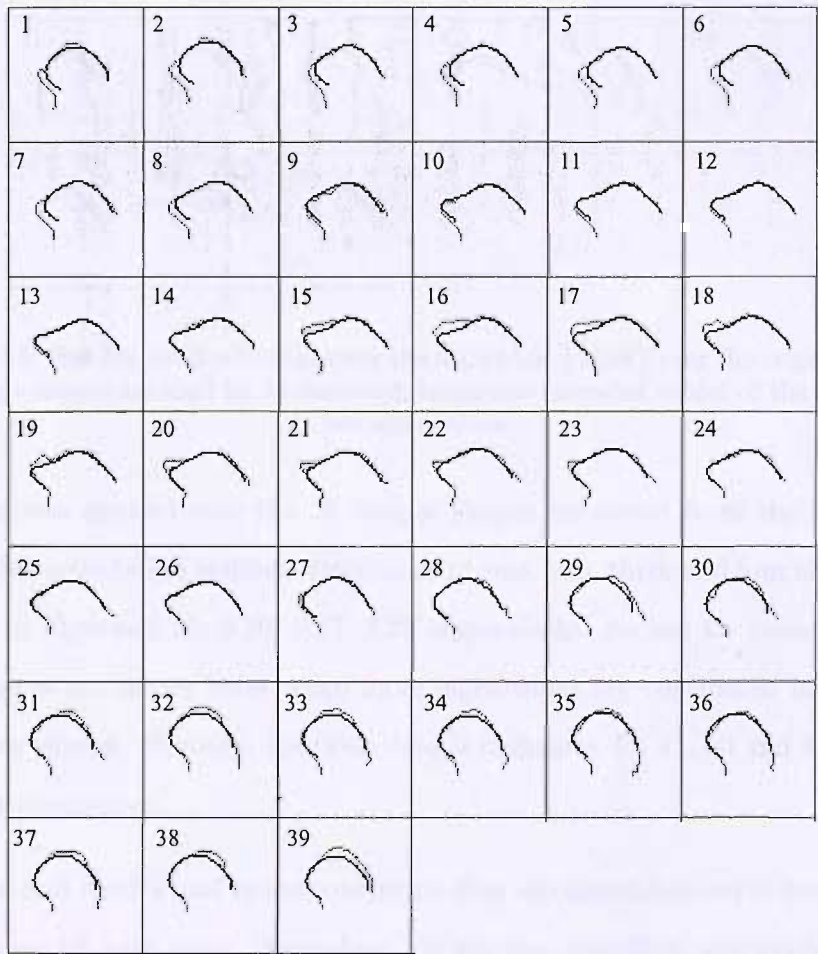


FIGURE 8.23: Results obtained of applying the algorithm ASDHT over the original sequence of 39 tongue shapes labelled by Mohammad, using the extended model of the tongue with one eigenvalue.

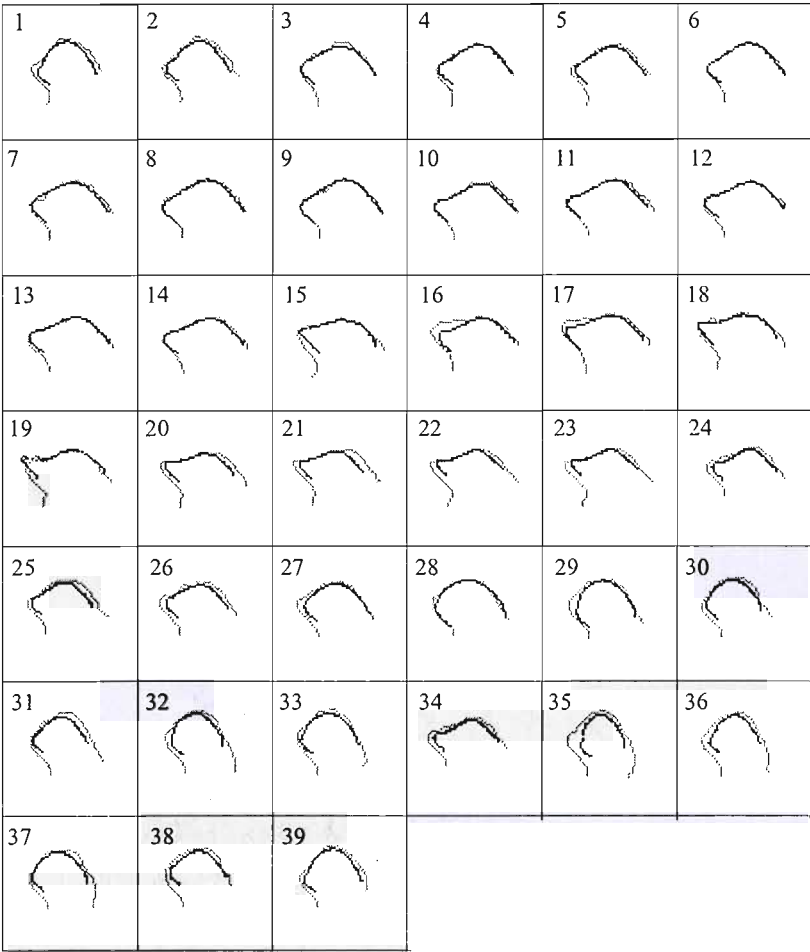


FIGURE 8.24: Results obtained of applying the algorithm ASDHT over the original sequence of 39 tongue shapes labelled by Mohammad, using the extended model of the tongue with two eigenvalues.

The algorithm was applied over the 15 tongue shapes extracted from the non-speech and sustained images collected in Halifax. Results using one, two, three and four eigenvalue model are presented in Figures 8.25, 8.26, 8.27, 8.28 respectively. As can be visually appreciated, the tongue shapes are better fitter when more eigenvalues are considered in the generation of the matching shapes. However, spurious details in frames 46, 47, 50 and 51 are not fitted even with four eigenvalues.

These experimental results lead to the conclusion that the algorithm can fit better any tongue shape using a model with more eigenvalues. When the algorithm was applied to full edge images some the tongue shapes were wrongly fitted in the skull edges as shown in Figure 8.29. Here, although multiple peaks were present the best match was chosen for each edge image to

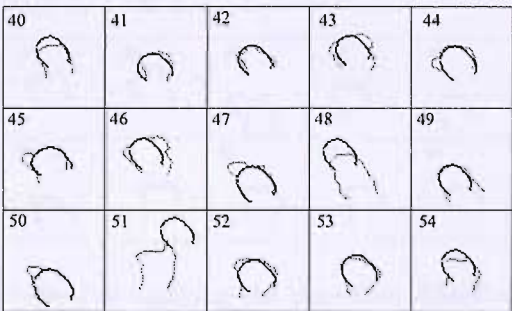


FIGURE 8.25: Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with one eigenvalue.

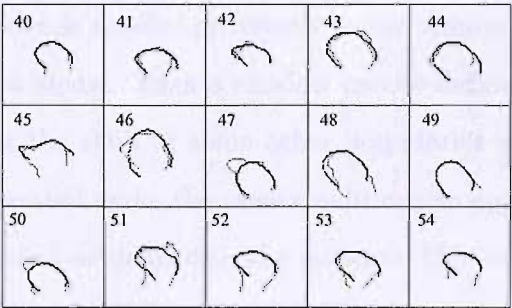


FIGURE 8.26: Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with two eigenvalues.

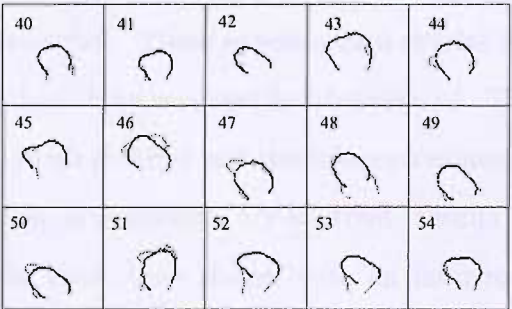


FIGURE 8.27: Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with three eigenvalues.

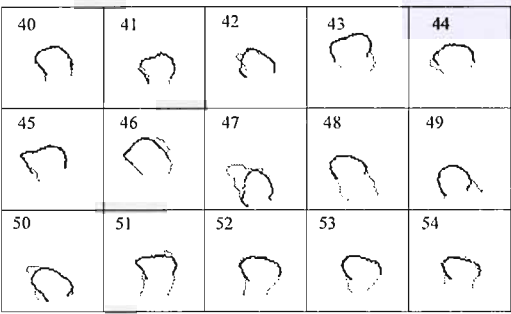


FIGURE 8.28: Results obtained of applying the algorithm ASHT over the set of 15 tongue shapes, extracted from the non-speech and sustained images collected in Halifax, using the extended model of the tongue with four eigenvalues.

be displayed. However, although the best match was selected most of the tongue shapes were wrongly fitted in the skull boundaries, some tongue shapes for subject 02 and subject 04 were well fitted. The localisation of the tongue shape can be improved if a model of the vocal tract with one eigenvalue is applied previously to the sequence, this will provide a good approximation of the tongue shape. Then a window can be defined to restrict the searching area of the tongue avoiding the skull or some other boundaries with similar shapes. If the sequences of images are generated under the same conditions an approximation of the position of the tongue can be predicted or defined in the system. This recommendations avoid any initialisation step because the aim of this project is an automatic extraction.

8.5 Conclusions

In this chapter the results and evaluation of the reconstruction performed over the data collected in Halifax were presented. These experimental results reveals less deformation of the vocal tract articulators than those reported by Mohammad. This could be a result of the inter row sampling rate of 120 ms (8.3 Hz) and the token pronounced, which does not require big changes in the vocal tract articulators. By contrast, results presented by Mohammad shows more variation on the vocal tract shape, with an inter-row sampling rate of 16 ms (62.5 Hz) and a token forcing large movements of the articulators.

Noisy and blurred effects are more evident in the present results. The magnetic field strength, the FOV, the spatial resolution and the pulse sequence used were discussed as possible causes.

