

Doctor of Philosophy

**A Novel Lip Geometry Approach for Audio-Visual
Speech Recognition**

Mohd Zamri bin Ibrahim

B030491

School of Electronic, Electrical and Systems Engineering

Loughborough University

United Kingdom

October 2014

DECLARATION

This thesis describes the results of research undertaken in the School of Electronic, Electrical and Systems Engineering, Loughborough University, United Kingdom. This research was supported by scholarship from the Malaysian Government under ‘Skim Latihan Akademik IPTA (SLAI)’ scheme.

The results and analysis presented in this thesis are my own original work, accomplished under the supervision of Dr. David J. Mulvaney except where otherwise acknowledge. This thesis has not been submitted for any other degree.

Mohd Zamri bin Ibrahim
School of Electronic, Electrical and Systems Engineering
Loughborough University
United Kingdom
17th October 2014

ACKNOWLEDGEMENTS

In the name of Allah, the Most Gracious and the Most Merciful

Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis.

First of all I would like to thank my excellent supervisor, Dr. David J. Mulvaney, for his supervision and constant support. His invaluable help of constructive comments and suggestions throughout the experimental and thesis works have contributed to the success of this research. Not forgotten, my special appreciation to Dr. Sekharjit Datta and Prof. Robert I. Damper for their support and knowledge regarding this topic.

My deepest gratitude goes to my beloved parents; Mr. Ibrahim bin Sharif and Mrs. Jalilah binti Abdul Ghani for their endless love, prayers and encouragement. Also not forgetting, my wife, Dr. Norashikin binti Mat Zain and my son, Muhammad Afiq Daniel for their love and care.

Last but not least, sincere thanks to all my friends especially Dr. Bashar Badr for his kindness and moral support during my study. Thanks for the friendship and memories. A special thank you goes to Segun Aina, Thomas Ottmann and Ali Taufiq for helping me with the recording studio setup and data collection for my experiments. To those who indirectly contributed in this research, your kindness means a lot to me. Thank you very much.

ABSTRACT

By identifying lip movements and characterizing their associations with speech sounds, the performance of speech recognition systems can be improved, particularly when operating in noisy environments. Various methods have been studied by research groups around the world to incorporate lip movements into speech recognition in recent years, however exactly how best to incorporate the additional visual information is still not known. This study aims to extend the knowledge of relationships between visual and speech information specifically using lip geometry information due to its robustness to head rotation and the fewer number of features required to represent movement. A new method has been developed to extract lip geometry information, to perform classification and to integrate visual and speech modalities. This thesis makes several contributions. First, this work presents a new method to extract lip geometry features using the combination of a skin colour filter, a border following algorithm and a convex hull approach. The proposed method was found to improve lip shape extraction performance compared to existing approaches. Lip geometry features including height, width, ratio, area, perimeter and various combinations of these features were evaluated to determine which performs best when representing speech in the visual domain. Second, a novel template matching technique able to adapt dynamic differences in the way words are uttered by speakers has been developed, which determines the best fit of an unseen feature signal to those stored in a database template. Third, following on evaluation of integration strategies, a novel method has been developed based on an alternative decision fusion strategy, in which the outcome from the visual and speech modality is chosen by measuring the quality of audio based on kurtosis and skewness analysis and driven by white noise confusion. Finally, the performance of the new methods introduced in this work are evaluated using the CUAVE and LUNA-V data corpora under a range of different signal to noise ratio conditions using the NOISEX-92 dataset.

PUBLICATIONS

During the course of this study, the following refereed conference and journal papers were published.

- M.Z. Ibrahim and D.J. Mulvaney. *Geometry based Lip Reading System using Multi Dimension Dynamic Time Warping*. In IEEE International Conference on Visual Communications and Image Processing (VCIP), San Diego, USA, 27th - 30th November, 2012.
- M.Z. Ibrahim and D.J. Mulvaney. *Robust Geometrical-based Lip-reading using Hidden Markov Models*. In IEEE International Conference on Computer as a Tool (EUROCON), Zagreb, Croatia, 1st - 4th July, 2013.
- M.Z. Ibrahim and D.J. Mulvaney. *Geometrical-Based Lip-Reading using Template Probabilistic Multi-Dimension Dynamic Time Warping*. In Elsevier Journal of Visual Communication and Image Representation (JVCI), Accepted on 20th May 2014 for publication.
- M.Z. Ibrahim and D.J. Mulvaney. *A Lip Geometry Approach for Feature-Fusion based Audio-Visual Speech Recognition*. In IEEE International Symposium on Communications, Control, and Signal Processing (ISCCSP), Athens, Greece, 21st - 23rd May 2014.

TABLE OF CONTENTS

DECLARATION	ii
ACKNOWLEDGEMENTS	iii
ABSTRACT	iv
PUBLICATIONS	v
TABLE OF CONTENTS	vi
LIST OF ABBREVIATIONS	x
LIST OF TABLES	xi
LIST OF FIGURES	xiii
CHAPTER 1 INTRODUCTION	1
1.1 Motivation	3
1.2 Aim and objectives	6
1.3 Original contributions.....	7
1.4 Thesis organization.....	9
CHAPTER 2 AN OVERVIEW OF AUDIO-VIDEO SPEECH RECOGNITION SYSTEMS	10
2.1 Architecture	10
2.2 Audio and visual feature extraction.....	15
2.2.1 Audio feature extraction	15
2.2.2 Visual feature extraction.....	16
2.3 Audio and visual mapping.....	20
2.4 Audio-Visual speech corpora	22
2.5 Chapter summary.....	25
CHAPTER 3 LIP GEOMETRY FEATURE EXTRACTION USING SKIN COLOUR FILTER, BORDER FOLLOWING METHOD AND CONVEX HULL	27
3.1 Introduction	27
3.2 Software, experimental setup and database.....	28
3.3 Face and mouth detection	30
3.4 Lip segmentation	32
3.5 Lip geometry features.....	37
3.6 Lip recognition analysis	41

3.6.1 Qualitative assessment.....	41
3.6.2 Quantitative assessment.....	42
3.7 Lip reading performance evaluation.....	44
3.7.1 Lip reading system	44
3.7.2 Evaluation of Gaussian mixture function and HMM architecture	47
3.7.3 Effect of head rotation and brightness changes.....	50
3.8 Chapter summary.....	54
CHAPTER 4 SHAPE BASED LIP READING SYSTEM USING TEMPLATE PROBABILISTIC MULTI DIMENSION DYNAMIC TIME WARPING.....	55
4.1 Introduction	55
4.2 Methodology.....	56
4.2.1 Lip dynamic information.....	58
4.2.2 Dynamic time warping	60
4.2.3 Multi dimension dynamic time warping.....	62
4.2.4 Novel template probabilistic approach	63
4.3 Results and discussion.....	68
4.3.1 Performance using a single feature.....	69
4.3.2 Performance using multiple features	69
4.3.3 Comparison with state-of-the-art techniques.....	70
4.4 Chapter summary.....	74
CHAPTER 5 LIP GEOMETRY APPROACH IN FEATURE-FUSION BASED AUDIO-VISUAL SPEECH RECOGNITION.....	75
5.1 Introduction	75
5.2 Methodology.....	76
5.2.1 Audio-visual feature fusion	78
5.2.2 Classification	78
5.3 Results and discussion.....	79
5.3.1 Performance under noise conditions	79
5.3.2 Digit performance analysis.....	83
5.3.3 Digit confusion analysis	86
5.3.4 Comparison with appearance based system	89
5.4 Chapter summary.....	91

CHAPTER 6 NOVEL ADAPTIVE FUSION IN AUDIO-VISUAL SPEECH RECOGNITION USING AUDIO STATISTICAL ANALYSIS	93
6.1 Introduction	93
6.2 Audio statistical analysis	95
6.2.1 Background of skewness and kurtosis.....	96
6.2.2 Initial investigation on statistical variation of the audio signals	98
6.3 Methodology.....	105
6.3.1 Architecture	105
6.3.2 Training phase	106
6.3.3 Testing phase	107
6.4 Results and discussion.....	110
6.4.1 Performance of skewness and kurtosis in adaptive fusion AVSR	110
6.4.2 Results of performances trend and modality chosen in adaptive fusion AVSR.....	112
6.4.3 Comparison with conventional method.....	117
6.5 Chapter summary.....	119
CHAPTER 7 LOUGHBOROUGH UNIVERSITY AUDIO-VISUAL SPEECH DATA CORPUS	121
7.1 Introduction	121
7.2 The design of the LUNA-V data corpus.....	122
7.2.1 Ethical clearance.....	122
7.2.2 Subject population	123
7.2.3 Sentence selection	123
7.2.4 Recording studio and hardware	124
7.2.5 Recording process	128
7.2.6 Audio post-processing	129
7.2.7 Video post-processing	131
7.3 Validation of feature extraction and AVSR	133
7.3.1 Lip geometry feature extraction	133
7.3.2 Speech model and evaluation set-up	138
7.3.3 Digit recognition results	144
7.3.4 Speech recognition results.....	150
7.4 Summary.....	155

CHAPTER 8 CONCLUSION AND FUTURE WORKS	156
8.1 Summary of the work in the thesis	156
8.1.1 Lip geometry feature extraction	156
8.1.2 Robustness of lip geometrical features to head rotation and brightness changes in lip reading system	157
8.1.3 Geometrical based lip reading system using TP-MDTW	158
8.1.4 Lip geometry approach to reduce the ‘curse of dimensionality’ in feature-fusion based AVSR	158
8.1.5 Novel adaptive fusion AVSR using skewness and kurtosis analysis	159
8.1.6 Development of the LUNA-V data corpus	159
8.2 Future work	160
8.2.1 Scale invariant features	160
8.2.2 Extending the LUNA-V data corpus	161
8.2.3 Using the Microsoft Kinect as the hardware interface	162
8.2.4 Using lip gestures for HCI	162
REFERENCES	163
Appendix A – Statistical analysis based on McNemar test	174
Appendix B – Digit performance in early integration simulated with white noise	177
Appendix C – Digit performance in early integration simulated with babble noise	182
Appendix D – Human participant documents	187
Appendix E – LUNA-V sentence selection	196
Appendix F – Confusion matrix for speech recognition	199

LIST OF ABBREVIATIONS

AAM	-	Active Appearance Models
AIFD	-	Affine-Invariant Fourier Descriptor
ASM	-	Active Shape Model
ASR	-	Automatic Speech Recognition
AVSR	-	Audio Visual Speech Recognition
BST	-	B-Spline Template
CMYK	-	Cyan, Magenta, Yellow and Black
DCT	-	Discrete Cosine Transform
DTW	-	Dynamic Time Warping
DWT	-	Discrete Wavelet Transform
FAP	-	Facial Animation Point
GVF	-	Gradient Vector Field
HCI	-	Human Computer Interface
HD	-	High Definition
HMM	-	Hidden Markov Model
HSV	-	Hue, Saturation and Value
LDA	-	Linear Discriminant Analysis
LUNA-V	-	Loughborough University Audio-Visual Data Corpus
MDTW	-	Multi-Dimension Dynamic Time Warping
MLLT	-	Maximum Likelihood Linear Transform
OF	-	Optical Flow
PCA	-	Principle Component Analysis
RGB	-	Red, Green and Blue
ROI	-	Region of Interest
SNR	-	Signal to Noise Ratio
SVM	-	Support Vector Machines
WER	-	Word Error Rate

LIST OF TABLES

Table 2.1 Equal error rates (EER) and identification rates (IR) for integration strategies (best performing strategies are highlighted) [27]	15
Table 2.2 The Phone to viseme mapping [65].....	21
Table 2.3 Comparison of popular AVSR data corpora	24
Table 3.1 Qualitative evaluation of the lip classification	42
Table 3.2 Quantitative evaluation of the lip classification	44
Table 3.3 List of phones used for English digit recognition	47
Table 4.1. Word accuracy for single lip geometry features using the TP-MDTW approach.....	69
Table 4.2 Word accuracy for candidate combinations of lip geometry features classified using TP-MDTW.....	70
Table 5.1 Confusion matrix for audio only recognition at 0 dB SNR using white noise.....	86
Table 5.2 Confusion matrix for audio-visual recognition at 0dB SNR using white noise.....	87
Table 5.3 Confusion matrix for audio only recognition at 0dB SNR using babble noise	88
Table 5.4 Confusion matrix for combination of audio and visual at 0 dB SNR using babble noise	88
Table 6.1 Performance of adaptive fusion AVSR in word recognition (%) using skewness and kurtosis analysis using Model 2R	111
Table 6.2 Performance of adaptive fusion AVSR in word recognition (%) using skewness and kurtosis analysis using Model 3R	111
Table 7.1 The sentences collected for the LUNA-V corpus.....	124
Table 7.2 Comparison of the qualitative evaluations of lip classification for the LUNA-V and CUAVE data corpora	136
Table 7.3 Comparison of quantitative evaluation of the lip classification for the LUNA-V and CUAVE data corpora	137
Table 7.4 List of phones present in sentence 1.....	139
Table 7.5 List of phones present in sentence 2.....	140
Table 7.6 List of phones present in sentence 3.....	140
Table 7.7 List of phones present in sentence 4.....	140

Table 7.8 Summary of word recognition for each modality under different environmental noise condition	154
Table E.1 The sentences collected for the LUNA-V corpus	196
Table E.2 Phone coverage	197
Table E.3 Viseme coverage	198
Table F.1 Confusion matrix for visual-only recognition.....	199
Table F.2 Confusion matrix for audio at clean environment.....	200
Table F.3 Confusion matrix for combination of audio and visual at clean environment.....	201
Table F.4 Confusion matrix for audio at 0 dB SNR using babble noise.....	202
Table F.5 Confusion matrix for combination of audio and visual at 0 dB SNR using babble noise	203
Table F.6 Confusion matrix for audio at 0 dB SNR using white noise.....	204
Table F.7 Confusion matrix for combination of audio and visual at 0 dB SNR using white noise	205

LIST OF FIGURES

Figure 1.1 Thesis organization	9
Figure 2.1 Illustration of possible levels of integration.....	12
Figure 2.2 DET curves for various integration strategies [27].....	14
Figure 2.3 Visual front-end processes	16
Figure 2.4 Visual speech features that utilize whether lip geometry, parametric or statistical lip models [15]	18
Figure 2.5 Quality of images in Audio-Visual Data Corpora [66].....	23
Figure 3.1 Block diagram of lip geometry feature extraction	28
Figure 3.2 Graphical user interface	29
Figure 3.3 Face detection using Viola-Jones object recognizer	31
Figure 3.4 Face and mouth detection process	31
Figure 3.5 Example of face and mouth detection process (a) face detection, (b) mouth detection, (c) face region separation and (d) mouth region detection in lower part of face. Operation when (e) a complex background is introduced and (f) when wearing a hat	32
Figure 3.6 Canny edge detection with different threshold and threshold linking.....	33
Figure 3.7 False boundaries with edge detection	34
Figure 3.8 Lip skin detection based using an HSV colour filter (a) original image, and after application of the filter using (b) Hue {5, 40}, (c) Hue {6, 40} and (d) Hue {7, 40}	36
Figure 3.9 Block diagram for lip segmentation process.....	36
Figure 3.10 Illustration of the convex hull process using an ‘origami cat’ diagram	37
Figure 3.11 Convex hull results for speaker ‘s04f’ (left), ‘s05f’ (centre) and ‘s01m’ (right), (a) input colour image, (b) binary lip image and the results of processing following (c) contour detection, (d) largest contour identification and (e) convex hull calculation	38
Figure 3.12 Automatic alignment using left and right vertices	39
Figure 3.13 Shape-based lip features obtained from a single video frame <i>i</i>	39
Figure 3.14 Lip contour detection using snakes method for speaker ‘s04f’	40
Figure 3.15 Lip contour detection using Gradient Vector Field (GVF) method for speaker ‘s04f’	40

Figure 3.16 Example of the grading classification used in the qualitative assessment	42
Figure 3.17 Comparison between manual annotation (top row), ASM technique (centre row) and convex hull technique (bottom row) for subject (a) ‘s28f’, (b) ‘s34f’ and (c) ‘s04f’	43
Figure 3.18 A block diagram for the shape-based lip reading system	45
Figure 3.19 3 state (top), 5 state (middle) and 7 state (bottom) word recognition models	46
Figure 3.20 Phone recognition model	46
Figure 3.21 Performance obtained using shape-based geometrical features for different numbers of Gaussian mixtures.....	48
Figure 3.22 Performance obtained using appearance-based DCT features for different number of Gaussian mixture.....	48
Figure 3.23 Performance of the lip reading systems using different HMM architectures.....	49
Figure 3.24 Comparison of the lip reading system performance using shape-based geometrical features and appearance-based DCT features during head rotation changes	50
Figure 3.25 Automatic head rotation involved in the shape-based geometrical feature extraction process, shown here for subject ‘s01m’ in the CUAVE database. Artificially rotated images (left) and corrected images (right).	51
Figure 3.26 Comparison of the lip reading system performance using shape-based geometrical features and appearance-based DCT features during brightness changes.....	52
Figure 3.27 Brightness information is not involved in geometrical feature extraction. The lip region extracted is the same following the darkening of the image by 20% (top), brightening by 20% (middle) resulting in the same features being extracted (bottom).....	53
Figure 3.28 Brightness information is preserved during RGB to grayscale conversion consequently affecting the appearance-based feature values. Examples are shown for image darkening by 20% (top) and brightening by 20% (bottom).....	53
Figure 4.1 Architecture of the proposed lip reading system	57
Figure 4.2 Dynamic lip information for digit “one” uttered by speaker ‘s01m’ in the CUAVE database.....	58
Figure 4.3 Dynamic lip information showing changes in lip height, lip width and the ratio of height to width for digit ‘five’ obtained from the CUAVE database.....	59
Figure 4.4 Cost matrix with the minimum-distance warp path.....	61
Figure 4.5 Example of distance values that need to be calculated when four reference templates are defined.	64

Figure 4.6 General structure of the classification operations used in the lip reading system	66
Figure 4.7 Models used in the lip reading classification	66
Figure 4.8 Performance of different classifiers using single and combinations of geometrical features	71
Figure 4.9 Confusion matrices using (a) OF, (b) DCT and (c) HWR features	72
Figure 4.10 Mean calculation times of pairwise comparisons of the features in TP-MDTW using samples from speaker ‘s01m’ in session 1. The error bars indicate ± 1 standard deviation for measurements obtained using 1, 10, 20, 40, 80 and 200 features.....	73
Figure 5.1 A block diagram for shape-based feature fusion AVSR system.....	77
Figure 5.2 Block diagram of the feature fusion for AVSR. The algorithm generates time-synchronous 39-dimensional audio $O_{a,t}$ and 5-dimensional visual feature $O_{v,t}$ vectors at 100 Hz rate	78
Figure 5.3 AVSR system performance using geometrical features when ‘babble noise’ is applied. Shown are the noisy audio (A), the visual only information (V), dynamic visual information with delta and delta-delta features ($V + \Delta V + \Delta\Delta V$), the combination of audio and visual (A + V) features and the combination audio and visual with delta and delta-delta features ($A + V + \Delta V + \Delta\Delta V$).....	80
Figure 5.4 AVSR system performance using geometrical features when ‘factory1 noise’ is applied.	81
Figure 5.5 AVSR system performance using geometrical features when ‘factory2 noise’ is applied.	82
Figure 5.6 AVSR system performance using geometrical features when ‘white noise’ is applied.....	82
Figure 5.7 AVSR system performance for digit ‘seven’ using geometrical features when ‘white noise’ is applied.	83
Figure 5.8 AVSR system performance for digit ‘seven’ using geometrical features when ‘babble noise’ is applied.....	84
Figure 5.9 AVSR system performance for digit ‘six’ using geometrical features when ‘white noise’ is applied.	85
Figure 5.10 AVSR system performance for digit ‘six’ using geometrical features when ‘babble noise’ is applied.....	85
Figure 5.11 AVSR system performance using DCT features with babble noise applied. The figure shows a comparison between noisy audio (A), visual only information using 16 DCT features (V_DCT16), 64 DCT features (V_DCT64) and 192 DCT features (V_DCT192), a combination of audio and 16 DCT visual features (A + V_DCT16), a combination of audio and 64 DCT visual features (A + V_DCT64) and a combination of audio and 192 DCT visual features (A + V_DCT192).....	90

Figure 5.12 AVSR system performance using PCA features with babble noise applied. The figure shows a comparison between noisy audio (A), visual only information using 64 PCA feature (V_PCA64), 128 PCA feature (V_PCA128) and 256 PCA features (V_256), a combination of audio and 64 PCA visual features (A + V_PCA64), a combination of audio and 128 PCA visual features (A + V_PCA128) and a combination audio and 256 PCA visual features (A + V_PCA256).....	91
Figure 6.1 Skewness distribution examples	96
Figure 6.2 Kurtosis distributions example.....	97
Figure 6.3 Histograms of digit ‘seven’ uttered by ‘s01m’ in ‘babble’ noise applied at a range of SNR values	99
Figure 6.4 Histograms of digit ‘seven’ uttered by ‘s01m’ in ‘factory1’ applied at a range of SNR values	99
Figure 6.5 Histograms of digit ‘seven’ uttered by ‘s01m’ in ‘factory2’ applied at a range of SNR values	100
Figure 6.6 Histograms of noises.....	100
Figure 6.7 Statistical parameters obtained under various levels of babble noise... ..	102
Figure 6.8 Statistical parameters obtained under various levels of factory1 noise.....	103
Figure 6.9 Statistical parameters obtained under various levels of factory2 noise.....	104
Figure 6.10 Architecture of the adaptive fusion system.....	106
Figure 6.11 The relationship between audio statistical parameters and modality chosen for model 2R	108
Figure 6.12 The relationship between audio statistical parameters and modality chosen for model 3R	109
Figure 6.13 Performance of adaptive fusion AVSR using kurtosis information when simulated under babble noise condition.....	113
Figure 6.14 Performance of adaptive fusion AVSR using kurtosis information when simulated under factory1 noise condition.....	114
Figure 6.15 Performance of adaptive fusion AVSR using kurtosis information when simulated under factory2 noise condition.....	115
Figure 6.16 Comparison of the AVSR performance the adaptive fusion approach and existing fusion methods.	119
Figure 7.1 Sony HXR-MC2000E video camera.....	125
Figure 7.2 Sony ECM-PS1 stereo microphone	125
Figure 7.3 Plan view of the recording studio setup	126
Figure 7.4 Example presentation of the text to be read by the participants	127
Figure 7.5 The principal direct lighting was supplied by daylight fluorescent lamps fitted with a diffusion canvas.....	127

Figure 7.6 Adjustable chair position	128
Figure 7.7 A highlighted section of ambient noise used for background noise removal.....	130
Figure 7.8 Audio signal produced following noise cancellation.....	130
Figure 7.9 Example of the alignment for word-level transcription.....	131
Figure 7.10 Video editing using Power Director 12.....	132
Figure 7.11 Sample frames from each of the 10 LUNA-V data corpus subjects...	132
Figure 7.12 Examples of face and mouth detection using the LUNA-V corpus....	134
Figure 7.13 Lip geometry feature extraction for speaker ‘v01m’ (left column), ‘v09f’ (centre column) and ‘v05m’ (right column), (a) input colour image, (b) binary lip image, and the results of processing following (c) contour detection showing the longest contour identification and (d) application of the convex hull and automatic alignment.....	134
Figure 7.14 Dynamic lip information for digit ‘one’ uttered by speaker ‘v01m’ for the LUNA-V database.....	135
Figure 7.15 Example of the grading classification used in the qualitative assessment	136
Figure 7.16 Comparison of the outcomes of manual annotation (top row) and the application of the convex hull technique (bottom row) for three subjects	137
Figure 7.17 7-state HMM word recognition model.....	139
Figure 7.18 Phone recognition models.....	139
Figure 7.19 Example of a triphone recognition model.....	142
Figure 7.20 Construction of an untrained word using trained phones	143
Figure 7.21 The recognition process involved the matching of words from the test file with models of trained words generated during training.....	143
Figure 7.22 Visual speech recognition results for the individual subjects in the LUNA-V and CUAVE data corpora	145
Figure 7.23 Performance of the geometrical-based AVSR system when ‘babble’ noise is applied and using the LUNA-V corpus. Shown are results when using audio-only data (A), visual-only data (V), dynamic visual information with delta and delta-delta features ($V + \Delta V + \Delta\Delta V$), a combination of audio and visual (A + V) features and a combination of audio, visual, visual delta and visual delta-delta features ($A + V + \Delta V + \Delta\Delta V$).....	146
Figure 7.24 Performance of the geometrical-based AVSR system when ‘factory1’ noise is applied and using the LUNA-V corpus.....	147
Figure 7.25 Performance of the geometrical-based AVSR system when ‘factory2’ noise is applied and using the LUNA-V corpus.....	147
Figure 7.26 Performance of the geometrical-based AVSR system when ‘white’ noise is applied and using the LUNA-V corpus	148

Figure 7.27 Direct comparison of the word accuracy AVSR system performance using the LUNA-V and CUAVE data corpora when 'babble' noise is applied.....	149
Figure 7.28 AVSR system performance using geometrical features when 'babble' noise was applied	151
Figure 7.29 AVSR system performance using geometrical features when 'white' noise was applied	151
Figure 7.30 AVSR system performance using geometrical features when 'factory1' noise was applied.....	152
Figure 7.31 AVSR system performance using geometrical features when 'factory2' noise was applied.....	152
Figure 8.1 Five possible scale invariance lip angle features for robust AVSR in environments where subjects are allowed some freedom of movement	161

CHAPTER 1

INTRODUCTION

Automatic speech recognition (ASR) systems are starting to become an integral part of human computer interfaces (HCI); for example Siri, marketed as the intelligent personal assistant for the iPhone 4S, is able to respond to spoken user requests [1]. In controlled environments, modern ASR systems are capable of producing reliable results, but in many real-world situations the intrusion of acoustic noise adversely affects recognition rates [2]. As many potential ASR users wish to use mobile devices in noisy environments such as vehicles, offices, airport terminals and train stations, solutions that provide reliable operation at high ambient noise levels will become increasingly important.

Humans are often able to compensate for noise degradation and uncertainty in speech information by augmenting the received audio with visual information. Such bimodal perception generates a rich combination of information that can be used in the recognition of speech. The fact that humans use bimodal perception is demonstrated by the ‘McGurk effect’, or as ‘hearing lips and seeing voices’ [3], in which, when a subject is presented with contradicting acoustic and visual signals, perception becomes confused, often resulting in a classification that is different from either the actual audio or visual signal. A well-known example is one of subjects viewing a video in which a speaker mouths ‘gah’, but which is dubbed with ‘bah’. Under such circumstances, most subjects report hearing the sound ‘dah’ [4].

People with hearing impairments may have a reduced ability to receive information in the audio domain and so will rely more heavily on the visual domain for speech recognition. The mechanism employed is often termed either ‘lip reading’ or ‘speechreading’. Lip reading is the ability to understand speech through information gleaned from the lower part of face, typically by following lip, tongue and jaw movement patterns. Speechreading includes lip reading information, but may provide additional means of understanding speech such as interpreting whole face expressions, gestures and body language [5]–[7], as well as employing environmental conditions, such as the specific characteristics of the speaker and the time and physical location at which the conversation took place [8].

When integrating lip reading or speechreading into an ASR system, one of the main issues to address is the selection of the visual features that will be the most advantageous in enhancing recognition performance. Research centres on two different types of feature, namely appearance-based and shape-based. Appearance-based features are used to model characteristics of the mouth region, typically capturing information related to spatial frequencies, whereas shape-based features extract geometrical measurements normally relating to measurements of the lips. In most research work, the area of the face that provides the information most relevant to ASR, namely the lips, is chosen, as this is likely to contain the visual information most closely related to the spoken sounds. Furthermore, the lip movements will normally be highly correlated with the speech sounds themselves, making the integration of visual features with speech features more straightforward.

A suitable method to perform the integration of speech and lip movement features is required in order to achieve good recognition results. Integration can take place either before the model information is processed (feature fusion) or after separate classification (decision fusion). However, which approach is the more effective remains a question yet to be resolved. In this thesis, both integration strategies are investigated under a number of acoustic noise conditions.

1.1 Motivation

Several approaches have been proposed for audio-visual speech recognition (AVSR) systems. The design of such systems depends on the choice of visual features, the classification approach and the speech database used. In [9], the results of visual ASR experiments involving the use of the IBM ViaVoice database were presented in their comparison of four types of visual features, namely discrete cosine transform (DCT) [10], discrete wavelet transform (DWT) [11], principal component analysis (PCA) [12], and active appearance models (AAM) [13]. A solution using hidden Markov models (HMMs) [14] as the classifier found that DCT-based visual features were the most promising for the recognition task.

In [15], both appearance and shape based visual features were obtained using PCA applied to facial animation parameters (FAPs) [16] obtained from outer and inner lip contours that in turn were found by tracking using a combination of a Gradient Vector Field (GVF) [17] and a parabolic template. The experiments showed that under challenging visual conditions (involving changes in head pose and lighting conditions), the lip reading performance of appearance-based visual features suffered. It was also shown that the features obtained from inner-lip FAPs did not provide as much useful information for lip reading as did those obtained from the outer-lip FAPs.

In [6], hue and canny edge detection [18] were used to segment the lip region and shape-based features, including lower and upper mouth width, mouth opening height and the distance between the horizontal lip line and the upper lip were extracted. These features were used in experiments to recognize 78 isolated words using an HMM classifier. Ten subjects from the Carnegie Mellon University database [19] were used to evaluate the performance of the system, with a best classification performance of 46% accuracy being attained when all the geometrical information and difference (delta) features were included and when operating in speaker-dependent mode. The performance was found to fall to 21% in the speaker independent case.

In [20], the lip region was located using a Bayesian classifier [21] that held estimates of the Gaussian distributions of face, non-face and lip classes in the red, green and blue colour space. The researchers then obtained visual features, namely the affine-invariant Fourier descriptors (AIFDs) [22], the DCT, the rotation-corrected DCT (rc-DCT) and the B-Spline template (BST) [20]. The results obtained using the appearance-based features, DCT and rc-DCT, were better than those achieved using the shape-based features, AIFDs and BST, and the authors concluded that this was due to their greater sensitivity to lip shape.

In [23], the authors proposed an appearance-based lip reading approach that generated dynamic visual speech features, termed the Motion History Image [24], that were classified using an artificial neural network. The approach captured movement in image sequences and generated a single grayscale image to represent the whole image sequence using accumulative image subtraction techniques. However, this approach proved highly sensitive to environmental changes. In addition, information about the timing of movements was lost following the combination of sequences into a single image, resulting in a consequential degradation of performance. In [25], the authors reported a technique that computed the optical flow (OF) of lip motions in a video data stream. The statistical properties of the vertical OF component were used to form feature vectors suitable for training a support vector machine classifier. However, as is the case for OF methods in general, the performance was adversely affected in practical cases due to its sensitivity to scaling and rotation of the images.

The literature suggests that appearance-based features are generally able to produce better classification results as they carry more information, but also because of the complexity of extracting accurate geometrical features when using shape-based approaches [20]. However, the appearance-based features exhibit a greater sensitivity to environmental condition changes such as illumination and head pose [15]. In general, there is a need to develop an approach that is reliable; one possible approach is to investigate approaches to improve the performance of shape-based

methods while maintaining their advantage of their inherent robustness in the face of changing environmental conditions.

Although the performance of an AVSR system relies heavily on the choice of visual features, classification approach and the database used, the fusion strategy adopted to combine the audio and visual modalities has a very significant effect on recognition performance. Several fusion approaches have been proposed in the literature, but these can be categorized into two major groups, namely feature fusion and decision fusion. Feature fusion for AVSR has been previously used [9], [26], [27], and have the benefit that they model the dependencies between audio and visual speech information directly. However, this approach suffers in two respects. Firstly, due to the both types of information being combined at early stage into single vector and before the classification itself, if either the audio or visual information become corrupted then so does the entire vector. Secondly, Lavagetto [28] demonstrated that acoustic and visual speech production are not synchronous, at least at a feature based level. It was shown that, during an utterance, visual articulators such as the lips, tongue and jaw perform movements both before the start and after the end of an acoustic utterance. This time delay is known as the voice-onset-time [29], defined as the time delay between the movement of the vocal folds for the voiced part of a voiced consonant or subsequent vowel and the burst sound coming from the plosive part of a consonant.

The literature has widely reported superior results for decision-fusion AVSR systems compared to those obtained for feature fusion [6], [9], [15], [27]. Decision fusion allows the synchronous classification of the audio and visual modalities and has the flexibility to allow the relative weightings of the modalities to be altered for final classification. However, a major drawback of this approach is that the fusion itself normally only takes place at the end of the utterance being recognized, which, compared to the feature-fusion case, can lead to a delay in generating the classification result and so make interactive sessions appear unnatural.

In the research community, opinions remain divided as to which is the more effective of the two fusion strategies in terms of speech recognition performance. Decision fusion generally appears to be favoured for in the implementation of an AVSR system under noisy environmental conditions, for the following two reasons. Firstly, decision fusion allows the modelling of AVSR systems asynchronously, since the audio and visual information are processed independently. Secondly, as decision fusion often delivers partial classification decision outcomes, it is able to provide a basis for their ranking and collation. Adaptive weights can then be applied to adjust the relative contributions of each partial outcome for making a final decision.

1.2 Aim and objectives

The aim of the research in this thesis is to improve the performance of automatic speech recognition systems by incorporating dynamic visual information from the mouth region. The objectives of this research are listed below.

- Develop an automatic feature extraction technique that is able to extract lip geometry information from the mouth region.
- Analyse the classification performance using a range of lip geometry features and determine which individual feature or which combination of features performs the best in representing speech in the visual domain.
- Design a state-of-art audio-visual speech recognition system using dynamic geometry features obtained from the lip shape.
- Evaluate the robustness of the audio-visual speech recognition system in noisy environments using a range of candidate integration strategies.

1.3 Original contributions

Several contributions to the field of AVSR have been made in the research work and are listed as below.

- A new method has been established that is able to extract automatically lip geometry information such as height, width, ratio, area and perimeter from the mouth region by utilizing a skin colour filter, a border following technique and the convex hull approach. This method is more reliable and requires less computation in extracting lip geometry features compared to conventional methods which generally use either the active contour or the active shape model. The results of this work were presented at IEEE Visual Communications and Image Processing Conference in San Diego, USA in November 2012 [30]. Details of the work can be found in Chapter 3.
- A demonstration has been produced of the robustness of the new lip geometrical features when affected by head rotation and brightness changes. The performance of the geometrical-based method remained consistent, while the appearance-based approach was adversely affected by the changes in environmental conditions. The results of this work were presented at the IEEE EUROCON 2013 conference in Zagreb, Croatia in July 2013 [31]. Details of the work can be seen in Chapter 3.
- A novel template probabilistic multi-dimension dynamic time warping (TP-MDTW) technique has been introduced to calculate the probability of each template being the best match to an unseen example based on the similarity with templates in a database. The assumption is that a template having the greatest similarity to other templates should be recognized as the most probable to occur and those templates having least similarity are less likely to occur. The results of this work have been accepted by the Journal of Visual Communication and Image Representation (Elsevier). Details of the work are in Chapter 4.

- A solution has been proposed to the ‘curse of dimensionality’ issue in the feature fusion based AVSR system and has been achieved by obtaining a small set of simple and efficient geometrical features that have a highly descriptive information content for the recognition task. The results of this work were presented at the IEEE International Symposium on Communications, Control, and Signal Processing 2014 in Athens, Greece in May 2014[32]. Details of the work can be found in Chapter 5.
- A novel adaptive fusion method has been introduced to select decision outcomes from the audio and video modalities by assessing the audio noise content using skewness and kurtosis values. The proposed system is able to select a preferred classification modality dependent on the estimated audio noise in the system. Compared to conventional feature-fusion and decision-fusion methods, the proposed method is able to follow closely the better performer from audio-only and video-only modalities across all levels and types of noise. Details of the work are presented in Chapter 6 and a journal paper is in preparation.
- A new data corpus termed the Loughborough University audio-visual (LUNA-V) speech corpus has been developed, whose video is of higher definition than those currently made available by other researchers. The corpus consists of 10 speakers each uttering 10 isolated digits and five sentences, with the sentence design adopted from the CUAVE and TIMIT databases. The new data corpus allows the validation of the method developed earlier in the thesis, not only by having a second source of images, but also by being able to assess whether features obtained to a better resolution can improve recognition performance. The LUVA-V data corpus has been made available to other researchers in the field. Details of the work can be found in Chapter 7 and a journal paper is in preparation.

1.4 Thesis organization

This thesis is organized as follows. Chapter 2 provides background information on the techniques and approaches used in this work, including audio-visual architectures, visual front-ends and classification methods. Chapter 3 introduces a new method to extract lip geometry features from video sequences and its performance is evaluated under simulated changes in environment conditions that arise from head movements and variations in image illumination. Chapter 4 presents a novel lip-reading technique that is able to adapt to the dynamic differences in the manner words are uttered by speakers using template training probabilities. Chapter 5 describes the effect of scalability in feature fusion AVSR using lip geometry features and a comparison is made with results obtained using appearance-based features. Chapter 6 presents a novel decision fusion based AVSR system that measures the quality of the audio under a range of different signal to noise ratio conditions. Chapter 7 details the new data corpus developed in order to validate the methods proposed. Chapter 8 concludes this thesis and presents possible future paths of research. Figure 1.1 shows an AVSR system and how the thesis is organized.

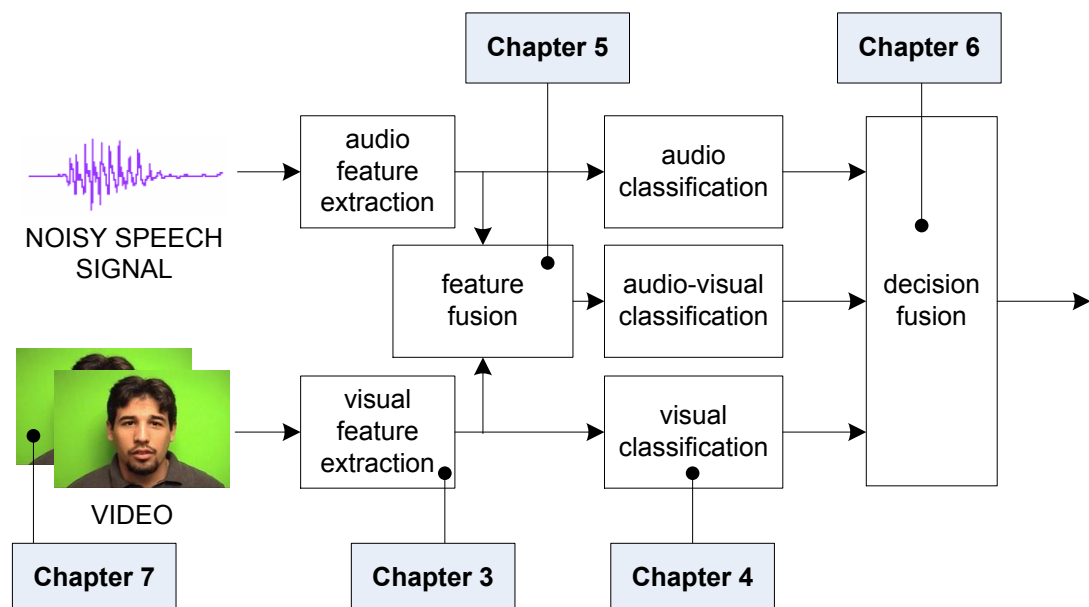


Figure 1.1 Thesis organization

CHAPTER 2

AN OVERVIEW OF AUDIO-VIDEO SPEECH RECOGNITION SYSTEMS

The human perception of the world is inherently multisensory since the information provided is multimodal. The perception of a spoken language is not an exception. In addition to the auditory information, there is visual speech information provided by the facial movements as a result of moving the articulators during speech production [33]. The complementary nature of visual and auditory perceptions in speech comprehension has influenced several ASR research teams to concentrate on augmenting speech recognition by adding a following visual channel. This chapter details the background and the terminology used in audio-visual speech recognition system. Firstly, the basic architecture of AVSR system is discussed in Section 2.1. Secondly, Section 2.2 discusses the audio and visual feature extraction techniques. Thirdly, mapping between a basic unit of audio and video is discussed in Section 2.3. Finally, the summary of audio-visual data corpora is addressed in Section 2.4.

2.1 Architecture

There remains a lack of clarity regarding the use of terminology pertaining to the levels of integration in AVSR. It is widely agreed [12], [34], [35] that the acoustic and visual speech modalities can be integrated either at the early or late processing and these two integration paradigm are termed feature-fusion and decision-fusion respectively. However, the interpretation of fusion strategies varies markedly depending on perspective and the task at hand.

For example, in continuous audio-visual speech applications, Dupont and Luetttin [12] interpret decision-fusion as combining scores at the sentence level. However, in earlier literature [34], [36], [37], specifically for the tasks of isolated word recognition, decision-fusion is interpreted as combining scores at the word unit level. Similar ambiguity exists in defining feature-fusion. feature-fusion approach suggests that visual speech information is converted to a vocal tract function [34] with the acoustic and visual transfer functions being averaged during integration. Alternatively feature-fusion can be interpreted [12], [34], [35] as the concatenation of acoustic and visual stimuli for processing as a single observation. In addition to feature-fusion and decision-fusion, a further level of integration, namely middle integration (MI) [27], has been defined to allow for a varying degree of temporal dependence between acoustic and visual modalities during testing, whilst still allowing the benefits of training the modalities independently. Although distinct from feature-fusion and decision-fusion, MI can be thought equivalent to decision-fusion for some specific cases. Thus it can be summarized that AVSR architecture can be categorized into three broad levels of integration as depicted in Figure 2.1 [27].