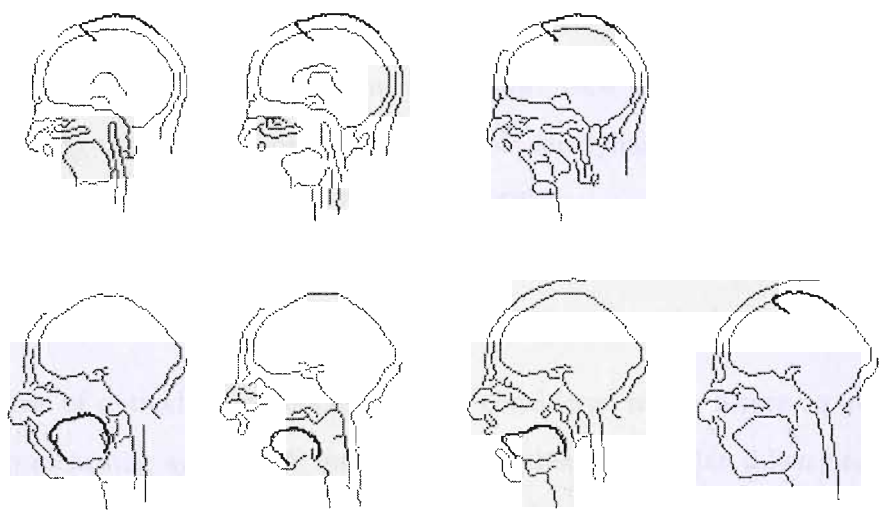


FIGURE 8.29: Results obtained of applying the algorithm ASHT over the set of full edge images for subject 01 (first row) subject 02 (second row), subject 03 (third row) and subject 04 (fourth row).

Non-speech and sustained images collected by this scanner appeared to be satisfactory for these studies. However, when these images are collected the magnetic field distortions are compensated by an optimisation process to achieve an acceptable image quality. These magnetic field distortions are altered when the articulators move and the compensation is no longer adequate for the new conditions.

However, these aspects are still under review. New experiments must be carried out before concluding that a 4T scanner is not suitable for dynamic speech studies. These experiments must include the acquisition of data with different pulse sequences, mainly trying to acquire one plane at a time, i.e. sequential row acquisition. A head and neck coil should be used to define a FOV with more emphasis in the vocal tract articulators.

The use of optical microphones must be considered in the future to avoid limitations in the audio recording system. A mechanism to achieve a regular token pronunciation should be implemented as well; for example, recording an audio file with a subject repeating the token and playing this audio file during the scanning session to guide the subject on a more regular pronunciation of the token may also be valuable help.

Although dynamic reconstructed images could not be used to test the ASDHT algorithm, the non-speech and sustained images were used to test the ASHT algorithm. The algorithm failed to match the tongue shape due to anatomical differences, therefore, the tongue shapes from this set of images were outlined and included in the set of training tongues to train the model using this new set. The variation of the tongue increased compared to the first model generated.

Using the extended model the ASDHT algorithm was tested on the set of 39 original tongue shapes. Although results were not as good as those obtained using the first tongue model, the tongue shape was well fitted. This probably be caused because less variation was covered with one and two eigenvalues of the extended model. Consequently, the additional set of 15 tongue shapes extracted from the set of non-speech and sustained images collected in Halifax were fitted. These tongue shapes were analysed with the ASHT algorithm instead of the ASDHT because the changes in shape and position from one frame to another were considerably bigger.

This led to the conclusion that the algorithm can be applied to different sequences and the model can be retrained as many times as necessary to detect new tongue shapes to cover not only anatomical and experimental differences but also different movements of the tongue for different tokens and phonemes.

Chapter 9

Conclusions and Future Work

9.1 Introduction

This thesis has focused on automatic extraction of articulatory shape from magnetic resonance image sequences of subjects during speech production. A combination of two feature extraction techniques, active shape models and Hough transform was presented. Active shape models were used to generate a model which was used as a shape description for the Hough transform. Images were analysed individually and results were satisfactory, however, when multiple candidate peaks were present in the accumulator space generated by the new form of the Hough transform there was no criteria to choose the optimal solution. Then, the problem was addressed as a global analysis of the sequence was performed. The dynamic Hough transform was adapted for our purposes, defining new constraints to limit changes in centroid position, scaling factor and shape parameters. Section 9.2 discusses the main ideas derived from the present work and Section 9.3 suggests possible future directions on this research subject.

9.2 Discussion

The automatic extraction of arbitrary shapes was addressed by combining two feature extraction techniques. Active shape models were used to generate a shape model which is used as shape description in a new form of the Hough transform: the active shape Hough transform (ASHT). In the evidence gathering process that characterises the Hough transform an accumulator is generated, where the number of parameters to consider in the searching process depends on the level of invariance required. This algorithm was tested on MRI sequences obtained by Mohammad (1999) using the SMDRI method. The algorithm was tested on the extraction of the tongue shape from the MRI sequences using one and two eigenvalue models. The set of searching parameters was composed by the centroid position, shape parameters (value of parameters for one and two eigenvalue models) and scaling factor.

Although this algorithm shows good results, it can go wrong when selecting the optimal shape and position due to the existence of multiple candidates, to noise or to missing information. As a solution a global analysis of the sequence was performed by using the dynamic Hough transform. This method focuses on the definition of an optimal path solution within the accumulator space generated by the application of the ASHT to each image in the sequence.

The computational cost was reduced by implementing the algorithms in the C language and by applying the coarse to fine dynamic programming formulation reported by Raphael (2001). However, large accumulator spaces tended to be generated by this algorithm when the variation in shape deformation involves the use of more than two eigenvalues. This disadvantage can be reduced by using a parallel implementation.

The tongue was extracted from the so-called ground-truth data set, unseen and synthetic sequences using one and two eigenvalue models. In these tests, nine different steps were used for the shape parameters and the scaling factor. The number of steps will increase the dimensions of the accumulator space. The image sequences of the ground-truth and unseen data sets correspond to two male native speakers of British English repeating the same nonsense token /pasi/, and images from both subjects were collected using the same

scanner. Anatomical differences were covered by using a different range of values for the scaling factor.

The differences between the target and the fitted shapes were measured using a chamfer distance metric. Experimental results revealed that tongue shapes were fitted better when a two eigenvalue model was used. Obviously, the error obtained from tests performed over the ground-truth data set were less than those obtained from the unseen data set. However, results obtained from the unseen data were encouraging by the fact that the associated deterioration in performance does not increase dramatically.

Synthetic sequences were generated using a three eigenvalue model to generate the tongue shapes. These experiments were performed to test the algorithm over a known set of shapes. In general, results obtained with two eigenvalue model were better than those obtained with a one eigenvalue model.

Then, the sequence of tongue shapes corresponding to the ground truth data set was contaminated with different levels of noise. The algorithm gave consistent results with 20% of noise added. Although the algorithm offers good fits for some tongue shape in the sequence when higher levels of noise are added, it can go terribly wrong in some frames.

Although the algorithm was tested using a model of the tongue shape, it was proven that the model can be easily extended as information for the lips was included in a new model which included information for tongue and lips. However, this model offered higher modes of deformation so that for covering 90% of the variation (as done for the tongue model), a larger number of eigenvalues must be used. This model for lips and tongue was tested over the same ground-truth data set and the unseen data sequences referenced before.

The generation of a model which includes information of more than one articulator offers better results in the localisation of the shape; however, the shape may not be fitted as well as using a separate model for each articulator. This may be caused by: differences in anatomical dimensions of the articulators and the higher variation in the composed shape. The scaling factor employed for each articulator may vary from subject to subject, as it was observed in the results obtained from fitting the tongue and lip model in the unseen data set. The tongue

shape was better fitted than the lips. The relative distance between the lips and the tongue could not be controlled with the scaling factor, and the range used for scaling was good for the tongue shape and not good for the lips.

The achievement of a multiplanar vocal tract dynamics demands a suitable multiplanar data set. Although the sequences of images generated by Mohammad included the information for three sagittal planes, it was not enough for fully describing the vocal tract dynamics. Thus, the 4T facilities provided by the Institute for Biodiagnostics (Atlantic) Neuroimaging Research Laboratory were used to collect more data.

Although information of seven sagittal planes were collected from two males and two females with different native speaking language using similar procedures to those followed by Mohammad, the reconstructed image sequences were not as good as expected. Three different sets of data were collected: audio data, MR images and gating pulses. Although the same nonsense word used by Mohammad was initially considered in these experiments, because of some limitations imposed by the intercom of the console room (which was cutting off audio information for frequency ranges above 2.5 kHz), the non sense token used in the experiments was changed to /pasa/. In the audio data recorded the noise generated by the scanner was reduced by applying a minimum mean squared error and minimum statistics noise estimation filter. The synchronisation of the gating pulses and the audio data revealed the existence of a delay between both sets of data and an initial offset of 74.5 ms. The delay was assumed to be caused by a slightly different sampling rate for the gating pulses and the initial offset by an inversion recovery time defined in the acquisition process, and a delay in the transmission of the sound.

The pulse sequence used in the image acquisition was different from that reported by Mohammad. The number of images to be reconstructed was defined as the number of collected images. The number of reconstructed frames was defined by Mohammad as a function of the T_R value, since this value was the actual inter-row collection time. The new pulse acquisition sequence defines an inter-row time of 120 ms. However, the actual collection time is that referred as acquisition time, at . If this value is considered in the number of expected frames,

this becomes very large. For this reason, the number of expected frames was defined as the same number of collected raw MR images.

Images reconstructed from the data collected in Halifax were very noisy and blurred. These images were not as good as expected. In general, information of the vocal tract articulators was very poor. The best sequence reconstructed corresponds to the subject 04, and even this sequence has not enough information of the vocal tract articulators to make further work useful. This may be caused by magnetic field distortions due to air-tissue susceptibility differences, which were compensated during the static and sustained collections by an optimisation process. However, this optimisation was altered when the articulators moved. These distortions become bigger when the magnetic field strength of the scanner increases. In our case, the magnetic field increases 8 times in comparison with the scanner used by Mohammad. These aspects are still under review and new experiments must be carried out before concluding that this 4T scanner is not suitable for dynamic speech studies.

Although the new reconstructed images could not be used to test the ASDHT algorithm, static and sustained images were used to test the algorithm in an individual analysis using the ASHT. The tongue shapes from this set were extracted, interpolated and smoothed to generate a consistent sets of points to be included in the original training set. The variation of the tongue increased compared to the first model generated and instead of using two eigenvalues to cover the 90% of the variation a total of 6 eigenvalues must be used for the extended model to cover the same variation.

The ASDHT algorithm was tested over the set of 39 original tongue shapes. Although results were not as good as those obtained using the first tongue model, the tongue shape was well fitted. It may be that the variation covered with one and two eigenvalues was less than in the previous model. However, the corresponding error did not increase dramatically, so that this encouraged the idea of including more variation within the tongue model. Consequently, the additional set of 15 tongue shapes were fitted better than in the previous model results. The algorithm can be applied to different unseen sequences and the model can be retrained as many times as necessary to detect new tongue shapes to cover anatomical differences and different movements of the tongue for different phonemes and tokens.

Some obvious ways that the method can fail are: the number of eigenvalues used is insufficient and the discretisation steps for the eigenvalues may have been too coarse. Although it is standard practice to use eigenvalues sufficient to cover 90% of the variance in the training data, our data could be limited or unrepresentative.

A disadvantage of the method is its highly computational cost. This is what prevented us from using more eigenvalues and less coarse discretisation. Although the resulting shapes were scaled to fit the target shapes, as the shape is described as a discrete set of points the rotation and scaling of the shape by larger factors may not be implemented optimally.

First, we have only considered tongue shape in one plane. It would be very useful to extend the work to extract 3D tongue shape as well as the moving shapes of other articulators. Ultimately, extraction of full 3D vocal tract shape would be invaluable. Second, our method has only been applied to SDMRI sequences for the spoken words, /pasi/ and /pasa/. At the moment, we tested the model developed for the /pasi/ token over /pasa/ images. However, experimental and anatomical differences were presented to assure this model would generalise to other words and speech sounds.

9.3 Future Work

This section presents the possible directions in which this thesis may be continued. Currently, a new method for the automatic extraction of arbitrary shapes from image sequences has been described. Furthermore, the method has been shown immunity to noise and resilience to missing data using a global analysis. However, there remains plenty of scope for improving this method and for analysing the collection of data using the SDMRI method with different scanners and pulse acquisition sequences.

9.3.1 Improving speed

This method has a disadvantage of high computational cost. The use of parallel computing can improve the performance of this algorithm. Different approaches have

been reported. Ercan and Fung (2003) described a study of higher-level image processing algorithms implemented by parallel computing using a standard PC. Improvements in the HT performance were reported. Using data parallelism Dantas et al. (2001) reported an implementation of the HT algorithm for a scintillating fibre tracker. A parallel hardware architecture dedicated for complex two- and three-dimensional video processing was presented by Meribout et al. (2002). Reduction of the memory cost can be improved by avoiding hold al the accumulators during the search of the optimal solution.

The tongue model should be tested over sequences of images for different phonemes and tokens. The extension of these results to multiplanar sequences should be carried out. However, although this method has been applied to MRI sequences for extracting vocal tract articulator shapes, i.e. the tongue, this should be applied to other sequences for tracking arbitrary shapes.

9.3.2 Improving Data Acquisition

The collection of data for reconstructing images using the SDMRI method must be studied in more detail.

In order to draw some conclusions from the data acquisition performed in Halifax using a 4T scanner further experiments should be performed. These experiments must include the acquisition of data with different pulse sequences, trying to acquire one plane at a time, which covers the sequential row acquisition.

A head and neck coil should be used to define a FOV with more emphasis in the vocal tract articulators since no brain information is needed for these studies. This will provide a better tracking of the vocal tract articulators.

The use of better acoustic microphones with better noise reduction techniques must be considered in the future to avoid limitations in the audio recording system and distortions caused by magnetic susceptibility.

A mechanism to achieve a regular token pronunciation should be implemented as well: recording an audio file with a subject repeating the token, and playing this audio file during the scanning session to guide the subject on a more regular pronunciation of the token may also be valuable help.

Thus several directions for further research have been suggested. Unfortunately due to time constraints these remain as areas for future work.

Appendix A

Magnetic Resonance Imaging

A.1 Introduction

Magnetic Resonance Imaging (MRI) is a medical imaging modality commonly used not only for medical purposes but also as a safe method to acquire images with good soft tissue contrast for research applications. Advances in MRI technology have been oriented to overcome some of its disadvantages and have lead to an extensive variety of scanners, receiver and transmitter coils, pulse sequences and software. Its major disadvantage is the long scanning time, which is critical in research areas involved with the study of the dynamic behaviour of an organ: for example, the heart in cardiac MRI and the vocal tract articulators in speech production. Scanners have been designed with different magnetic strength fields to improve their spatial resolution; additionally, pulse sequences have been developed to improve either spatial or temporal resolutions. Manufacturers offer an extensive variety of receiver and transmitter coils for imaging specific areas of the human body; moreover, a variety of applications have been developed to address specific research areas, such as cardiac imaging which is continuously being developed.

This appendix will give a short introduction to the foundations of MRI theory and it is divided as follows: first, a general description of the basic principles of MRI is presented; then, concepts involved in MRI acquisition, such as pulse sequences, are explained; this is

followed by a discussion of the advantages and disadvantages offered by this method; and finally, artefacts and safety considerations are mentioned.

A.2 Basic Principles

MRI is a medical imaging technique based on the magnetic properties of elements. Within an atom, each proton possesses a magnetic moment, called a dipole magnetic moment denoted by μ , and an angular momentum, which causes it to spin on its axis as shown in Figure A.1(a).

Hydrogen, which is very abundant within the human body, is the most commonly-imaged element. Under normal conditions, the protons are randomly aligned; consequently, if all the dipole magnetic moments are summed up, the resulting net magnetisation M will be zero. However, under the presence of an external magnetic field, B_0 , the dipole magnetic moments precess either parallel or anti-parallel to this external field. The precessing effect is shown in Figure A.1(b). This constitutes two different energy states: a low energy state, denoted by E_1 , which is associated with the protons aligned parallel to B_0 ; and a high energy level, represented by E_2 , which corresponds to the protons aligned antiparallel to this magnetic field, as shown in Figure A.2(a). Now, if the moments are added up, the net magnetisation will grow from 0 to a state of equilibrium, represented by M_0 , along the direction of the external magnetic field B_0 and the protons will precess with a frequency known as Larmor frequency ω_0 given by:

$$\omega_0 = \gamma B_0, \quad (\text{A.1})$$

where γ is the gyro-magnetic ratio (Hashemi and Bradley, 1997). For hydrogen protons, $\gamma/2\pi = 42.6 \text{ MHz/T}$ (Pope, 1999).

A schematic representation of a 3D scenario is shown in Figure A.2(b). Here, the external field B_0 is applied towards the longitudinal axis represented by the z -axis. The protons precess around B_0 out of phase, and consequently there is not a component of M_0 in the transversal plane.

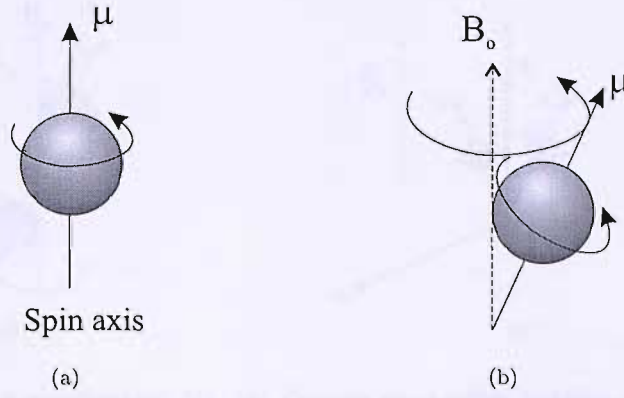


FIGURE A.1: Proton properties. (a) Single proton spins with a magnetic moment μ , (b) Precessing effect around the magnetic field B_0 , with a Larmor frequency ω_0 .

The application of RF pulses with the same Larmor frequency excite the protons making them resonate and consequently generating an electromotive force (emf). This force is regarded as the MRI signal (Pope, 1999). The application of an RF pulse introduces an additional magnetic field B_1 which is much smaller than the magnetic field B_0 . When an RF pulse is applied in the transverse plane (xy), the protons which are still precessing around B_0 will gradually precess in phase; then, these will sum up a component of the net magnetisation on the transversal plane M_{xy} , causing the net magnetisation to flip through an angle θ . This angle is determined by the strength of the RF pulse B_1 and its duration τ as follows,

$$\theta = \gamma B_1 \tau. \quad (\text{A.2})$$

When the RF pulse is turned off, the protons start to precess out of phase again. Then the component M_{xy} will decrease rapidly and the component on the longitudinal axis will slowly recover. Note that each proton spins much faster on its own axis than when it precesses around the axis of the external magnetic field (Hashemi and Bradley, 1997). Consequently, two concepts are introduced: longitudinal relaxation time or spin lattice relaxation time T_1 , and transverse relaxation time or spin spin relaxation time T_2 . The relaxation time T_1 represents the time for the longitudinal component of M_0 to recover, whereas The decaying rate for the transversal component M_{xy} is characterised by the relaxation time T_2 . The protons return to precess out of phase as a result of two important factors and the consequent

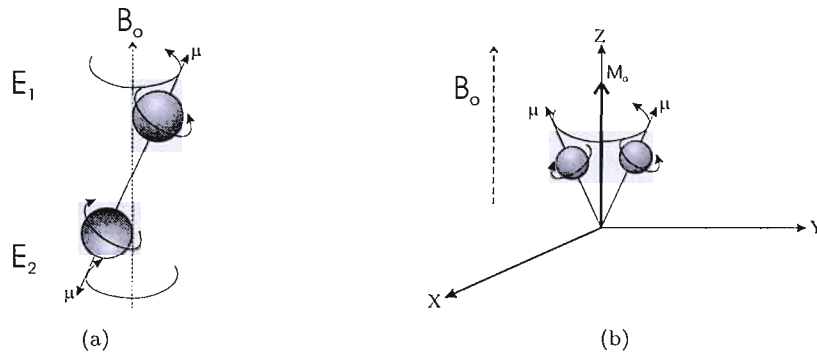


FIGURE A.2: Net magnetisation M_0 . (a) Protons align either parallel or antiparallel to the magnetic field B_0 , constituting two energy states, the higher state, E_1 , with all the protons aligned parallel to B_0 , and the lower state, E_2 , with all the protons aligned anti-parallel to B_0 . (b) The protons are precessing out of phase, therefore the net magnetisation M_0 lies in the longitudinal axis.

decrease of the transversal magnetisation component M_{xy} after the RF signal is turned off. The first factor is the interaction among spins, which is inherent in the properties of the tissue. The second one is the external magnetic field inhomogeneity, which **causes** protons in different locations to precess at slightly different frequencies due to the slightly different magnetic field strength. T_2 decay is inherent in the proximity of spins, as a defined property of the tissue. There is another T_2^* time depending not only on the spin-spin interactions but also in the external magnetic field inhomogeneity (Hashemi and Bradley, 1997).