- Feature fusion, in which acoustic and visual speech stimuli are synchronized and merged in some manner (e.g. concatenation or averaging of vocal tract functions) for joint learning and classification. This approach assumes there is direct dependence between the acoustic and visual modalities at the lowest levels of human speech perception.
- Middle integration (MI) attempts to integrate the acoustic and visual speech modalities at a somewhat higher level than feature-fusion and attempts to learn and classify acoustic and visual speech cues independently. However, during the classification of an utterance there may be temporal dependence between modalities.
- Decision fusion assumes complete independence between acoustic and visual speech modalities. During the classification process there is no interaction between the modalities, with only the final classifier confidence scores being combined. In this approach, temporal coherence between speech modalities is lost, except at the anchor points at which decisions are combined (typically the word unit level).

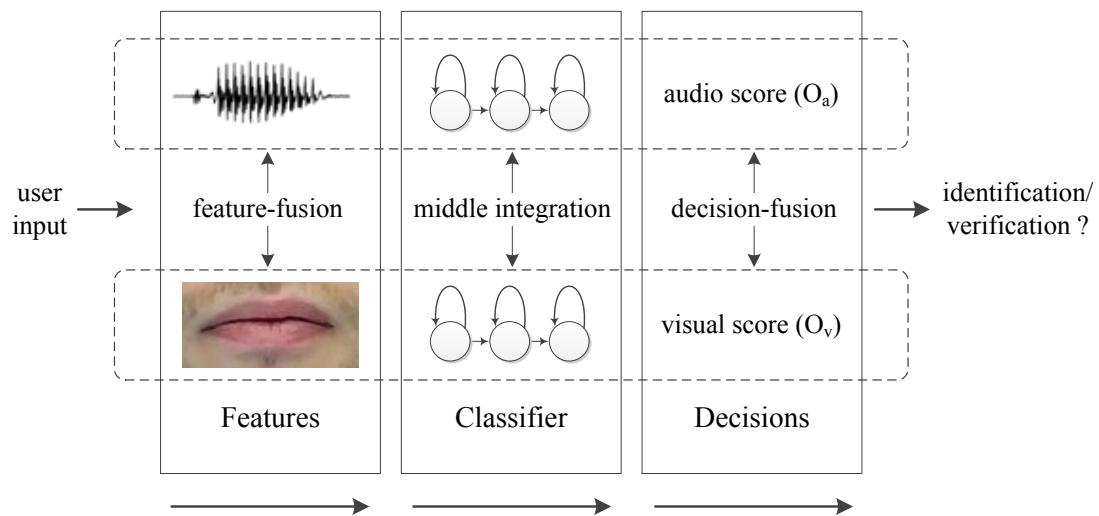


Figure 2.1 Illustration of possible levels of integration

Feature-fusion techniques for audio-visual speech recognition [34], [36] have been previously used, and have benefit as they model the dependencies between acoustic and visual speech modalities directly. However, feature-fusion approaches suffer in two aspects. First, due to the inherent low-level classification either the acoustic or visual speech modalities become corrupted then so does the entire speech modality. Second, assuming HMM classifier are employed there is the need to ensure that the acoustic and visual speech modalities are synchronized at the state level, a requirement that is difficult to fulfill in practice.

Lavagetto [28] demonstrated that acoustic and visual speech stimuli are not synchronous, at least at a feature based level. It was shown that, during an utterance, visible articulators begin and complete their trajectories asynchronously, exhibiting both forward and backward coarticulation with respect to the acoustic speech wave. In particular, visual articulators such as the lips, tongue and jaw perform movements both before the start and after the end of an acoustic utterance. This time delay is known as the voice-onset-time (VOT) [12], defined as the time delay between the movement of the vocal folds for the voiced part of a voiced consonant or subsequent vowel and the burst sound coming from the plosive part of a consonant. McGrath and Summerfield [38] found an audio lead of less than 80ms or a lag of less than

140ms could not be detected during speech, but if the audio was delayed by more than 160ms it no longer contributed useful information, signifying the importance of a reasonable degree of asynchrony and synchrony in continuous audio-visual speech perception.

Decision-fusion is able to avoid these problems. For automated isolated word applications, decision-fusion strategy have reported superior results compared to feature-fusion [34], [36], [37], [39]. Decision-fusion allows for the asynchronous classification of speech and can emphasize or deemphasize the importance of a modality in classification depending on the relative quality of the two signals. However, decision-fusion has not proved to be as effective in continuous speech applications where integration is attempted at length of speech greater than the word unit level [12]. Waiting until the end of the spoken utterance before combining modalities, as the decision-fusion strategy was perceived by Dupont and Luetttin [12], introduces a time delay that is not suitable for natural communication. In order to overcome this, some form of synchrony is required.

MI is the third option and is able to provide synchronization while still providing a framework for guarding against corruption in either modality. The MI integration strategy can be naturally model by multi-stream HMMs [27] as they provide relative independence between streams statically with a loose temporal dependence dynamically. If the streams are supposed to be entirely synchronous and represented by HMMs with the same topologies, they may be accommodated simply. However, it is often the case that the streams are not synchronous, that they do not even have the same frame rate and it might be necessary to define models that do not have the same topology. MI based approaches have been used to great success in continuous audio-visual speech applications [12], [39], [40].

In order to determine which format integration perform the best, Lucey *et al.* [27] performed experimental work to compare various integration strategies. In this work, HMMs were used to model audio-visual utterances while the M2VTS database was employed for all experiments. HMMs are excellent for modeling

bimodal speech as they provide a natural way to stochastically capture the temporal fluctuations of speech in each modality and are able to naturally incorporate the different levels of integration. Training for the feature-fusion strategy involved the synchronization of the acoustic and visual features. Both acoustic and visual features were concatenated into one feature vector, which was used to train a single joint audio-visual HMM. For the MI and decision-fusion strategies, two separate independent acoustic and visual HMMs were trained independently. Figure 2.2 shows a detection error tradeoff (DET) curve that used to represent the tradeoff between the false acceptance (FA) and false rejection (FR) errors by varying the threshold. Overall performance of the verification system can be determined by equal error rate (EER) of the system where this is the point where the FA and FR error rates are equal. It shows that the superiority of the decision-fusion strategy over all other integration strategies. For the identification task it can be seen in Table 2.1 where it shows that decision-fusion strategy (av-product) and middle integration (av-async) outperformed other strategies with the an EER of 1.15% and an identification rate of 96.57%.

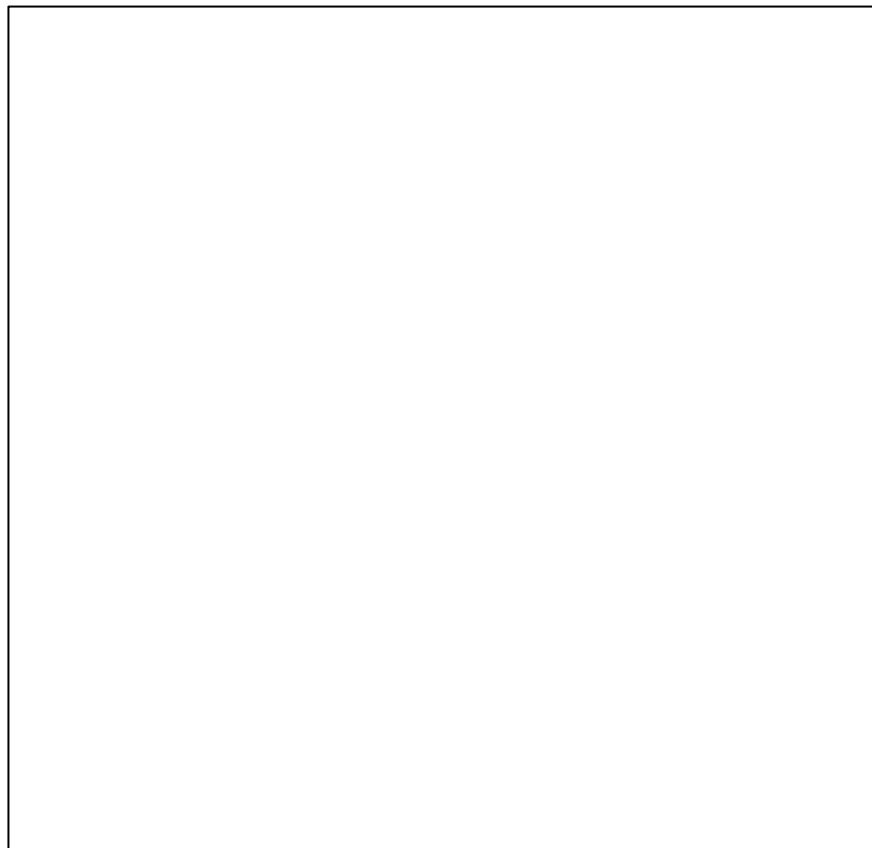


Figure 2.2 DET curves for various integration strategies [27]

Table 2.1 Equal error rates (EER) and identification rates (IR) for integration strategies (best performing strategies are highlighted) [27]

Modality	Integration	EER (%)	IR (%)
audio (a)	none	2.00	90.28
video (v)	none	7.43	78.00
av	feature-fusion	4.85	83.43
av (async)	middle	1.15	96.57
av (sync)	middle	1.74	91.14
av (sum)	decision-fusion	1.43	92.28
av (product)	decision-fusion	1.15	96.57

2.2 Audio and visual feature extraction

Feature extraction play very important part in the AVSR system research field. If the extracted features carefully chosen from the audio signal and speaker's image, it is expected that relevant information would perform an efficient recognition.

2.2.1 Audio feature extraction

To help prepare the incoming audio signal for feature extraction stage, pre-processing techniques such as signal filtering and audio enhancement must be made in advance. Several results have been reported in the literature regarding the audio feature extraction techniques [14]. Mel-frequency cepstral coefficients (MFCCs) [41] and linear prediction coefficients (LPCs) [42] represent the most commonly used audio features in last few decades. There are still on-going research in the field of robust audio features and such features will not be considered in this work.

MFCCs is very popular and has been shown to outperform others feature extraction techniques as revealed in [41]. MFCCs are derived from a Mel-frequency where this frequency axis is warped according to the Mel-scale, which approximate the human auditory system's response. The dynamic features which are first (delta-MFCCs) and second time-derivatives (delta-delta-MFCCs) of cepstral coefficients is now commonly employed to improve speech recognition performance [43], [44].

2.2.2 Visual feature extraction

Before being applied to the classifier for training or recognition purposes, visual information need to be pre-processed to generate relevant data that describes certain characteristics suitable for classification purposes. These pre-processing stages of the video data are known as the visual front-end. It involves the detection of the mouth regions and the extraction of visual lip features as shown in Figure 2.3.

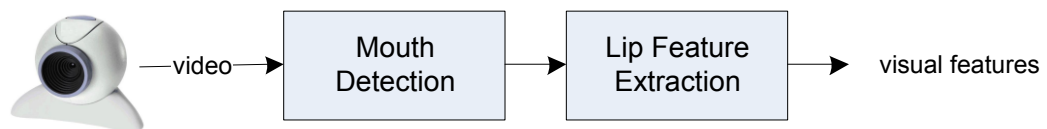


Figure 2.3 Visual front-end processes

The first task to be solved in visual feature extraction process consists of the automatic detection and tracking of the mouth region of interest (ROI). Mouth detection is applied when there is no previous information of the mouth position available. This will occur in the first frame of a sequence or whenever the mouth has not been correctly located in the previous frame, otherwise mouth tracking will proceed using knowledge of the mouth position from the previous frame and the position of the mouth remains little changed. Mouth tracking tends to be preferred to mouth detection in subsequent frames as it is generally more reliable and less computationally intensive.

A number of reported systems use traditional image processing techniques for mouth detection, such as color segmentation, edge detection, image thresholding, template matching or motion information in image sequences [45], taking advantage of the fact that many local facial sub-features are approximately rigid and contain well define edges. An alternative popular approach is to use a cascade of weak classifiers that are trained using the AdaBoost technique [46]. Once the ROI is located, a number of algorithms can be used to obtain lip features.

The choice of lip visual features and their robust extraction from a video sequence is important consideration in the design and implementation of AVSR. Several sets of visual features have been proposed in the literature in recent years as shown in Figure 2.4, and these can be grouped into three categories [47].

- Appearance-based features, such as the transformed vectors of the mouth region into pixel intensities.
- Shape-based features, such as geometric or model-based representations of the lip contour.
- Features that are a combination of both appearance and shape.

The appearance-based features from the mouth ROI are normally considered informative for speechreading. Such a ROI is typically defined as a rectangular window containing the mouth and possibly including additional parts of the lower face, such as the jaw and cheeks [48], or even the entire face [47]. A series of ROI is normally captured in order to represent video speech sequence information [11].

By concatenating the ROI pixel values, a feature vector is obtained that is expected to contain the required visual speech information. However, the dimensionality of the ROI vector becomes prohibitively large for successful statistical modeling of the classes of interest, such as sub-phonetic classes using hidden Markov models [14] for AVSR. For example, in the case of a 128×128 pixel greyscale ROI, one feature vector has total of 16384 dimensions, and for a colour mouth region image in RGB space, the same image size has 49152

dimensions. Consequently, transformation to lower dimensionality is required in order to reduce the number of features for computational purposes. The method for implementing such a transformation is typically borrowed from the image compression and pattern classification literature. Examples of such transforms are principal component analysis (PCA), also known as ‘eigenlips’, used in the literature for speechreading [11], [12], [49], [50], the discrete cosine transform (DCT) [11], [50]–[52], the discrete wavelet transform (DWT) [11], linear discriminant analysis (LDA) [48] and the maximum likelihood linear transform (MLLT) [48], [53]. The advantage of appearance-based features are that they allow visual speech representation in real-time systems, however their performance degrades under significant head-pose and illumination variations [54].

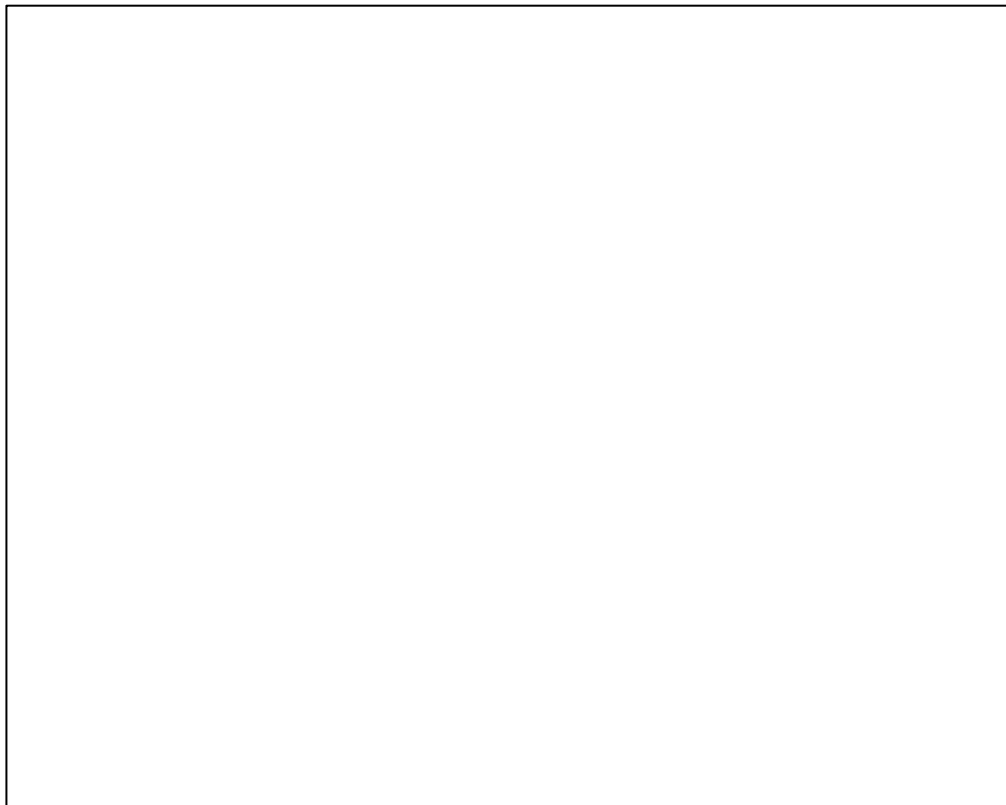


Figure 2.4 Visual speech features that utilize whether lip geometry, parametric or statistical lip models [15]

In contrast to appearance-based features, shape-based feature extraction assumes that most speechreading information is contained in the shape (inner and outer contours) of the speaker's lip, or, more generally, in the contours extracted from the face as a whole [47]. As a result, such features achieve a compact representation of visual speech using low-dimensional vectors and are largely invariant to head pose and illumination. However, in order to ensure good performance, their extraction requires robust lip-tracking, which often proves difficult and computationally intensive in realistic scenarios.

Shape-based visual features can be defined as either geometric or model-based. Geometric features that are meaningful to humans can be extracted from the inner and outer contours of the lip, such as the height, width, perimeter, and area within the contour. Such features contain significant visual speech information, and have been successfully used in speechreading [6], [55]. Alternatively, model-based visual features can be obtained in conjunction with one of the parametric or statistical lip-tracking algorithms. Some popular methods for this task are edge detection [6], [20], active contour or 'snakes' [56], templates [57], gradient vector field (GVF) [17], active shape model (ASM) [13] and active appearance models (AAM) [13]. A 'snake' is an elastic curve represented by a set of control points and is used to detect important visual features, such as lines, edges, or contours. The snake control point coordinates are iteratively updated, converging towards a minimum of the energy function, defined on basis of curve smoothness constraints and a matching criterion to the desired features of the image [56]. Templates are parametric curves that are fitted to the desired shape by minimizing an energy function, defined similarly to snakes. In contrast, ASMs are statistical models obtained by performing PCA on vectors containing the coordinates of a training set of points that lie on the shapes of interest, such as the eyes, nose, and mouth contours. Such vectors are projected onto a lower dimensional space defined by the eigenvectors that correspond to the largest PCA eigenvalues, representing the axes of genuine shape variation. AAMs are an extension to ASMs with two more PCAs, where the first captures the appearance variation around the ROI and the second PCA is built on concatenated weighted vectors of the shape and appearance representations. AAMs thus remove the redundancy that would arise due to shape

and appearance correlation and they create a single model that compactly describes shape and the corresponding appearance deformation. ASMs and AAMs can be used for tracking lips or other shapes by means of the algorithm proposed in [13], in which, given small perturbations from the actual fit of the model to a target image, a linear relationship exists between the difference in the model projection and image and the required updates to the model parameters. Fitting the models to the image data can be done iteratively, as in [47], or by the downhill simplex method, as in [12].

A number of drawbacks have been observed for shape-based methods in application to the image fitting process. For example, there is no convergence criterion for ASM during the fitting process and poor results can be obtained [58]. Also, in AAM, the PCA reconstruction error is used as a distance measure for the evaluation of alignment quality although this may not be a good discriminant for such a purpose, for example, regions that do not contain a face can look like a face when projected into the PCA face subspace [58].

2.3 Audio and visual mapping

The smallest structural unit that distinguishes meaning that present in speech is phoneme [59]. Phonemes are not the physical segments themselves, but an abstract concept, defined by the ability to distinguish one word of the lexicon from another in a minimal sense. Hence, it is an equivalence class of sounds rather than a sound itself. On the other hand, phones are the speech sounds that refer to the instances of phonemes in the actual physical segments. However, since only a small part of the vocal tract is visible, not all phones can be disambiguated solely by video information. Visually distinguishable units are called visemes [60]–[62] and consist of phone clusters derived from human speechreading studies, or are generated using statistical techniques [63], [64].

An example of a phone-to-viseme mapping is depicted in Table 2.2. Although some visemes are well-defined, such as the bilabial viseme consisting of the phone set $\{/p/, /b/, /m/\}$ but there is no universal agreement about the exact partitioning of phones into visemes. There are many acoustic sounds that are visually ambiguous and can be grouped into the same class that represents a viseme. For example, the $/p/, /b/,$ and $/m/$ phones are all produced by a closed mouth shape and are visually indistinguishable, and so form a single viseme group. Similarly, $/f/$ and $/v/$ both belong to the same viseme group that represents a mouth in which the upper teeth are touching the lower lip. Viseme information can be captured either from single mouth image or a sequence of several images that capture the movements of the mouth. Some vowels that cannot be distinguish in single image, required a sequence of images for their identification.

Table 2.2 The Phone to viseme mapping [65]

Viseme class	Phones in cluster
Silence	$/sil/, /sp/$
Lip-rounding based vowels	$/ao/, /ah/, /aa/, /er/, /oy/, /aw/, /hh//uw/, /uh/, /ow//ae/, /eh/, /ey/, /ay//ih/, /iy/, /ax/$
Alveolar-semivowels	$/l/, /el/, /r/, /y/$
Alveolar-fricatives	$/s/, /z/$
Alveolar	$/t/, /d/, /n/, /en/$
Palato-alveolar	$/sh/, /zh/, /ch/, /jh/$
Bilabial	$/p/, /b/, /m/$
Dental	$/th/, /dh/$
Labio-dental	$/f/, /v/$
Velar	$/ng/, /k/, /g/, /w/$

In both the acoustic and visual modalities, most vowels are distinguishable. However, this is not true for some consonants; illustrated by the common use of the phonetic alphabetic to confirm individual letter during telephone conversation. In many cases, such confusions in the auditory modality can be distinguished in the visual modality. For example ‘C’ can be easily distinguished from ‘U’ by the visual cue of a closed mouth rather than an open mouth. Therefore, for speech

understanding, if information about lip movements are extracted from the video of a speaker, it can be utilized to improve overall performance.

2.4 Audio-Visual speech corpora

In AVSR research, the quality of the data corpora chosen can greatly influence the research results obtained. As the AVSR field is still in its infancy and it takes additional time and resources to develop a multi-modal audio-visual database, the number of available multi-modal audio-visual data corpora is small compared to the number of single modal audio only datasets.

Currently there are two main application areas in which audio-visual datasets are put to use, namely audio-visual speech recognition (TULIPS1, AVletters, AVOZES, CUAVE, VidTIMIT, IBM LVCSR and DUTAVSC) and speaker recognition (M2VTS and VidTIMIT). From the point of view of speech recognition, the common limitations that may be exhibited by audio-visual datasets are listed below [66].

- The recordings contain only a small number of respondents. This greatly reduces the generality of the results as it is unlikely that there is substantial variation in the sex, ethnicity and age of the respondents.
- The utterances are usually very limited. The datasets often contain only isolated words or digits or even only the letters of the alphabet rather than continuous speech, giving poor coverage of the set of phones and visemes of a language.
- The quality of the recording is very poor. The images may not have been captured using modern high definition equipment or the scene may be poorly or unevenly illuminated. Poor quality images make the task of isolating the lips, skin and face regions more difficult.
- The datasets are not publicly available. Many datasets that are reported in scientific papers are not open to the public for example ‘IBM LVCSR’, making independent verification of results impossible.

Figure 2.5 shows the mouth regions captured from a number of popular publicly available datasets [66]. In each case the mouth area has been determined manually such that the bounding box touches the lower and upper lips and the left and right corners of the mouth. Subjectively, the image of best resolution and most even illumination is that of the DUTAVSC dataset, whereas in the VidTIMIT dataset the image resolution and contrast is relatively poor.

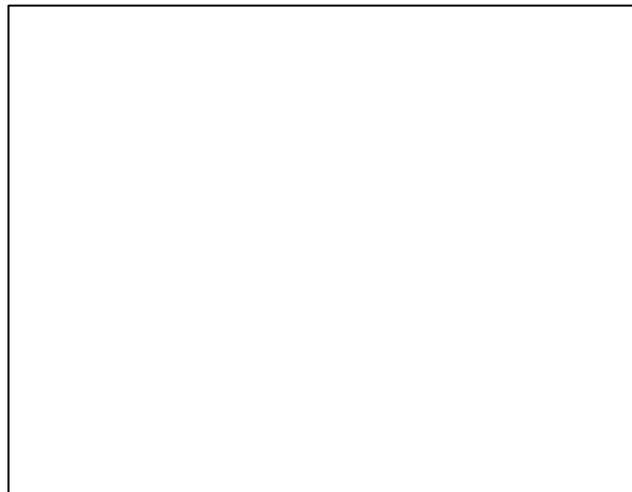


Figure 2.5 Quality of images in Audio-Visual Data Corpora [66]

Table 2.3 lists the most commonly used AVSR data corpora and, for each corpus, gives the audio and video characteristics as well as an indication of the contents of the database. The structure of the table is presented so as to emphasize the quality of the datasets with respect to video, audio and language. In all cases, only the sex of the speaker is recorded.

Table 2.3 Comparison of popular AVSR data corpora

Corpus	Subjects	Audio Quality	Video Quality	Reference
Tulips1	7 male 5 female	11.1kHz, 8bits controlled audio	100x75, 30fps mouth region	[67]
AVletters	5 male 5 female	22kHz, 16bits controlled audio	80x60, 25fps mouth region	[68]
AVOZES	10 male 10 female	48kHz, 16bits controlled audio	720x480, 29.97fps entire face, stereo view	[69]
CUAVE	19 male 17 female	44kHz, 16bits controlled audio	720x480, 29.97fps passport view	[20]
VidTIMIT	24 male 19 female	32kHz, 16bits controlled audio	512x384, 25fps upper body	[70]
AVICAR	50 male 50 female	48kHz, 16bits car environment	720x480, 29.97fps 4 cameras from different angles	[71]
DUTAVSC	7 male 1 female	48kHz, 16bits controlled audio	384x288, 25fps lower face view	[72]
M2VTS	25 male 12 female	48kHz, 16bits controlled audio	286x350, 25fps passport view	[73]

Note. All data corpora are in English language except DUTAVSC in Dutch and M2VTS in French

2.5 Chapter summary

This chapter has given an overview of techniques presented in the literature on the various aspects of AVSR system. First, an overview has been presented on the architecture of AVSR system and the investigation into which fusion strategies are most efficient for audio and visual modality. Although decision-fusion has an advantage due to its asynchronous classification and the ability to emphasize or deemphasize the importance of each modality in classification, it is not proved to be as effective in continuous speech recognition.

Next, the literature on audio and visual feature extraction techniques have been presented. While well-established parameters exist for the audio (features from spectrums analysis), it is not clear which features describe the best in visual domain. The choice of visual features, either appearance-based or shape-based and their robust extraction from a video sequence is important consideration in the design and implementation of AVSR system. Finally, the discussion included an overview of audio-visual data corpora has been presented. The quality of the data corpora that chosen can greatly influence the recognition results. It should be noted that common data corpora to enable close comparison of various results reported are still not available in the field of AVSR.

In conclusion, AVSR system has been explored by a number of researchers in the last few decades. The literature review has shown that several areas in AVSR field need further investigations. Among them are the robust shape-based feature extraction of visual speech information from lip region and the investigation of the appropriate fusion of audio and visual features. These two areas are investigated in this study. A novel lip geometry feature extraction technique is presented which is based on skin colour filter, border following method and convex hull. Geometrical features are robust to the environmental changes (head rotation and brightness changes) and proved to have less suffering from the ‘curse of dimensionality’ (or scalability) issue that is often a bane in feature-fusion based. A novel adaptive fusion system based on audio statistical analysis is introduced. Skewness and kurtosis

analysis is used to measure the quality of the audio based on its distribution and automatically choose between audio and visual modality for best recognition. As current data corpora contain lack of visual quality where most of them having poor resolution for shape-based feature extraction, the new LUNA-V data corpus is created. Using high definition video recording, the quality of the image captured able to provide rich information about the face features especially mouth area that would help improving visual recognition system.

CHAPTER 3

LIP GEOMETRY FEATURE EXTRACTION USING SKIN COLOUR FILTER, BORDER FOLLOWING METHOD AND CONVEX HULL

This chapter details the process that has been implemented to extract the lip geometry features from the mouth region. Firstly, a brief justification of the approach taken is given in Section 3.1. Secondly, the methodology including the software developed, experimental setup and database structure are discussed in Section 3.2. Thirdly, the new approach to extract lip geometry using skin colour filter, border following method and convex hull are detailed in Section 3.3 to Section 3.5 respectively. Finally, the performance of this approach is addressed in Section 3.6 and 3.7.

3.1 Introduction

The literature suggests that those AVSR approaches that adopt appearance-based features produce better experimental results than those based on shape-based features, principally because the former method carries more information and the latter method has the drawback of the practical difficulty of obtaining good quality geometrical features from the lip shape [20]. Generally, appearance-based features are not reliable as their performance is highly sensitive to changes in environmental conditions such as illumination and head pose [15]. The motivation for the work described in this chapter is the development of a reliable approach that is able to extract lip geometry yet is able to generate performance results that are similar to those achieved by appearance-based methods. There are three operational stages that need to be carried out in order to extract geometrical features from the lip region. Firstly, a method is required to locate the position of the mouth in an image;

secondly a method is required to segment the lip from the non-lip region and thirdly suitable geometrical lip features need to be selected. These components for lip geometry feature extraction are shown in Figure 3.1.

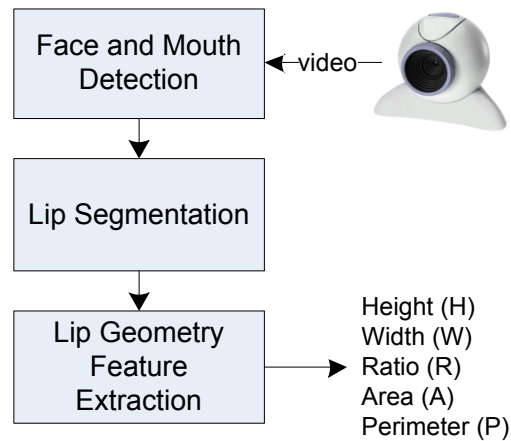


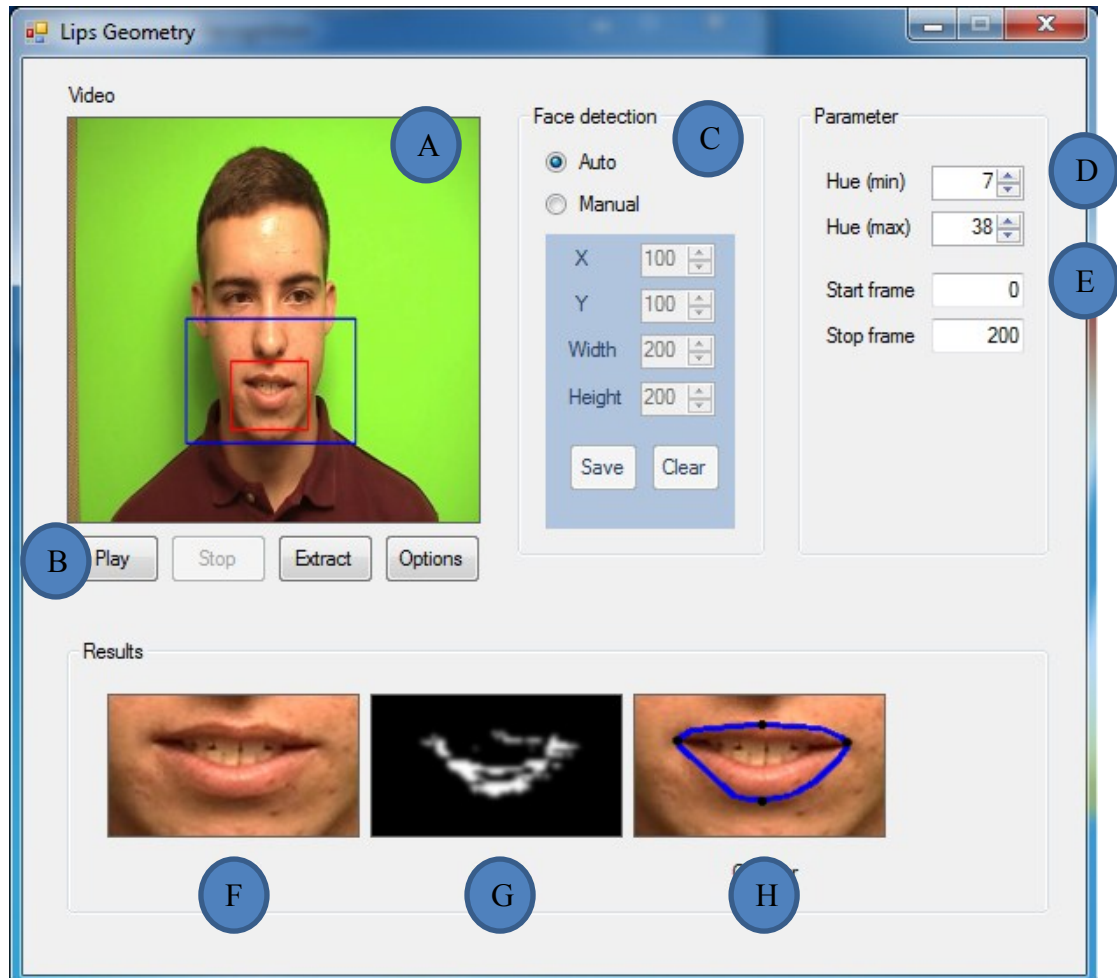
Figure 3.1 Block diagram of lip geometry feature extraction

3.2 Software, experimental setup and database

The software for this work was developed using Microsoft Visual C# 2010 [74] and utilized the open source image processing library OpenCV [75]. In order to analyse the effectiveness for this method in practical applications, the CUAVE corpus database [20] was used. This database consists of 36 individuals, where 19 speakers are male and 17 speakers are female. Some of the speakers have glasses, facial hair or wear hats. The speakers also have a variety of skin and lip tones, as well as a range of face and lip shapes. Lighting was controlled and a green background was used to allow chroma-keying of new backgrounds. The videos were recorded at 29.97 frame/s at a resolution of 720 x 480 and encoded in MPEG-2 files at data rate of 5000 kbit/s.

In order to visualize the data provided by the experiments, a new data visualization tool needed to be developed, as no-off-the shelf program was found to be suitable. Figure 3.2 shows a screenshot of the program that has been developed in

this work and it is able to display the video, face and mouth region, lip contour and lip geometry features extracted as well as the parameters to adjust the face and mouth detection, and skin colour (hue) threshold.



- A – Video from CUAVE database
- B – Video control panel
- C – Face detection options
- D – Skin detection options
- E – Length of video to process
- F – Mouth region
- G – Skin segmentation
- H – Lip shape

Figure 3.2 Graphical user interface

3.3 Face and mouth detection

The speaker images as acquired from the CUAVE database are cropped to the mouth region using a face-detection process followed by a mouth-detection process. Many techniques are available for extracting the face and mouth region, as discussed in Section 2.2.1 and the method chosen for the current work involves a machine learning technique integrated with a human knowledge base.

The method adopted here was the Viola-Jones object recognizer [46], [76], that uses simple rectangular Haar features [76] and is applied to each image in a wide range of translations and at many different scales. To select specific Haar features, the AdaBoost [77] technique is used to train a weak classifier. Single strong object classifiers can then be formed by cascading such weak classifiers as shown in Figure 3.3. The advantage of having weak classifiers operating in cascade is that early processing can isolate regions likely to contain objects, thereby allowing greater concentration of effort to be brought to bear on these regions in subsequent operations. Also note that an accelerated computation can be achieved by adopting integral images in order to reduce multiplicative operations to those involving only addition and subtraction. This technique was chosen because of its high detection accuracy and ability to minimize computation time [78]. The approach was applied in two stages, first to obtain the face region and secondly the mouth region was found from the lower half of the face in which it is assumed the mouth is located as shown in Figure 3.4. The calculation time needed to isolate the mouth region was effectively halved by this approach but also the risk of false detection that can arise due to the inadvertent classification of the eyes as a mouth is reduced. These would improve the efficiency of system with respect to processing time and resources used.

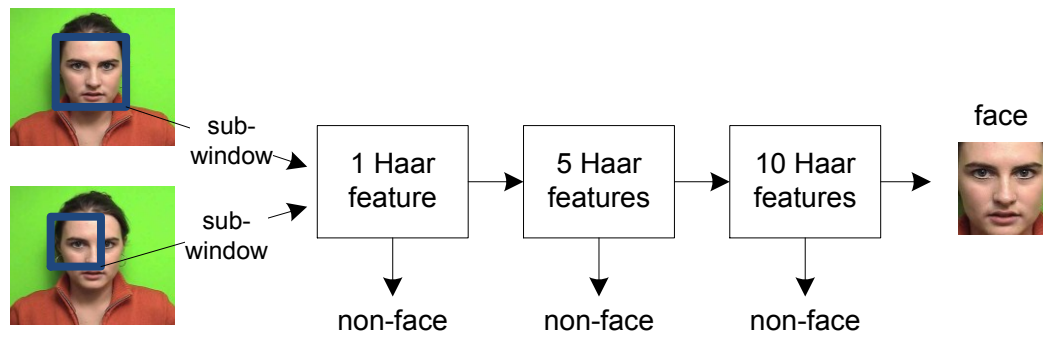


Figure 3.3 Face detection using Viola-Jones object recognizer

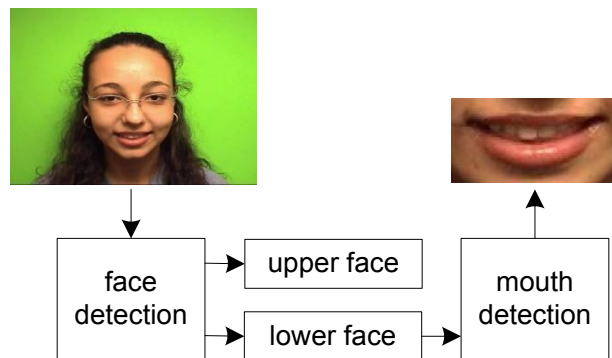


Figure 3.4 Face and mouth detection process

Figure 3.5 shows sample results of the face and mouth detection process. Figure 3.5(e) shows an image that has been superimposed using chroma-keying to replace the green background with a more complex one. Although there are three faces visible in the image, the algorithm assumes that the largest face region frame is to be selected since the target speaker is likely to be the one nearest to the camera. Figure 3.5(f) shows the detection process is able to continue to perform correctly for speaker wearing a hat.

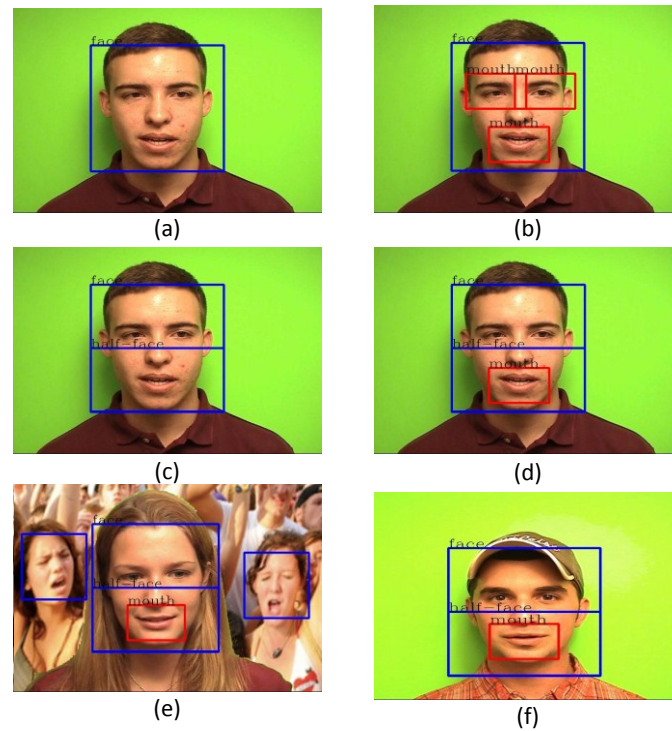


Figure 3.5 Example of face and mouth detection process (a) face detection, (b) mouth detection, (c) face region separation and (d) mouth region detection in lower part of face. Operation when (e) a complex background is introduced and (f) when wearing a hat

3.4 Lip segmentation

Once the mouth region has been isolated, there is a wide choice of algorithms that can be applied to extract features and a large number of alternative features have been investigated in the literature. In early stage of the experimental work, canny edge detection was investigated [6], [20] and the threshold and threshold linking adjusted to extract geometrical features such as height, width and ratio. Figure 3.6 shows examples of edges generated from the lip region using different values of the canny edge detector parameters. The threshold parameter affects the visibility of edges; the higher the threshold, the fewer the number of visible edges that will be found, while the threshold linking parameter determines the length in term of pixels above which edge joining polygons are retained [18].

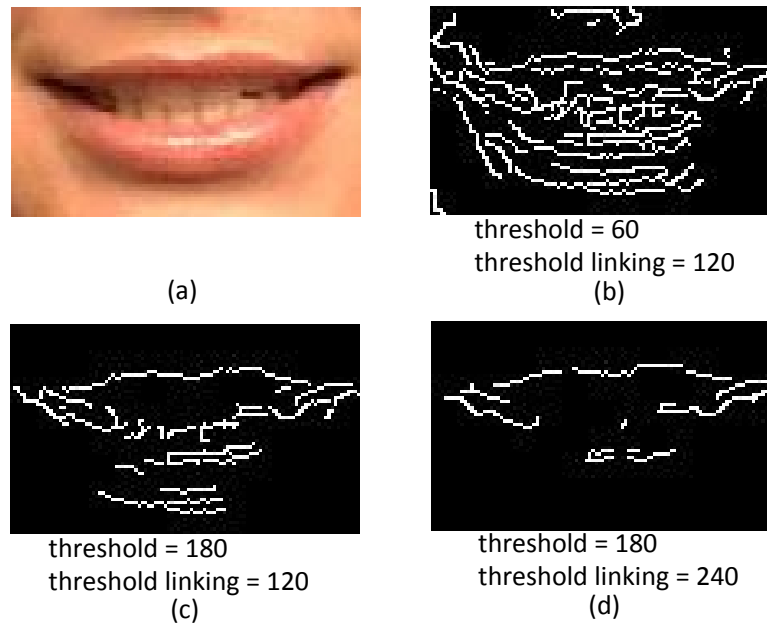


Figure 3.6 Canny edge detection with different threshold and threshold linking

The parameters shown in Figure 3.6(c) were the most appropriate for extracting the lip geometry features from the CUAVE database. However, this technique suffers from a number of drawbacks. Firstly, the user needs to set the edge threshold parameter manually for individual speakers. Secondly, if there are additional elements in the image, for example the moustache as seen in Figure 3.7(a), edges will often be identified in such regions, thus increasing the difficulty of extracting the lip geometry features, as shown in Figure 3.7(b). Thirdly, not all the lip boundaries are apparent in the image, for example the lower part of outer lip in Figure 3.7(c) does not have a visible boundary separating it from the skin region, thus it will be difficult for the Canny edge detection to identify the edge, as shown in Figure 3.7(d). Consequently, Canny edge detection, while able to extract part of the boundary between lip and non-lip regions is unlikely to be part of a reliable scheme for segmenting the non-lip region in an AVSR system.

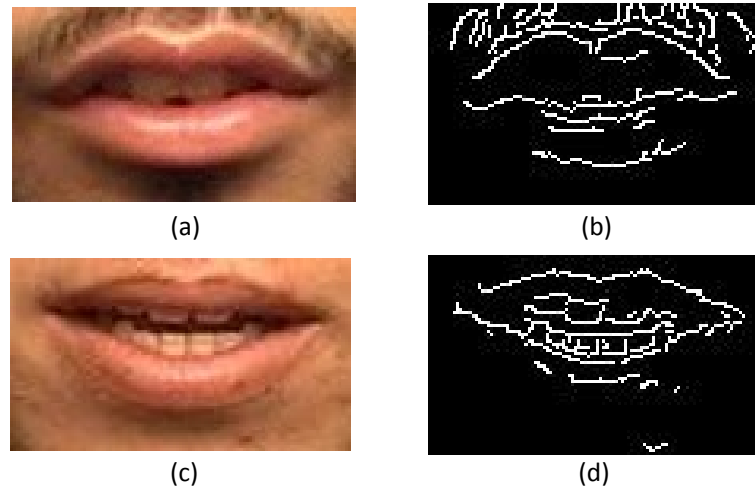


Figure 3.7 False boundaries with edge detection

As it is often the case that a non-lip area is covered by skin, this can be utilized in order to segment the lip area simply by assuming it is of a non-skin colour. In this work, the lip and non-lip regions are segmented using skin detection, with the aim of removing as many skin-coloured pixels as possible from the images in order to narrow the focus on the remaining lip-coloured regions. For this purpose an appropriate colour space needs to be selected from those available, including red, green and blue (RGB), cyan, magenta, yellow, and black (CMYK), and hue, saturation, and value (HSV). This work has adopted the HSV colour model for segmentation as this model comes closest to mimicking how humans perceive skin colour [79], [80]. The CUAVE images are in RGB colour space, but the transformation to HSV makes the source images invariant to high intensities of white and ambient light as well as the surface orientation relative to the light source [81]. Below are the equations used to transform RGB to HSV colour space [75].

$$V = \max(R, G, B) \quad (3.1)$$

$$S = \begin{cases} [V - \min(R, G, B)]/V & \text{if } V \neq 0 \\ 0 & \text{if } V = 0 \end{cases} \quad (3.2)$$

$$H = \begin{cases} 60(G - B)/S & \text{if } V = R \\ 120 + 60(B - R)/S & \text{if } V = G \\ 240 + 60(R - G)/S & \text{if } V = B \end{cases} \quad (3.3)$$

if $H < 0$ *then* $H = H + 360$

where

$R = \text{red}$, $G = \text{green}$, $B = \text{blue}$, $H = \text{hue}$, $S = \text{saturation}$, $V = \text{value}$

The skin colours of subjects tend to cluster in a small region of the colour space provided that the images are obtained under consistence illumination conditions [20]. Hence, one of most straightforward and common methods for lip region segmentation is to define the skin colour cluster decision range for one or more of the colour space components. Only those image pixel values that fall within the predefined range are assumed to be skin pixels. In the work performed in [80], pixel values in the range $H = [0, 50]$ and $S = [0.23, 0.68]$ were found to be well suited for discriminating the skin of Far Eastern and Caucasians subjects found in the M2VTS database [73]. In the current work, appropriate threshold values were found by examining the images in CUAVE video database in HSV colour space and different component ranges were then investigated, as shown in Figure 3.8. Using these threshold values, a binary image was then generated in which those portions of the image falling within the threshold were made black and the remainder made white.

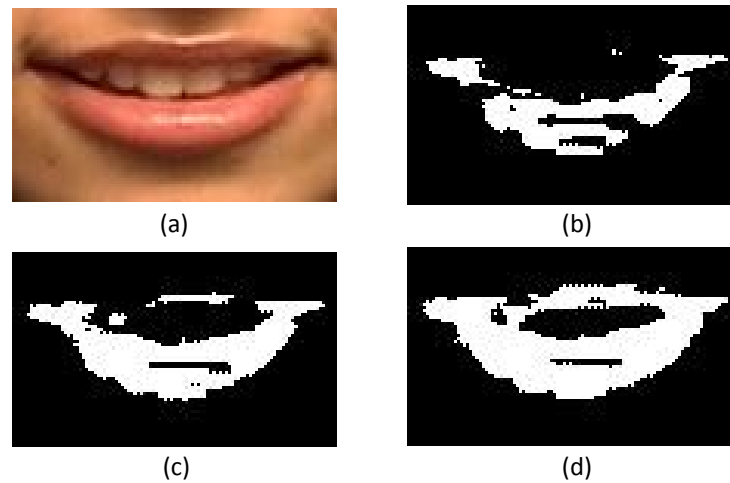


Figure 3.8 Lip skin detection based using an HSV colour filter (a) original image, and after application of the filter using (b) Hue {5, 40}, (c) Hue {6, 40} and (d) Hue {7, 40}

Following the investigations using images from the CUAVE video database, a suitable solution was found in which only those pixels with a hue channel value ranging from 7 and 40 inclusive were marked as being skin. The images were then subjected to a morphological process (erode and dilate) to minimize the noise (salt and pepper) in the image and a smoothing process (down-sample and up-sample) was applied to soften the image. At the end of this process a binary image was formed showing only the lip and non-lip area as depicted in Figure 3.9.

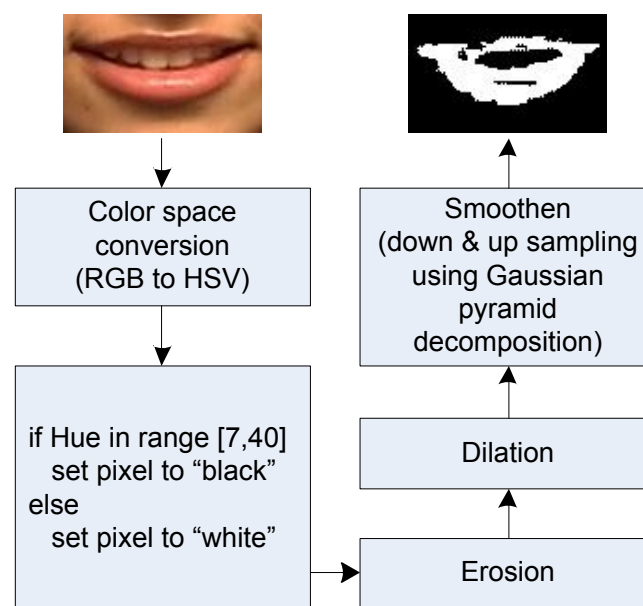


Figure 3.9 Block diagram for lip segmentation process

3.5 Lip geometry features

To the binary image containing the lip region, a contour extraction algorithm was applied using a border following technique [82]. Experiments showed that, from the collection of contours produced, it is the largest that regularly most closely matched the actual lip outline. Although this contour generally followed the outline of the lips, because it is generated as a simple polygon with many non-intersecting edges it remained a poor representation of the actual lip shape. However, a complex polygon such as that shown in Figure 3.10(a) can be reduced to a simpler convex polygon using the convex hull algorithm [83]. This algorithm determines the convex polygon of smallest area such that it contains all the vertices of the original polygon, as shown in Figure 3.10(b). Figure 3.10(c) is the single final polygon. The results showed that the convex hull solution was able to extract the outline shape of the lips to an accuracy sufficient for the estimation of good quality lip geometrical features such as height, width, area and perimeter.

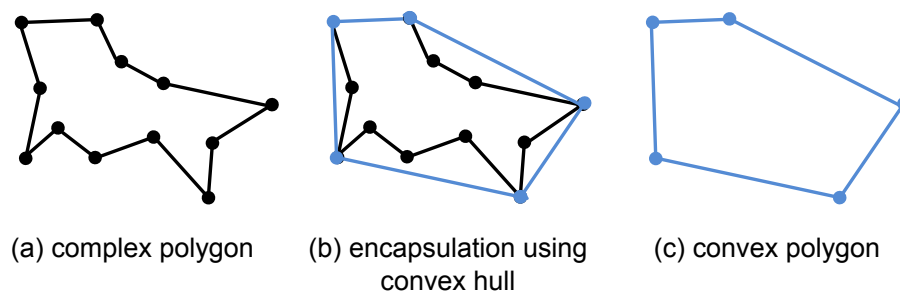


Figure 3.10 Illustration of the convex hull process using an ‘origami cat’ diagram

Figure 3.11 shows the results of lip geometry extraction from images of three different speakers. Binary lip images as shown in Figure 3.11(b), are generated by the process shown in Figure 3.9. A border following technique [82] was used to generate a collection of contours with different size and shape as shown in Figure 3.11(c). The experiments show that the largest contour generated by this process will most closely contain the lip area as shown in Figure 3.11(d). Using the largest contour, it can be seen that the convex hull approach is able to generate a

close approximation to the actual lip shape in the original image as shown in Figure 3.11(e).

In order to obtain good quality height and width information for later classification purposes, the lip shape must be consistently aligned along a horizontal axis. As the images used in this work involved only frontal pose, alignment can be achieved by rotation so that the left and right vertices of the lip contour are at the same vertical position in the image, as shown in Figure 3.12. The vertical and horizontal dimensions of the bounding box that encapsulates the entire shape represents the lip height and width respectively while the area and perimeter can be extracted from the convex polygon.



Figure 3.11 Convex hull results for speaker ‘s04f’ (left), ‘s05f’ (centre) and ‘s01m’ (right), (a) input colour image, (b) binary lip image and the results of processing following (c) contour detection, (d) largest contour identification and (e) convex hull calculation

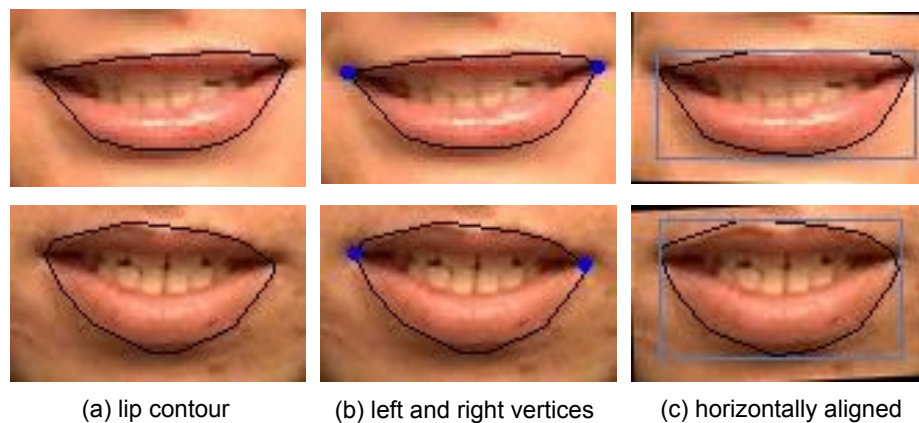


Figure 3.12 Automatic alignment using left and right vertices

Figure 3.13 shows a single video frame to illustrate the five shape-based features that were obtained in this work, namely height, width, ratio (height/width), area and perimeter. The perimeter is that of the polygon generated from the convex hull operation while the area is the region bounded by this polygon.

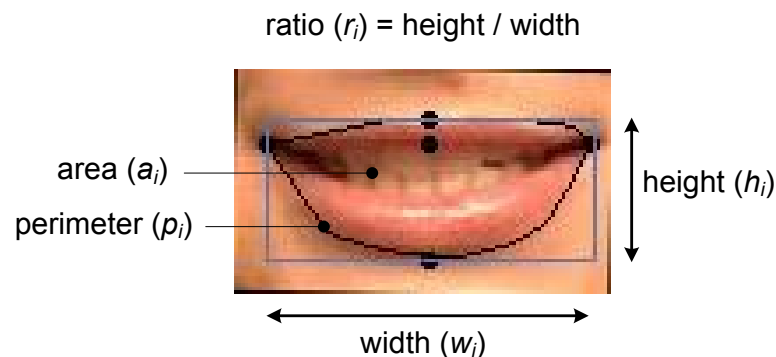


Figure 3.13 Shape-based lip features obtained from a single video frame i

To demonstrate the performance of the new approach, a comparison was carried out with both the active contours ('snakes') technique [84] and the GVF technique used to extract shape-based information from the lips [15]. The results of the experiment are shown in Figures 3.14 and 3.15. For this example, both techniques were unable to convergence to the correct lip shape, despite repeating the experiments using a range of different parameters. The main problem encountered was that the pattern of external energy generated was not closely related to the shape

of the mouth. In addition, the calculation time taken to converge to a final result took longer than the convex hull method and the outcome was unreliable, depending greatly on the initialization region and the termination criteria set during detection. The results obtained from the convex hull method also performed visually better than the BST technique used in [20] when applied to the same data corpus.

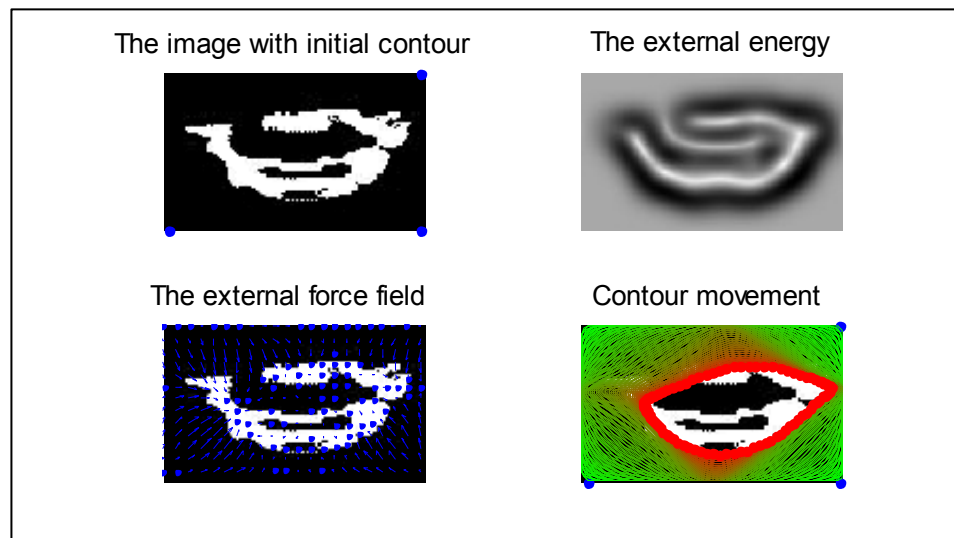


Figure 3.14 Lip contour detection using snakes method for speaker ‘s04f’

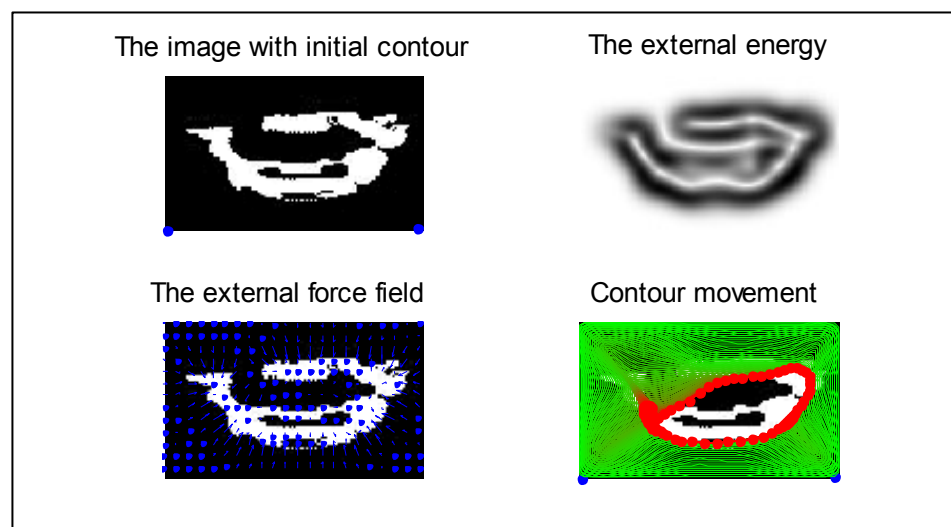


Figure 3.15 Lip contour detection using Gradient Vector Field (GVF) method for speaker ‘s04f’

3.6 Lip recognition analysis

To measure the effectiveness of the new method presented in this work, both a qualitative and a quantitative evaluation were conducted to assess the accuracy of the lip extraction used to generate geometrical features. A data set containing 180 face images with different expressions was randomly selected (five for each speaker) from the CUAVE database for this evaluation.

3.6.1 Qualitative assessment

For the qualitative assessment a visual inspection is normally conducted; for example in [85] four benchmark points were defined on the lips (left, right, top and bottom) and were used to grade performance as ‘wrong’, ‘poor’, ‘fair’, ‘good’ or ‘perfect’. In this work, the four grades used to define performance were ‘wrong’, ‘poor’, ‘satisfactory’ and ‘good’, each a subjective assessment of how closely the lip region has been isolated as well as quality of fit of the four benchmark points (left, right, top and bottom of the lips). Figure 3.16 shows examples of grading and classification carried according to the scheme shown in Table 3.1. The number of images that resulted in good visual lip segmentation quality is over 75 % with no image wrongly segmented (as the mouth region was found in all cases). For the same samples, the lip geometry segmentation obtained as a result of applying GVF and snakes was similar in quality to the examples shown in Figures 3.14 and 3.15 respectively. The segmentation provided by GVF and snakes was consistently qualitatively categorized as ‘poor’. Although the grading is subjective, the segmentation performance difference is clearly apparent in these results and only the proposed method is suitable for extracting lip geometry information for use in lip reading or AVSR systems.

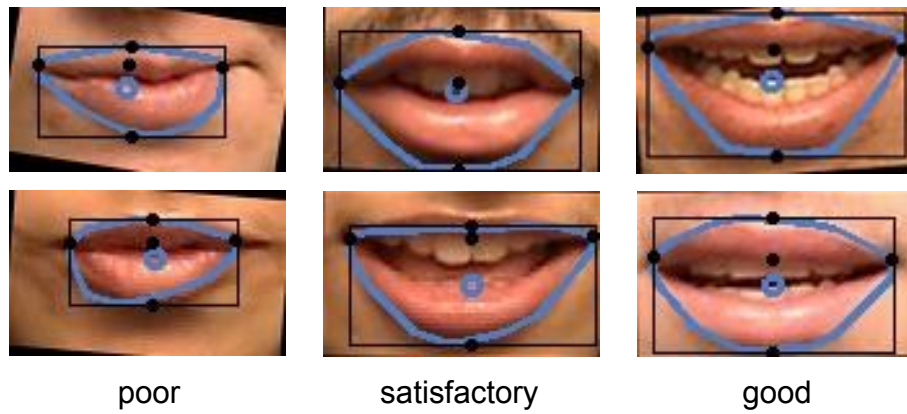


Figure 3.16 Example of the grading classification used in the qualitative assessment

Table 3.1 Qualitative evaluation of the lip classification

Grade	Good	Satisfactory	Poor	Wrong
Classification (%)	75	18.9	6.1	0

3.6.2 Quantitative assessment

The quantitative evaluation estimates the accuracy of the contour produced based on specific points of that contour. The method used for the quantitative evaluation was based on that described in [86], where four key lip points (top, bottom, left and right) are defined in both the original image and the convex hull contour, and the distance between corresponding points (in pixel units) defines an error that is normalized according to the distance between the mouth corners. For comparison purposes, the ASM technique proposed by Cootes *et al.* [87] was also used to determine the same four lip points in the original image. The ASM technique has a tendency not to be able to escape from local minima and to produce unreliable results under certain initial conditions. So, for fair comparison, the shape model was pre-initialized using prior information regarding the location of the eyes and mouth obtained from the Viola-Jones object recognizer used in our system. Figure 3.17 shows the lip outline recognition performance for the convex hull and ASM methods when compared with a manual annotation of the lip outlines. In Figure 3.17(a) and

Figure 3.17(b), the shape produce by the convex hull matches the outline of the lips, while the ASM results show errors compared to the expected shape both for the top and the bottom lip. Figure 3.17(c) shows an example in which the two techniques exhibit similar performance.

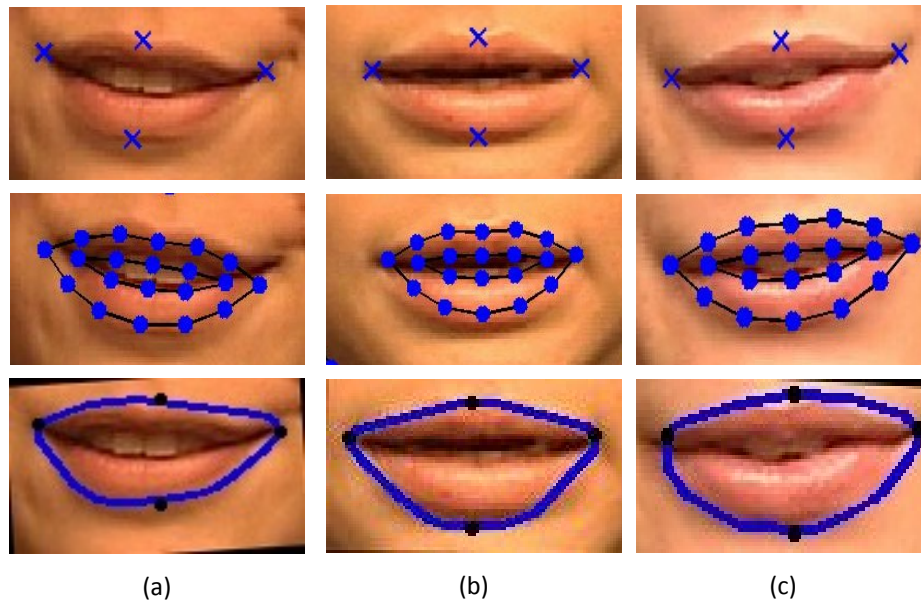


Figure 3.17 Comparison between manual annotation (top row), ASM technique (centre row) and convex hull technique (bottom row) for subject (a) ‘s28f’, (b) ‘s34f’ and (c) ‘s04f’

Table 3.2 shows a quantitative evaluation of the convex hull and ASM techniques for 180 image samples. It shows that, compared to a manual annotation of the lips, the convex hull method has an error of 14.42% in height and 6.85% in width. The fact that the height error is larger is likely to be due to a small number of the subject images in the database having a shadow under their bottom lip because of the illumination angle (an example of this can be seen for the first ‘satisfactory’ speaker in Figure 3.16), thus increasing the uncertainty in the determination of the height measurement. Nevertheless, the quantitative results also confirm that the lip extraction approach based on the convex hull method is suitable for determining features. The results shown in Table 3.2, demonstrate that the ASM approach is significantly worse than the Convex Hull method at extracting the outline of the lips.

Table 3.2 Quantitative evaluation of the lip classification

Method	Relative errors (%)	
	Height	Width
Convex hull	14.42	6.85
ASM	21.95	10.09

3.7 Lip reading performance evaluation

To further assess the effectiveness of the new method presented in this work, a lip reading experiment was conducted to investigate the accuracy of the lip geometrical features extracted. The CUAVE database consists of five sessions, where, in each session, the subject speaks the words ‘zero’ to ‘nine’. In the investigations, data from sessions 1, 2 and 3 (30 samples) were employed for training and data from sessions 4 and 5 (20 samples) were used for testing. Only the first 10 speakers from the database were used in this work, making a total of 500 samples that were actually used for demonstrating the utility of the proposed approach.

3.7.1 Lip reading system

The components of the new shape-based approach for lip reading system using geometrical features are shown in Figure 3.18. The system can be divided into three phases, namely preprocessing, feature extraction and classification.

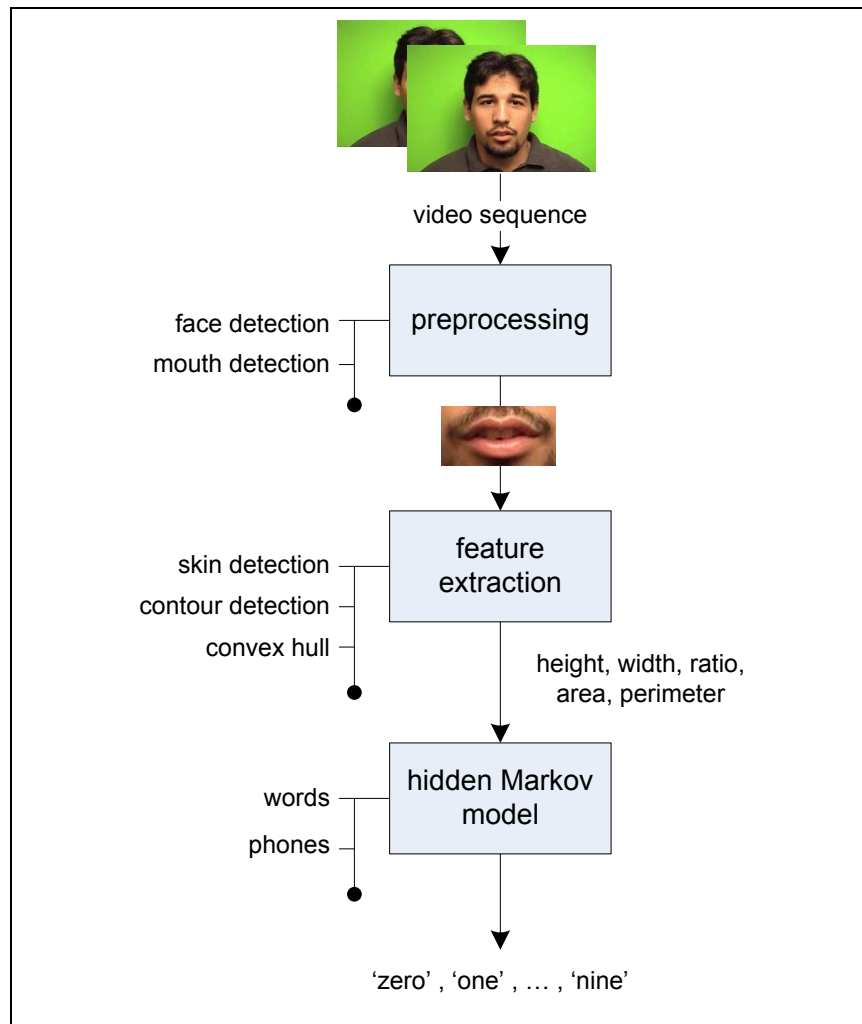


Figure 3.18 A block diagram for the shape-based lip reading system

The method adopted for preprocessing and feature extraction have been explained in Sections 3.3 to 3.5. As for the classification, HMM was used due to its popularity that has followed from many successful applications in the statistical modeling of audible speech [88]. Since lip reading also attempts to perform speech recognition, albeit in the visual domain, HMMs are often also used for the visual modeling. In this work, four different types of HMM architecture based on word and phone models were used. Three different word models were implemented with 3, 5 and 7 states as shown in Figure 3.19, while the phone model has a variable number of states dependent on the content to be classified, as shown in Figure 3.20 and Table 3.3. Each Markov state is modelled using solely Gaussian functions with diagonal covariance. The widely-used HTK library was used for the implementation [89].

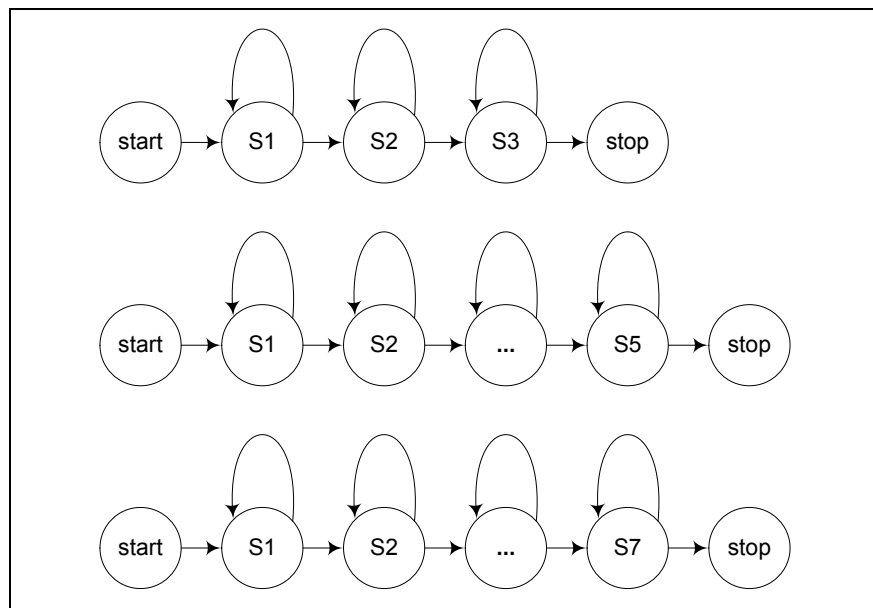


Figure 3.19 3 state (top), 5 state (middle) and 7 state (bottom) word recognition models

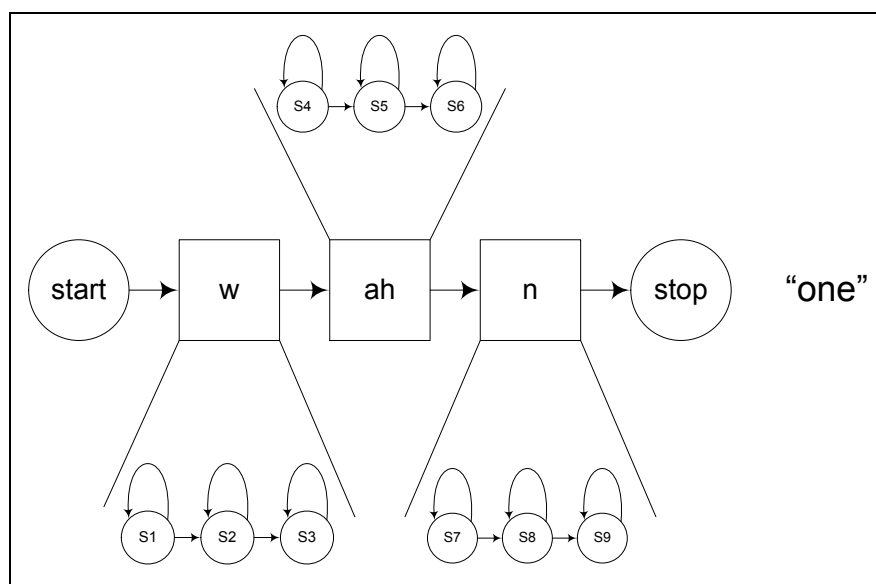


Figure 3.20 Phone recognition model

Table 3.3 List of phones used for English digit recognition

Digit	Phones
0	z - iy - r - ow
1	w - ah - n
2	t - uw
3	th - r - iy
4	f - ow - r
5	f - ay - v
6	s - ih - k - s
7	s - eh - v - ih - n
8	ey - t
9	n - ay - n

Training was carried out using only the original database video sequences [20]. To investigate classification performance under conditions where head rotation and subject illumination levels change, a number of new sequences were created. To further demonstrate the effectiveness of the new approach, the results were compared with an appearance-based technique similar to those used in many current practical realizations. In the implemented appearance-based approach, each image of the mouth region was resized to 64 x 64 and two-dimensional DCT visual features were extracted from the mouth region. To reduce the computational effort, only 16 DCT low-frequency coefficients were kept in order to represent each image.

3.7.2 Evaluation of Gaussian mixture function and HMM architecture

In these experiments, the head pose and the illumination of each subject in the database are controlled in the sense that the original source video sequences were used in which there are only minimal head movements and no changes in lighting. For each Markov model, a state can be modelled using 1-mixture, 2-mixture or 3-mixture Gaussian mixture function. It is important that an appropriate Gaussian mixture function is selected for a particular application, as this will influence system performance. Figure 3.21 shows the performance of the shape-based lip reading system for 3, 5 and 7 state HMMs. Figure 3.22 shows the corresponding results using appearance-based DCT lip-reading system.

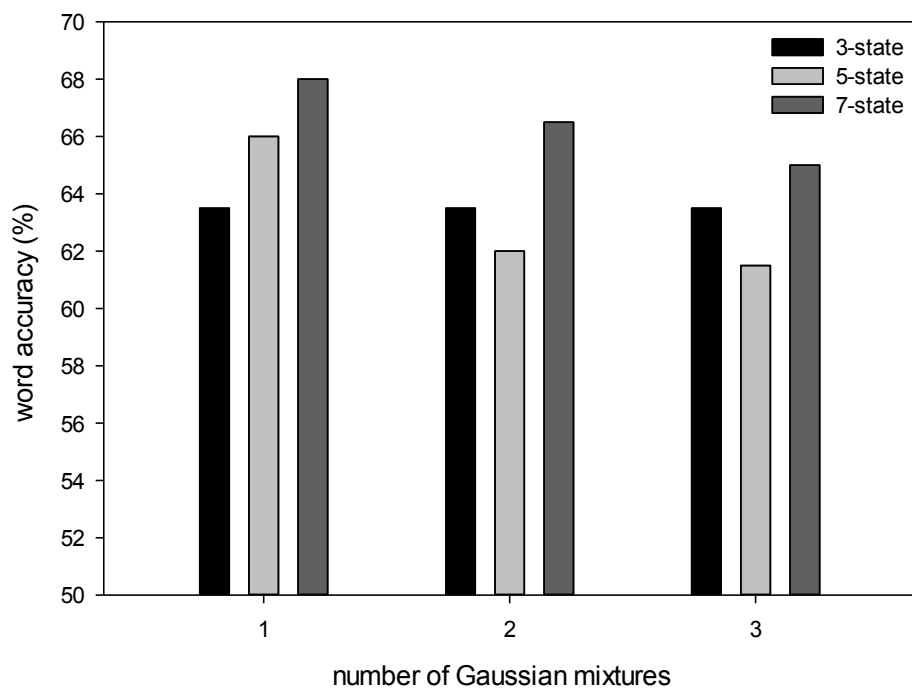


Figure 3.21 Performance obtained using shape-based geometrical features for different numbers of Gaussian mixtures

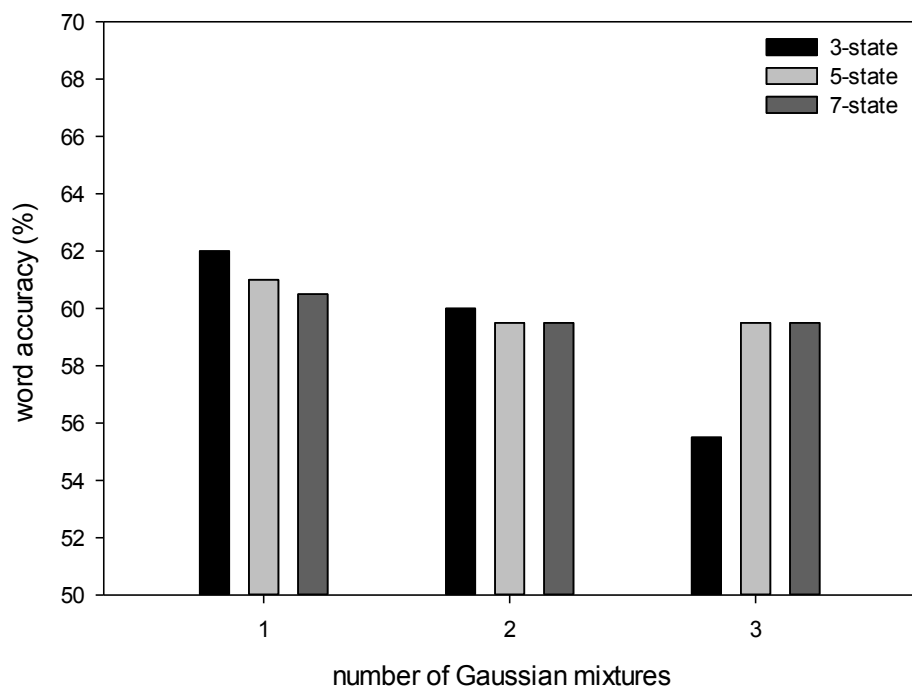


Figure 3.22 Performance obtained using appearance-based DCT features for different number of Gaussian mixture

For both the shape-based and appearance-based implementations, the results showed that the performance worsened as more Gaussian mixture functions were added and that using only a single Gaussian function for each Markov state was found to be the most suitable for the lip reading system. Consequently, the results in the following sections were obtained for HMM architectures that used a single Gaussian mixture function.

Figure 3.23 shows the performance of the lip reading system using both shape-based and appearance-based features. Using the HMM word recognition model, the performance of the system using shape-based features improved as the number of HMM states was increased, while the appearance-based results worsened as more states were added. It can be seen that for both the word and phone recognition models, the shape-based systems achieved a recognition performance of around 68% and consistently performed better than the corresponding appearance-based system

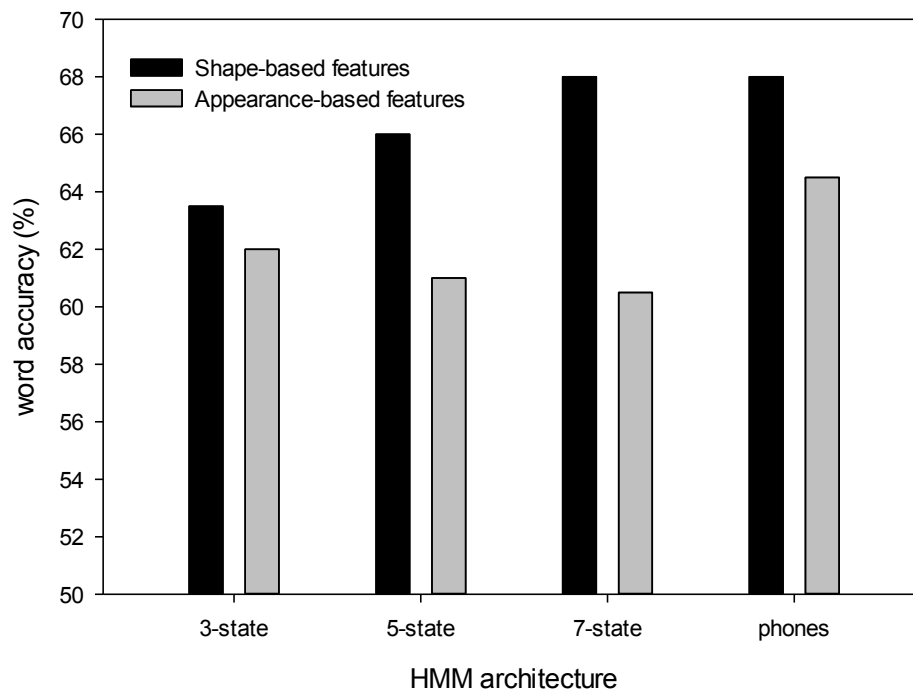


Figure 3.23 Performance of the lip reading systems using different HMM architectures

3.7.3 Effect of head rotation and brightness changes

In real-world conditions, speakers tend to move their head position while talking. As a first attempt to move towards performing lip reading in more natural situations, this work has investigated the robustness of the new shape-based approach to head rotational movements. To generate data suitable for performance testing, images from the video sequences were artificially rotated (in the plane of the image) by $\pm 20^\circ$ in increments of 5° . Figure 3.24 shows the performances of the lip reading systems during head rotation changes.