The transverse magnetic component M_{xy} will decay in time inducing a current on the receiver coil and generating a free induction decay (FID) signal. This is the MRI signal.

A.3 Slice selection

The application of magnetic fields and the generation of the FID signal gives information about the entire body, making no distinction whatsoever among different parts of the body and tissues. Hence, the concept of gradient fields is introduced. A gradient field is used to select a slice of the body by means of exciting this section in such a way that information for small voxels (volume elements) can be acquired.

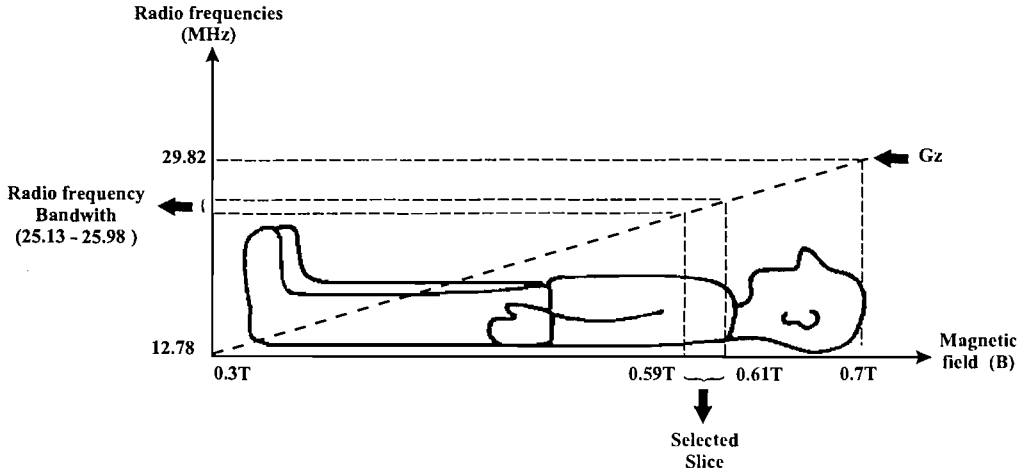


FIGURE A.3: Slice selection. The use of a gradient magnetic field G_z determines the bandwidth of radio frequencies exciting the desired protons. G_z varies from 0.3 to 0.7 T defining the desired slice, the strength of the magnetic field for such a slice (0.59 to 0.61 T) the radio frequency bandwidth, (25.13 – 25.98 MHz), which is defined by equation A.1. Reproduced from Hashemi and Bradley (1997)

A gradient coil generates a magnetic field that varies along the body. This magnetic field is applied with a flipped angle so that different intensities of such a field are applied to different parts of the body; this is denoted by G_z . As shown in Figure A.3, the body is magnetised with different magnetic field intensities along it. If the section of interest comprises part of the chest, the selection of the RF pulses will be based on the range of magnetic fields corresponding to such a section. The determination of the range of radio frequencies to be applied is straightforward when the range of intensities of the magnetic field is known by using the Larmor frequency (Equation A.1).

The slice selection procedure starts with the application of the external magnetic field B_0 . Consequently, all protons will precess at the Larmor frequency ω_0 as shown in Figure A.4(a). Then, the gradient magnetic field G_z is applied and the bandwidth of radio frequencies is determined for the slice of interest; the corresponding RF pulse is applied to cause the longitudinal magnetisation M_0 to flip into the transversal plane with an angle of θ . Thus, the RF pulse and the gradient magnetic field G_z are turned off, and all the volumes in the selected slice will have the same frequency ω_0 as illustrated in Figure A.4(b). Then, the corresponding gradient magnetic fields are applied to generate information along the x and y directions. The gradient G_y , or phase encoding gradient, is applied to modify the phase on each row of the selected slice. As shown in Figure A.4(c), all the volumes precess at different

frequencies; however, when this gradient is turned off, they go back at the same frequency ω_0 with a phase shift. The gradient G_x , or frequency encoding gradient, is used to modify the frequency on each column of the selected slice. The resulting matrix of the selected slice will be composed of voxels with different frequencies and phases. The composite signal is stored in a row of the K -space. This process is repeated until the matrix is full, following a sequential row order as illustrated in Figure A.4(e).

The K -space matrix, which is in the Fourier space, is represented by two spatial frequencies axis, K_x and K_y . The final image is generated when the inverse Fourier transform is applied as shown in Figure A.5.

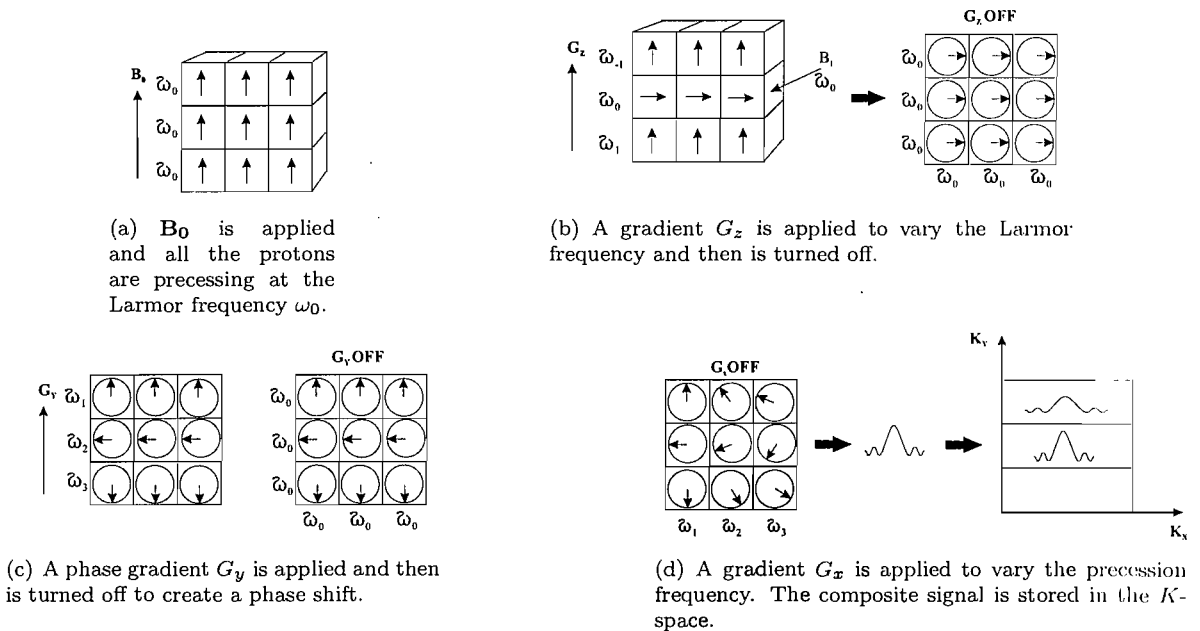
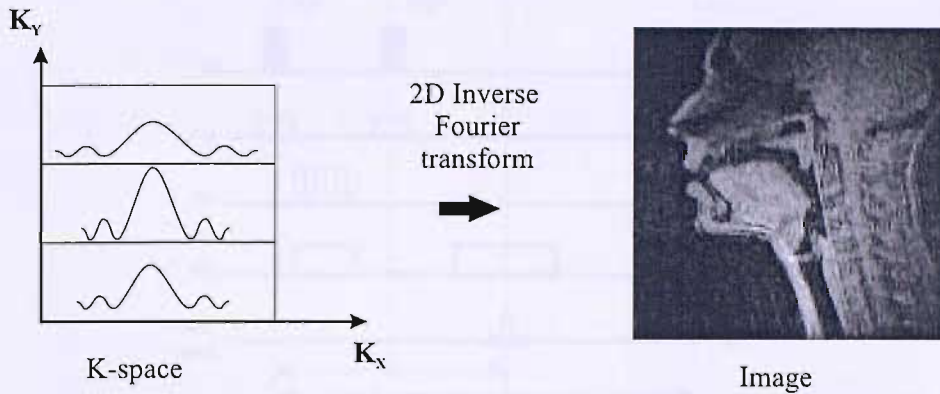


FIGURE A.4: Procedure for slice selection.

A.4 Fourier Imaging

An example of a pulse sequences for Fourier imaging is shown in Figure A.6; this is a pulse sequence diagram for a spin echo sequence. This illustrates the application of radio frequency pulses and gradient magnetic fields, whereby for every rF pulse a flip angle is specified. The gradient G_z selects the slice of interest. The frequency encoding gradient, G_x , is assumed to

FIGURE A.5: Image reconstruction of the sampled K -space matrix.

be the readout gradient. The phase encoding gradient, G_y , determines the resolution of the image, according to the phase encoding steps employed. The MRI signal S of the slice $z = z_0$ is given by:

$$S(t_x, t_y) = M_0 \iint \rho(x, y; z = z_0) \exp[-i\gamma(xG_x t_x + yG_y t_y)] dx dy, \quad (\text{A.3})$$

where M_0 is the equilibrium net magnetisation and $\rho(x, y; z = z_0)$ is the spin density function which includes a dependence on T_1 and T_2 implicitly; $t_x = n\Delta T_x$ and $t_y = n\Delta T_y$ where ΔT_x and ΔT_y are the corresponding increments of the x and y gradient pulse lengths. Equation A.3 is very similar to the 2D Fourier transform. Thus, the final image ρ , can be reconstructed by applying the inverse Fourier transform to the acquired data $S(t_x, t_y)$ (Cho et al., 1993, Chap. 10).

The Fourier transform is used for Cartesian pulse sequences. There are pulse sequences such as radial, spiral and polar sequences, where the image reconstruction consists of transforming information towards the Cartesian space and applying the inverse Fourier transform. A method of transforming from any other coordinate system to the Cartesian one commonly used is interpolation.