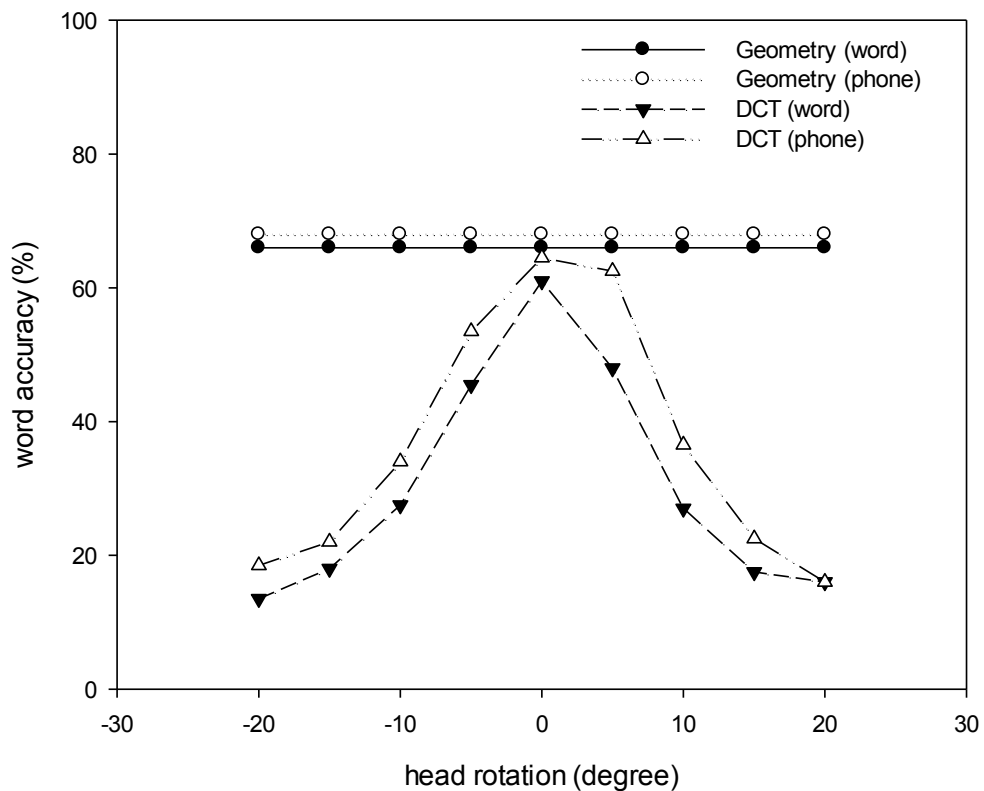


Figure 3.24 Comparison of the lip reading system performance using shape-based geometrical features and appearance-based DCT features during head rotation changes

It can be seen that the performance of the appearance-based system is adversely affected by head rotation, but that the shape-based system was largely unaffected. This difference in robustness is likely to be due to there being no in-built mechanism in the generation of the DCT features that is able to correct for head alignment. Conversely, the computation of the geometrical features includes a process of extracting the lip outline in which the contour is effectively rotated (within certain angular limits) until it is horizontally aligned, as shown in Figure 3.25. This mechanism substantially corrects for minor rotations of the face within the plane of the image.

The brightness level of a subject in an image is affected by their location relative to light sources. To evaluate the performance of the proposed system under a range of simulated lighting conditions, the brightness of the images is artificially changed (that is, after image capture) by $\pm 20\%$ in increments of 4%. Figure 3.26 shows the performances of the lip reading systems investigated in this work when subjected to such brightness changes.



Figure 3.25 Automatic head rotation involved in the shape-based geometrical feature extraction process, shown here for subject ‘s01m’ in the CUAVE database. Artificially rotated images (left) and corrected images (right).

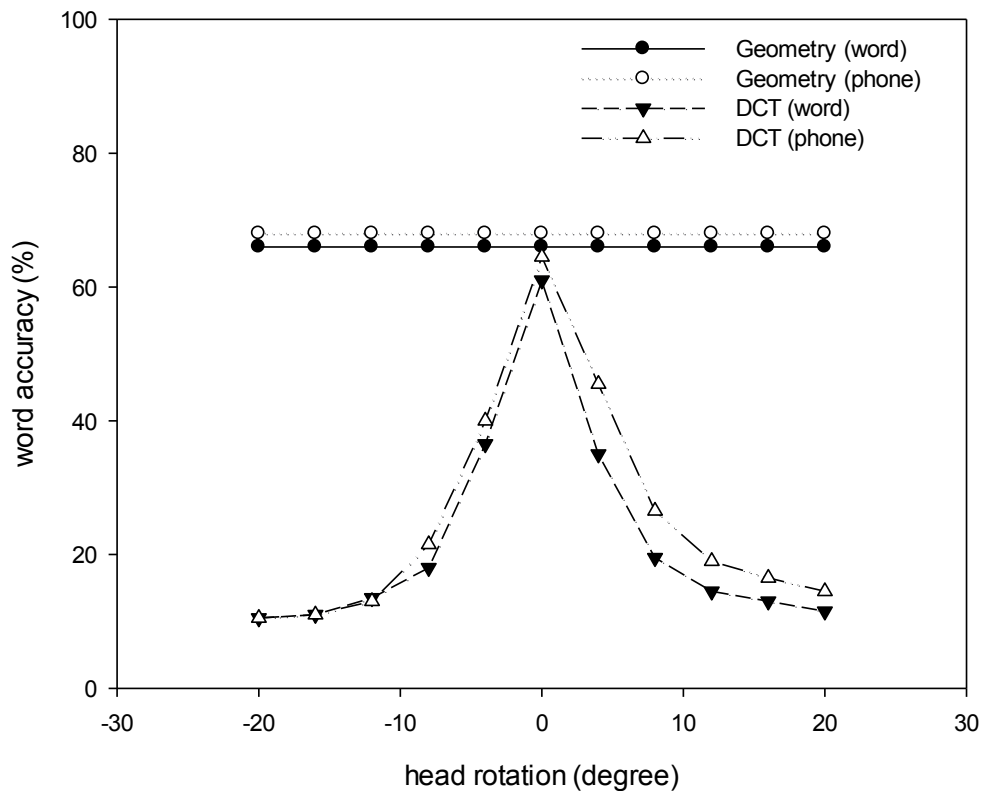


Figure 3.26 Comparison of the lip reading system performance using shape-based geometrical features and appearance-based DCT features during brightness changes

For the shape-based approach, the brightness changes have little or no effect on the recognition performance, whereas that of the appearance-based based approach is adversely affected. The difference in performance is largely due to the different color spaces used to extract the geometrical-based and appearance-based features. In the new method of this work, the extraction of lip geometry features is performed in the HSV color space and brightness information was not involved in the hue filter used to detect the skin region, as shown in Figure 3.27. In contrast, as part of the process to extract the appearance-based features, the original image in RGB color space is converted to 8-bit grayscale image (preserving brightness information) before the DCT algorithm is applied, as shown in Fig. 3.28.

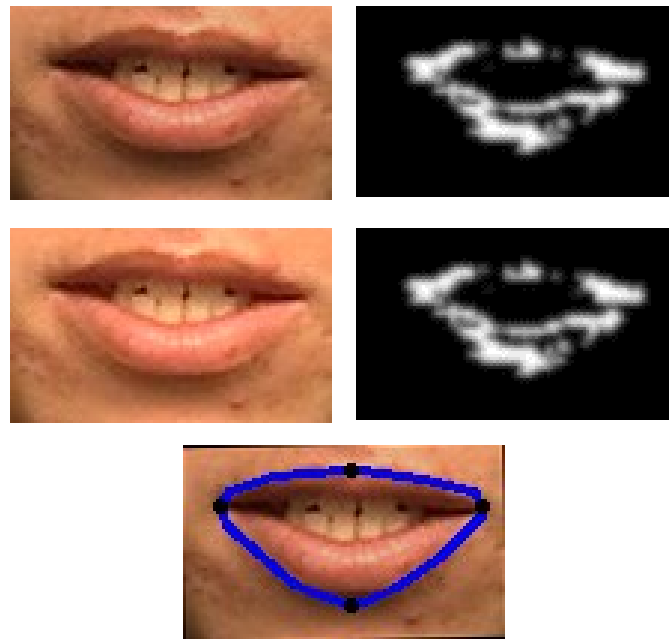


Figure 3.27 Brightness information is not involved in geometrical feature extraction. The lip region extracted is the same following the darkening of the image by 20% (top), brightening by 20% (middle) resulting in the same features being extracted (bottom).



Figure 3.28 Brightness information is preserved during RGB to grayscale conversion consequently affecting the appearance-based feature values. Examples are shown for image darkening by 20% (top) and brightening by 20% (bottom).

3.8 Chapter summary

This chapter has described and evaluated a process to obtain lip geometry information from video sequences in the CUAVE database. The new approach uses a skin colour filter, a border following method, applies a morphological process and obtains the final contour by computing the convex hull. The proposed method provides an improvement in performance compared with the methods described in the literature, specifically the active contour ('snake'), GVF and ASM techniques. To measure the effectiveness, analysis techniques using both qualitative and quantitative assessment have been carried out.

In addition, performance evaluation based on lip reading to recognize the English digits 0 to 9 as presented in the CUAVE database, showed that the shape-based geometrical approach performed better than appearance-based DCT approaches. All classification was carried out using HMM classifiers; indeed four types of HMM architecture were investigated and it was found that a phone-based model performed better than a word-based approach. An analysis of the robustness of the proposed system was also carried out to simulate changes in both head pose (by rotation of the image) and illumination (by adjusting the brightness). The results showed that the new geometrical-based method generates features that are robust to these environmental changes, whereas the resulting changes in the DCT features of the appearance-based method severely affected performance.

In comparison with the state-of-the-art appearance-based methods, the new shape-based lip reading system exhibits better word recognition accuracy and robust to rotation and brightness effects, making it more suitable to be incorporated in multi-modal speech recognition systems for use in noisy environments. The next four chapters give more details about the implementation and the performance of the shape-based geometrical features in the AVSR experiments.

CHAPTER 4

SHAPE BASED LIP READING SYSTEM USING TEMPLATE PROBABILISTIC MULTI DIMENSION DYNAMIC TIME WARPING

This chapter explains the approach taken in the classification of visual speech using lip geometry features extracted from the mouth region as input. A brief introduction to visual feature classification in AVSR is given in Section 4.1 and a novel approach developed in this work to classify lip geometry, namely the Template Probabilistic Multi-Dimension Dynamic Time Warping (TP-MDTW) method, is detailed in Section 4.2. Finally, the performance of the proposed approach is addressed in Section 4.3.

4.1 Introduction

As explained details in literature (Section 2.3), the lip shapes, positions and movements relevant to speech in the visual domain are described as viseme [60]. Since only a small part of the vocal tract is visible when we speak, only partial physical information is available regarding the generation of visemes and not all can be mapped to a unique phone [48], the basic unit of speech in the audio domain [59].

A viseme may be represented by a time sequence of lip shapes, but the actual set of lip shapes and their durations are dependent on the speaker. For example, although it would be expected that the visual representation of the word ‘hello’ may vary between speakers (inter class), there are also likely to be differences if the word is spoken again by the same speaker (intra class), for example if on the second

occasion the individual circumstances of the speaker changes, perhaps they now shout the word or they find themselves in a stressful situation.

As the application domain is the same, lip reading classification techniques are borrowed from those applied in the audio speech recognition (ASR) field and, consequently, Dynamic Time Warping (DTW) [90], [91] and HMMs [11], [19], [68], are popular. Moreover, by using a method common to both the audio and visual aspects of speech, there is the potential for a more straightforward combination of results obtained from separate audio and visual investigations and such integration has often been carried out using machine learning techniques, such as Time Delay Neural Networks (TDNN) [92], Support Vector Machines (SVM) [93] and AdaBoost [94].

4.2 Methodology

The geometric-based approach for the proposed lip reading system described in this chapter is shown in Figure 4.1. The system can be divided into the following four stages.

- The face and then the mouth regions are extracted from the images contained in the video sequences of the speakers.
- The mouth region is segmented into lip and non-lip areas.
- A new approach is applied that uses border following and convex hull computation to extract the lip geometry and to generate shape-based features.
- A novel technique termed TP-MDTW is used to classify dynamic geometry information.

The component parts of stages 1 to 3 have been described in detail in the previous chapter and so only the final stage of the proposed system will be explained in the following sections. Lip geometry features, including height, width, ratio, area, perimeter and various combinations of these features were evaluated to determine which performs the best when representing speech in the visual domain using the novel TP-MDTW technique. The recognition performance of the proposed system has been assessed in the recognition of the English digits 0 to 9 as spoken by the subjects in the video sequences available in the CUAVE database. For comparison purposes, two additional separate classification methods, namely the conventional DTW and HMM classifiers have been implemented. The software for this work was developed using Microsoft Visual C# 2010 [74] and utilized the open source image processing library, OpenCV [75].

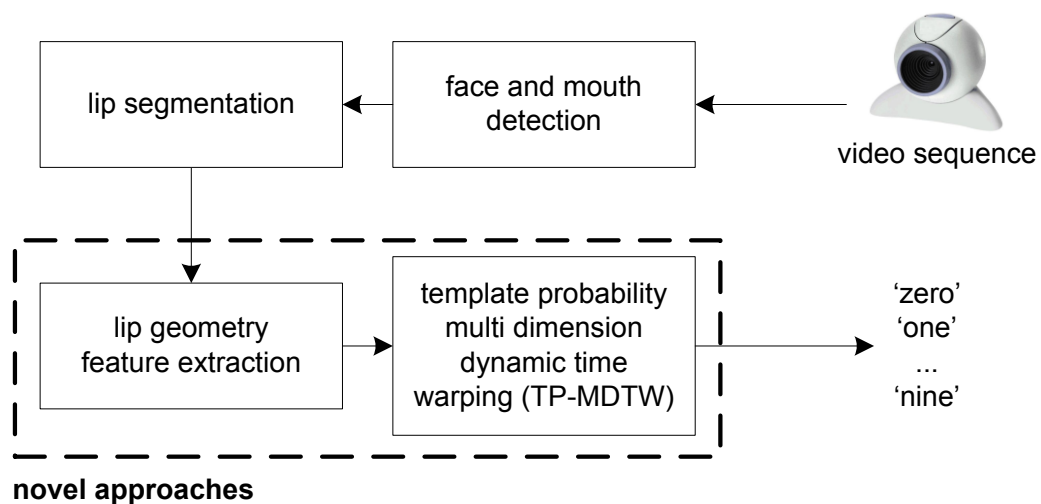


Figure 4.1 Architecture of the proposed lip reading system

4.2.1 Lip dynamic information

The geometrical information obtained regarding the lips is provided to the lip reading process, but the performance can be substantially improved using information not just from single images, but from dynamic information generated from a sequence of feature values obtained during the speech utterance. By finding the convex hull contour for each video frame, a time-series of feature values can be derived, as shown in Figure 4.2. This information is stored as speaker models (templates) during training for later use in the lip reading system in which the models are compared with test series.

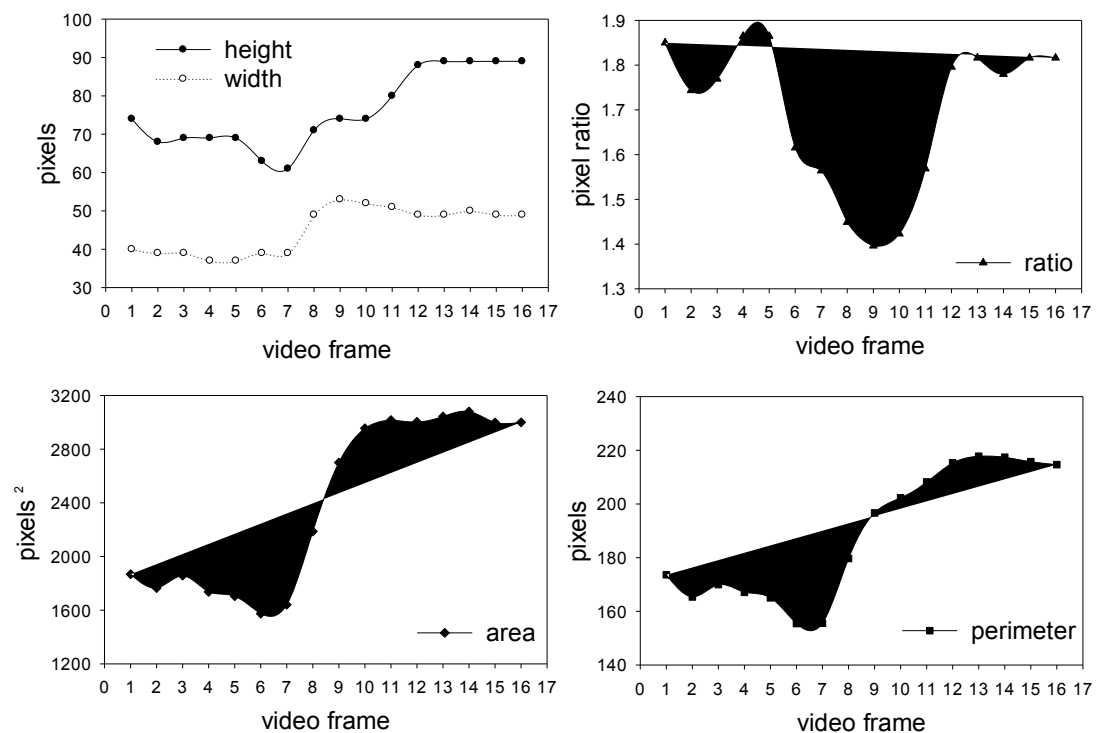


Figure 4.2 Dynamic lip information for digit "one" uttered by speaker 's01m' in the CUAVE database

Figure 4.3 shows the time-domain changes in the height, width and ratio of the lips when a number of different speakers uttered the digit ‘five’. The results confirm that a similar underlying time-varying pattern was produced by all the speakers, indicating the potential of this method in its ability to identify the words being spoken.

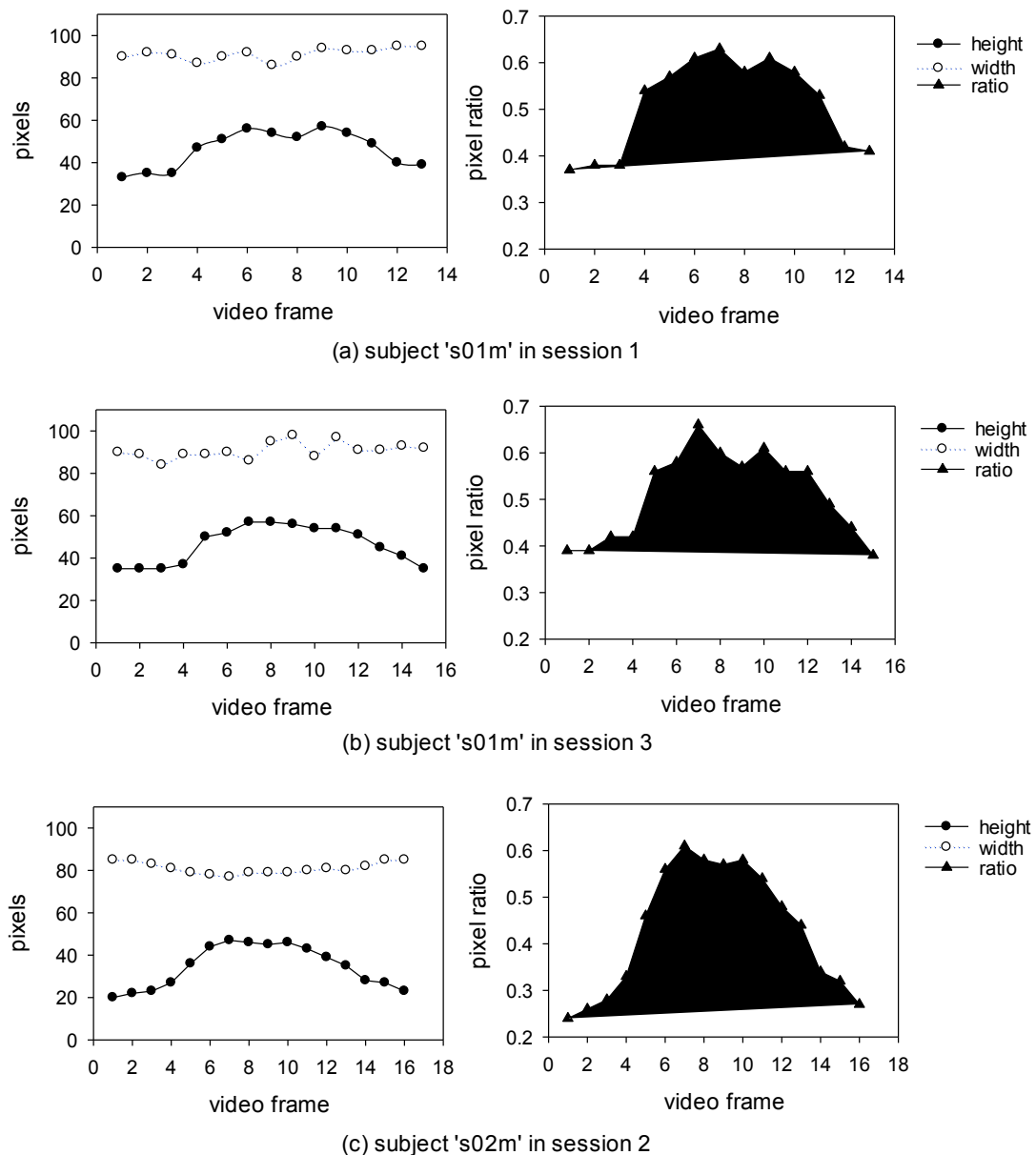


Figure 4.3 Dynamic lip information showing changes in lip height, lip width and the ratio of height to width for digit ‘five’ obtained from the CUAVE database

Throughout this work the following notation is used to represent lip geometry features in the time domain. If \mathbf{G}_i are the geometrical features obtained from video frame i , then

$$\mathbf{G}_i = [h_i \ w_i \ r_i \ a_i \ p_i]^T \quad (4.1)$$

where

$$h = \text{height}, w = \text{width}, r = \text{ratio}, a = \text{area}, p = \text{perimeter}$$

and the geometrical features obtained from a video sequence can be denoted as a matrix \mathbf{G} , given by

$$\mathbf{G} = [\mathbf{G}_1 \ \mathbf{G}_2 \ \dots] \quad (4.2)$$

4.2.2 Dynamic time warping

The representation of the lip reading problem in the current work is now described in terms of the time series sequence shown in Equation 4.2. As discussed in the previous section, DTW and HMM are the two main alternatives for time series classification and have been widely used in both speech recognition and lip reading. In addition, DTW can also be thought of as a special case of HMM, with each point along the reference signal representing a hidden state and transition probabilities restricted to prevent movements backwards in time.

DTW utilizes dynamic programming to generate candidate stretched and compressed sections in sequences of feature vectors, in order to find an alignment between two time-series that minimizes distortion [95] and in doing so produces a suitable warping function that minimizes the total distance (normally Euclidean) between an unknown sample and the reference template. While a DTW-based lip reading system has been proposed previously [90], [91], [96], to the best of the

author's knowledge, the problem has not been addressed using multi-dimensional DTW and reference template probabilities.

The version of DTW utilized in this work follows the approach found in [97] and [98]. Here, a two-dimensional M by N cost matrix \mathbf{D} is constructed, where each of the $D(i,j)$ values is the minimum distance warped path at time i for the time series \mathbf{x} and time j for time series \mathbf{y} , where $\mathbf{x} = (x_1, \dots, x_i, \dots, x_M)$ and $\mathbf{y} = (y_1, \dots, y_j, \dots, y_N)$ are time series. M and N are the size of the time series \mathbf{x} and \mathbf{y} respectively. The value at $D(M,N)$ will contain the minimum distance warped path between the time series \mathbf{x} and \mathbf{y} . Figure 4.4 shows an illustration of a cost matrix and a minimum-distance warp path traced from $D(1,1)$ to $D(M,N)$.

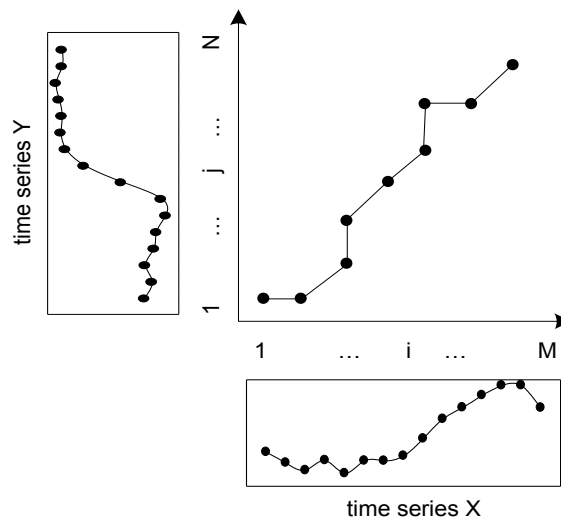


Figure 4.4 Cost matrix with the minimum-distance warp path

If the absolute difference between any two pairs of values in the time series \mathbf{x} and \mathbf{y} is given by

$$d(i, j) = |x_i - y_j| \quad 1 \leq i \leq M, \quad 1 \leq j \leq N, \quad (4.3)$$

then the values in the cost matrix \mathbf{D} are defined as follows

$$D(1,1) = d(1,1) \quad (4.4)$$

$$D(1,j) = D(1,j-1) + d(1,j) \quad 2 \leq j \leq N \quad (4.5)$$

$$D(i,1) = D(i-1,1) + d(i,1) \quad 2 \leq i \leq M \quad (4.6)$$

$$D(i,j) = d(i,j) + \min \begin{cases} D(i,j-1) \\ D(i-1,j-1) \\ D(i-1,j) \end{cases} \quad 2 \leq i \leq M, 2 \leq j \leq N \quad (4.7)$$

DTW as defined in the literature is only applicable to problems requiring single feature alignment. To provide alignments in the current work that can include up to five shape-based features, the conventional approach would be to apply DTW operations to each feature separately and then subsequently select the one exhibiting the shortest distance (lowest error). However, as the features are obtained independently and there is no synchronization between them, the results obtained from the initial experiments were unsatisfactory. To improve performance an alternative approach was sought.

4.2.3 Multi dimension dynamic time warping

In this work, an extension termed the multi-dimensional DTW (MDTW) was made to the DTW algorithm to allow it to operate with multiple features simultaneously while providing synchronization between the time series. The experimental results for MDTW showed a marked improvement in lip reading accuracy compared to those obtained using DTW. In the MDTW method, the two time series \mathbf{x} and \mathbf{y} must first be reconstructed as multi-dimensional matrices where each column represents one series of the \mathbf{G}_i features. These matrices can be generated from Equation 4.2 and can be re-written as

$$\mathbf{x} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,M} \\ x_{2,1} & \cdots & x_{2,M} \\ \vdots & \ddots & \vdots \\ x_{K,1} & \cdots & x_{K,M} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_{1,1} & \cdots & y_{1,N} \\ y_{2,1} & \cdots & y_{2,N} \\ \vdots & \ddots & \vdots \\ y_{K,1} & \cdots & y_{K,N} \end{bmatrix}, \quad (4.8)$$

where M is the length of the reference video sequence, N is the length of the unknown video sequence and K is the number of geometrical features evaluated. In this work, Equation 4.3 also needs to be modified in order to operate in the multi-dimensional case and is given by

$$d(i, j) = \sum_{k=1}^K \left| \frac{M \cdot x_{k,i}}{\sum_{m=1}^M x_{k,m}} - \frac{N \cdot y_{k,j}}{\sum_{n=1}^N y_{k,n}} \right| \quad 1 \leq i \leq M, 1 \leq j \leq N. \quad (4.9)$$

Then the minimum distance warped path between the two multi-dimensional time series \mathbf{x} and \mathbf{y} using MDTW is given by

$$MDTW(\mathbf{x}, \mathbf{y}) = D(M, N). \quad (4.10)$$

4.2.4 Novel template probabilistic approach

In the initial experiments, it was observed that a few of the templates used for training were never selected for the subsequent lip reading matching operations. This was found to occur because the training set contained examples of sequences spoken in a manner not found elsewhere in the training set, although they may potentially form a good match to test examples. To overcome this issue, a novel Template Probabilistic MDTW (TP-MDTW) technique was introduced to calculate the probability of each template being the best match to an unseen example based on its similarity to other templates in the database. The assumption is that a template

with a greater similarity to other database templates should be recognized as more probable to occur.

To understand the operation of TP-MDTW, consider a system with a set of R reference templates $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, \dots, \mathbf{T}_R$. An example is shown in Figure 4.5 for the case when $R=4$.

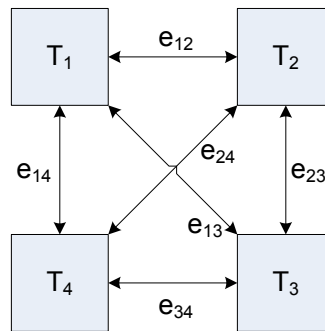


Figure 4.5 Example of distance values that need to be calculated when four reference templates are defined.

The similarity between the templates can be calculated using any suitable distance measurement technique. In the current implementation, MDTW is used to find the distance between e_{ij} the features in the reference templates. For any pair of unknown and reference templates \mathbf{T}_i and \mathbf{T}_j in the form given in Equation 4.8, the distance e_{ij} between the features is found using MDTW and is given by

$$e_{ij} = MDTW(T_i, T_j) \quad 1 \leq i, j \leq R. \quad (4.11)$$

Since the distances between pairs of templates is the same regardless of the starting point

$$\begin{aligned} e_{ij} &= e_{ji} & j &\neq i \\ e_{ij} &= 0 & j &= i \end{aligned} \quad (4.12)$$

To find the similarity between any one template and the remainder, the cumulative distance to other templates can be computed by

$$\alpha_i = \sum_{j=1}^R e_{ij} \quad 1 \leq i \leq R \quad (4.13)$$

Based on the cumulative distance between the templates, the probability of a template being the best match to an unseen example can be calculated using

$$P(T_i) = \frac{\frac{1}{\alpha_i}}{\sum_{g=1}^R \left[\frac{1}{\alpha_g} \right]} \quad 1 \leq i \leq R \quad (4.14)$$

where

$$\sum_{i=1}^R P(T_i) = 1 \quad (4.15)$$

To make best use of the alternative approaches that are available at this stage, these have been implemented as models able to process inputs (template probability, reference template and unknown sample), as shown in Figure 4.6 and Figure 4.7. This work introduces four lip reading models whose operations were designed based on a series of preliminary experiments.

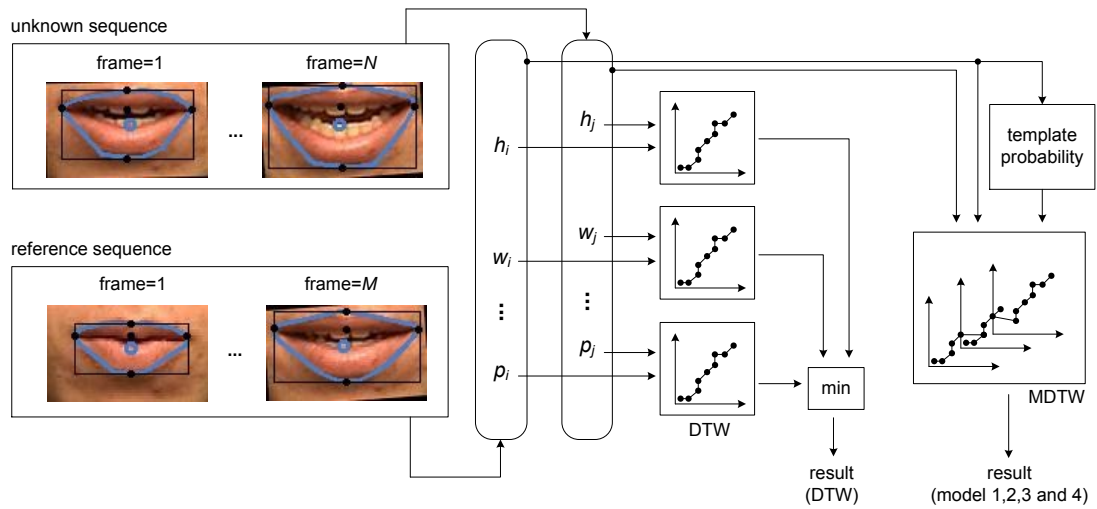


Figure 4.6 General structure of the classification operations used in the lip reading system

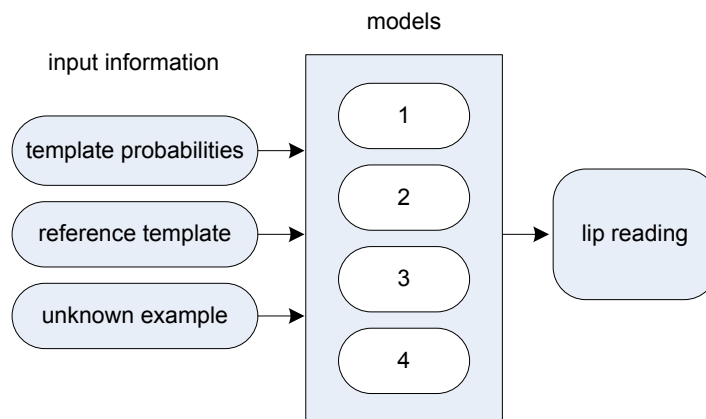


Figure 4.7 Models used in the lip reading classification

The objective of model 1 is to determine the accuracy of the system assuming there is a uniform probability in the distribution for each template as shown in Equation 4.16.

$$P_1(T_i) = \frac{1}{R} \quad 1 \leq i \leq R \quad (4.16)$$

The output for model 1 for an unknown example ϕ is

$$O_1(\phi) = \sum_{i=1}^R [P_1(T_i) \cdot MDTW(\phi, T_i)]. \quad (4.17)$$

Model 2 selects the template of largest probability of the all reference templates in the database by using Equation 4.18.

$$P_2(T_i) = \begin{cases} 1 & \text{if } P(T_i) = \max_{1 \leq k \leq R} \{P(T_k)\} \\ 0 & \text{otherwise} \end{cases} \quad 1 \leq i \leq R \quad (4.18)$$

The assumption is that this template gives the closest representation to the features from the unknown example. The output for model 2 is given by

$$O_2(\phi) = \sum_{i=1}^R [P_2(T_i) \cdot MDTW(\phi, T_i)]. \quad (4.19)$$

The objective of model 3 is to measure performance based on the template probability that is calculated from the accumulated distance from other templates as found by Equation 4.14. The output of model 3 is given by

$$O_3(\phi) = \sum_{i=1}^R [P(T_i) \cdot MDTW(\phi, T_i)]. \quad (4.20)$$

Model 4 measures performance based on the template probability calculated from the accumulated distance, but also gives emphasis to the reference template that has the highest probability. Its performance calculation is given by

$$O_4(\phi) = \frac{1}{2} \sum_{i=1}^R \left[(P(T_i) + P_2(T_i)) \cdot MDTW(\phi, T_i) \right]. \quad (4.21)$$

Clearly, the number of models need not be limited to these four and further models could be developed to implement additional operations.

4.3 Results and discussion

In order to investigate the effectiveness of the proposed approaches in practical applications, the CUAVE corpus database was used to provide examples of speech and video sequences of the speakers [20]. The database consists of 7000 utterances of connected and isolated digits spoken by 36 individuals, where 19 speakers are male and the remainder are female. The speakers also have a range of skin and lip tones as well as a variety of face and lip shapes and a number of the subjects wore additional visual items such as glasses, facial hair, and hats. Lighting was controlled and a green background was employed to allow custom video backgrounds to be added using chroma-keying if required. The video sequences were recorded at a resolution of 720 x 480 in MPEG-2 format at 29.97 frames/s and encoded at a data rate of 5,000 kbit/s.

The CUAVE database consists of five sessions, in each of which the subject speaks the words ‘zero’ to ‘nine’. In the investigations, data from sessions 1, 2 and 3 (30 samples) were employed for training and the data from sessions 4 and 5 (20 samples) were used for testing. All 36 speakers in the database were investigated in this work, making a total of 1800 samples for use in demonstrating the utility of the new approach. Two investigations have been designed to measure the performance of the lip reading system, namely classification using single features and classification using multiple features.

4.3.1 Performance using a single feature

The classification performances using the TP-MDTW technique for the five lip geometrical features operating individually, namely height, width, ratio of height to width, area and perimeter, are shown in Table 4.1. Model 1 provides the ground truth regarding the accuracy of the system, employing a uniform probability distribution for each template. Of the models investigated, it can be seen that model 4 (based on the template probability calculated from accumulative distance) provides the best single feature classification. It can be seen that the lip area feature produced the best performance, providing almost 62% correct lip reading identification using model 4.

Table 4.1. Word accuracy for single lip geometry features using the TP-MDTW approach

Model type	Geometry features				
	Height	Width	Ratio	Area	Perimeter
Model 1	53.33	40.28	44.17	57.36	48.75
Model 2	53.75	39.44	44.44	59.58	52.64
Model 3	54.31	41.94	45.00	59.72	52.64
Model 4	56.53	42.78	48.06	61.81	55.83

4.3.2 Performance using multiple features

Investigations of the classification results obtained using various combinations of lip geometry features were carried out to improve further the lip reading system performance. Since there are five lip geometrical features, there are 120 possible combinations of features exist. In order to simplify the feature combinations, lip geometry features were split into two groups, vector-based features (height, width and ratio) and scalar-based features (area and perimeter), and

possible feature combinations are shown in Table 4.2. The results of the classification using TP-MDTW shown in Table 4.2 demonstrate that the combination of height, width and ratio information gave the best performance, providing a classification performance of up to 70.69% correct when using model 4.

Table 4.2 Word accuracy for candidate combinations of lip geometry features classified using TP-MDTW

Model type	Geometrical feature combinations					
	HW	HWR	AP	HWRA	HWRP	HWRAP
Model 1	55.56	61.39	57.36	57.78	54.58	58.19
Model 2	56.81	62.78	59.44	60.83	57.64	61.11
Model 3	57.50	65.42	59.72	59.44	56.67	59.86
Model 4	60.42	70.69	61.94	62.64	58.75	63.33

Note. H=height, W=width, R=ratio, A=area, P=perimeter

4.3.3 Comparison with state-of-the-art techniques

To demonstrate the performance of TP-MDTW, a comparison with existing DTW and HMM classifiers was made using model 4. Left-right HMM models with eight states were used to develop word models, each state having an observation probability distribution modelled by a single Gaussian with diagonal covariance. The same lip images as those used to extract lip geometry for TP-MDTW were also used to generate the DTW and HMM results and the recognition results are shown in Figure 4.8. Compared to DTW and HMM, the classification results show a significant improvement, and the combination of height, width and ratio (HWR) performed the best, with 70.69% of the classification being successful using model 4; the corresponding figures being 55.97% for HMM and 52.22% for DTW. From the results, it can be seen that the simple measures of height and width and their ratio were sufficient to represent the lip dynamic information and proved suitable for the lip reading system.

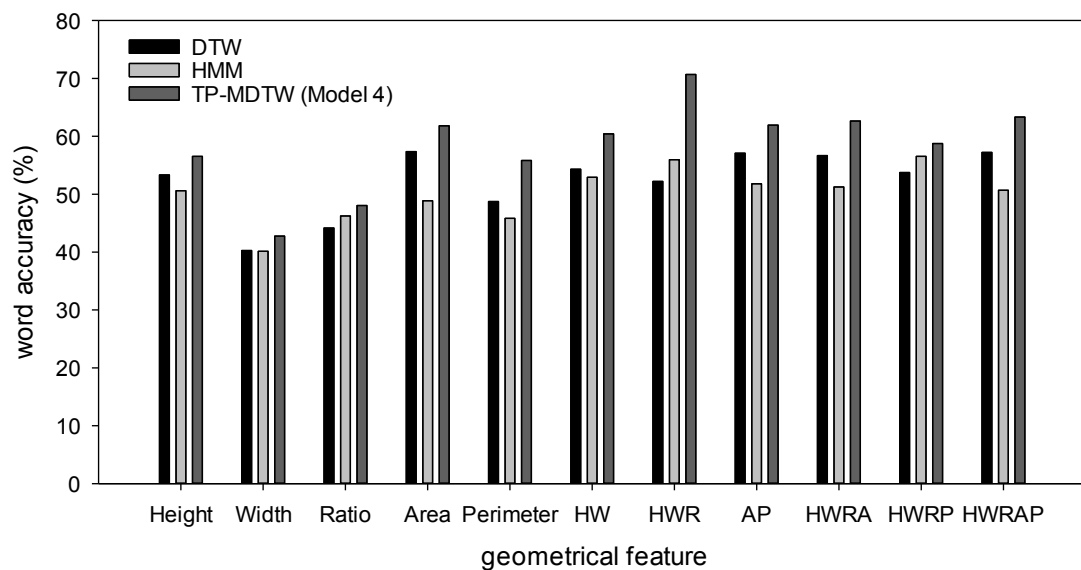


Figure 4.8 Performance of different classifiers using single and combinations of geometrical features

In order to analyse the recognition results, hypothesis testing based on McNemar's test [99] was used to decide whether the differences in performance of two algorithms when applied to the same database are statistically significant. The null hypothesis used for this evaluation is that for two algorithms investigated, the distribution of the outputs from both algorithms are equally likely. IBM SPSS Statistic [100] was used to run the McNemar's test based on the performance of two algorithms, generating a standard statistical report showing that whether to accept or reject the null hypothesis at certain confidence level. It was found that the performance of TP-MTDW using Model 4 compared to the baseline DTW classifier is significant at the 0.001 level (reject null hypothesis). Moreover, a comparison between TP-MTDW using Model 4 and the HMM classifier also exhibited a significance in performance difference at a level of 0.001 (reject null hypothesis). The full statistical report generated by IBM SPSS software can be found in Appendix A.

To assess the performance of the new shape-based approach with respect to motion-based and appearance-based techniques, both OF and DCT methods were implemented in the manner described in [25] and [101]. The same lip images used to extract the lip geometry were supplied to the OF and DCT implementations, producing respectively recognition rates of 26.94% and 57.92%. Using McNemar's test, the shape-based results exhibited a significant difference from the OF and DCT implementations, both at the 0.001 level (reject null hypothesis). The full statistical report generated by IBM SPSS software can be found in Appendix A.

Figure 4.9 shows a direct comparison of the confusion matrices for digits '0' to '9' obtained for OF, DCT and HWR features classified using TP-MDTW (model 4). It is important to note that the recognition results obtained in this work only use three geometrical features (height, width and ratio), while OF and DCT require 8192 features and 16 features respectively.

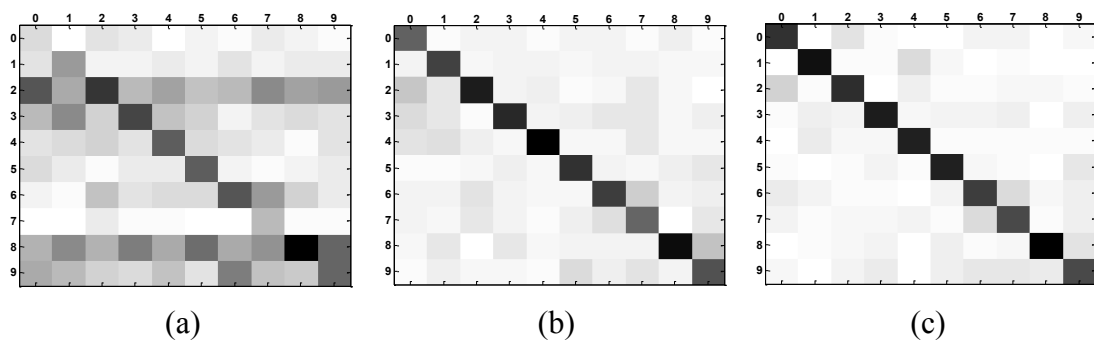


Figure 4.9 Confusion matrices using (a) OF, (b) DCT and (c) HWR features

For any classification approach, an important issue is that of scalability. It is known that additional calculation time will be needed to accommodate applications that may require additional features or dimensions and it is important to demonstrate that TP-MDTW is capable of such an extension. The OF motion-based features was used to estimate the effect of scalability, with the number of features being increased in stages to 200 and each timing measurement carried out 10 times. Figure 4.10 shows the time needed for pair-wise comparison of the features in TP-MDTW and it can be seen that the time taken to complete the underlying operations is only 0.442

ms, with an additional $3.477 \mu\text{s}$ needed for each additional feature. It is clear that the calculation time increases approximately linearly with the number of features. For the application of TP-MDTW to the lip reading application reported in this work, classification involves just three shape-based features and it took 0.45 ms to complete the operation. In comparison, the OF motion-based method with 8192 features took approximately 29 ms, while the DCT appearance-based with 16 features took around 0.49 ms. As appearance-based and motion-based methods generally use far more features than shape-based methods, the latter approach is generally much less computationally intensive.

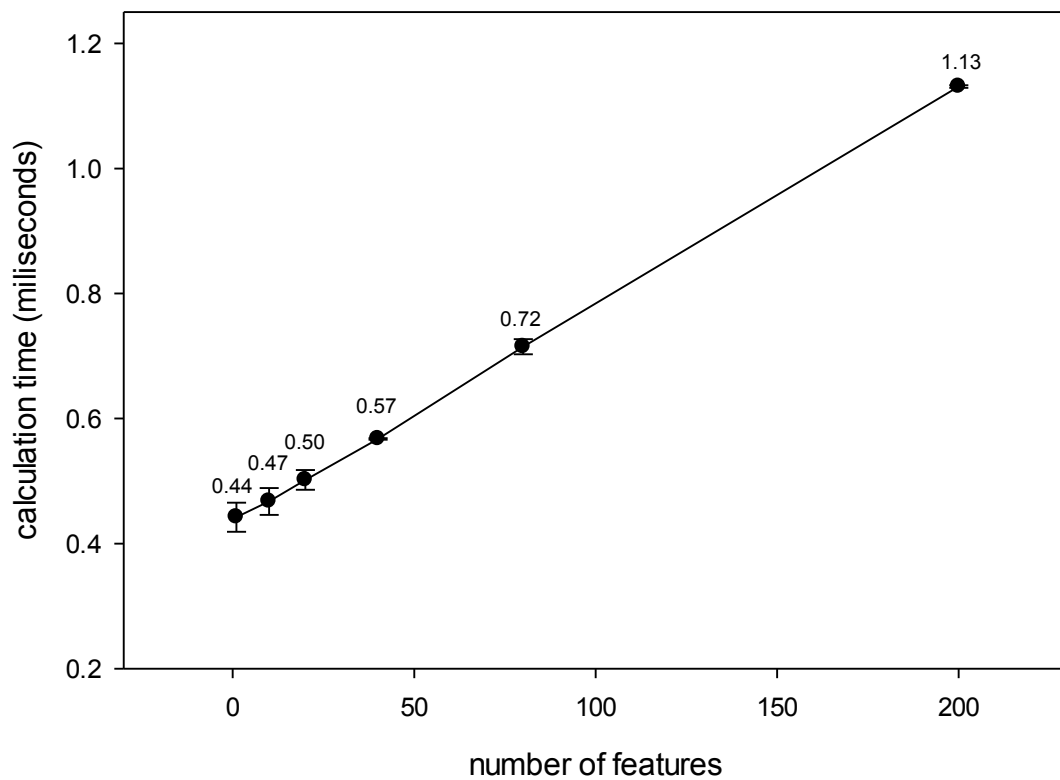


Figure 4.10 Mean calculation times of pairwise comparisons of the features in TP-MDTW using samples from speaker ‘s01m’ in session 1. The error bars indicate ± 1 standard deviation for measurements obtained using 1, 10, 20, 40, 80 and 200 features.

4.4 Chapter summary

The work in this chapter has proposed and evaluated a lip reading system using dynamic information obtained from lip geometry. In tailoring the operation of DTW to multiple features, this new method was able to correlate all the features simultaneously and thus maximize the synchronization between them. Five geometrical features were extracted from each image in the CUAVE database, which are height, width, ratio, area and perimeter and these were used as input to TP-MDTW, a novel technique to classify lip dynamic geometry information using the probability of matching a reference template in the database while performing multi-dimensional DTW. Four models have been proposed to work with this technique.

This work also investigated the performance of a range of lip geometry features operating individually and a number of feature combinations. The experiments showed that the proposed method provides a high performance lip-feature extraction technique and that TP-MDTW is a promising classifier for lip reading. The potential exists to further enhance the current system by including additional models designed to augment those already present in TP-MDTW, with the aim of improving the performance and robustness of the lip reading system as well as providing a high-performing visual component in an audio-visual speech recognition system.

CHAPTER 5

LIP GEOMETRY APPROACH IN FEATURE-FUSION BASED AUDIO-VISUAL SPEECH RECOGNITION

This chapter describes a feature-fusion audio-visual speech recognition (AVSR) system that extracts lip geometry from the mouth region using a combination of skin colour filter, border following and convex hull, and classification using a Hidden Markov Model. By defining a small number of highly descriptive geometrical features relevant to the recognition task, the approach avoids the poor scalability (often termed the ‘curse of dimensionality’) that is frequently associated with feature-fusion AVSR methods. A brief introduction to feature-fusion based classification in AVSR is given in Section 5.1 and a lip geometry approach developed in this work is described in Section 5.2. Finally, the performance of the new approach and comparisons with conventional appearance-based methods, namely DCT and PCA techniques when operating under simulated ambient noise conditions is addressed in Section 5.3.

5.1 Introduction

That visual lip movements are successfully used by people with hearing impairments in order to aid the understanding of speech, promises the potential of being able to improve the robustness of automatic speech recognition in environments where substantial audible ambient noise is present. Studies available in the literature have shown a close correlation between the information present in speech signals and the lip movements themselves, and consequently the addition of visual information has been a line of investigation followed by a number of researchers in their efforts to improve machine perception of the spoken word [102].

Whenever two modalities are to be considered jointly, the question arises as to the processing stage at which the modalities' information content should be fused. In the case of speech and visual speech modalities, fusion can take place either at the feature level (often termed early integration) or the decision level (often termed late integration).

In this work, the feature-fusion approach is adopted and the features extracted for each modality are combined into a common vector to be used by the recognition system. An advantage of this type of fusion is the straightforward extension of techniques already developed for audio-only speech recognition to include the visual aspects, although in a practical implementation, the modalities need to be synchronized and interpolation used to correct for the different frame rates. The main drawback with this approach is that, due to the large number of visual features often acquired, the combined feature vector often becomes considerably longer. In order to reach satisfactory convergence, a substantial increase is required in both the number of training vectors and the training time of the recognition models. In the literature, this problem is known as 'the curse of dimensionality' [103], [104].

5.2 Methodology

The components of the shape-based approach for the AVSR system described in this chapter are shown in Figure 5.1. The system can be divided into the following three stages, namely visual feature extraction, integration and classification. The method adopted for visual feature extraction have been explained in Sections 3.3 to 3.5.

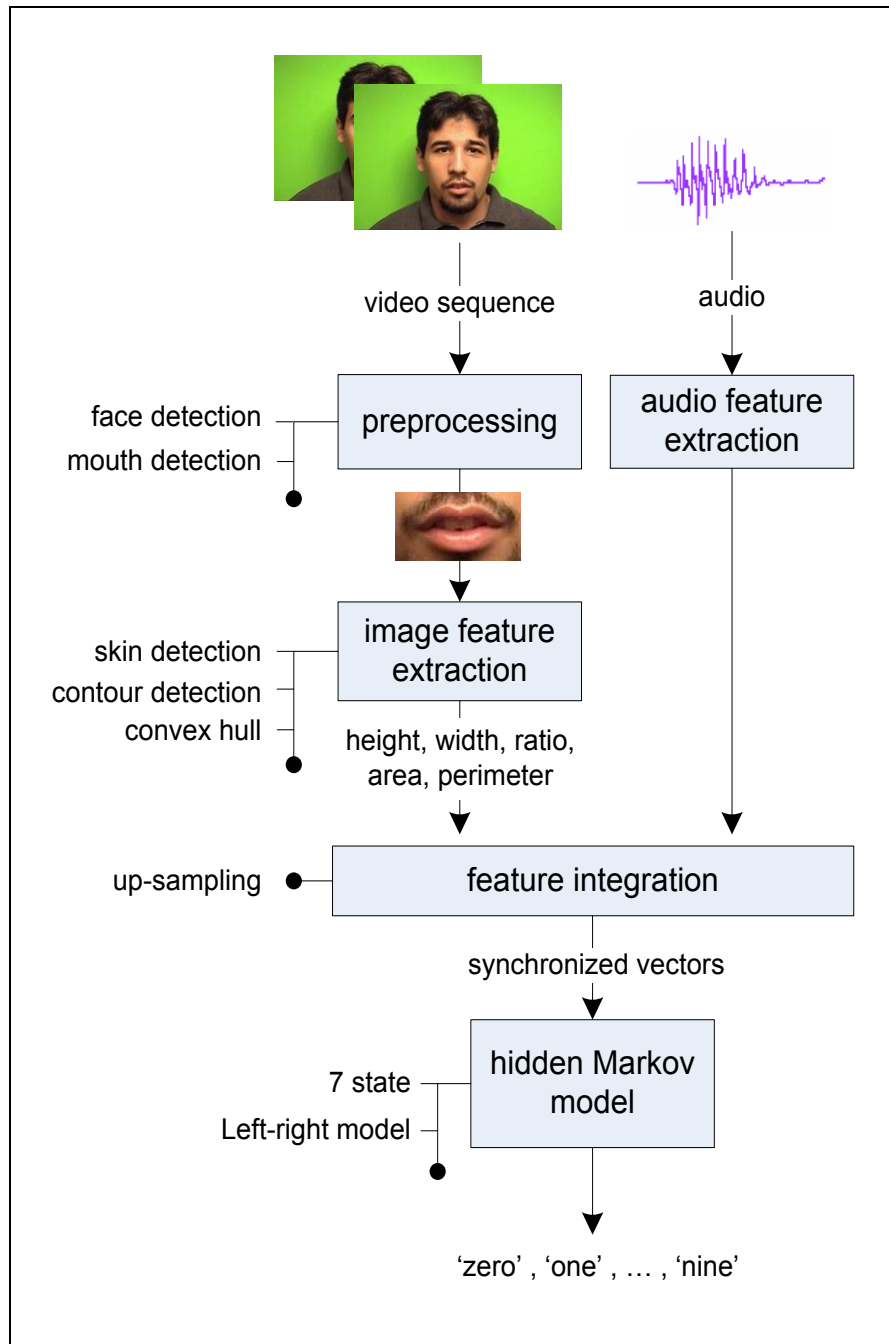


Figure 5.1 A block diagram for shape-based feature fusion AVSR system

5.2.1 Audio-visual feature fusion

The HTK library [89] was used to extract the Mel-frequency cepstral coefficients (MFCC) features and their first and second derivatives, resulting in a feature vector of 39 dimensions. To achieve feature integration, the visual and audio feature extraction rates must be equalized. In the current work, the video frame rate is 29.97Hz, whereas the audio MFCC feature rate is 100Hz. Equalization involved linear interpolation of the visual features to match the audio frame rate. Finally, the audio and visual features are combined and the resulting synchronized feature vector of dimensionality 44 (39 audio features and five visual features) is used for training and testing.

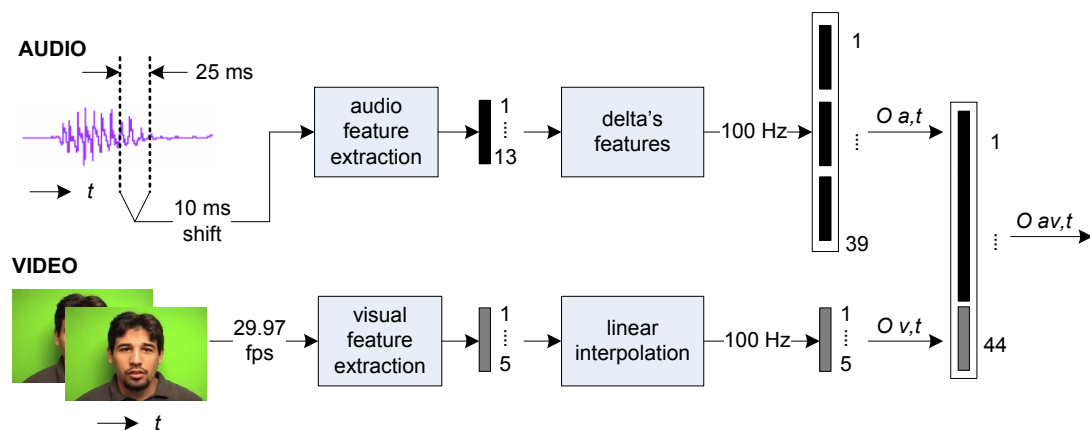


Figure 5.2 Block diagram of the feature fusion for AVSR. The algorithm generates time-synchronous 39-dimensional audio $O_{a,t}$ and 5-dimensional visual feature $O_{v,t}$ vectors at 100 Hz rate

5.2.2 Classification

HMMs are a popular and successful approach to the statistical modelling of audible speech [88]. In this work, 7-state left-to-right models were used with each Markov state being modelled using solely Gaussian functions with diagonal covariance. The models are trained using the Baum-Welch algorithm [14], [105] with recognition performed using the Viterbi algorithm [14], [105].

The CUAVE corpus database [20] was used to investigate the effectiveness of the proposed approaches. The CUAVE database consists of five sessions, where, in each session, the subject speaks the words ‘zero’ to ‘nine’. In this work, data from sessions 1, 2 and 3 (30 samples) were employed for training and data from sessions 4 and 5 (20 samples) were used for testing. All 36 speakers from the database were used in this work, yielding a total of 1800 samples for demonstrating the utility of the proposed approach. For the noise experiments, the training was carried out using the raw speech data and only the test data had noise introduced.

5.3 Results and discussion

This section presents the results obtained for the geometrical-based AVSR system and their comparison with appearance-based AVSR approaches. Both systems are also exposed to simulated variations in environmental conditions that arise from the introduction of audible noise. The software was developed using Microsoft Visual C# 2010 [74] and utilized both the open source image processing library, OpenCV [75] and the Hidden Markov Model Toolkit (HTK) speech processing library [89].

5.3.1 Performance under noise conditions

This section presents results obtained for an AVSR system that utilizes lip geometry information in an attempt to improve the speech recognition rate in noisy environments. The experiments conducted under the influence of ‘babble’, ‘factory1’, ‘factory2’ and ‘white’ noise. These types of noise are part of NOISEX-92 dataset [106] and have been added to the speech signals obtained from CUAVE data corpus, such that specific signal-to-noise ratios (SNRs) are attained.

Figure 5.3 shows the performance of the geometrical-based AVSR system in terms of word accuracy rate as a function of SNR. The audio signal was disturbed by adding ‘babble noise’ from the NOISEX-92 database (100 people speaking in a canteen) provided at a range of intensities such that the SNR lies in the interval 25 to -10dB. It can clearly be seen that as the noise level increases (here specifically at SNRs below 10dB), using combined information from the audio and visual modalities gives the best classification performance of the tests executed. For instance, at an SNR of 0dB the improvement in performance is more than 25% compared with the corresponding audio-only figure.

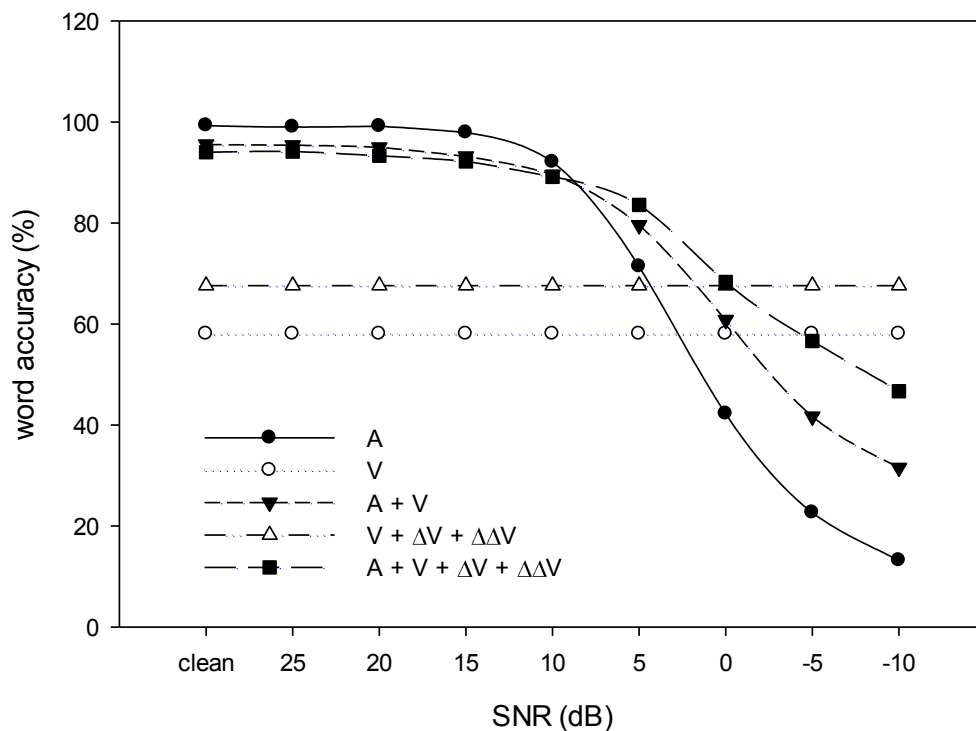


Figure 5.3 AVSR system performance using geometrical features when ‘babble noise’ is applied. Shown are the noisy audio (A), the visual only information (V), dynamic visual information with delta and delta-delta features ($V + \Delta V + \Delta\Delta V$), the combination of audio and visual ($A + V$) features and the combination audio and visual with delta and delta-delta features ($A + V + \Delta V + \Delta\Delta V$)

The classification performance provided by the visual modality was found to be improved if motion information is incorporated by the addition of the first-order (delta) and second-order difference (delta-delta) geometrical features. Such features are commonly used in audio speech recognition and Figure 5.3 shows the improvements in both audio and visual recognition that results from the inclusion of the difference features; for example at an SNR of -5dB the performance can be seen to have been improved by more than 30% with respect to audio-only recognition.

Further tests of the geometrical-based AVSR system were carried out using ‘factory1’, ‘factory2’ and ‘white noise’ datasets from NOISEX-92. Factory1 noise was recorded near to plate-cutting and electrical welding equipment, while factory2 noise was recorded in a car production hall. White noise was acquired by sampling a high-quality analog noise generator operating in the range 0 Hz to 16 kHz. The results obtained for these three types of noise were similar to those of the babble noise, where the combined audio-visual performance at a SNR of -5dB improved by more than 30%.

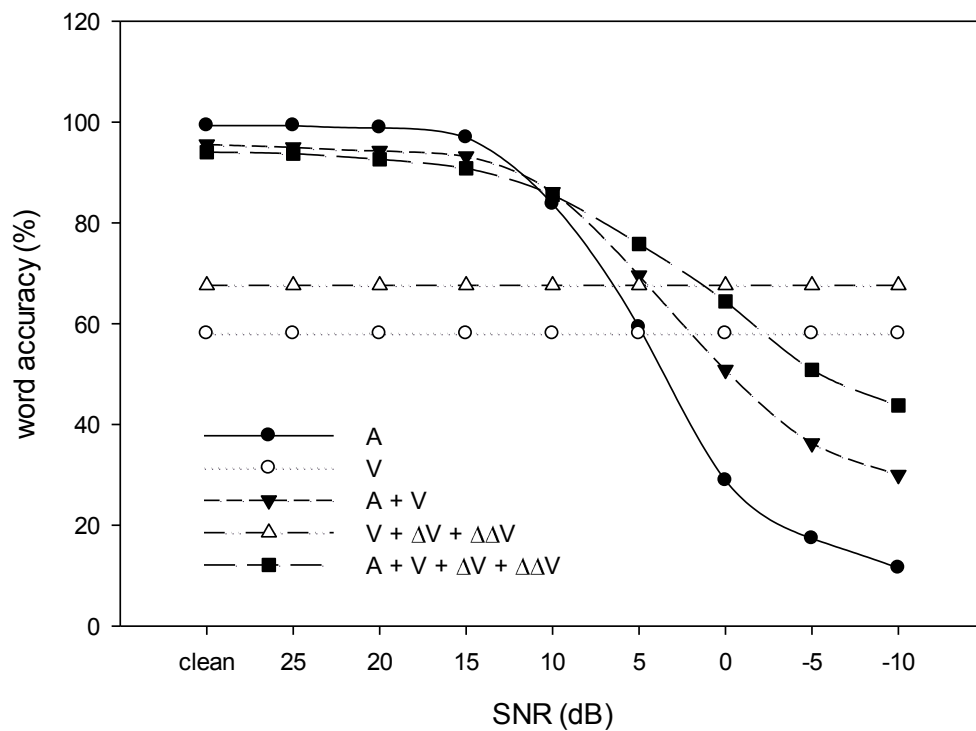


Figure 5.4 AVSR system performance using geometrical features when ‘factory1 noise’ is applied.

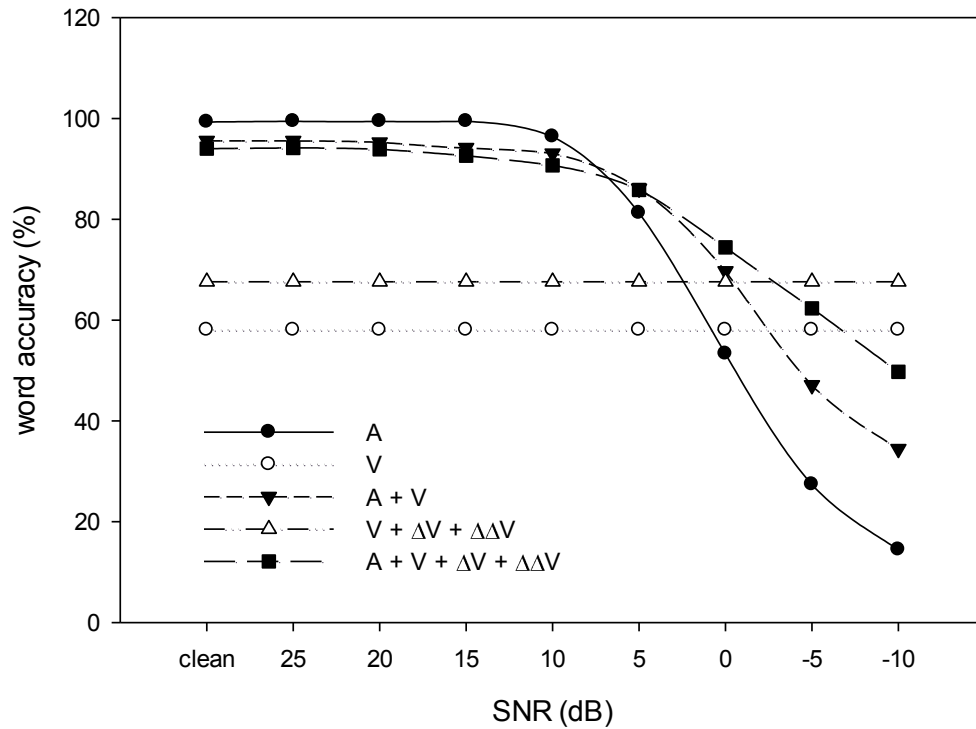


Figure 5.5 AVSR system performance using geometrical features when 'factory2 noise' is applied.

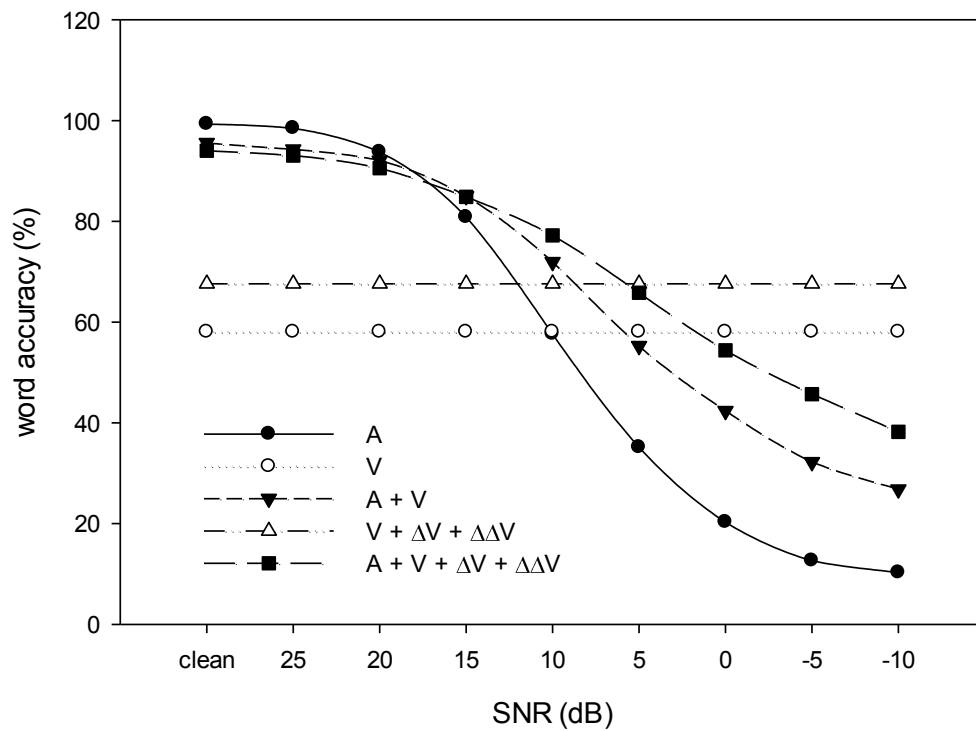


Figure 5.6 AVSR system performance using geometrical features when 'white noise' is applied.

5.3.2 Digit performance analysis

Due to wide variability in the lip movement involved in articulation, not all English digit recognition can be substantially improved by audio-visual integration. For example, when pronouncing the word ‘six’, only small movements of the lips are involved, while in the production of the word ‘seven’ considerable lip movements are required. In general, the greater the lip movements required to generate the word, the better an AVSR system is likely to perform. Figure 5.7 and Figure 5.8 show the recognition performance of the new system in identifying the digit ‘seven’ when simulated under ‘white’ and ‘babble’ noise. The graphs show that the performance when using only visual information is 75% and the combination of the audio-visual information improved the performance by more than 40% relative the audio-only results at SNRs below 0dB.

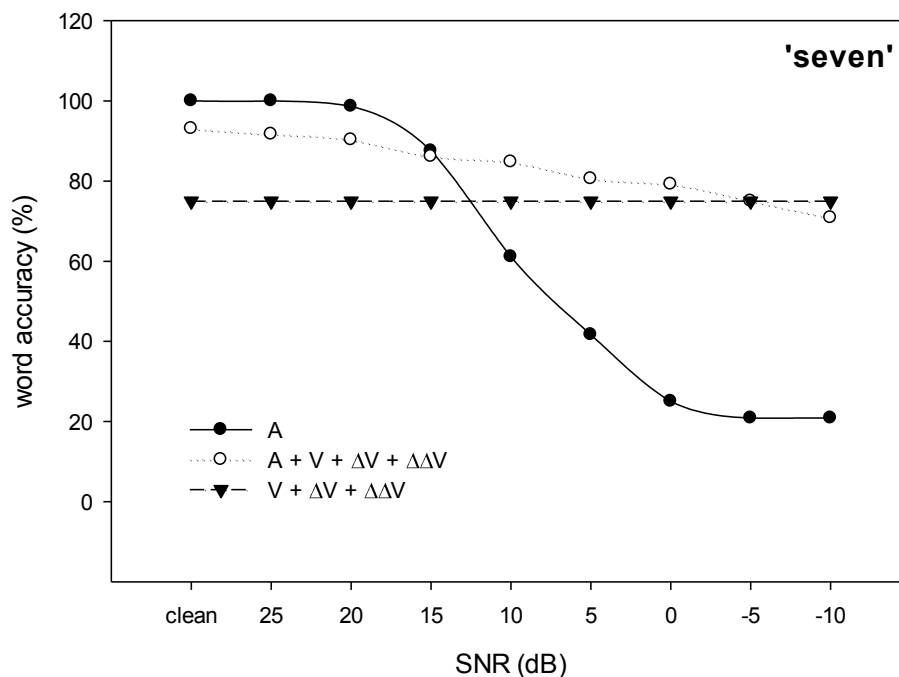


Figure 5.7 AVSR system performance for digit ‘seven’ using geometrical features when ‘white noise’ is applied.

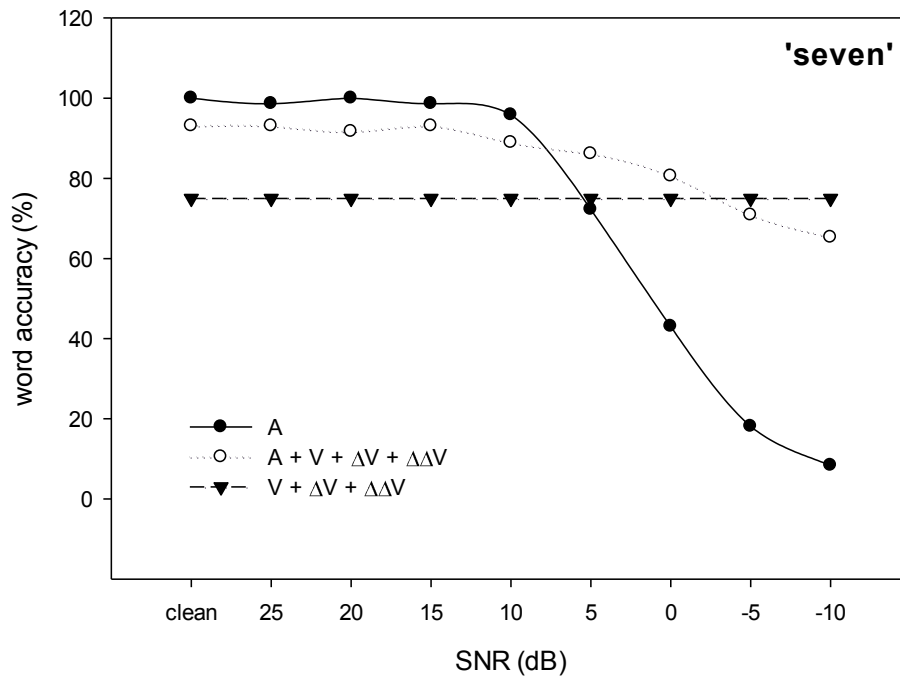


Figure 5.8 AVSR system performance for digit ‘seven’ using geometrical features when ‘babble noise’ is applied.

The contrast in performance for the audio-visual recognition of the digit ‘six’ can be seen in Figure 5.9 and Figure 5.10, again when ‘white’ and ‘babble’ noise are added. The recognition performance using video can be seen to be 50%, such poor performance being due to the minimal movement of the lip when ‘six’ is pronounced. The audio-visual results are also adversely affected, with only up to a 30% improvement compared to audio-only recognition for SNRs less than 0dB. The classification performances of the remaining digits investigated in this work can be found in Appendix B and Appendix C.

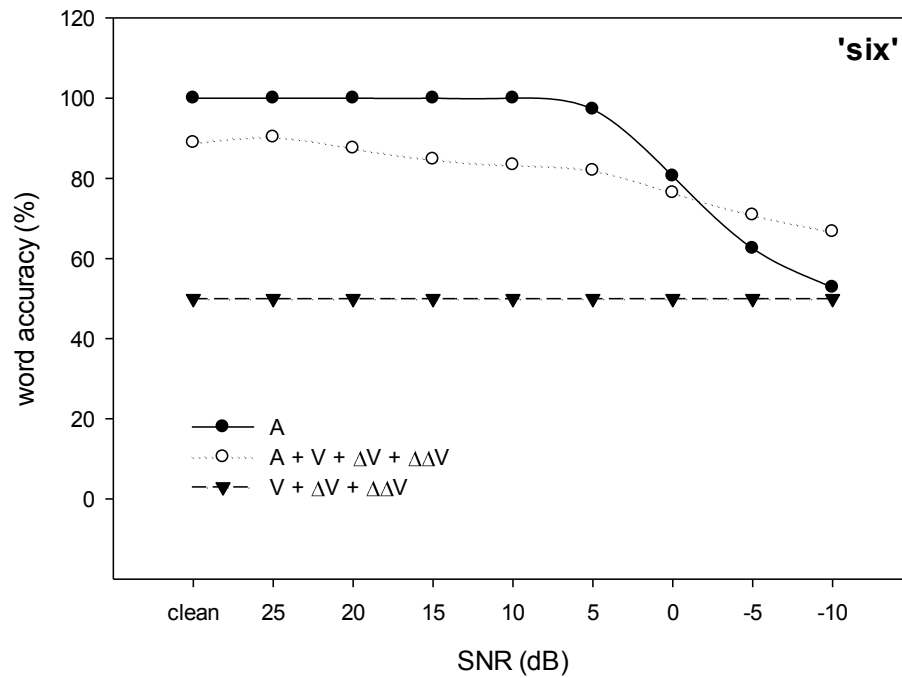


Figure 5.9 AVSR system performance for digit 'six' using geometrical features when 'white noise' is applied.

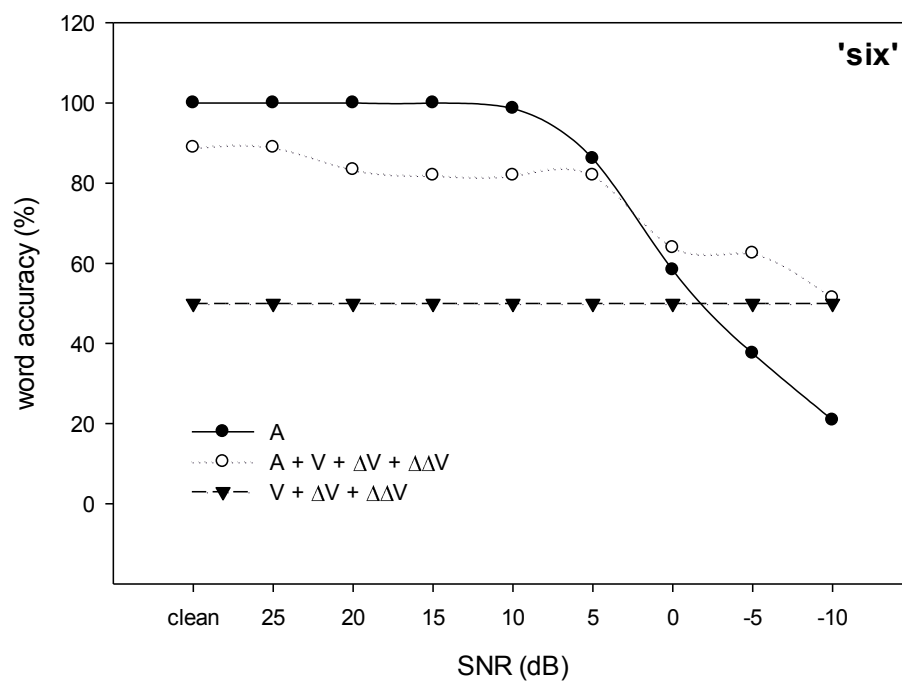


Figure 5.10 AVSR system performance for digit 'six' using geometrical features when 'babble noise' is applied.

5.3.3 Digit confusion analysis

To understand further the digit recognition of the new shape-based AVSR system, a confusion analysis is now described in order to identify under what circumstances misclassification occurs.

Table 5.1 shows the confusion matrix for the audio-only speech recognition and Table 5.2 for the audio-visual system, including lip geometry features and its derivative at 0dB SNR with white noise added. The white noise was again obtained from the NOISEX-92 database. As seen in confusion matrix in Table 5.1, the similarity between white noise and digit six will bias the recognizer in favour of this digit. This is because digit six begins and ends with sibilant sounds. The accuracy of the digit six classifying correctly is the highest among other achieving up to 80%.

Table 5.1 Confusion matrix for audio only recognition at 0 dB SNR using white noise

Actual	Predicted									
	zero	one	two	three	four	five	six	seven	eight	nine
zero	20.8	2.8			2.8	4.2	45.8	18.1	5.6	
one	15.3	4.2	1.4		5.6	9.7	26.4	19.4	8.3	9.7
two	8.3	2.8	5.6	1.4		1.4	59.7	13.9	5.6	1.4
three	8.3	2.8	2.8	2.8		4.2	59.7	11.1	5.6	2.8
four	18.1	2.8			4.2	5.6	44.4	16.7	6.9	1.4
five	8.3	4.2			4.2	25.0	26.4	20.8	5.6	5.6
six	2.8						80.6	12.5	4.2	
seven	5.6	1.4		1.4	1.4	8.3	51.4	25.0	5.6	
eight	4.2	1.4	6.9	2.8			62.5	8.3	12.5	1.4
nine	9.7	2.8			4.2	8.3	25.0	20.8	6.9	22.2

Note. Correct predicted percentages have been pooled over participants and digit contexts.

By combining the visual and speech information in noisy environments, the new system achieved better performance than audio alone as shown in Table 5.2. It can be seen that the recognition of all digits improves considerably, except digit six which worsens slightly by 4.2%. The substantial differences in the results for digits ‘one’ and ‘seven’ are probably due to the greater movement of the lips required in the articulation; in contrast digit ‘six’ and ‘nine’ that require only small lip movements in their articulation, showed only small changes in recognition performance.

Table 5.2 Confusion matrix for audio-visual recognition at 0dB SNR using white noise

Actual	Predicted									
	zero	one	two	three	four	five	six	seven	eight	nine
zero	47.2		4.2			6.9	23.6	15.3		2.8
one	4.2	68.1	2.8	1.4	2.8	11.1	5.6	1.4		2.8
two	18.1	2.8	41.7	4.2		2.8	11.1	18.1		1.4
three	9.7	8.3	1.4	37.5	1.4	5.6	20.8	11.1		4.2
four	16.7	16.7		1.4	38.9	11.1	5.6	5.6		4.2
five		1.4				65.3	8.3	16.7		8.3
six	6.9		1.4			2.8	76.4	11.1	1.4	
seven	2.8					5.6	12.5	79.2		
eight	4.2		1.4			8.3	16.7	8.3	56.9	4.2
nine	9.7	2.8				11.1	18.1	22.2	2.8	33.3

Note. Correct predicted percentages have been pooled over participants and digit contexts.

Further tests of the shape-based AVSR system were carried out using the ‘babble’ noise datasets from NOISEX-92. Table 5.3 shows the confusion matrix for the audio-only speech recognition system when simulated using babble noise (100 people speaking in a canteen) at a SNR of 0dB. It can be seen that digits ‘zero’, ‘one’ and ‘nine’ are highly effected by babble noise compared to the remaining digits.

Table 5.3 Confusion matrix for audio only recognition at 0dB SNR using babble noise

Actual	Predicted									
	zero	one	two	three	four	five	six	seven	eight	nine
zero	72.2	2.8			4.2	4.2	1.4	8.3	1.4	5.6
one	6.9	62.5			1.4	2.8	2.8			23.6
two	41.7	9.7	20.8	1.4		1.4	4.2	8.3		12.5
three	40.3	16.7	4.2	8.3	1.4	8.3	2.8	5.6		12.5
four	30.6	36.1			9.7	8.3		4.2		11.1
five	8.3	23.6				51.4				16.7
six	19.4	4.2					58.3	12.5		5.6
seven	29.2	9.7				1.4	4.2	43.1		12.5
eight	27.8	12.5	4.2	4.2		2.8	2.8	5.6	23.6	16.7
nine	5.6	16.7				4.2		1.4		72.2

Note. Correct predicted percentages have been pooled over participants and digit contexts.