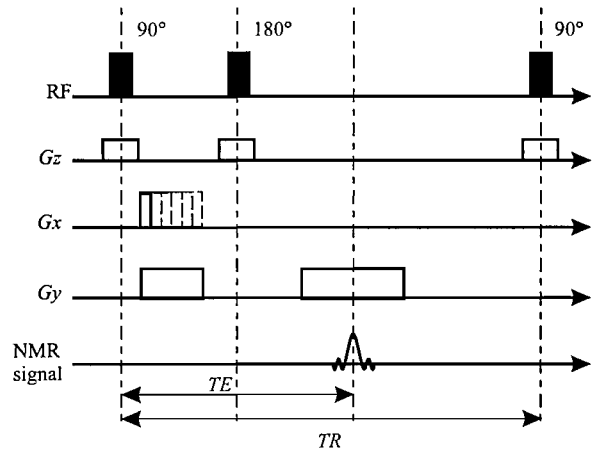


FIGURE A.6: Pulse sequence diagram for the spin-echo sequence for Fourier imaging.

A.5 Imaging Techniques

There are different categories of imaging techniques, appropriate to suit different applications. For example, high speed imaging is used to examine or investigate the dynamic behaviour of an internal organ such as the heart. On the other hand, high spatial resolution techniques are used, for example, in spectroscopy research.

A common trade off is between the spatial and temporal resolution. For dynamic MRI (DMRI) is important not only the acquisition time but also the quality of the final image. Different techniques have been developed in recent years to improve the acquisition time with no sacrifice of the spatial resolution whatsoever.

Spin echo (SE) and gradient echo (GE) sequences are the two basic pulse sequences in MRI acquisition. Variations and extensions of these pulse sequences have been developed to improve either spatial or temporal resolution and to collect multiplanar or volumetric images.

Spin echo sequences (SE) are based on the application of at least two RF pulses: one 90° pulse and one or more 180° pulses. The application of 180° pulses aims to rephase the spins and to reduce the external magnetic field homogeneities. Hence, SE sequences offer good spatial resolution. A SE echo sequence is illustrated in Figure A.6. First a 90° RF pulse is applied to flip the net magnetisation M_0 into the transversal plane. As mentioned before,

the protons tend to spin out of phase due to the slight differences in the magnetic field of each proton. A 180° RF pulse is applied in order to cause a rephasing. After this pulse is applied the FID signal is acquired at the echo time T_E . Then, the next sequence is applied at time T_R .

By contrast, gradient echo (GE) sequences are usually used to reduce the scanning time using partial flip angles; the T_R time is reduced and consequently the total scanning time. In general, the scanning time is defined as:

$$T_S = T_R \times P_e,$$

where T_S is the total scanning time, T_R is the repetition time, and P_e is the number of phase encodings (rows) to be acquired. Therefore, if within every T_R period more than one RF pulse is applied and the associated flip angle is large then, the T_R value will be relatively long. Otherwise, the recovery time T_1 will be short and so does will be T_R .

A.6 Scanners

A general scanning system consists mainly of a room where the scanner is sited and a room with a console to control the process. At this console, images are visualised, stored and processed. The console room includes an intercom that enables communication with the patient inside the magnet room.

An MRI scanner usually comprises a main magnet, magnetic field gradients, transmitter and receiver coils. The main magnet, considered as the heart of the system, defines the strength of the scanner. This strength is measured in tesla units (T). There are some characteristics relating to the nature of the magnet strength. For example, homogeneity defines the uniformity of the magnetic field. Magnetic field gradients are built into the magnet, and the transmitter and receiver coils are designed according to the region of interest, such as knees, shoulders, head and neck among others (McRobbie et al., 2003).

A.7 Artefacts

There are extensive classifications of MRI artefacts. Hashemi and Bradley (1997) present a description of artefacts related to the gradient, external magnetic fields, magnetic susceptibility among others. McRobbie et al. (2003) present a shorter classification of motion, inhomogeneity and digital imaging artefacts. Westbrook and Kaut (1998) present a description of the most common artefacts and the solutions to minimise or eliminate them.

For the purposes of this work, motion, susceptibility and homogeneity artefacts are introduced. Motion artefacts, may be present for causes such as breathing, involuntary movement of the head and due to the natural speech process involved during the acquisition. The result of the movement of the vocal tract articulators is to generate in the final image a band of distortion along the region where it is located. Ghosts are produced on the phase encoding direction since two consecutive phase encoding samples (rows) are acquired within T_R intervals, which makes it a long interval of time.

Susceptibility refers to the degree of magnetisation of the tissue. Air has no susceptibility to magnetic fields; susceptibility differences among elements may generate distortions in the magnetic fields. Such distortions may be reduced by an optimisation process. The magnitude of these increases as the strength of the applied field increases.

Homogeneity refers to the uniformity of the magnetic field; this can be distorted by ferromagnetic parts and by the homogeneity properties of the scanner itself; However, scanners usually have a good correction of inhomogeneities by using of shimming coils.

A.8 Safety

There are many important safety procedures to follow when a patient is taken inside the magnet room. First, an interview takes place where the subject is questioned about any possible situation whereby small metallic residues may be trapped inside his or her body. The more obvious situations are recent surgery, tattoos, metallic implants, or possession of heart pace makers. Additionally, the questionnaire must determine whether the patient

suffers from claustrophobia or not. A metal detector is usually passed along the body of the patient to detect the presence of metallic objects. Some metallic parts, such as screw drivers, could be attracted with such a force that can convert it into a projectile with dangerous consequences, while small pieces of metal might move and cause internal damage. Inside the magnet, the subject should wear a coat and hearing protectors because of the noisy environment. Additionally, the subject is supplied with microphone and headphones so as to be all the time in contact with the operator in the console room. This provides security and support for the patient and a means for reporting any eventuality if necessary. There are no known hazards or secondary effects from exposure to magnetic fields of MRI scanners. A review of safety of strong, static magnetic fields is presented by Schenck (2000).

Any concise survey in MRI theory is likely to be incomplete and therefore a list of complementary readings is recommended. Pope (1999) presents an illustrative introduction to MRI. Hashemi and Bradley (1997) explain in detail the MRI basics; McRobbie et al. (2003) present a different approach explaining the process backward, from the final image to the basic MRI principle, introducing the mathematical details of the technique. Cho et al. (1993) present a mathematical explanation of the technique.

Appendix B

Southampton Dynamic Magnetic Resonance Imaging

B.1 Introduction

Southampton Dynamic Magnetic Resonance Imaging (SDMRI) technique, as described by Mohammad (1999), is used to obtain information about the dynamic behaviour of the vocal tract articulators. This method consists of first acquiring MR images and audio data of a subject repeating a token. These data are then synchronised to reconstruct a sequence of images of the vocal tract dynamics. This method improves the acquisition rate of the MRI scanner which is used, similar to those obtained with cardiac gating; however SDMRI works offline by post-processing the collected data with total control over the reconstruction procedure.

Information about most of the vocal tract articulators is collected at the same time, non invasively. This allows a better understanding of the elements involved in the speech production system and the interaction between them. Furthermore, the subject does not need to be phonetically trained, thus providing a more natural speech environment.

This appendix will present the basic theory of the SDMRI method and a brief introduction to the results reported by Mohammad (1999).

B.2 Basic Theory

The SDMRI method is divided into three stages: acquisition, synchronisation and reconstruction. Data acquisition involves the acquisition of audio and MR images. During the synchronisation every row in the raw images, which are defined in the k-space, is associated with a phase value. Finally, in the reconstruction stage problems of missing and oversampled rows are addressed.

B.2.1 Data Acquisition

During the acquisition stage, audio data and MR images are collected simultaneously. The subject, who does not require to be phonetically trained, is asked to repeat a non-sense word. This word has no meaning but is chosen to demonstrate a phonetic effect, e.g. /pasi/.

Preliminary tests must be performed to define acquisition procedures and parameters. Audio tests are conducted to determine the properties and limitations of the recording mechanisms and to be certain that the audio information recorded is sufficient to achieve a phoneme segmentation as accurate as possible. Thus, it is necessary to verify the range of frequencies transmitted and recorded by the audio recording system. For example, where the console intercom is used to collect the audio, it is necessary to determine whether the transmitted frequencies include or not the phoneme formant frequencies. The audio gain must also be adjusted so that there is no clipping in the recorded signal.

In order to achieve the desired tissue contrast, field of view (FOV), and scanning time, image acquisition tests must be conducted; this will determine the pulse sequence and the corresponding parameters.

Time is critical in speech research. Experiments must be defined considering the possible fatigue of the subject due to a large number of repetitions of the token. These may cause the decay in audio volume and irregular durations in token repetitions. The total time the subject stays inside the scanner may also be an issue, as it is physically uncomfortable and noisy.

During the acquisition the subject is asked to repeat a token as regular as possible, then the scanning process can start at any time.

The initial collected matrices, called raw matrices, are defined in the Fourier space named as the k-space. These matrices are composed of rows acquired at different times and phases of the pronunciation of the token. However, the final images must be composed of rows with the same phase or range of phases. The phase definition and final matrix arrangement are addressed in the synchronisation stage.

B.2.2 Synchronisation

The key of SDMRI is the synchronisation of the audio data with the MR images. The acquisition time of each row is determined and a phase value is assigned according with its occurrence in the corresponding token pronunciation. The first step consists of manually measuring the token and phone duration. Then, the acquisition time is determined and subsequently the phase for each row is defined.

The process of defining the phase for each row is illustrated in Figure B.1. Here a MR image is shown to illustrate the process; however, as mentioned before, the raw data is a k-space matrix. Each data acquisition consists of collecting images while a word is repeated, in this example /pasi/. The scanner may start the image acquisition at any point of the token pronunciation. In Figure B.1, the scanner starts approximately when 17% of the third token has been pronounced, and finishes at 65% of the pronunciation of the fifth token. The final matrix has 128 rows (0 to 127).

Once the phase for each row is defined, synchronised matrices are generated; these will be composed of rows with the same phase or range of phase as illustrated in Figure B.2.