The performance of the AVSR system was able to improve recognition performance compared with the audio-only system as shown in Table 5.4. Recognition accuracy of all digits improved, especially digits ‘three’ and ‘five’ by 45.8% and 40.3% respectively. Those digits showing only a small improvement in performance were digits ‘six’ and ‘nine’, a similar result to that obtained using white noise. This demonstrates that the improvement offered by AVSR depends largely on the quality of the visual information presented to the system.

Table 5.4 Confusion matrix for combination of audio and visual at 0 dB SNR using babble noise

Actual	Predicted									
	zero	one	two	three	four	five	six	seven	eight	nine
zero	80.6	1.4			1.4	1.4	4.2	8.3		2.8
one		90.3	1.4		1.4	1.4	2.8			2.8
two	26.4	2.8	48.6	4.2	1.4		1.4	12.5		2.8
three	6.9	13.9	1.4	54.2	1.4	4.2	2.8	11.1		4.2
four	8.3	37.5		1.4	45.8	4.2		1.4		1.4
five		2.8				91.7	1.4			4.2
six	15.3					1.4	63.9	16.7		2.8
seven	12.5					2.8	2.8	80.6		1.4
eight	8.3	1.4		2.8		2.8	2.8	4.2	52.8	25.0
nine	4.2	2.8				9.7	1.4	6.9		75.0

Note. Correct predicted percentages have been pooled over participants and digit contexts.

5.3.4 Comparison with appearance based system

A common alternative to the geometrical-based visual features used in the current work is to adopt appearance-based features that extract information directly from the whole mouth region [101]. Results are now presented that compare the performance of the appearance-based approach with the new geometrical-based approach described in the previous section.

To generate the appearance-based features, the mouth region was first resized to a bounding box of 64x64 pixels, resulting in a feature vector of 4096 dimensions for each mouth image. To reduce the features to a number more manageable for processing, transformation to a lower dimensionality is required and this is normally achieved using an image compression or pattern classification technique. Examples of such transforms are the DCT and PCA. In the experiments performed in this work, the DCT was applied using a method similar to [101] and the information contained in the 16, 64 and 192 lowest frequency coefficients were kept. The PCA was also applied using an approach similar to that found in [27], and appearance-based feature vector lengths of 64, 128 and 256 were generated.

The same set of lip images used to extract lip geometry was provided for the appearance-based implementation. Figure 5.11 and Figure 5.12 show the classification results for appearance-based AVSR using the DCT and PCA feature reduction methods respectively, both with the addition of babble noise. It can be seen that as the SNR falls below 5dB, the combined information gives better results than the audio alone. These results accord with work reported by Lucey *et al.* [27], in which the authors used the same integration strategy.

In comparison with the shape-based results shown in Figure 5.3, that were simulated under the same noise conditions, the appearance-based results show a poorer classification performance especially when the SNR is above 0dB. As mentioned earlier in the introduction (section 5.1), the main drawback with feature-fusion based AVSR is that, due to the large number of visual features often acquired, the combined feature vector often becomes considerably longer than the audio-only vector. DCT and PCA required large numbers of features to represent lip information (16, 64, 128, etc.) while the new shape-based approach uses only five features to represent lip information. Figure 5.11 and Figure 5.12 demonstrate that the more dominant visual information is in the feature vector, the more closely the recognition follow video-only modality performance. In the literature, this problem is known as the ‘curse of dimensionality’ [103], [104].

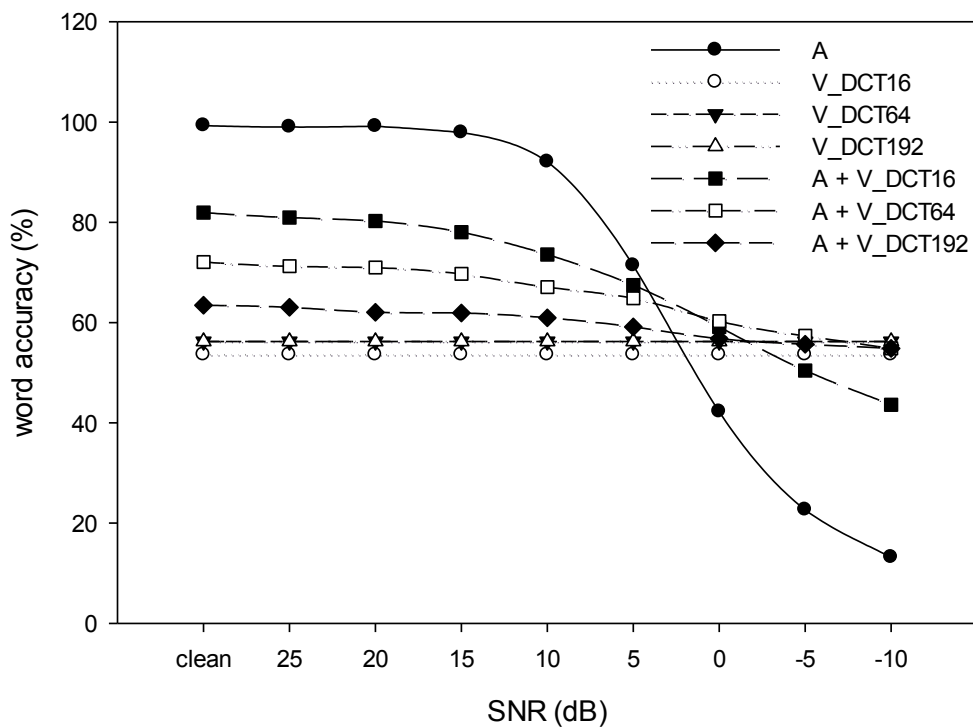


Figure 5.11 AVSR system performance using DCT features with babble noise applied. The figure shows a comparison between noisy audio (A), visual only information using 16 DCT features (V_DCT16), 64 DCT features (V_DCT64) and 192 DCT features (V_DCT192), a combination of audio and 16 DCT visual features (A + V_DCT16), a combination of audio and 64 DCT visual features (A + V_DCT64) and a combination of audio and 192 DCT visual features (A + V_DCT192)

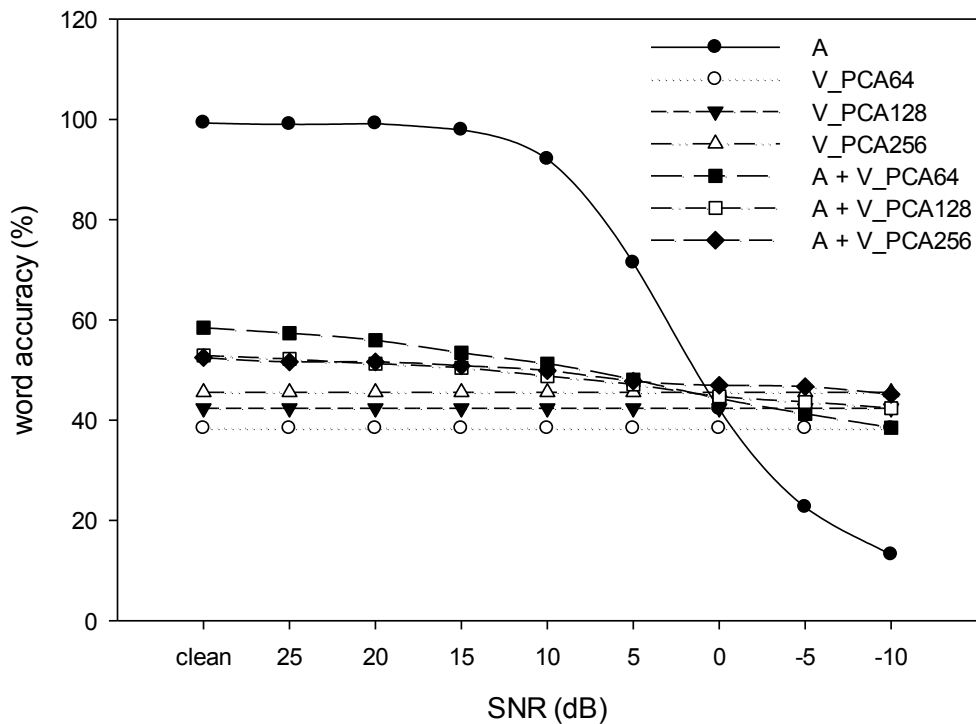


Figure 5.12 AVSR system performance using PCA features with babble noise applied. The figure shows a comparison between noisy audio (A), visual only information using 64 PCA feature (V_PCA64), 128 PCA feature (V_PCA128) and 256 PCA features (V_PCA256), a combination of audio and 64 PCA visual features (A + V_PCA64), a combination of audio and 128 PCA visual features (A + V_PCA128) and a combination audio and 256 PCA visual features (A + V_PCA256)

5.4 Chapter summary

A range of experiments has been carried out to demonstrate that the geometrical features established in this work have information content that is highly relevant for the recognition task, while also suffering little from the ‘curse of dimensionality’ (or scalability) issue that is often a bane in feature-fusion based AVSR systems. The results were compared to those obtained using conventional appearance-based feature methods, namely the Discrete Cosine Transform and Principal Component Analysis techniques, when operating under a range of different signal to noise ratio conditions. Experimental results show that the new geometrical-

based features outperform appearance-based features in terms of recognition accuracy, even though a significantly reduced number of features is used.

This work has also investigated the performance of a shape-based AVSR system applied to digit recognition under noisy conditions and the results presented using confusion matrices. The experiments showed that the recognition of individual digits exhibits a performance that depends substantially on the magnitude of the movement of the lips required in its articulation. The potential exists to further enhance the current AVSR system by using a decision-fusion approach where recognition is performed separately for each of the modalities, with the partial results from each sub-process being combined to produce the final classification. This topic will be considered in the next chapter.

CHAPTER 6

NOVEL ADAPTIVE FUSION IN AUDIO-VISUAL SPEECH RECOGNITION USING AUDIO STATISTICAL ANALYSIS

Augmenting speech signals with information from lip movements is known to improve recognition performance under such conditions, but the question remains as to how best to combine the audio and visual data. In this chapter, a statistical tool is used to assess the quality of the speech signal in order to automatically select between the audio modality, the visual modality or a combination of the two, so as to achieve the best recognition performance. A brief introduction to fusion strategy in AVSR is given in Section 6.1 and the work involves an analysis of the effectiveness of the statistical tool in identifying the noise content of speech signals using skewness and kurtosis values of their probability distributions in Section 6.2. A novel adaptive fusion developed in this work that uses these statistical moments is detailed in Section 6.3. Finally the performance of the new method is then demonstrated for speech signals across a range of environmental noises and is addressed in Section 6.4.

6.1 Introduction

AVSR studies have often been inspired by how humans are assumed to perform bimodal perception [3], [54], [60]. This has led to the implementation of two separate approaches in which fusion can take place either before information is processed by the recognizer, termed feature fusion, or after the audio and visual information have been separately classified, known as decision fusion [27]. As explained details in literature (Section 2.1), in feature fusion, the raw data are combined, typically by simple concatenation or by using a sum of weighted values. This approach assumes that there is synchronization and direct correspondence between the audio and visual information at the lowest levels of human speech

perception. In the decision fusion process, the audio and visual signals are treated independently, requiring two separate recognition systems whose outputs are mathematically combined, typically using a weighted summation or product. This approach allows more freedom to the user and avoids audio-visual asynchrony issue at word unit level.

Which is the more effective of the two AVSR fusion strategies under noisy audio conditions remains to be resolved. However, as decision fusion often delivers different partial classification decision outcomes, this approach is able to provide a suitable method for ranking and collation. This can be easily achieved by using an adaptive weight to adjust the relative contribution of each partial outcome in making a final decision.

In this work, a new approach, termed ‘adaptive fusion’ is introduced in which the decision to base the speech recognition on audio-only information, visual-only information, or a combination of the two, is automatically determined according to the prevailing quality of the audio signals being received. For example, in an environment where the audio quality is good, in that it is little affected by noise, the AVSR system is likely to perform best if audio recognition dominates and visual information is not used. When the speech quality is significantly degraded by the presence of audio noise, a better recognition performance is likely to be achieved by considering visual rather than audio signals. In intermediate cases where noise only partially affects recognition, better performance may be achieved by taking as input a suitable mixture of audio and visual information. Assuming that the visual signal is not affected by noise and that the speakers’ lips are visible and can be detected to a given resolution in the plane of the image.

There are two novelties in the work described. Firstly, a method is proposed that uses statistical analysis to determine the relative content of audio noise present in speech signals. Secondly, an adaptive fusion AVSR system has been developed whose assessment of the quality of the audio is used to select an appropriate fusion approach with the aim of improving AVSR performance. A simple weight

assignment is applied to the audio and visual recognizers to generate suitable fusion. Determining proper weight to each fusion output is crucial to get best of the system representation.

6.2 Audio statistical analysis

The estimation of the quality of the speech signal, or SNR has been widely investigated for decades and has been successfully implemented in the field of speech enhancement [107] and speech recognition [108]. It is easier to compensate the effects of the noise in the audio when having a knowledge of the SNR. There are several methods to estimate the value of SNR. One of the method is based on measurement of the energy in certain frequency band [109]. The disadvantage of this method is that the limitation of the usage with high level noise. Another method is based on the analysis of the noise spectral as investigated by [108]. However spectral analysis required higher computational costs. Apart from energy and spectral analysis, there is also an approach that is based on audio statistics analysis, for example in [110], the kurtosis value has been used to estimate the SNR level in each frequency band. This technique requires less computational cost and easy to implement.

In this study, the method chosen to estimate the SNR is based on audio statistical analysis due to the easy implementation. The statistical value can be obtained directly from waveform samples rather than from energy and spectral coefficients. In this section, the background of statistical analysis for measuring the characteristics of the audio signal are described in section 6.2.1 and initial experiments to investigate the possibility of implementing statistical analysis to adaptive fusion system are described in section 6.2.2.

6.2.1 Background of skewness and kurtosis

The first two statistical moments, mean and variance, provide information on the location and variability of a set of data. The third and fourth statistical moments, termed skewness and kurtosis respectively, provide additional information on the shape of frequency distribution. The skewness of the distribution relates to its symmetry, where in a symmetrical distribution the ordinate at the mean value divides the distribution into two equal halves as shown in Figure 6.1. When the distribution is clearly non-symmetrical the distribution of the data is considered 'skewed'. If the tail extends to right and the mass of the distribution is concentrated on the left of the figure, then the distribution is said to be positively skewed. Conversely, if the tail extends to the left and the mass of the distribution is concentrated on the right of the figure then the distribution is said to be negatively skewed [111], [112]. For a normal distribution, the skewness value is 0 (symmetrical).

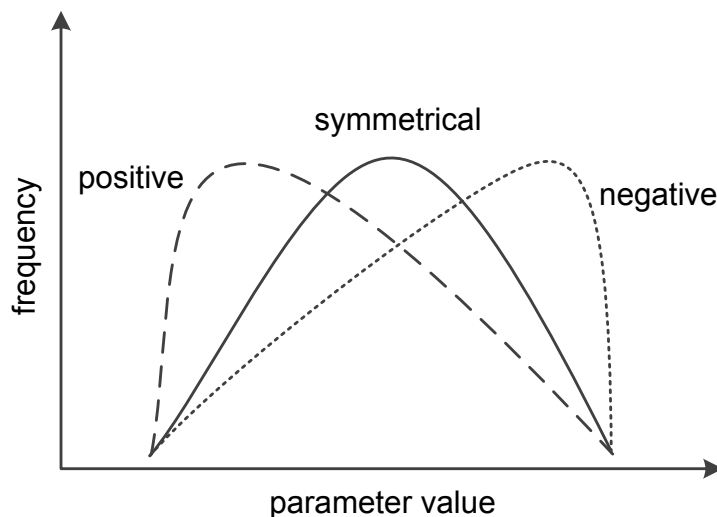


Figure 6.1 Skewness distribution examples

Consider $\mathbf{x} = [x_1, x_2, \dots, x_n]$ as sample audio value and μ is the mean of \mathbf{x} , then the skewness of the audio distribution can be calculated as [113]

$$s = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^3}{\left(\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2} \right)^3} \quad (6.1)$$

Kurtosis is often used to quantify how pronounced is the main peak of the statistical distribution. If the peak is more pronounced than that of a normal distribution, it is termed leptokurtic, if equally pronounced it is mesokurtic and platykurtic if less pronounced, as shown in Figure 6.2 [112]. For a normal distribution (mesokurtic), the kurtosis value is 3 [111], while leptokurtic and platykurtic have a value more than 3 and less than 3 respectively.

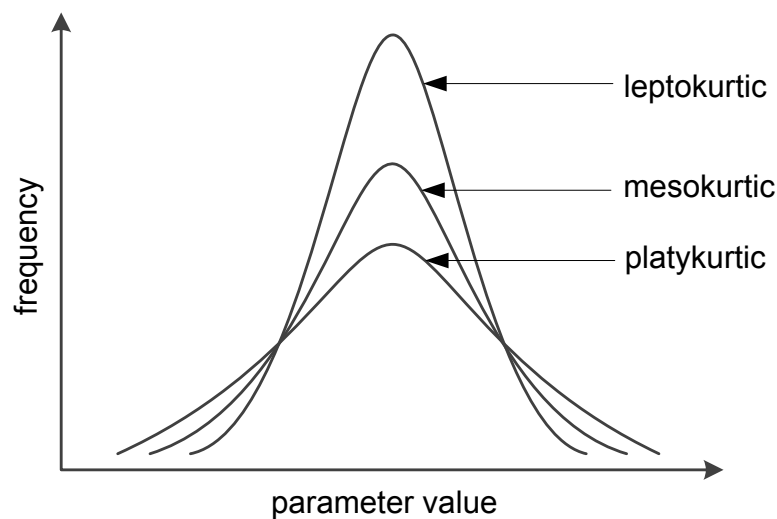


Figure 6.2 Kurtosis distributions example

Again, consider $\mathbf{x} = [x_1, x_2, \dots, x_n]$ as sample audio value and μ is the mean of \mathbf{x} , then the kurtosis of the audio distribution can be calculated as [113]

$$k = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 \right)^2} \quad (6.2)$$

6.2.2 Initial investigation on statistical variation of the audio signals

To ensure that statistical analysis can be used to determine the characteristic of the audio in noisy environment, a simple experiment was conducted by analysing the distribution of the audio data when it is corrupted by the artificial noise. A histogram was used to display the distribution of the audio data so that the information about the central tendency and dispersion as well as symmetrical and peakedness can be seen easily.

An example is now given in which an audio file from the CUAVE database was used [20], specifically the digit ‘seven’ uttered by speaker ‘s01m’ in session 1. The experiment was conducted under the effects of ‘babble’, ‘factory1’ and ‘factory2’ noise obtained from the NOISEX-92 dataset [106] and were each separately added to the speech signal to produce specific values of SNR. Figures 6.3 to Figure 6.5 show examples of the frequency distributions of speech signals when noise at SNRs of 0dB and -10dB were added.

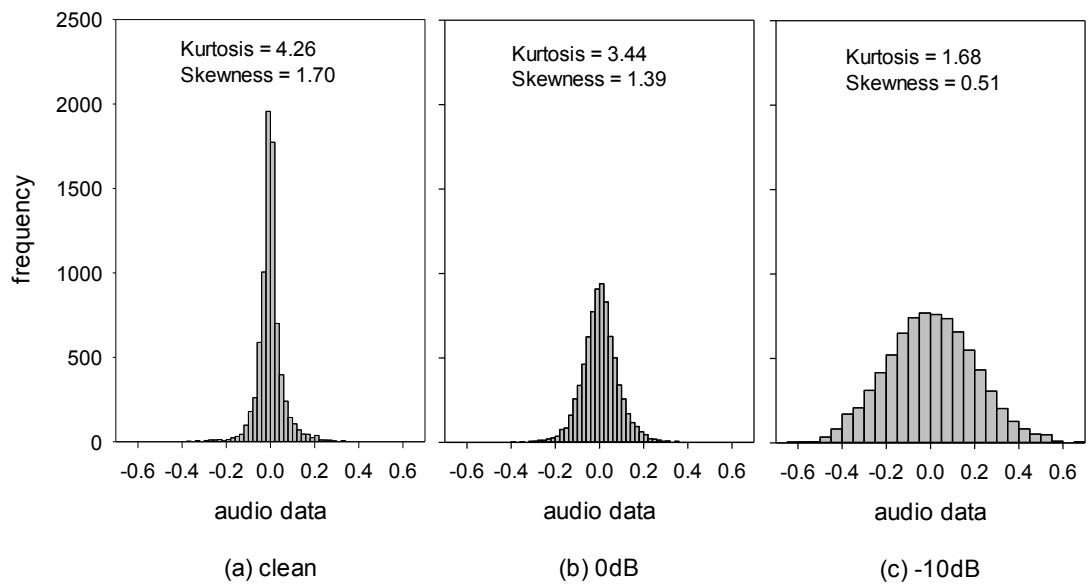


Figure 6.3 Histograms of digit 'seven' uttered by 's01m' in 'babble' noise applied at a range of SNR values

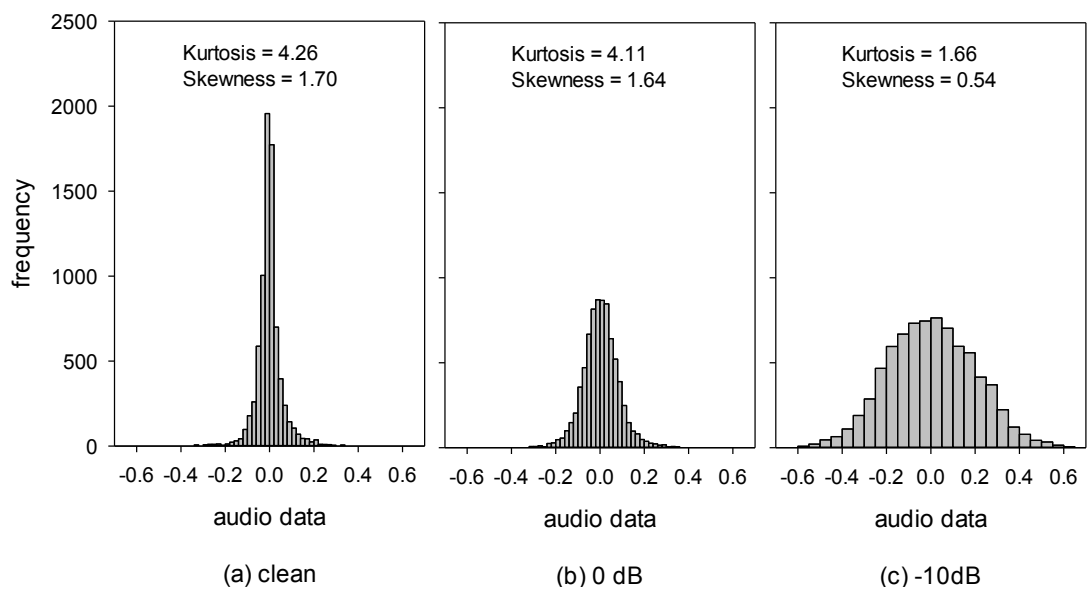


Figure 6.4 Histograms of digit 'seven' uttered by 's01m' in 'factory1' applied at a range of SNR values

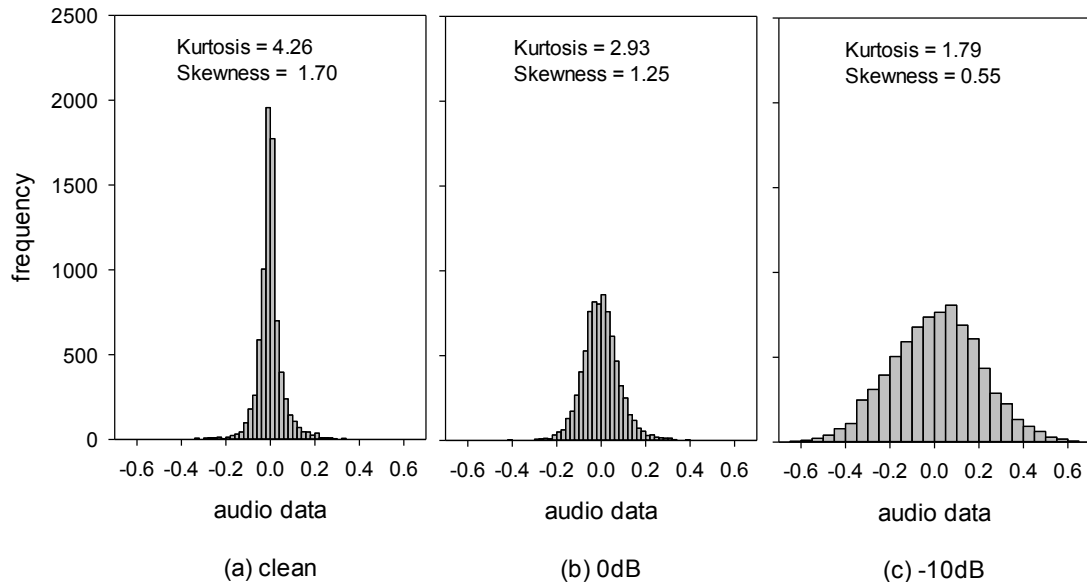


Figure 6.5 Histograms of digit ‘seven’ uttered by ‘s01m’ in ‘factory2’ applied at a range of SNR values

It can clearly be seen that as the noise content in the audio signal is increased, values of skewness and kurtosis both decrease and the shape of the distribution alters significantly and closer to that of a noise distribution as shown in Figure 6.6.

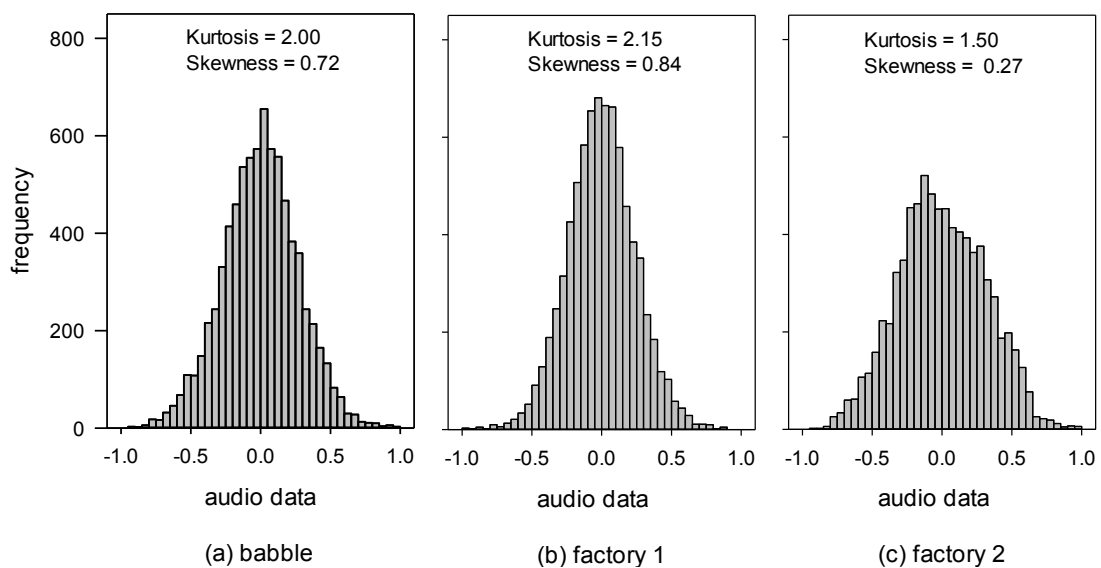
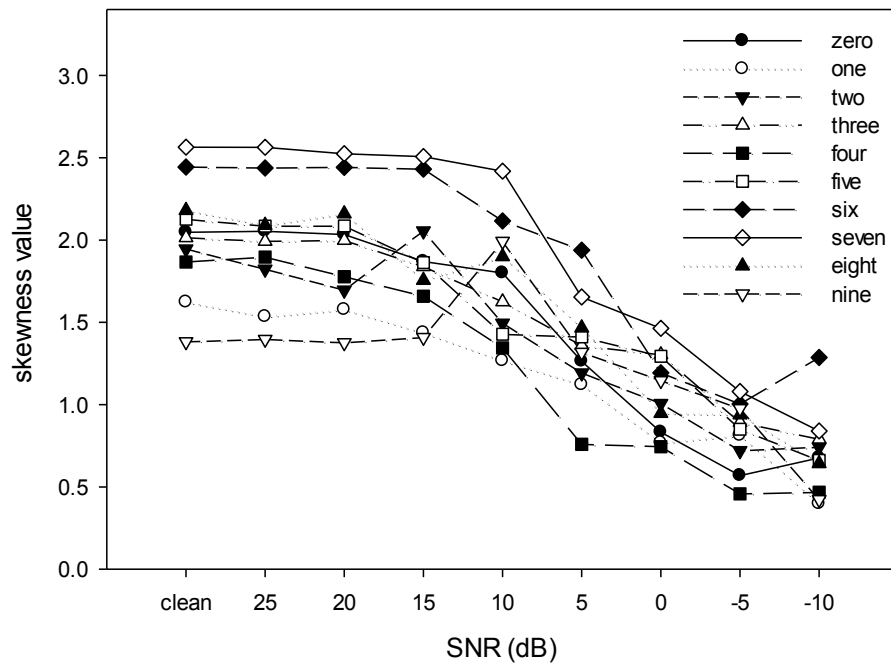


Figure 6.6 Histograms of noises

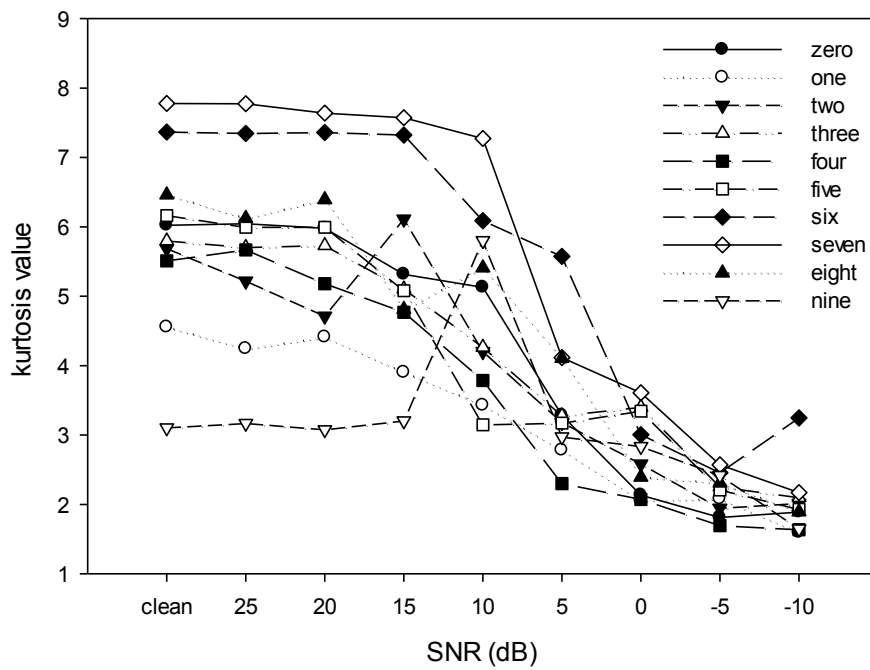
Further experiments are needed to demonstrate that the values of skewness and kurtosis are affected by the increased noise content and the distribution tends to be closer to the noise distribution regardless of the test signals used.

For example, the skewness and kurtosis values for speaker 's01m' in the CUAVE database are shown in Figures 6.7 to Figure 6.9 for 'babble', 'factory1' and 'factory2' noise respectively. It can be seen that the values of skewness and kurtosis generally drop towards its respective noise as shown in Figure 6.6 as more noise is added to the speech as expected by the application of the CLT. The trends of both the skewness and kurtosis values appear to be significant in indicating noise content. It can also be seen that, for specific SNR values, the differences between the skewness and kurtosis appear to be reduced as the SNR is increased and this is probably due to the distributions more closely becoming that of a noise distribution.

To obtain a suitable combination of audio and visual signals, a weight value needs to be determined from the skewness or kurtosis parameters. In providing the results in Figures 6.7 to Figure 6.9, the different noise sources have different effects on the manner in which skewness and kurtosis change and no simple relationship between these values and the SNR could be determined. As a solution is required that is able to operate regardless of the type of noise that is present, a single type of noise was used in training, namely white noise, and the selections from skewness and kurtosis to the audio and visual weightings developed. Clearly, the performance of such a solution is likely to be worse than a system that is able to analyze and model the specific noise type found in each individual case, and such an approach may form the basis of an adaptive fusion AVSR solution. If training using white noise is adequate in providing suitable skewness and kurtosis thresholds for each digit, it would potentially be possible to select an appropriate modality or a suitable mixture of modalities according to audio noise content.

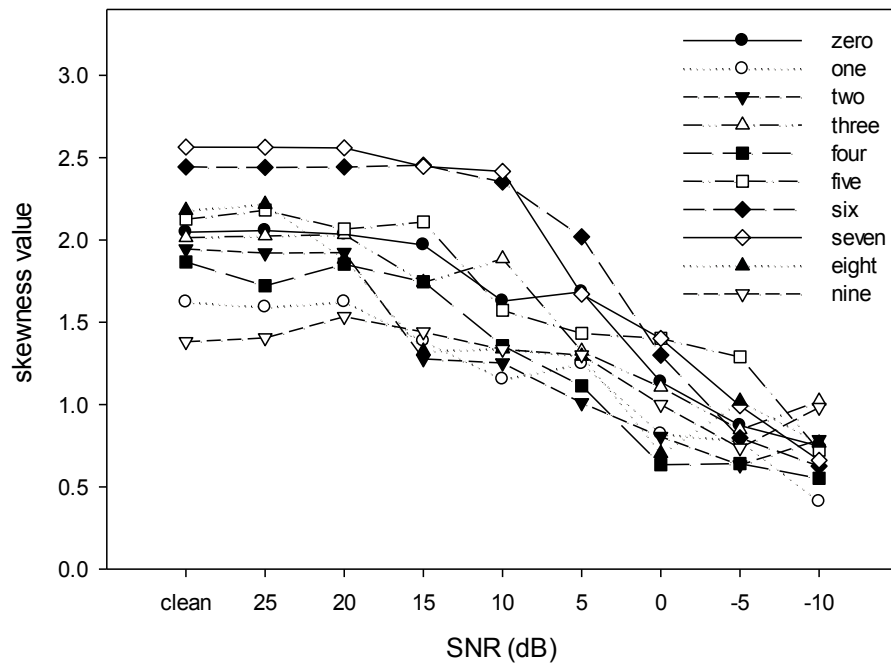


(a) skewness

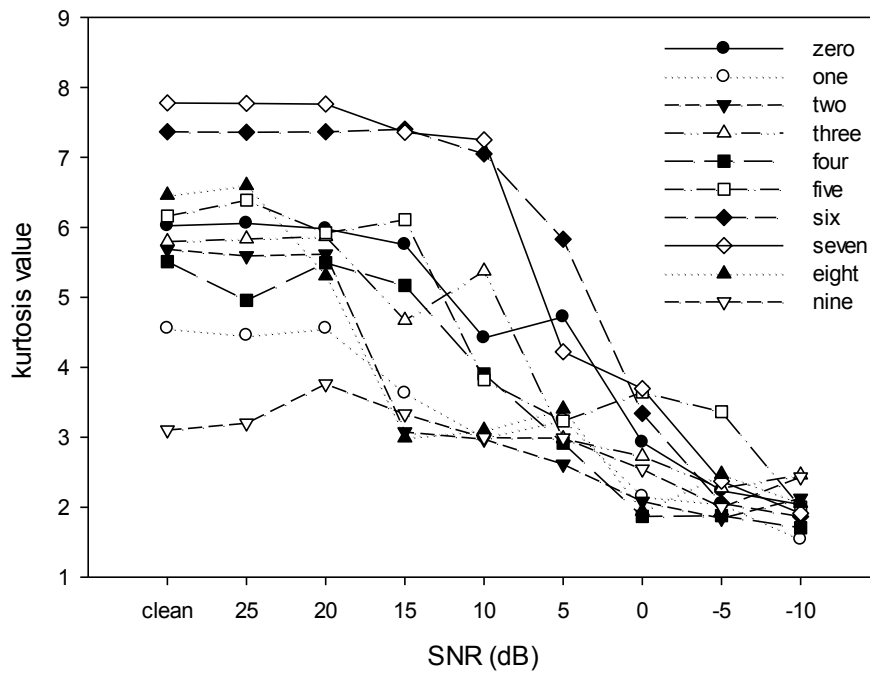


(b) kurtosis

Figure 6.7 Statistical parameters obtained under various levels of babble noise

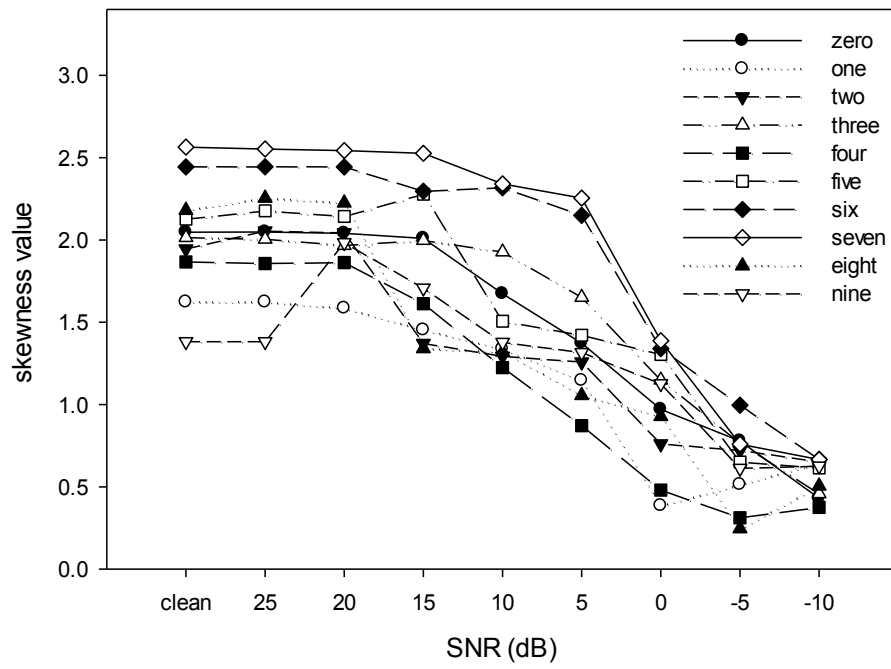


(a) skewness

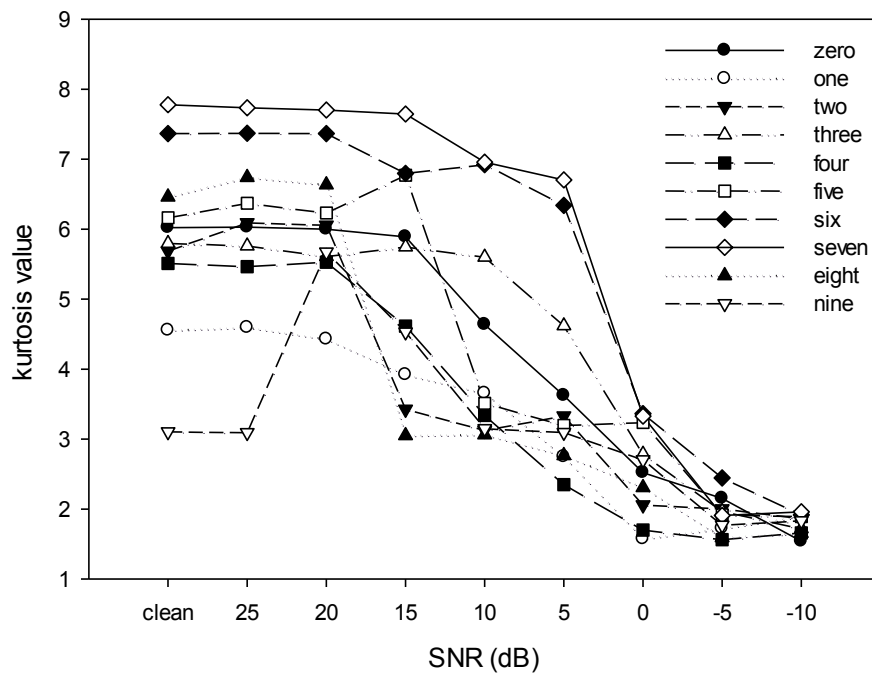


(b) kurtosis

Figure 6.8 Statistical parameters obtained under various levels of factory1 noise



(a) skewness



(b) kurtosis

Figure 6.9 Statistical parameters obtained under various levels of factory2 noise

6.3 Methodology

The research question is clearly how to quantify the noise content of the audio signal so that a suitable weight value can be determined. The method used here is to gather statistical information from the audio signal and assess whether it is characteristic of speech signals or of speech signals affected by noise, and, if the latter, the degree of noise contamination. For training purposes, white noise at various energies has been added to clean audio and the skewness and kurtosis values are obtained. By trial and error, the skewness and kurtosis values are mapped to weight values that are found to give the best AVSR performance. The architecture of the proposed systems is described in detail in the next section and the subsequent section describes the experimental procedure.

6.3.1 Architecture

The new adaptive fusion system uses a combination of feature fusion and decision fusion methods, as shown in Figure 6.10. Here the output from feature fusion (O_{av}) and partial outputs from decision fusion (O_a) and (O_v) are combined at the final stage with an appropriate weight calculated from the statistical properties of the mixture of speech signal and environmental noise as described. The audio and video inputs are processed independently to extract relevant features for classification. Since feature fusion is included in the implementation, the audio and visual feature rates are equalized through an interpolation process.

The system can be divided into two phases, namely training and testing. The method adopted for audio-visual feature extraction and classification have been explained in details in sections 3.3 to 3.5 and section 5.2.1 to section 5.2.2. The method for adaptive fusion based on the statistical analysis of the audio signal will now be described in greater detail in following section.

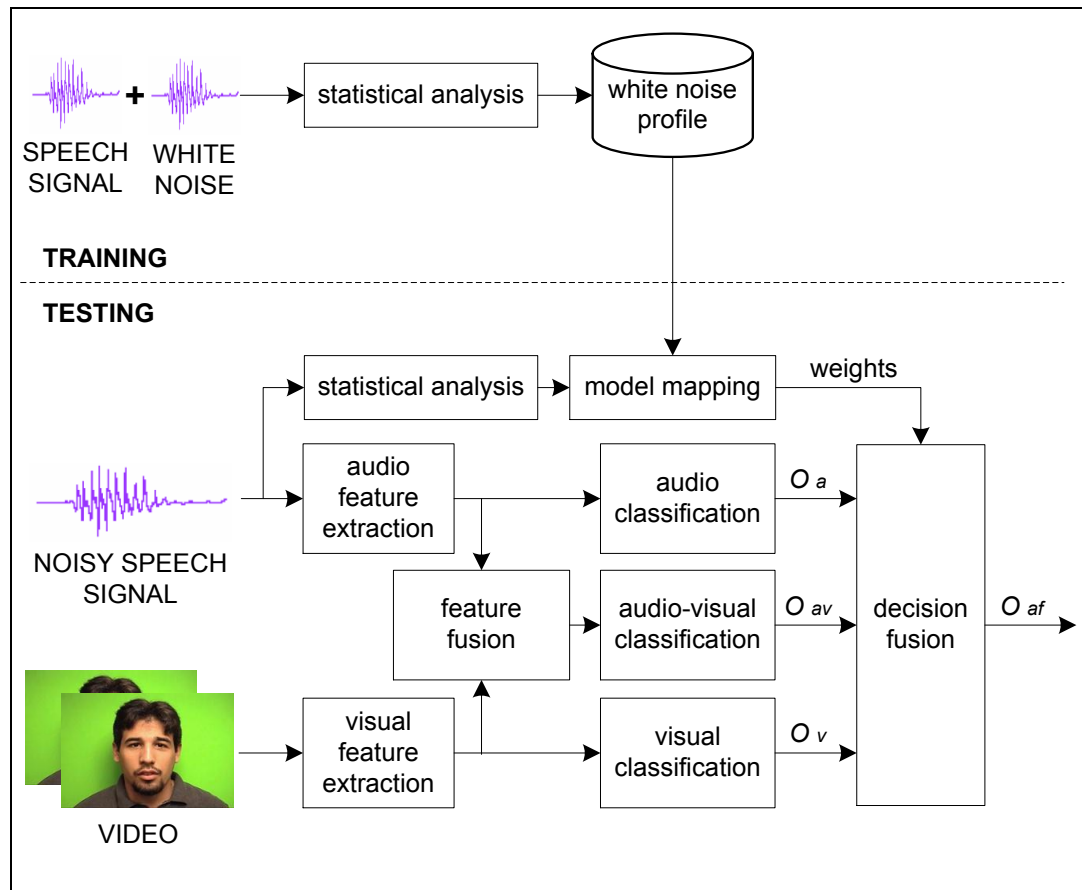


Figure 6.10 Architecture of the adaptive fusion system

6.3.2 Training phase

The CUAVE database consists of five sessions, where, in each session, the subjects speak the words ‘zero’ to ‘nine’. In this work, data from sessions 1, 2 and 3 (30 samples) were employed for training and data from sessions 4 and 5 (20 samples) were used for testing. All 36 speakers from the database were used, yielding a total of 1800 samples for demonstrating the utility of the proposed approach.

Training was carried out using the speech data from session 1, 2 and 3 that had white noise [106] introduced. Statistical analysis has been applied to all the training speech data across all SNRs and the statistical parameters (skewness and kurtosis) were stored in the database as ‘white noise profile’. During testing phase, ‘model selection’ will request these parameters to determine appropriate weight for adaptive fusion system.

6.3.3 Testing phase

Two selection models were considered, a ‘two region’ (2R) model and a ‘three region’ (3R) model as shown in Figures 6.11 and 6.12 respectively. These selection models were devised based on empirical understanding of how they work effectively based on the results in section 5.3.1.

When the quality of the audio is equivalent or better than that obtained for SNRs greater than 0dB under white noise conditions, the adaptive fusion is weighted to receive only the audio modality features, otherwise only the visual modality features are used. The threshold value (0dB) chosen here was determined empirically to optimize performance of the system across all the audio noises investigated in this work. Model 2R implemented using either skewness or kurtosis values and the operation can be represented as follows.

$$O_{af} = \begin{cases} O_a & \text{for } s \geq \lambda_s \\ O_v & \text{for } s < \lambda_s \end{cases} \quad (6.3)$$

$$O_{af} = \begin{cases} O_a & \text{for } k \geq \lambda_k \\ O_v & \text{for } k < \lambda_k \end{cases} \quad (6.4)$$

where s is the measured value of skewness and k is the measured value of kurtosis. In this application, the parameters λ_s and λ_k are the values of the skewness and kurtosis at 0dB respectively when the data was simulated under white noise conditions during training.

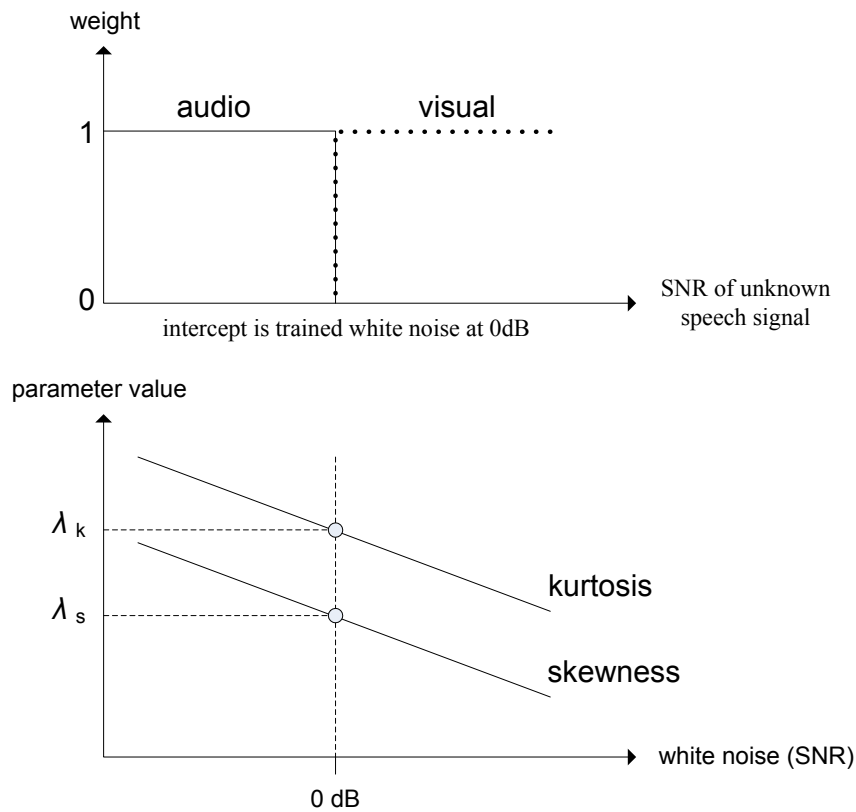


Figure 6.11 The relationship between audio statistical parameters and modality chosen for model 2R

Model 3R as shown in Figure 6.12 is an extension of model 2R but with an additional modality option defined during training when the SNR is in the range 3dB to -3dB by using the output from feature fusion output (O_{av}). The threshold value (± 3 dB) chosen here was also determined empirically to optimize performance of the system across all the audio noises investigated in this work. The operation of model 3R can be represented as follows.

$$O_{af} = \begin{cases} O_a & \text{for } s \geq \lambda_{su} \\ O_{av} & \text{for } \lambda_{sl} \leq s < \lambda_{su} \\ O_v & \text{for } s < \lambda_{sl} \end{cases} \quad (6.5)$$

$$O_{af} = \begin{cases} O_a & \text{for } k \geq \lambda_{ku} \\ O_{av} & \text{for } \lambda_{kl} \leq k < \lambda_{ku} \\ O_v & \text{for } k < \lambda_{kl} \end{cases} \quad (6.6)$$

where s is the measured value of skewness and k is the measured value of kurtosis. Parameters λ_{su} and λ_{ku} are respectively the upper limit values of skewness and kurtosis at 3dB while λ_{sl} and λ_{kl} are respectively the lower limit values of skewness and kurtosis at -3dB. All these parameters are obtained when data is simulated under white noise condition during training session.

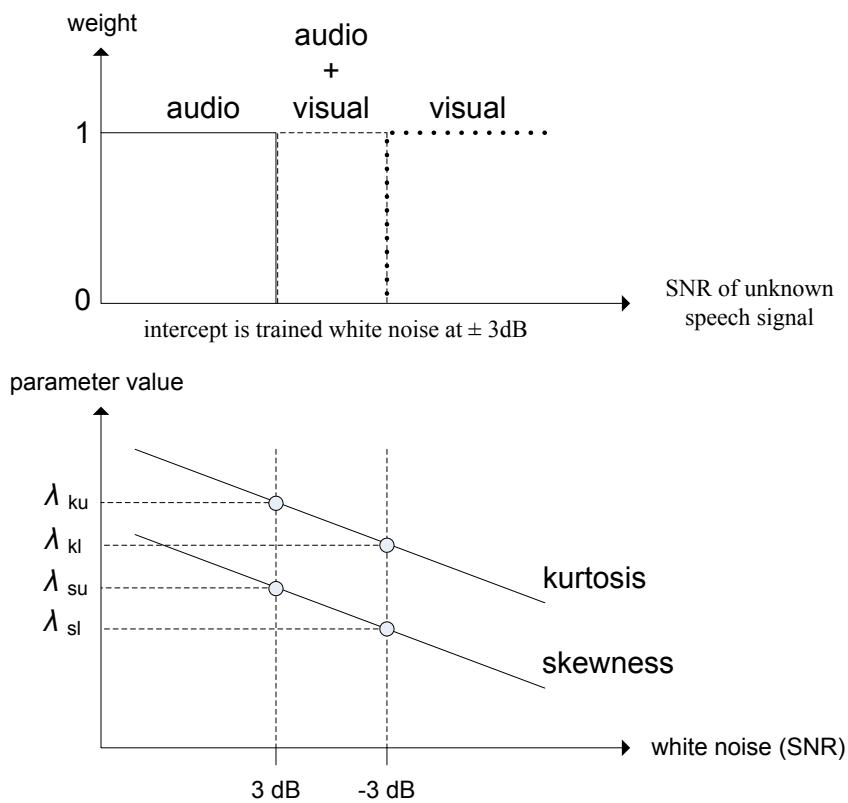


Figure 6.12 The relationship between audio statistical parameters and modality chosen for model 3R

6.4 Results and discussion

This section presents the results obtained for the adaptive fusion AVSR system and their comparison with conventional feature-based and decision-based fusion approaches under the influence of ‘babble’, ‘factory1’ and ‘factory2’ noise that are part of NOISEX-92 dataset [106]. Three experiments were conducted to measure the performance of the proposed methods. Firstly, the performance of the skewness and kurtosis analysis will be analysed across all SNRs using both model selection, model 2R and model 3R. Then, details about the performances trend and modality chosen using model 2R and model 3R in comparison with audio-only and visual-only recognition are investigated. Finally, comparison with conventional fusion strategy was made with the best selection model.

6.4.1 Performance of skewness and kurtosis in adaptive fusion AVSR

The performance of the parameters skewness and kurtosis are considered in terms of facilitating an adaptive fusion AVSR system. The purpose is to compare skewness and kurtosis to determine which is the better for the purpose of adaptive fusion.

Tables 6.1 and 6.2 show the performances of models 2R and 3R respectively when using both skewness and kurtosis. It can be observed for both models that kurtosis performs better for SNRs above 15dB, but below 10dB both the parameters perform equally well. The superiority of kurtosis at higher SNRs is probably due to it being a more sensitive indicator as the obtained range of values is greater for the CUAVE investigation, as can be seen in Figures 6.7 to 6.9.

Table 6.1 Performance of adaptive fusion AVSR in word recognition (%) using skewness and kurtosis analysis using Model 2R

Noise	Method	Clean	25dB	20dB	15dB	10dB	5dB	0dB	-5dB	-10dB
Babble	Skewness	98.6	98.6	98.5	97.5	91.9	74.6	56.1	61.1	64.3
	Kurtosis	98.9	98.9	98.8	97.6	91.4	72.2	55.7	61.4	64.3
Factory 1	Skewness	98.6	98.8	98.1	95.4	85.7	66.5	55.8	61.7	65.1
	Kurtosis	98.9	98.9	98.2	95.6	85.1	65.3	56.4	62.8	65.3
Factory 2	Skewness	98.6	98.8	98.6	98.3	95.4	81.7	64.6	65.8	67.5
	Kurtosis	98.9	99.2	98.8	98.6	95.1	82.1	64.3	65.4	67.4

Note. The highlight value represent highest performance in specific SNR value

Table 6.2 Performance of adaptive fusion AVSR in word recognition (%) using skewness and kurtosis analysis using Model 3R

Noise	Method	Clean	25dB	20dB	15dB	10dB	5dB	0dB	-5dB	-10dB
Babble	Skewness	98.2	98.5	98.2	96.9	91.5	78.3	63.6	65.6	65.7
	Kurtosis	98.6	98.6	98.2	97.1	91.4	78.3	65.1	66.4	66.0
Factory 1	Skewness	98.2	98.5	97.5	95.4	85.8	71.4	65.6	65.3	64.6
	Kurtosis	98.6	98.5	97.8	95.8	85.3	72.6	64.9	64.7	65.1
Factory 2	Skewness	98.2	98.3	98.2	97.6	93.8	83.2	70.4	67.2	67.6
	Kurtosis	98.6	98.5	98.3	98.3	93.8	83.6	71.3	67.6	67.5

Note. The highlight value represent highest performance in specific SNR value

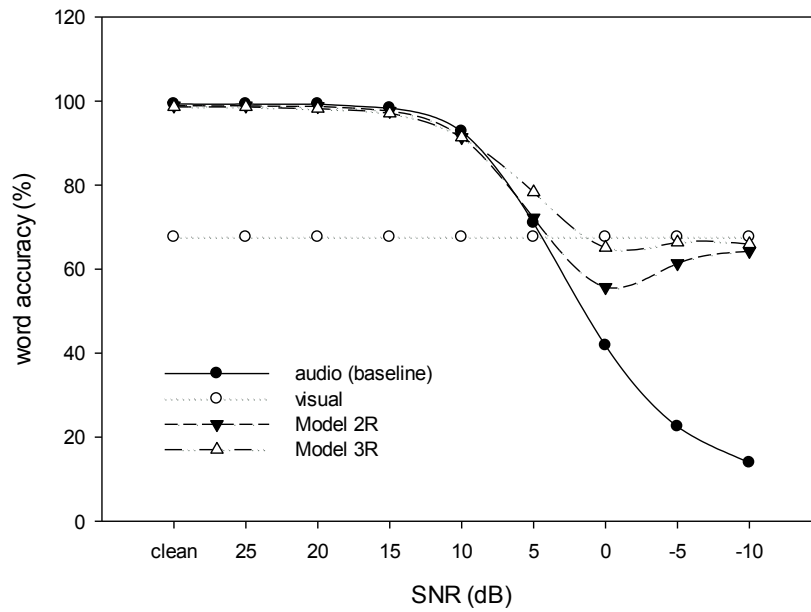
In the next section, details of the proposed system performance in comparison to audio-only and visual-only recognition will be described. As there is little difference in performance between skewness and kurtosis, the arbitrary choice was made between them for further investigation. The work concentrates on kurtosis, but not exclusively.

6.4.2 Results of performances trend and modality chosen in adaptive fusion AVSR

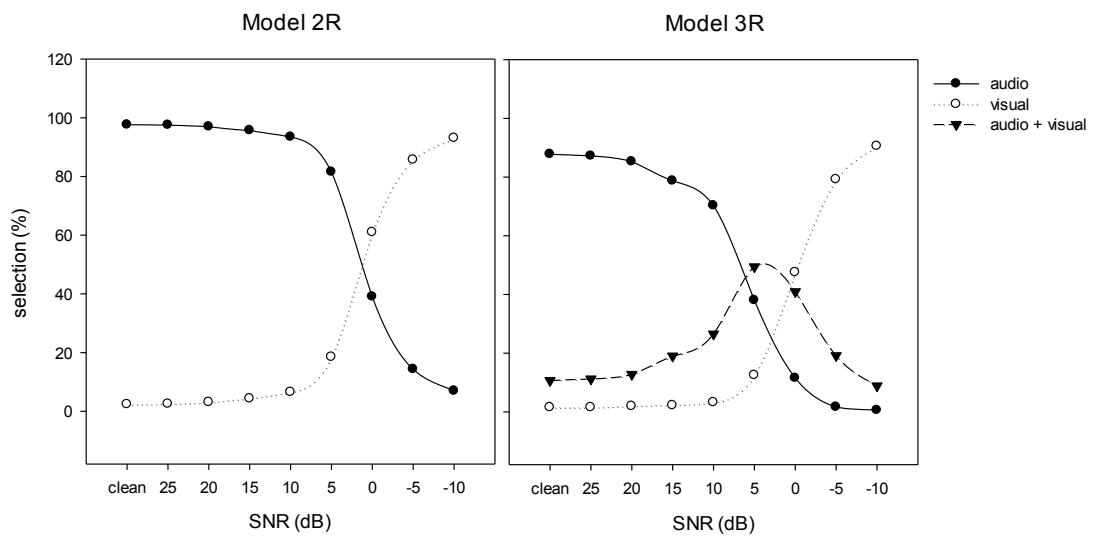
Figure 6.13(a), 6.14(a) and 6.15(a) show the performances trend of model 2R and model 3R in comparison with audio-only (baseline) and visual-only recognition, when simulated with babble, factory1 and factory2 noise. For a visual-only system, the accuracy is the same for all SNRs because the visual modality is not affected by audible noise. For the audio-only performance, the recognition performance of the system is severely degraded for SNR values below 10dB and falls to an accuracy of around 20% at an SNR of -10dB.

The trend for audio-only performance is similar for all noise sources, the main difference being how sharply the word accuracy falls when noise is introduced. There is also a small difference in the value at which the audio or visual curves cross; for babble noise, the intersection occurs at 6dB, for factory1 at 6dB and for factory2 at 2dB. It would seem that factory1 noise has a greater effect on the recognition performance and factory2 the least effect.

It can be seen that both models 2R and 3R produce significantly better results compared to the baseline audio model when the SNR is below 5dB, with the models providing a performance at all SNRs that is close to the better of that exhibited by the audio-only and visual-only modalities. Below 5dB, it can be seen that Model 3R has a better recognition performance than Model 2R, due to the additional mixed modality it provides at SNRs between -3dB and 3dB.

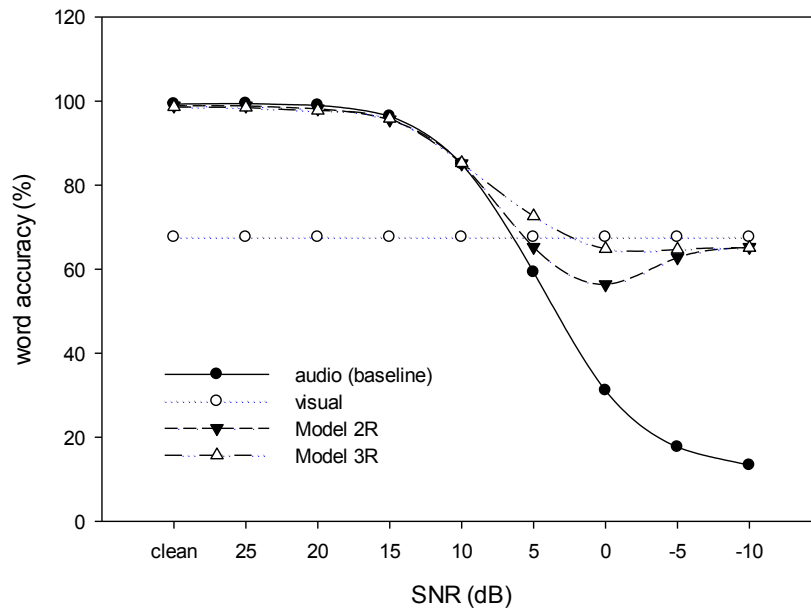


(a) performances trend

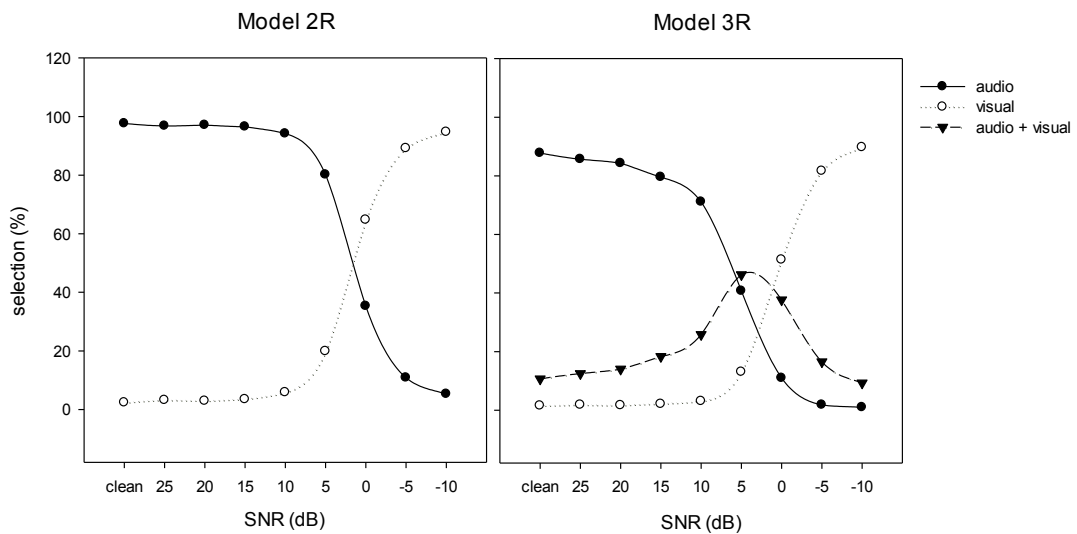


(b) modality chosen

Figure 6.13 Performance of adaptive fusion AVSR using kurtosis information when simulated under babble noise condition

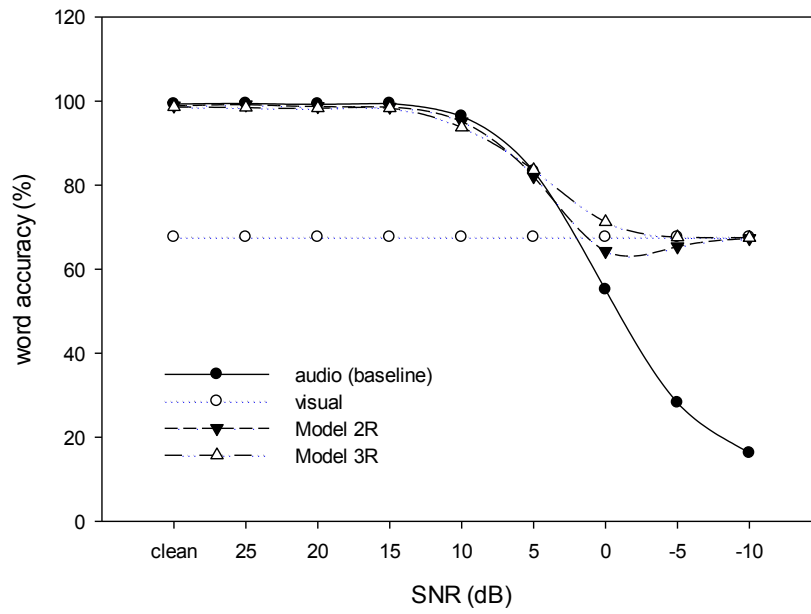


(a) performances trend

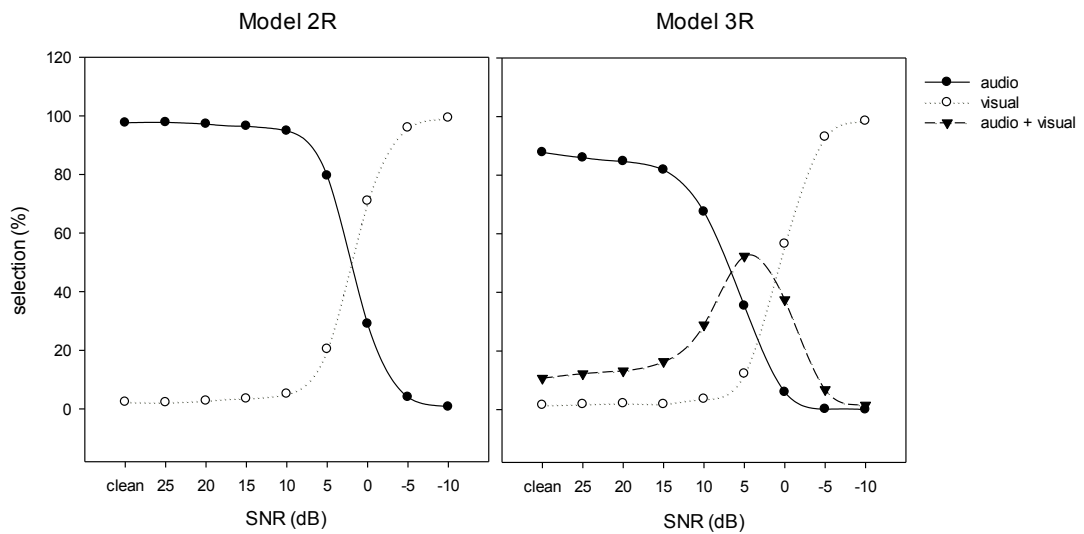


(b) modality chosen

Figure 6.14 Performance of adaptive fusion AVSR using kurtosis information when simulated under factory1 noise condition



(a) performances trend



(b) modality chosen

Figure 6.15 Performance of adaptive fusion AVSR using kurtosis information when simulated under factory2 noise condition

Figure 6.13(b), 6.14(b) and 6.15(b) show the selection of modality during the recognition when simulated with babble, factory1 and factory2 noise for the proposed Model 2R and Model 3R.

In Model 2R only two modalities are involved in recognition. At higher SNRs, the adaptive fusion will normally choose the audio modality as a final output, but as more noise is induced in the audio signal, the visual modality will be more often selected. Model 3R has an additional modality with the aim of improving performance at mid-range values of SNR at which point the integrated audio-visual modality generally performs better than either the audio or visual modalities acting alone.

It was observed that for Model 3R, when both audio and visual modalities intersect together, audio-visual modality achieved its highest selection based on interpolation studies. By combining three different type of modalities the overall system improved significantly especially for SNR between -3dB to -3dB as what demonstrated by Model 3R.

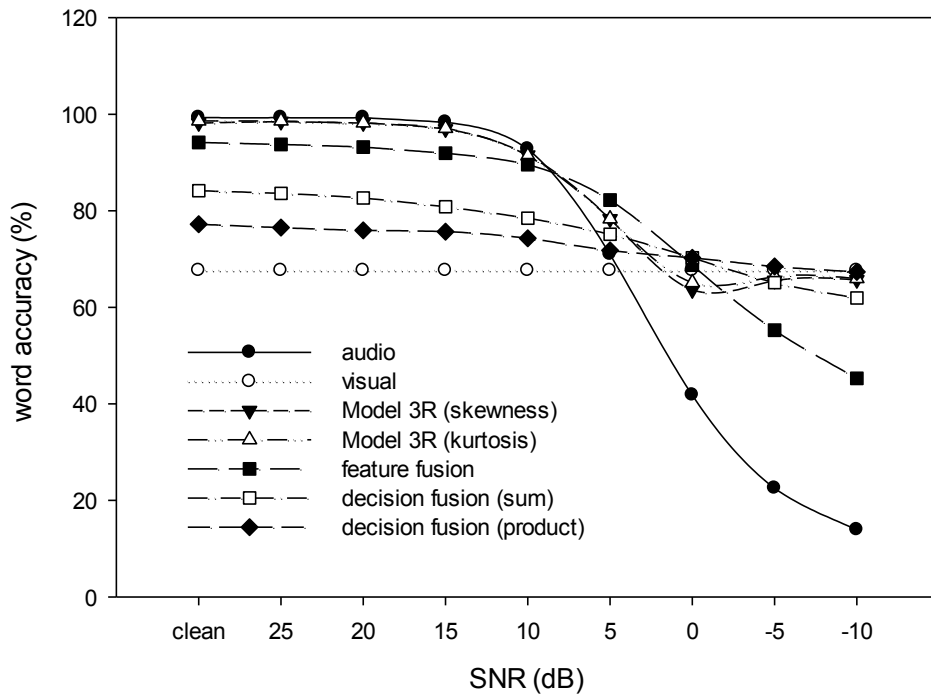
It has also been demonstrated the use of white noise in calculating suitable thresholds for skewness and kurtosis and applied to unknown noise (babble, factory1 and factory2) is suitable for the CUAVE database. In a specific application, the exact nature of the noise will be unknown, white noise proved sufficient in obtaining a good estimate in the experiments conducted. Overall, the improvement found when using the proposed method compared to the baseline system is very promising; indeed simply by being able to select the visual modality automatically brings an improvement in word recognition performance from around 15% to over 65% at an SNR of -10dB.

6.4.3 Comparison with conventional method

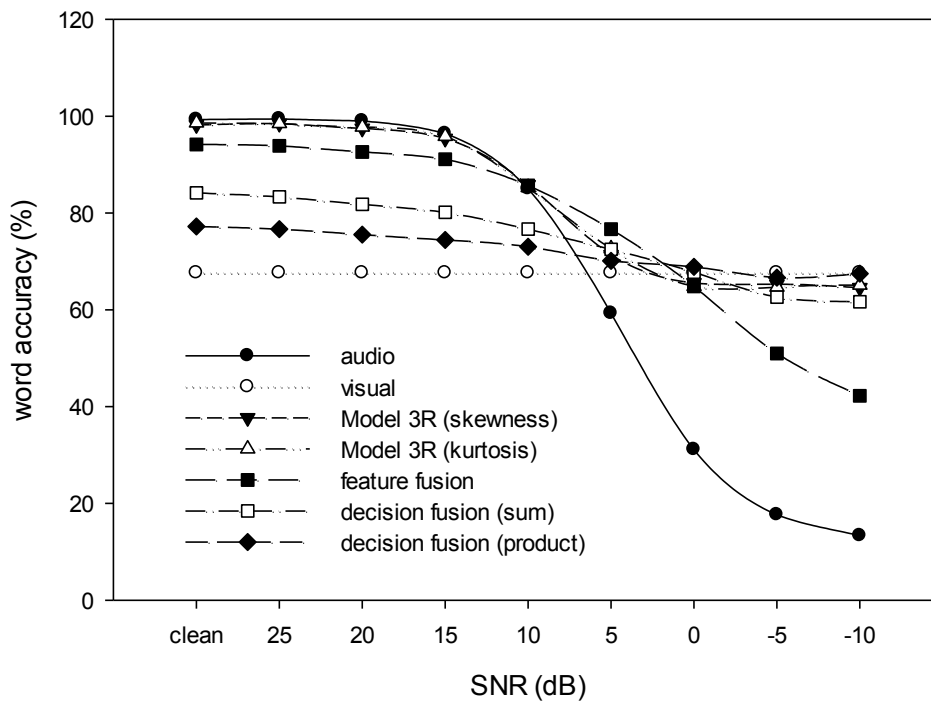
To access the performance of the new adaptive fusion AVSR system with respect to conventional fusion strategies, feature-fusion and decision fusion approaches that are based on sum and product operations were implemented in the manner described in [26] and [27]. The same audio and visual information were supplied to the feature-fusion and decision-fusion implementations and its comparison between the best model selection (model 3R) are shown in Figure 6.16.

At higher SNRs, model 3R achieved a significant performance improvement compared to the other approaches (apart from the audio-only results). This is due to model 3R directly choosing the audio modality, whereas the other solutions include the visual modality in their calculations which has the effect of degrading performance due to its relatively poor performance at high SNRs. At lower SNRs the performance of Model 3R is similar to the results produced by decision fusion using product operations and those from the visual modality.

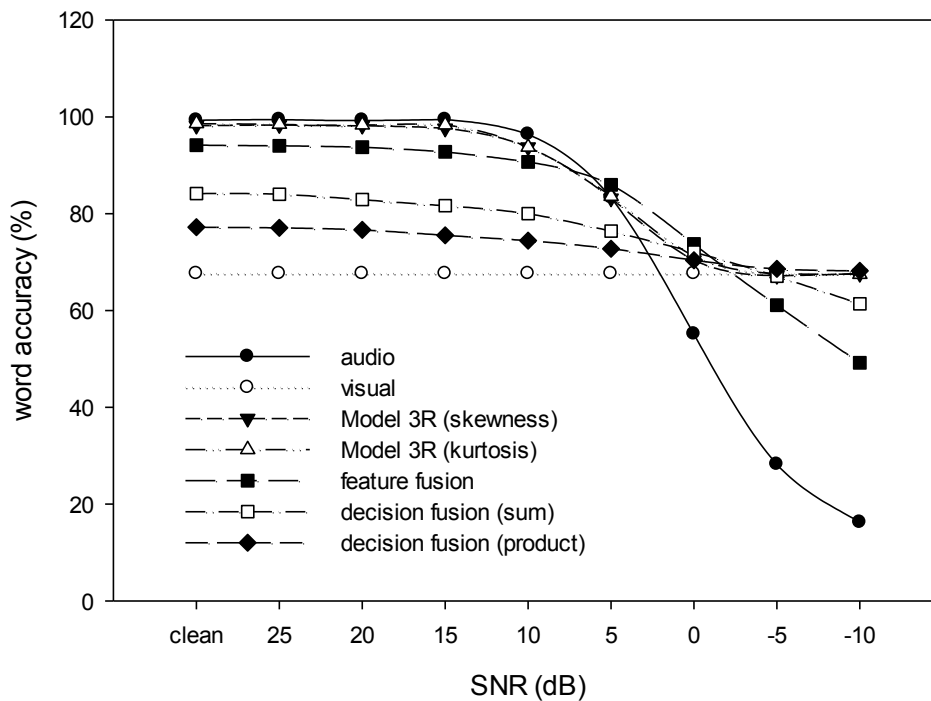
Overall, it can be concluded that the proposed adaptive fusion method enhances the fusion recognition performance compared to the conventional methods. The new adaptive fusion method is able to follow closely the better performer from the audio-only or video-only modalities for any given SNRs. Also the generalization of the selection models introduced for various levels and types noises is verified. Although the new adaptive fusion are trained only with -3dB, 0dB and -3dB data containing statistical parameter from the white noise, it successfully works under a variety of different untrained noise conditions.



(a) babble



(b) factory1



(c) factory2

Figure 6.16 Comparison of the AVSR performance the adaptive fusion approach and existing fusion methods.

6.5 Chapter summary

In AVSR, each modality has its best performance in specific SNR regions: audio modality at higher SNRs, visual modality at lower SNRs and an integration of audio and visual modalities at intermediate SNRs. In this work, the aim has been to enhance the robustness of an AVSR system in noisy environments by altering the selection of the modalities in a single system. A novel adaptive fusion AVSR system has been proposed and constructed whose modality is dependent on the statistical distribution of the audio input. The statistical parameter ranges that select the modality have been obtained by training under white noise conditions and verified using various noise sources.

Although the effectiveness of the proposed adaptive fusion with the conventional fusion method across a broad range of SNR values have been shown, the potential exists to further enhance the current system by including statistical or reliability information from visual modality to augment those already present in the proposed method with the aim of improving the performance and robustness of the AVSR system in real practical application. For example, when the lighting (illumination) condition during recognition is not as good as in testing data, or the application is unable to capture the fully lip region, the performance of the visual speech recognition will decrease, thus the overall AVSR system also degrades. Therefore the problem of measuring the confidence of visual information needs to be addressed so that a robust AVSR can be realized in real-world application.

CHAPTER 7

LOUGHBOROUGH UNIVERSITY AUDIO-VISUAL SPEECH DATA CORPUS

This chapter introduces a new data corpus that has been specifically acquired to provide additional data not only to validate the AVSR approach proposed in the thesis, but to investigate whether recognition performance can be improved by using high-definition visual data. As far as the author is aware, this is the first high-definition AVSR corpus that has been made freely available to the research community, giving the potential to generate more accurate identification and measurements of visual features. A brief introduction to the new data corpus is given in Section 7.1, the contents of the data corpus are detailed in Section 7.2 and the performance of the AVSR approaches developed in this thesis when applied to the new corpus are described in Section 7.3.

7.1 Introduction

The choice of data corpus can greatly influence the research results obtained. Although there is a number of existing AVSR data corpora available, many have limited features in terms of recording quality, number of participants and word coverage. In addition, some corpora referenced in the literature are not made available to researchers, making the performance results quoted difficult to verify. The CUAVE database has been used in this thesis as it has good audio and visual quality, a reasonable number of participants and it is freely available. Many researchers have used this database, allowing performance comparisons of a range of AVSR approaches to be made and independently verified. Generally, the results published in the literature [32], [101] demonstrate that that the better the visual accuracy, the better the overall AVSR system performance.

Recent advances in video recording and compression technologies has opened new opportunities for AVSR. By using high definition (HD) video recordings, the quality of the images captured can be substantially improved and more detailed information can be acquired of facial features, including the mouth. The motivation in developing the new corpora is that no HD databases are currently available and their potential to improve AVSR recognition is an important area to explore.

The new data corpus that has been developed is called the Loughborough University Audio-Visual data corpus or LUNA-V. It contains sentences that include the English digits from ‘zero’ to ‘nine’ and recorded in five sessions. Additional sentences were also spoken by the participants to make the database useful to researchers investigating phone recognition.

7.2 The design of the LUNA-V data corpus

This section describes the collection process and the content of the LUNA-V corpus. The audio and video post-processing that were carried out on the collected data are also introduced.

7.2.1 Ethical clearance

Data collection from human participants requires prior ethical approval to ensure the safety, dignity and well-being of both the subjects and the researcher. For this data corpus, ethical approval has been obtained from the Loughborough University Ethics Approvals (Human Participants) Sub-Committee.

To demonstrate that research is being conducted openly and without deception, information about the nature of the research and the process of data gathering was supplied in written form by means of an information sheet and a consent and release form signed by the participants. Copies of these forms can be found in Appendix D.

7.2.2 Subject population

The data corpus contains the utterances of 10 speakers, one female and nine male. The participants were undergraduate students at Loughborough University, all have been brought up in England and spoke English as their first language, albeit with a variety of regional accents. Participants were asked to read the sentences in their own natural style and with no instruction regarding pronunciation. All were asked to read the sentences carefully with a short pause between words to help facilitate the audio and video labelling procedure during post-processing. To protect the identity of the participants, each of them has been given a special code to represent them in the data corpus, for example first participant identity is ‘v01m’, second is ‘v02m’ and so on. The word ‘v’ represents LUNA-V participant while the word ‘m’ represents gender, ‘m’ for male or ‘f’ for female.

7.2.3 Sentence selection

The database consists of two separate parts. The first part contains the English digits ‘zero’ to ‘nine’, each spoken five times by each speaker. In the second part, several sentences from the well-known TIMIT data corpus have been adopted, extending the use of LUNA-V to the investigation of speech recognition. A list of the sentences recorded in the LUNA-V data corpus is given in Table 7.1.

Table 7.1 The sentences collected for the LUNA-V corpus

sentence	content
digits	zero, one, two, three, four, five, six, seven, eight, nine
1	She had your dark suit in greasy wash water all year
2	Each untimely income loss coincided with the breakdown of a heating system part
3	The easygoing zoologist relaxed throughout the voyage
4	The same shelter could be built into an embankment or below ground level

The sentences in the TIMIT data corpus contain 61 phones and it is a common practice to convert these to the standard 39 phone set proposed by Lee and Hon [114]. Sentences 1 to 4 in the LUNA-V corpus were selected from those available in TIMIT because of their coverage of these 39 phones. Full details of the phone and viseme coverage of the LUNA-V corpus can be found in Appendix E.

7.2.4 Recording studio and hardware

For capturing the audio and visual information, a Sony HXR-MC2000E HD semi-professional video camera as shown in Figure 7.1 was used. The camera has a 1/4" Exmor R CMOS sensor, a maximum resolution of 1920x1080 pixels and records in AVCHD format [115]. It has 64GB of internal storage and an external memory card slot should additional storage be required. The device is also equipped with a Sony ECM-PS1 stereo microphone as shown in Figure 7.2.



Figure 7.1 Sony HXR-MC2000E video camera



Figure 7.2 Sony ECM-PS1 stereo microphone

A plan view of the studio can be seen in Figure 7.3. The video camera and microphone were arranged near the speaker in order to obtain a close-up high resolution view of the speaker's face and a suitably high SNR for the audio signal, respectively. The distance between the microphone and the speaker was 30cm, while the distance between video camera and the speaker was 100cm. Automatic image

focus and the white balance of the camera were selected as these were found to produce a good colour balance under the lighting conditions used.