The number of synchronised matrices or final frames is determined as follows

$$N_{EF} = \bar{d}_{tok}/T_R, \quad (\text{B.1})$$

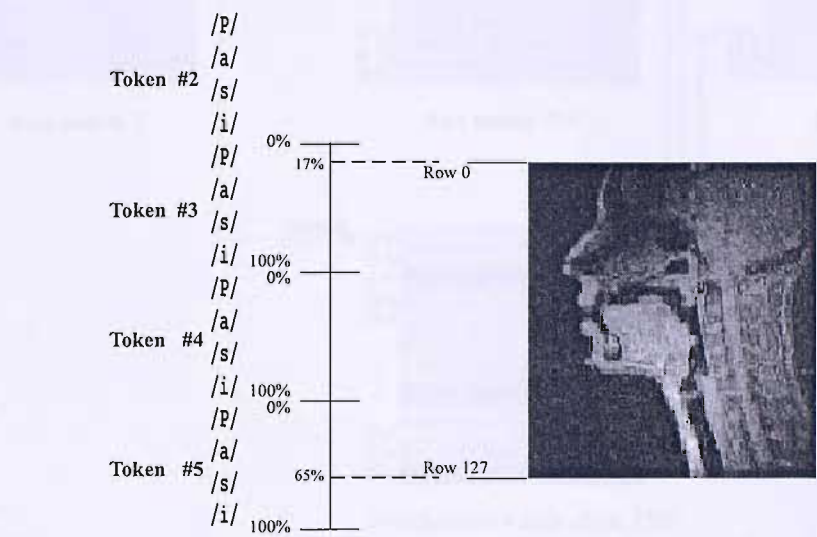


FIGURE B.1: Data acquisition. Example of the image acquisition, while the subject is repeating the word /pasi/. The scanner starts approximately at 17% of the pronunciation of the third token, and finishes at 65% of the fifth one.

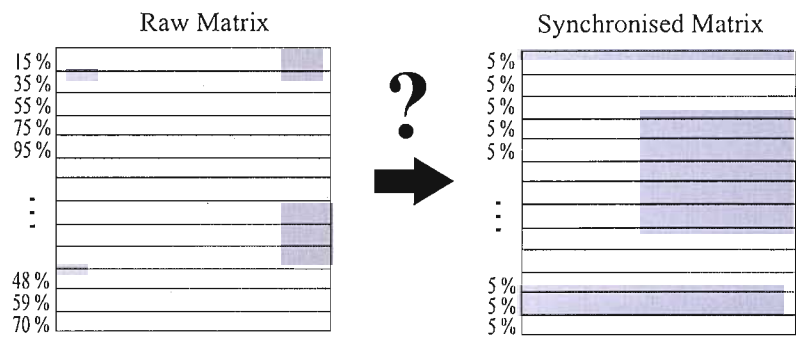


FIGURE B.2: Data synchronisation. Raw images are composed by rows with different phases. Matrices with the same phase or range of phases must be defined.

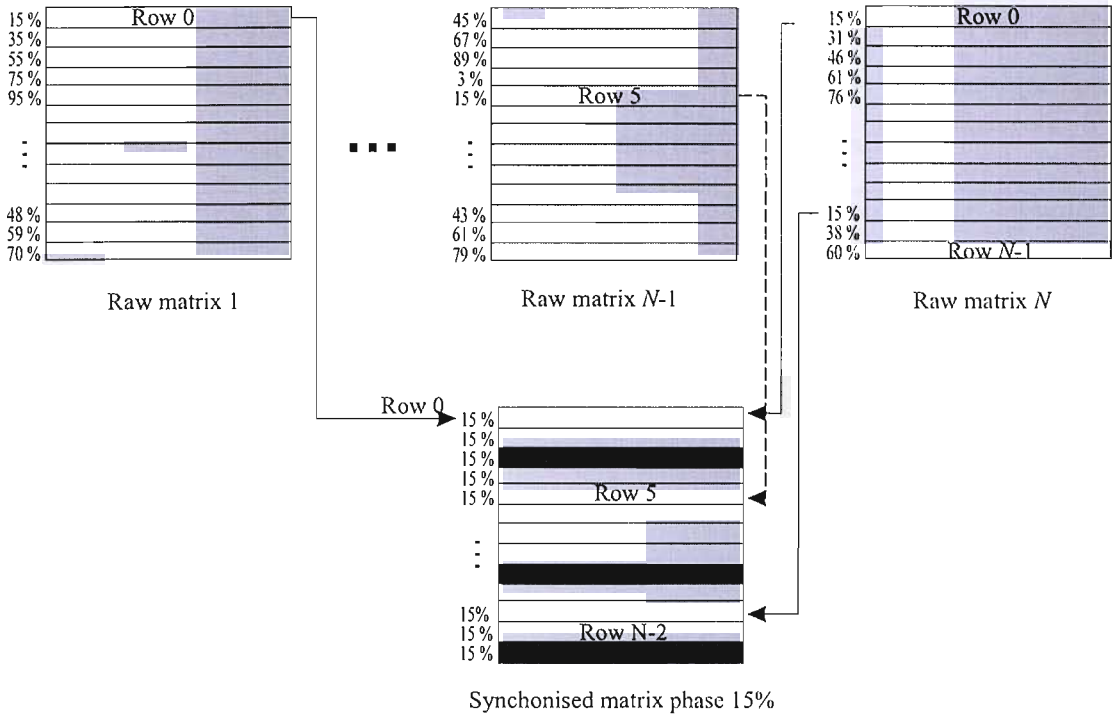


FIGURE B.3: Generation of matrices with the same phase. A matrix for the phase of 15% is generated. Copying rows from the raw matrices with the phase value of 15%. However, this process will generate matrices with oversampled rows (Row 0), and missing rows (Row $N - 1$).

where N_{EF} is the number of expected frames, \bar{d}_{tok} is the duration of the token and T_R is the repetition time.

The synchronisation process will generate a set of N_{EF} matrices consisting of rows with the same phase but with missing and oversampled rows. These problems are addressed in the reconstruction stage.

B.2.3 Reconstruction

Two problems are presented in the reconstruction stage; these concern the missing and oversampled rows, as can be observed in Figure B.3. Oversampled rows in a synchronised matrix are characterized by the presence of two or more rows within the same range of phase. A solution to this problem is to simply average them.

By contrast, the problem concerning the missing data rows is solved using the criterion of borrowing rows from the immediate neighbours. As presented in Algorithm 3, If both

neighbours can contribute to the missing row, an average of them is used. When only one neighbour contributes to the missing row, it is just copied. This process continues until either there are no more missing rows or information in such rows is not available.

Algorithm 3 Pseudo code of the borrowing criterion used to overcome the missing row problem.

```

do
  missing = 0;
  for cFrame = 1 : 1 : TotalImages
    for cRow = 1 : 1 : TotalRows
      if sum(frame(cRow, :, cFrame)) == 0
        missing = missing + 1
        prev = cFrame - 1;
        next = cFrame + 1;
        if prev < 1
          prev = TotalImages;
        if next > TotalImages
          next = 1;
        if sum(frame(cRow, :, prev)) = 0 & sum(frame(cRow, :, next)) = 0
          frame(cRow, :, cFrame) = (frame(cRow, :, prev) + frame(cRow, :, next))./2;
        elseif sum(frame(cRow, :, prev)) = 0
          frame(cRow, :, cFrame) = frame(cRow, :, prev);
        elseif sum(frame(cRow, :, next)) = 0
          frame(cRow, :, cFrame) = frame(cRow, :, next);
      end
    end
  end
while missing

```

The generation of the final images is achieved by applying the inverse Fourier transform. For more details refer to Appendix A.

B.3 Mohammad Results

Mohammad (1999) reported four experiments using two subjects. First subject SG a 25 year old man, native speaker of British English was used in the first two experiments; in the last two experiments were conducted with a 27 year old man, native speaker of British as well. For analysis and discussion purposes of the present work a special emphasis is made in the fourth experiment. Images were collected using a 0.5T Signa GE scanner and a fast RF-spoiled gradient echo sequence. Three sagittal planes were acquired with an image resolution of

128×196 and 5mm thickness with slice separation of 6mm; a total of 24 frames were acquired with a T_R of 16 ms. Table B.1 summarises the acquisition parameters.

| Parameter | Fourth Experiment |
|-------------------------------|-------------------------------|
| Subject | PG |
| Age | 27 |
| Native Language | British English |
| Scanner | 0.5T Signa GE scanner |
| Pulse sequence | Fast RF-Spoiled Gradient-Echo |
| Number of slices | 3 |
| Slice thickness | 5mm |
| Number of initial frames | 24 |
| Slice separation | 6mm |
| T_R | 16ms |
| $Token$ | /pasi/ |
| Phase encodings (rows) | 128 |
| Frequency encodings (columns) | 196 |
| Plane orientation imaged | Sagittal |

TABLE B.1: Acquisition parameters used by Mohammad in the fourth experiment.

The acquisition sequence is illustrated in Figure B.4. As can be observed, the collection was made one plane at a time, where consecutive rows were separated by T_R intervals. Figure B.5 shows an audio file and its spectrogram for a single acquisition. Three sections, which are marked as b, c and d, can be easily detected; these represent the collection of the left, mid and right sagittal planes.

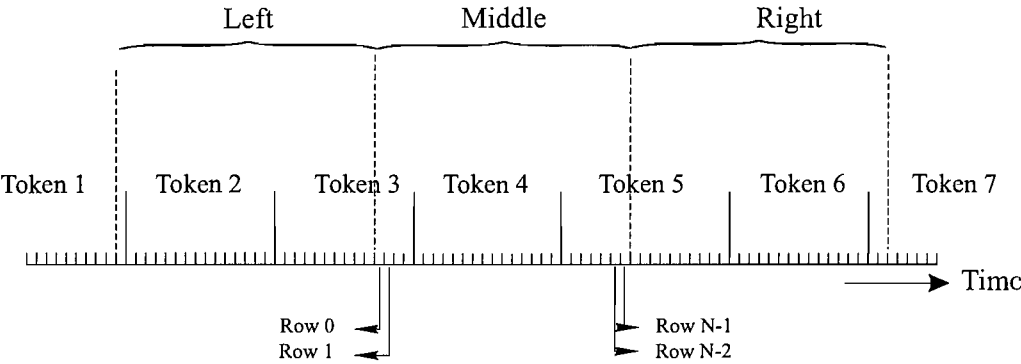


FIGURE B.4: Acquisition sequence used. Each plane is collected at a time. Rows for a specific plane are collected in a sequential order on intervals of T_R .

During the synchronisation, a phase value was assigned to each row of the raw matrices. First, each audio file was segmented so that each token and phoneme duration was measured;

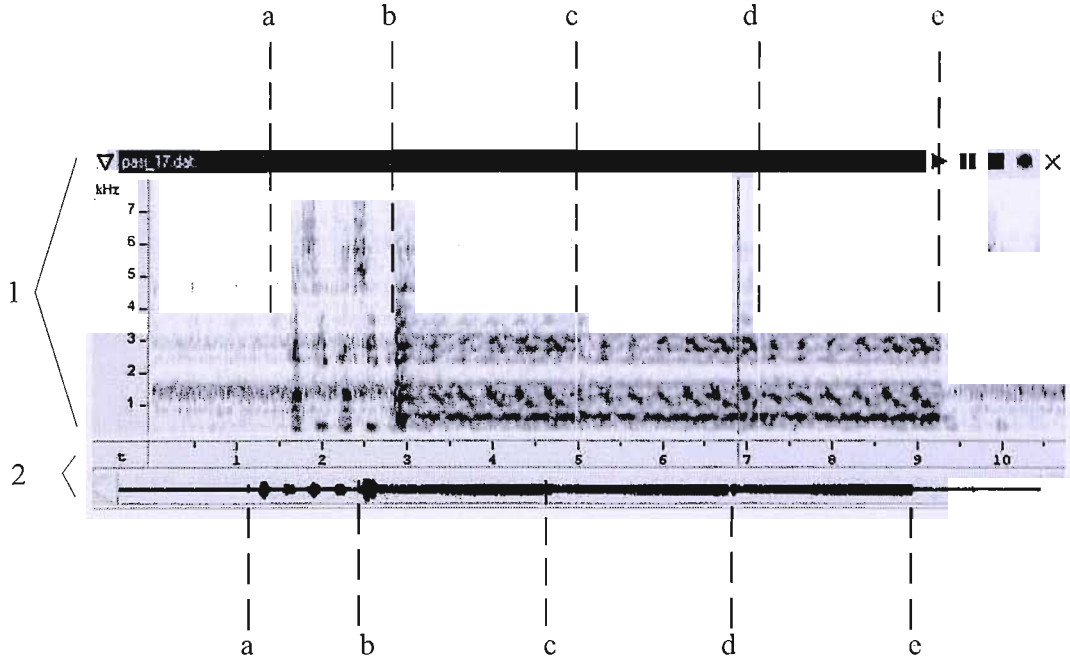


FIGURE B.5: Digitised audio data. In (a) the subject starts to pronounce /pasi/; then, the scanner starts to collect: (b) the left (c) the mid and (d) the left sagittal planes; the image acquisition finishes in (e).

an average of such durations was computed. Then, the acquisition time for the first row of each matrix was determined. It was assumed that the acquisition time for this row, denoted by T_{fr} , was defined by the time the scanner starts emitting its noise; this is marked as (b) in Figure B.5. Each phoneme had a range of phase assigned in 25% intervals as shown in Figure B.6. For example, the phase value for those rows collected during the /p/ pronunciation was defined in the interval of 0 to 0.25. The phase for the first row, ph_{fr} , was then calculated as:

$$ph_{fr} = 0.25(N_{ph}) + \frac{0.25(T_{fr} - T_{startPh})}{d_{ph}}, \quad (\text{B.2})$$

where N_{ph} is the phone number defined as illustrated in Figure B.6; the expression $N_{ph} * 0.25$ defines the initial value of the range of phases for the /ph/ phone; $T_{startPh}$ is the time when the phoneme pronunciation starts and d_{ph} is the duration of the phoneme. Subsequent row phases were defined as increments of the ph_{fr} value in ph_{step} steps given by:

$$ph_{step} = \frac{0.25}{k},$$

where k is the number of rows per phoneme defined by:

$$k = \frac{d_p}{T_R},$$

where d_p is the phoneme duration. Figure B.7 shows an example where the first row is acquired at 1.41s, during the pronunciation of the /a/ phoneme, and N_{ph} is defined as 1. Thus, using Equation B.2, the phase value for this row is defined as 0.38. The next step was to compose matrices with the same phase.

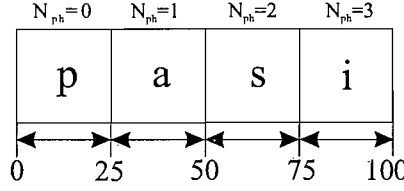


FIGURE B.6: Range of phases defined for the token /pasi/. Each range is defined as 25% and the number of phoneme is defined by N_{ph} .

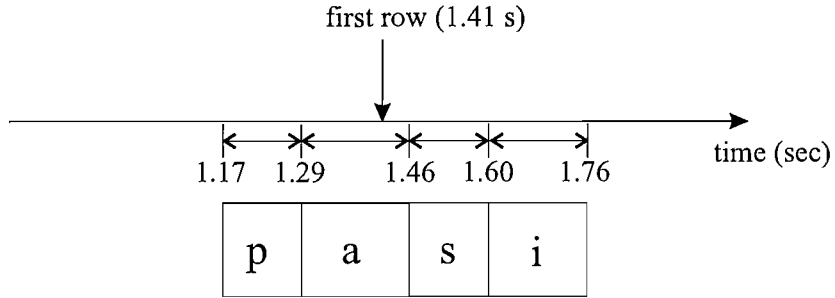


FIGURE B.7: Example of definition of phase for the first row. When the time the first row is determined, it can be deduced the phoneme that was pronounced at that time. The phase value is calculated using Equation B.2.

Using Equation B.1, the range of frequencies for each final frame was defined; thus, the matrices were composed with the corresponding rows. These matrices had missing and oversampled rows; however, this problem was overcome with the borrowing and averaging criterion mentioned before.

A reconstruction of multiplanar sagittal images of the vocal tract with an apparent sampling rate of 63 Hz was reported. This increased by a factor of 136 over the rate of the scanner and pulse sequence used. The sequence of 39 midsagittal reconstructed frames is presented in Figure B.8 and a summary of the results in Table B.2.



FIGURE B.8: Sequence of 39 midsagittal frames reconstructed by Mohammad for the subject PJ in the experiment 4.

| Parameter | Fourth Experiment |
|------------------------|-------------------|
| Subject | PG |
| Initial Frames | 24 |
| Reconstructed Frames | 39 |
| Sampling rate achieved | 63 Hz |
| Image resolution | 128 × 128 |

TABLE B.2: Results generated by Mohammad for the fourth experiment.

Bibliography

- Albers, J., J. Boese, C. Vahl, and S. Hagl (2003). In vivo validation of cardiac spiral computed tomography using retrospective gating. *Society of Thoracic Surgeons* 75, 885–889.
- Badin, P., G. Bailly, M. Raybaudi, and C. Segebarth (1998). A three-dimensional linear articulatory model based on MRI data. In *Proc. Third ESCA/COCOSDA International Workshop on Speech Synthesis*, Australia, pp. 249–254.
- Badin, P., G. Bailly, L. Reveret, M. Baciú, C. Segebarth, and C. Savariaux (2002). Three-dimensional linear articulatory modeling of tongue, lips and face, based on MRI and video images. *Journal of Phonetics* 30, 533–553.
- Baer, T., J. Gore, L. Gracco, and P. Nye (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America* 90(1), 799–828.
- Ballard, D. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition* 13(2), 111–121.
- Beautemps, D., P. Badin, and G. Bailly (2001). Linear degrees of freedom in speech production: Analysis of cineradio- and labio- film data and articulatory-acoustic modeling. *Journal of the Acoustical Society of America* 109(5), 2165–2180.
- Beautemps, D., P. Badin, G. Bailly, A. Galvan, and R. Laboissiere (1996). Evaluation of an articulatory-acoustic model based on a reference subject. In *4th Speech Production Seminar. 1st ESCA Tutorial and Research Workshop on Speech Production Modelling*, Autrans, pp. 45–48.

- Borgefors, G. (1988). A parametric edge matching algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10(6), 849–865.
- Caselles, V., R. Kimmel, and G. Sapiro (1997). Geodesic Active Contours. *International Journal of Computer Vision* 22(1), 61–79.
- Cho, Z., J. Jones, and M. Singh (1993). *Foundations of Medical Imaging*. New York, NY: John Wiley & Sons.
- Cohen, L. (1991). Active Contour Models and Balloons. *Computer Vision, Graphics, and Image Processing: Image Understanding* 53(2), 211–218.
- Cohen, L. and R. Kimmel (1997). Global Minimum for Active Contour Models: A Minimal Path Approach. *International Journal of Computer Vision* 24(1), 57–78.
- Cootes, T., C. Taylor, D. Cooper, and J. Graham (1995). Active shape models – Their training and application. *Computer Vision and Image Understanding* 61(1), 38–59.
- Dantas, A., J. de Seixas, and F. Franca (2001). Parallel Implementation of a Tract Recognition System Using Hough Transform. In *Lecture Notes in Computer Science*, Volume 1981, Berlin, Germany, pp. 467–480. VECPAR.
- DeLucia, J. and F. Kochman (2000). A new noniterative algorithm for computing acoustically constrained vocal tract area functions. *IEEE Transactions on Speech and Audio Processing* 8(2), 177–183.
- Demolin, D., M. George, V. Lecuit, T. Metens, A. Soquet, and H. Raeymaekers (1997). Coarticulation and articulatory compensations studied by dynamic MRI. In *Proc. 5th Eurospeech*, Rhodes, Greece, pp. 31–34.
- Demolin, D., S. Hassid, T. Metens, and A. Soquet (2002). Real-time MRI and articulatory coordination in speech. *Comptes Rendus Biologies*, 547–556.
- Dick, D., C. Ozturk, A. Douglas, E. McVeigh, and M. Stone (2000). Three-dimensional tracking of tongue motion using tagged MRI. In *International Society for Magnetic Resonance in Medicine*, Volume 8, Denver, CO, pp. 553.

- Engelke, W., T. Bruns, M. Striebeck, and G. Hoch (1996). Midsagittal velar kinematics during production of VCV sequences. *Cleft Palate Craniofac J* 33(3), 236–244.
- Engwall, O. (2000). Are static MRI measurements representative of dynamic speech? Results from a comparative study using MRI, EPG and EMA. In *Proc. International Conference on Spoken Language Processing*, Volume 1, Beijing, China, pp. 17–20.
- Engwall, O. (2003). Combining MRI, EMA and EPG measurements in a three-dimensional tongue model. *Speech Communication* 41, 303–329.
- Ephraim, Y. and D. Malah (1984). Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32(6), 1109–1121.
- Epstein, M. and M. Stone (2005). The tongue stops here: Ultrasound imaging of the palate (1). *Journal of the Acoustical Society of America* 118(4), 2128–2131.
- Ercan, M. F. and Y. Fung (2003). Parallel High-Level Image Processing on a Standard PC. In *Proc Lecture Notes in Computer Science*, Volume 2667, Berlin, Germany, pp. 752–760. ICCSA.
- Fletcher, S. (1992). *Articulation. A physiological approach*. San Diego: Singular Publishing Group.
- Gabioud, B. (1994). Articulatory models in speech synthesis. In *Fundamentals of speech synthesis and speech recognition*, Chichester, England, pp. 215–230. John Wiley & Sons.
- Grant, M., M. Nixon, and P. Lewis (2002). Extracting moving shapes by evidence gathering. *Pattern Recognition* 35, 1099–1114.
- Hardcastle, W. (1976). *Physiology of speech production*. New York, NY: Academic Press.
- Hashemi, R. and W. Bradley (1997). *MRI: The Basics*. Baltimore: Lippincott Williams and Wilkins.
- Horiguchi, S. and F. Bell-Berti (1987). The Velotrace: A Device for Monitoring Velar Position. *Cleft Palate J.* 24, 104–111.

- Hough, P. (1962). Method and means for recognizing complex patterns. *US Patent 3069654*.
- Howe, B., A. Gururajan, and H. Sari-Sarraf (2004). Hierarchical segmentation of cervical and lumbar vertebrae using a customized generalized Hough transform and extensions to active appearance models. *Image analysis and interpretation*, 182–186.
- Illingworth, J. and J. Kittler (1988). A survey of the Hough transform. *Computer Vision, Graphics and Image Processing* 44(1), 87–116.
- Kass, M., A. Witkin, and D. Terzopoulos (1988). Snakes: Active contour models. *International Journal of Computer Vision* 1(4), 321–331.
- King, S. and R. Parent (2005). Creating speech-synchronized animation. *IEEE Transactions on Visualization and Computer Graphics* 11(3), 341–352.
- Kiritani, S. (1986). X-ray Microbeam Method for Measurement of Articulatory Dynamics-Techniques and Results. *Speech Communication* 5, 119–140.
- Koninklijke Philips Electronics NV (2003). Magnetic Resonance Imaging. The New Intera 0.5, 1.0, & 1.5T. <http://www.medical.philips.com/main/products/mri/products/interafamily/intera/features/>.
- Lappas, P., J. Carter, and R. Damper (2001). Object tracking via the dynamic velocity Hough transform. In *Proceedings of IEEE International Conference on Image Processing*, Thessaloniki, Greece, pp. 371–374.
- Lappas, P., J. Carter, and R. Damper (2002). Robust evidence-based object tracking. *Pattern Recognition Letters* 23(1-2), 253–260.
- Ma, J. and R. Cole (2004). Animating visible speech and facial expressions. *The Visual Computer* 20, 86–105.
- Makowski, P., T. Sorensen, S. Therkildsen, A. Materka, H. Stodkilde-Jorgensen, and E. Pedersen (2002). Two-phase active contour method for semiautomatic segmentation of the heart and blood vessels from MRI images for 3D visualization. *Computerized Medical Imaging and Graphics* 26, 9–17.

- Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing* 9(5), 504–512.
- Masaki, S., M. K. Tiede, K. Honda, Y. Shimada, I. Fujimoto, Y. Nakamura, and N. Nimoyima (1999). MRI-based speech production study using a synchronized sampling method. *Journal of the Acoustical Society of Japan (E)* 20, 375–379.
- Matthews, I., T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(2), 198–213.
- McRobbie, D., E. Moore, M. Graves, and M. Prince (2003). *MRI From Picture to Proton*. Cambridge, UK: Cambridge University Press.
- Meribout, M., M. Nakanishi, and T. Ogura (2002). Accurate and Real-time Image Processing on a New PC-compatible Board. *Real-Time Imaging* 8, 35–51.
- Mohammad, M. A. (1999). *Dynamic Measurements of Speech Articulators Using Magnetic Resonance Imaging*. Ph. D. thesis, Department of Electronics and Computer Science, University of Southampton, Southampton, UK.
- Muller, E. and J. Abbs (1979). Strain gauge transduction of lip and jaw motion in the midsagittal plane: Refinement of a prototype system. *Journal of the Acoustical Society of America* 62(2), 481–486.
- Napadow, V., R. Kamm, and R. Gilbert (2002). A biomechanical model of sagittal tongue bending. *Journal of Biomechanical Engineering* 124, 547–556.
- Narayanan, S., A. A. Alwan, and K. Haker (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals. *Journal of the Acoustical Society of America* 101(2), 1064–1077.
- Narayanan, S., K. Nayak, S. Lee, A. Sethy, and D. Byrd (2004). An approach to real-time magnetic resonance imaging for speech production. *Journal of the Acoustical Society of America* 115(4), 1771–1776.

- Nash, J., J. Carter, and M. Nixon (1997). Velocity Hough Transform: A new technique for dynamic feature extraction. In *Proc. IEEE International Conference on Image Processing ICIP 97*, pp. 386–389.
- Nixon, M. and A. Aguado (2002). *Feature Extraction and Image Processing*. Oxford: Newnes.
- O’Shaughnessy, D. (2000). *Speech Communication: Human and Machine*. Piscataway New York: IEEE Press.
- Perkell, J., M. Cohen, M. Svirsky, M. Matthies, I. Garabieta, and M. Jackson (1992). Electromagnetic Midsagittal Articulometer Systems for Transducing Speech Articulatory Movements. *Journal of the Acoustical Society of America* 92(6), 3078–3096.
- Perona, P. and J. Malik (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE transactions on Pattern analysis and Machine Intelligence* 12(7), 629–639.
- Pope, J. (1999). *Medical Physics: Imaging*. Oxford, UK: Heinemann.
- Raphael, C. (2001). Coarse-to-fine dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(12), 1379–1390.
- Roerdink, J. and M. Zwaan (1993). Cardiac magnetic resonance imaging by retrospective gating: Mathematical modelling and reconstruction algorithms. *Journal of Applied Mathematics* 4, 241–270.
- Schenck, J. (2000). Safety of Strong, Static Magnetic fields. *Journal of Magnetic Resonance Imaging* 12(1), 2–19.
- Sethi, I. K. and R. C. Jain (1987). Finding trajectories of feature points in a monocular image sequence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 9(1), 56–73.
- Stone, M. (2005). A guide to analyzing tongue motion from ultrasound images. *Clinical Linguistics and Phonetics* 19(6-7), 455–502.
- Stone, M., E. Davis, A. Douglas, M. NessAiver, R. Gullapalli, W. Levine, and A. Lundberg (2001). Modelling the motion of the internal tongue from tagged cine-MRI images. *Journal of the Acoustical Society of America* 109(6), 2974–2982.

- Stone, M., D. Dick, A. Douglas, E. Davis, and C. Ozturk (2000). Modeling the Internal Tongue using Principal Strains. In *Proceedings of the 5th Speech Production Seminar*, Kloster-Seeon, Germany, pp. 133–136.
- Stone, M., M. H. Goldstein, and Y. Zhand (1997). Principal component analysis of cross sections of tongue shapes in vowel production. *Speech Communication* 22, 173–184.
- Stone, M. and A. Lundberg (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustical Society of America* 99(6), 3728–3737.
- Stone, M., T. H. Shawker, T. L. Talbot, and A. H. Rich (1998). Cross-sectional tongue shape during the production of vowels. *Journal of the Acoustical Society of America* 83(4), 1586–1596.
- Story, B. H., I. R. Titze, and E. A. Hoffman (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustical Society of America* 100(1), 537–554.
- Takemoto, H. and K. Honda (2006). Measurement of temporal changes in vocal tract area function from 3D cine-MRI data. *Journal of the Acoustical Society of America* 119(2), 1037–1049.
- Tezmoz, A., H. Saru-Sarraf, S. Mitra, R. Long, and A. Gururajan (2002). Customized Hough transform for robust segmentation of cervical vertebrae from X-ray images. In *Proceedings of the 5th IEEE Southwest Symposium on Image Analysis and Interpretation*, Kloster-Seeon, Germany, pp. 224–228.
- Thiel, E. and A. Montanvert (1992). Chamfer masks: discrete distance functions, geometrical properties and optimization. In *Proceedings of the 11th International Conference in Pattern Recognition*, pp. 244–247.
- Westbrook, C. and C. Kaut (1998). *MRI in Practice*. Oxford, UK: Blackwell Science.
- Williams, D. J. and M. Shah (1992). A fast algorithm for active contours and curvature estimation. *CVGIP:Image understanding* 55(1), 14–26.
- Xu, C. and J. Prince (1998). Snakes, shapes and gradient vector flow. *IEEE Transactions on Image Processing* 7(3), 359–369.

- Yezzi, A. (1997). A geometric snake model for segmentation of medical imagery. *IEEE Transactions on Medical Imaging* 16(12), 199–209.
- Zamora, G., H. Sari-Sarraf, and R. Long (2003). Hierarchical segmentation of vertebrae from X-ray images. In *Proceedings of the SPIE Medical Imaging: Image Processing*, pp. 1370–1381.

List of Authors's Relevant Publications

M.S. Avila-Garcia, J.N. Carter and R.I. Damper (2004). Extracting tongue shape dynamics from magnetic resonance image sequences. 'In Proc. International Conference on Signal Processing', Istanbul, Turkey, pp. 288-291.

M.S. Avila-Garcia, J.N. Carter and R.I. Damper (2005). Automatic extraction of tongue shape dynamics from magnetic resonance image sequences. 'In Proc. 10th International Conference on Speech and Computer', Patras, Greece, pp. 119-122.