The text to be read by the speakers was mounted immediately below the camera so that the eyes of speaker are directed close the camera's line of view (see Figure 7.4). The text was displayed in a cue card rather than on a monitor screen, which would have the potential to generate electrical noise that could be picked up by the recording equipment. A green screen was used as the background behind the speaker to offer the potential of chroma keying to corpus users. In order to minimize the studio changes needed between sessions, the recording equipment remained in a fixed position and compensation for the physical differences of the speakers was made simply by adjusting the height of the subject's chair.

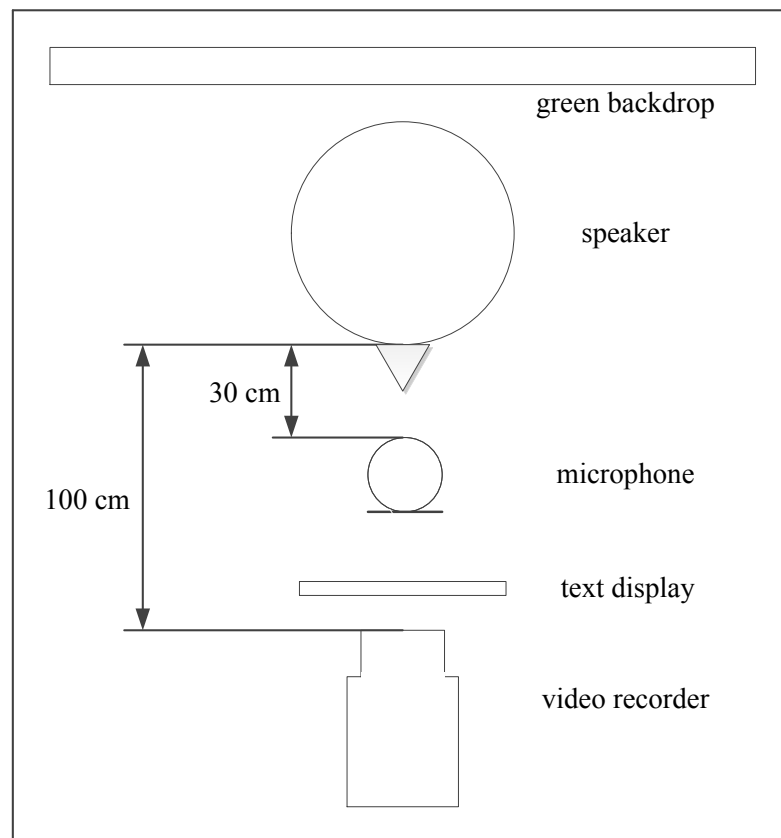


Figure 7.3 Plan view of the recording studio setup

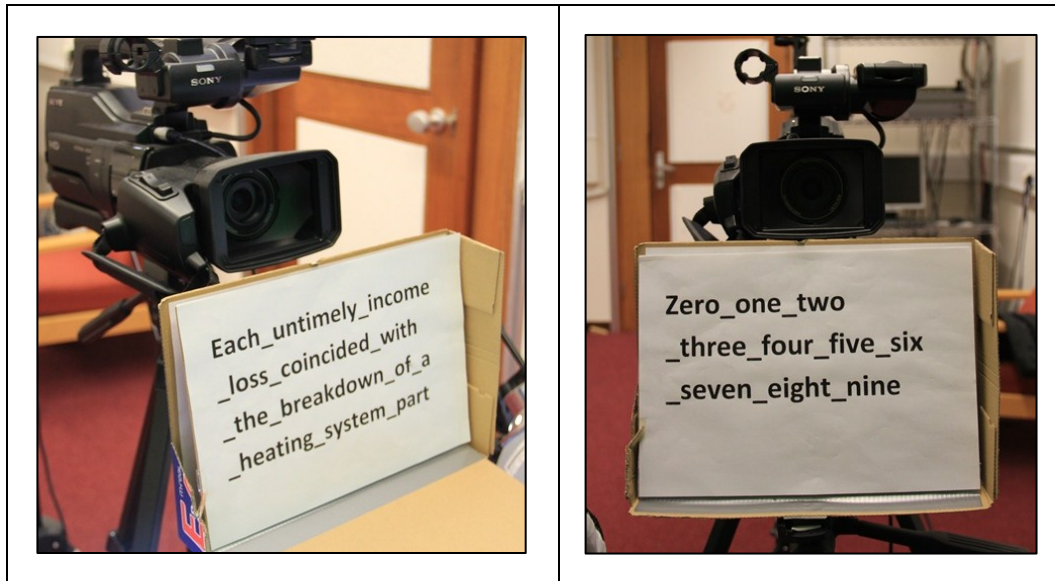


Figure 7.4 Example presentation of the text to be read by the participants

The recordings were conducted in a research laboratory not fitted with sound insulation material, but its remote location meant extraneous noise was rarely encountered. The room had no windows, allowing lighting conditions to be controlled. General room lighting was supplied by fluorescent lamps fitting in the ceiling and direct lighting was provided by three sets of stand-mounted daylight spiral-fluorescent 24W studio lamps each filtered by a diffusion canvas to provide a uniform output, as can be seen in Figure 7.5.

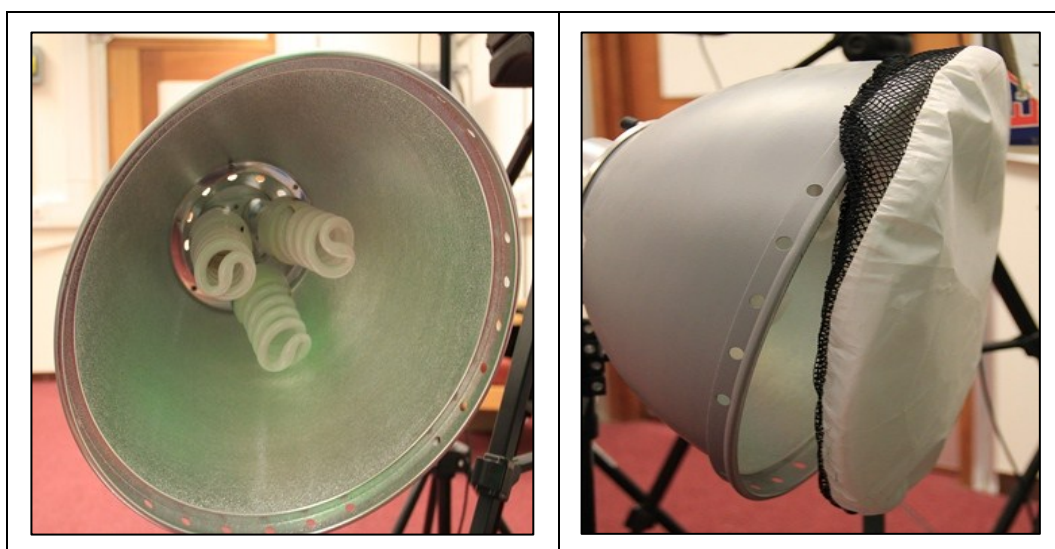


Figure 7.5 The principal direct lighting was supplied by daylight fluorescent lamps fitted with a diffusion canvas

7.2.5 Recording process

Following the completion of the ethical clearance documentation, the participant's chair was adjusted so that the speaker's face was directly in front of the camera, as can be seen in Figure 7.6. To assist the speakers in keeping a consistent position within the frame of view, a copy of the video being captured was displayed to the participants.



Figure 7.6 Adjustable chair position

Speakers were asked to utter sentences in their natural style, but with a short pause between words. No time limit was specified for the speaking of sentences or to complete the task as a whole. Speakers were allowed to repeat sentences if they felt they had made a mistake during recording. In addition to the recording technician, an assistant was always present to judge whether participants had made a mistake in reading the words, but had not realized such themselves. In such cases,

the mistake was highlighted to the participant and they were asked to re-read the sentence.

To generate sufficient data for both training and testing purposes, each speaker was asked to speak the English digits a total of five times and each of the sentences a total of three times. In each speaker session, a total of 182 words were recorded and the whole process typically took 20 to 30 minutes to complete.

7.2.6 Audio post-processing

Power Director version 12.0 [116] was used to extract audio information from the AVCHD video file, generating a standard quality stereo audio file at a sampling rate of 16kHz and at 16-bit resolution. This quality is sufficient for speech recognition as human speech production bandwidth is generally between 100Hz and 8kHz.

Audio noise sources identified during the recording process included ventilation systems, computing equipment and fluorescent lighting. To reduce their presence in the corpus audio file, a specific frequency filter was applied. As no in-built noise cancelling hardware was available in the video camera, this was performed in software during post-processing, in this case using the open source audio editor Audacity version 2.0.5 [117]. The noise cancelling process involves making a recording when only the ambient room noise is apparent, as shown in Figure 7.7, and the identified frequency bands have their strengths reduced in the filtered audio signal, as can be seen in Figure 7.8. Although the ambient noise is not significant (range from -40dB to -70dB across all the frequency), it is essential to remove the noise as it would help the labelling process. However this process might reduce the content of the audio signal with less significant effect.

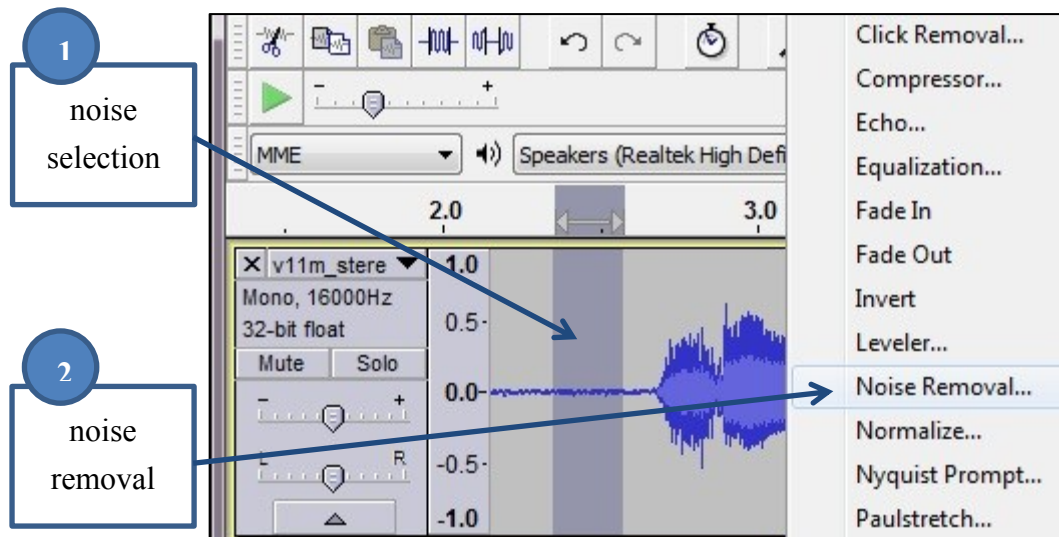


Figure 7.7 A highlighted section of ambient noise used for background noise removal

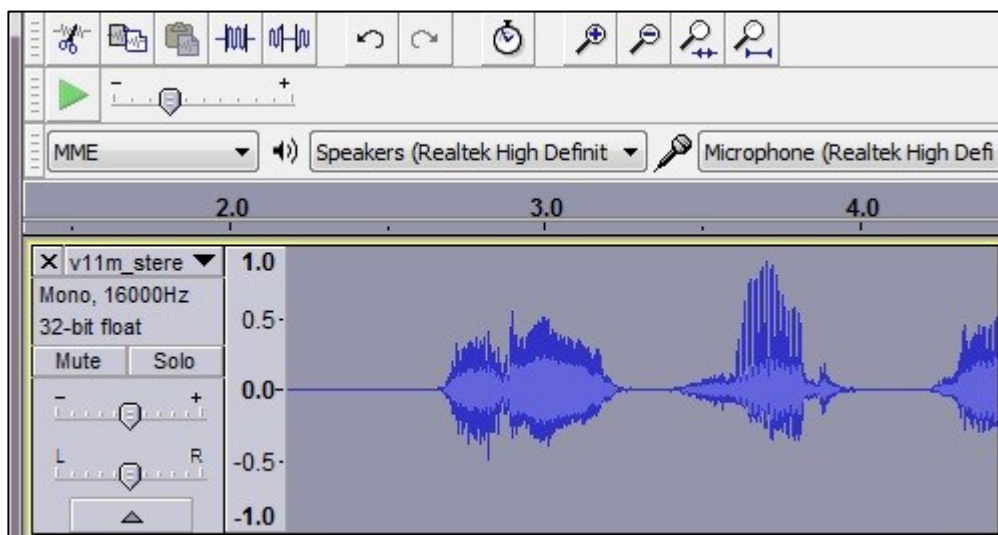


Figure 7.8 Audio signal produced following noise cancellation

Before the audio can be used for recognition purposes, it must be correctly labelled according to the utterances made. In combination with the audio output itself, Audacity's display of the amplitude of the audio signals was used as a visual aid to estimate the temporal extent of the spoken words. Each word was manually labelled as shown in Figure 7.9 and the information was stored in a separate text file that can be used as a word reference by researchers.

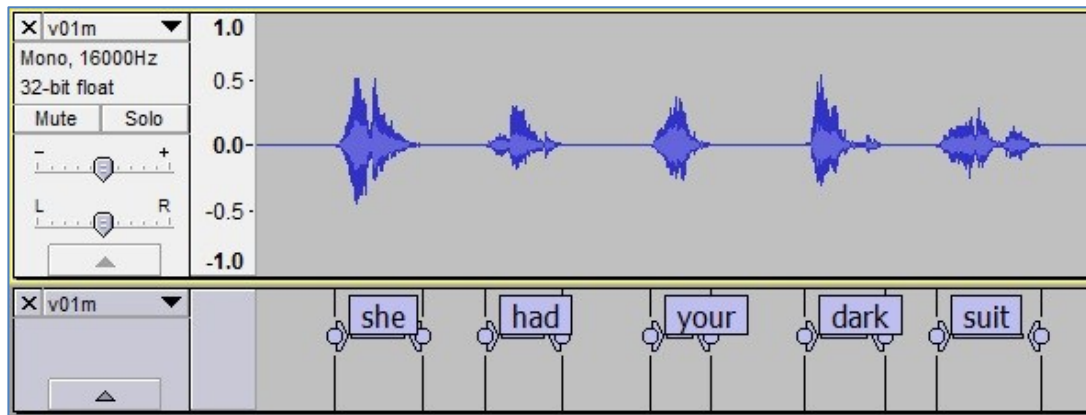


Figure 7.9 Example of the alignment for word-level transcription

7.2.7 Video post-processing

The AVCHD format video provided by the camera was converted to MPEG-2 format using Power Director; an example is shown in Figure 7.10. This format was chosen as it provides very good video quality at a reasonable file size and is widely decoded by media players. The HD video was stored at 25 frames/s, 20 Mbits/s at a resolution of 1280x720. Figure 7.11 shows sample frames taken from the converted MPEG-2 video file for each of the 10 speakers in the LUNA-V data corpus. A video file was created for each individual speaker, containing all the words spoken by that speaker (182 words). The file sizes for the 10 speakers range from 400 to 700 MBytes and the duration from 3 to 5 minutes. Note that the video file preserves the original unmodified audio signal sampled at 48kHz and without noise cancelling applied, allowing researchers access to this data if so desired.

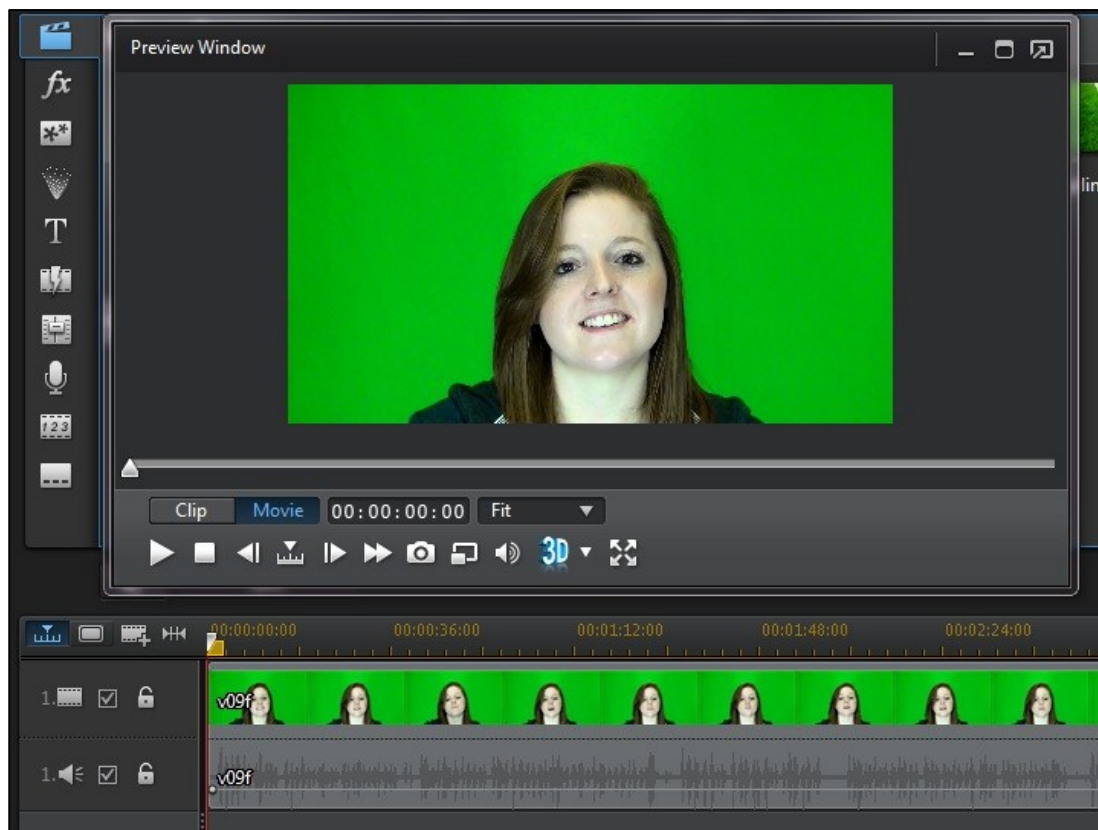


Figure 7.10 Video editing using Power Director 12



Figure 7.11 Sample frames from each of the 10 LUNA-V data corpus subjects

7.3 Validation of feature extraction and AVSR

Two experiments were conducted on the LUNA-V data corpus to verify the new approaches that have been proposed in this thesis, but that have previously used the CUAVE database, namely lip geometry feature extraction and AVSR. In the lip feature extraction experiment, the same approach as that proposed in Chapter 3 was used in order to extract the five lip geometrical features, namely width, height, ratio of width to height, area and perimeter. The AVSR experiment used early integration as described in Chapter 5 to assess the speech recognition accuracy of the lip geometrical-based approach when simulated under a range of environmental conditions arising from the introduction of audible noise.

7.3.1 Lip geometry feature extraction

Speaker images acquired from the LUNA-V video files were cropped to the mouth region by applying a face detection process followed by a mouth detection process using the Viola-Jones object recognizer [46], as shown in Figure 7.12. The performance obtained using the LUNA-V data corpus was similar to that obtained for the CUAVE database.

Figure 7.13 shows the results of lip geometry extraction from images of three different speakers. It can be seen that by applying an HSV colour filter, border following and convex hull techniques, a close approximation of the actual lip shape can be generated. The qualitative results obtained here using the LUNA-V data corpus are similar to that obtained for the CUAVE data corpus that were given in Figures 3.11 and 3.12.

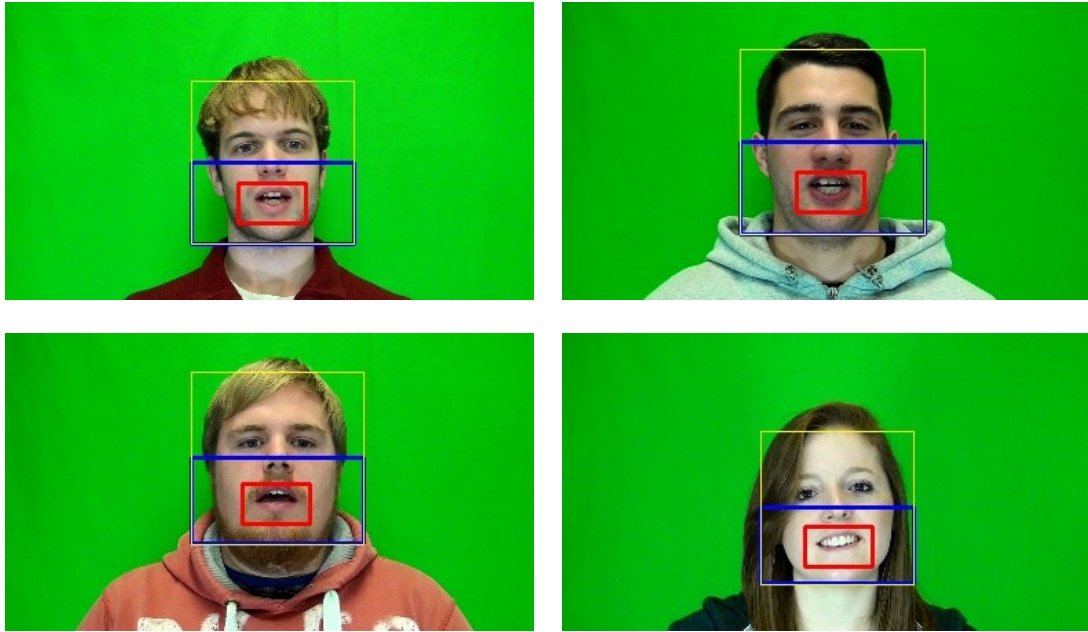


Figure 7.12 Examples of face and mouth detection using the LUNA-V corpus

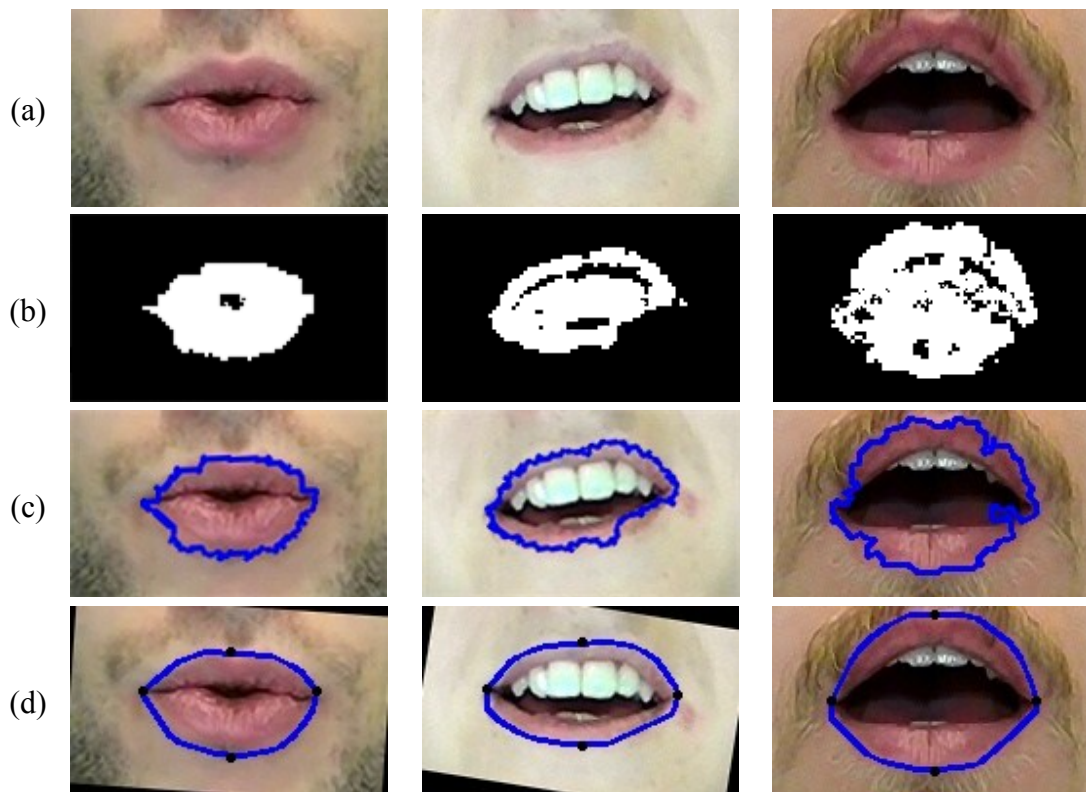


Figure 7.13 Lip geometry feature extraction for speaker 'v01m' (left column), 'v09f' (centre column) and 'v05m' (right column), (a) input colour image, (b) binary lip image, and the results of processing following (c) contour detection showing the longest contour identification and (d) application of the convex hull and automatic alignment

By finding the convex hull contour for each video frame, a time-series of feature values can be derived, as shown in Figure 7.14. Dynamic information generated from a sequence of feature values obtained during the speech utterance is used as the input to the AVSR system.

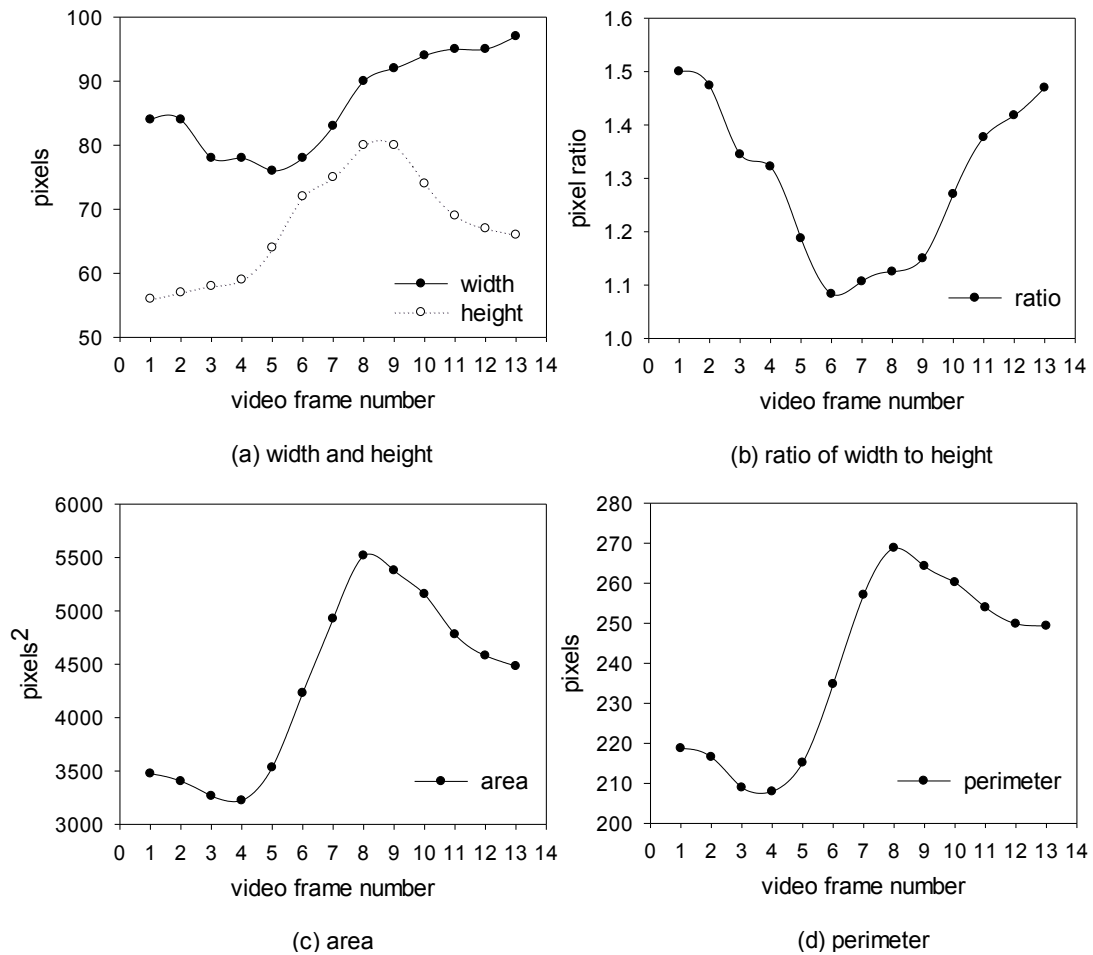


Figure 7.14 Dynamic lip information for digit ‘one’ uttered by speaker ‘v01m’ for the LUNA-V database

To measure the effectiveness of the new method presented in this work when applied to the LUNA-V data corpus, both a qualitative and a quantitative evaluation were conducted to assess the accuracy of the lip extraction used to generate geometrical features. A data set containing 50 face images with different expressions was randomly selected (five for each speaker) for this evaluation.

For the qualitative assessment of the LUNA-V data corpus, a similar visual inspection to that carried out for the investigation of CUAVE corpus was performed. Four categories of performance, namely ‘wrong’, ‘poor’, ‘satisfactory’ and ‘good’, were used as a single combined subjective assessment of how closely the lip region has been isolated and the quality of fit of four benchmark points (left, right, top and bottom of the lips). Figure 7.15 shows grading examples for the LUNA-V data corpus and a comparison of the classification performances for the CUAVE and LUNA-V data corpora is shown in Table 7.2. The number of images from the LUNA-V data corpus that resulted in good visual lip segmentation quality is more than 80%, with no image wrongly segmented (as the mouth region was found in all cases).



Figure 7.15 Example of the grading classification used in the qualitative assessment

Table 7.2 Comparison of the qualitative evaluations of lip classification for the LUNA-V and CUAVE data corpora

Database	Grade classification (%)			
	Good	Satisfactory	Poor	Wrong
LUNA-V	82.0	14.0	4.0	0
CUAVE	75.0	18.9	6.1	0

For the quantitative evaluation, the accuracy of the contour defined at specific points along its length is estimated. The method used for the quantitative evaluation was based on that described in [86], where four key lip points (top, bottom, left and right) are defined in both the original image and the convex hull contour and the sum of the distances between corresponding points (in pixel units) defines an error that is then normalized according to the distance between the mouth corners. Figure 7.16 shows an example demonstrating this method of calculating the lip outline recognition performance. The results of the comparison of the quantitative evaluation of the LUNA-V and CUAVE data corpora are shown in Table 7.3. and demonstrate that the relative errors when using the LUNA-V data corpus are reduced by more than 50% compared to the corresponding results obtained for the CUAVE corpus.

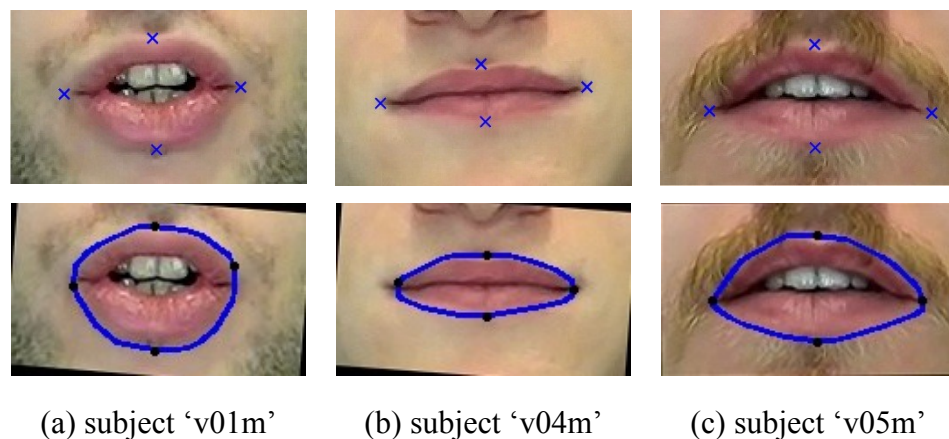


Figure 7.16 Comparison of the outcomes of manual annotation (top row) and the application of the convex hull technique (bottom row) for three subjects

Table 7.3 Comparison of quantitative evaluation of the lip classification for the LUNA-V and CUAVE data corpora

Database	Relative error (%)	
	Height	Width
LUNA-V	7.06	2.57
CUAVE	14.42	6.85

For both qualitative and quantitative evaluations, the improved segmentation performance and the difference in the error obtained when using the LUNA-V data corpus is clearly apparent and this is likely to be due to the improved accuracy of the contour generated from the higher intensity images in the LUNA-V data corpus, which is three times higher compared to CUAVE data corpus. Intensity of the image is calculated from total number of pixels available in the image. In addition, the videos in LUNA-V data corpus were recorded using latest back-illuminated CMOS image sensor technology [118] from SONY enabling to produce superb image quality. More importantly in the context of the thesis, the results also show that the proposed lip geometry feature extraction technique can be successfully applied to a second data corpus, increasing the confidence that the technique is generally applicable.

7.3.2 Speech model and evaluation set-up

Although the design of the AVSR system has been described in detail in Chapter 5, a number of changes have been made to the HMM model as, in addition to digit recognition, speech recognition has been made available by the use of the LUNA-V corpus.

Two new HMM systems have been implemented, the first has a word model and the second a phone model. The word model is a left-to-right Markov chain with 7 states as shown in Figure 7.17, while the phone model, shown in Figure 7.18, has a number of states that is dependent on the number of phones to be classified, as shown from Table 7.4 to Table 7.7 for the sentences in the LUNA-V corpus. Each Markov state is modelled using only Gaussian functions with diagonal covariance. The HTK library was used for the implementation [89].

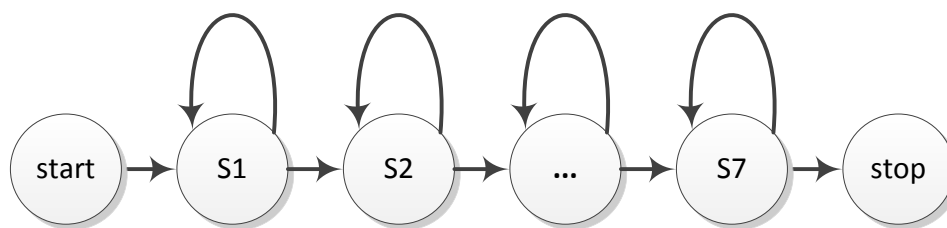


Figure 7.17 7-state HMM word recognition model

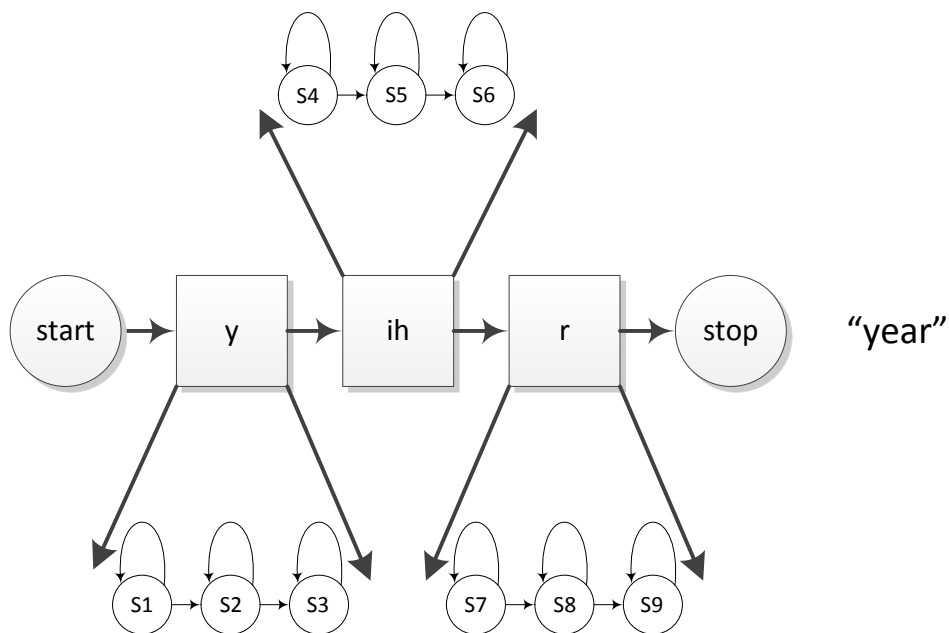


Figure 7.18 Phone recognition models

Table 7.4 List of phones present in sentence 1

Word	Phones
she	sh - iy
had	hh - ae - d
your	y - uh - r
dark	d - aa - r - k
suit	s - uw - t
in	ih - n
greasy	g - r - iy - s - iy
wash	w - aa - sh
water	w - aa - t - er
all	aa - l
year	y - ih - r

Table 7.5 List of phones present in sentence 2

Word	Phones
each	iy - ch
untimely	ah - n - t - ay - m - l - iy
income	ih - n - k - ah - m
loss	l - aa - s
coincided	k - ow - ih - n - s - ay - d - ih - d
with	w - ih - dh
the	dh - ah
breakdown	b - r - ey - k - d - aw - n
of	ah - v
a	ah
heating	hh - iy - t - ih - ng
system	s - ih - s - t - ah - m
part	p - aa - r - t

Table 7.6 List of phones present in sentence 3

Word	Phones
the	dh - ah
easygoing	iy - z - iy - g - ow - ih - ng
zoologist	z - ow - aa - l - ah - jh - ih - s - t
relaxed	r - ih - l - ae - k - s - t
throughout	th - r - uw - aw - t
the	dh - ah
voyage	v - oy - ih - jh

Table 7.7 List of phones present in sentence 4

Word	Phones
the	dh - ah
same	s - ey - m
shelter	sh - eh - l - t - er
could	k - uh - d
be	b - iy
built	b - ih - l - t
into	ih - n - t - uw
an	ae - n
embankment	ih - m - b - ae - ng - k - m - ah - n - t
or	aa - r
below	b - ah - l - ow
ground	g - r - aw - n - d
level	l - eh - v - l

For digit recognition, a separate word model was trained for each speech unit. Such an approach is feasible for the ten digits in the LUNA-V data corpus [119], [120], but a larger vocabulary application would require sharing of (phone) states between models to keep the calculation time and memory usage manageable. The LUNA-V data corpus contains five digit recognition sessions for each speaker. Data from sessions 1, 2 and 3 (30 samples) were employed for training and data from sessions 4 and 5 (20 samples) were used for testing, making a total of 500 samples for demonstrating the utility of the proposed approach.

For the speech recognition application, a phone model was used rather than a word model due to the former's ability to share parameters between models and so make more efficient use of computational resources. An advantage of phone models is that they can be trained using a reasonably small training data set when there are only a small number of alternative phones in the language, such as the 61 generally accepted as being present in English [114]. Phone models do have the drawback of individual phone recognition being strongly affected by context, that is, which phones are adjacent to the one currently under consideration. For example, phones may be pronounced clearly when at the beginning of words, but not when they are at the end. The effect of this problem can be mitigated by implementing a triphone model, in which each phone can be linked with their left or right neighbouring phones, or both [121]. Compared to the phone model, a triphone model may be able reduce word error rate by more than 50% [122], [123]. In this work, the conversion from monophone to triphone model, as shown in Figure 7.19, can be accomplished using the HTK library.

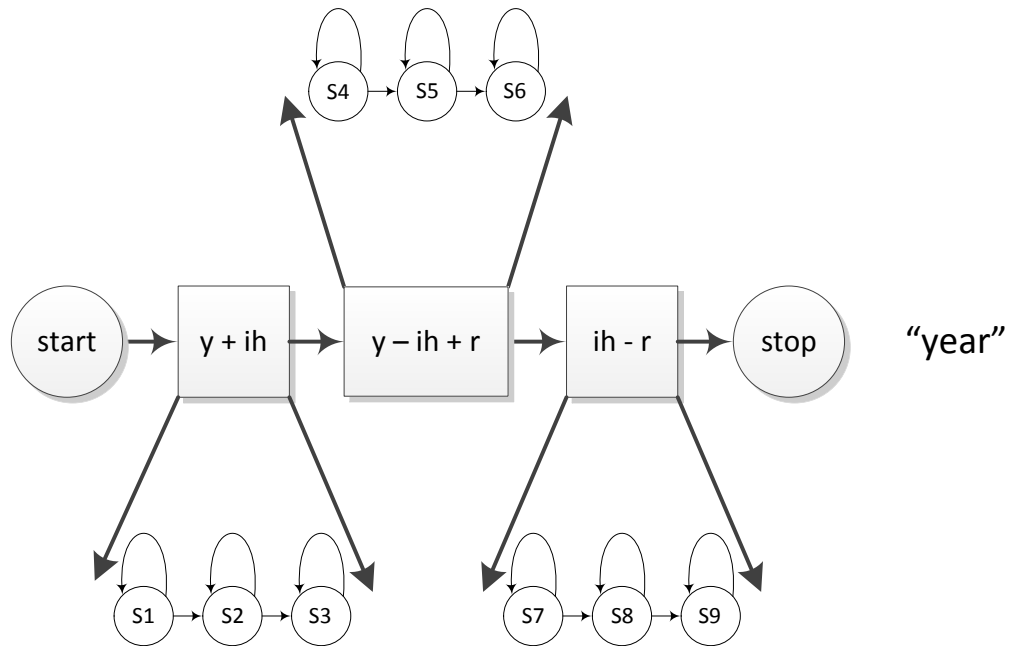


Figure 7.19 Example of a triphone recognition model

The four sentences in the LUNA-V data corpus were spoken three times by each of the subjects. Sentences 1, 2 and 3 as spoken three times by each of the subjects were employed for training. To increase further the coverage of phones, the first three sessions (of a total of five sessions) spoken by each of the subjects were included in the training set, bringing the total number of training word samples to 1230.

For testing purposes, sentence 4 spoken three times by each of the speakers (39 word samples) was used, making a total of 390 words in all. The recognition task uses ‘unseen’ data, as the words in sentence 4 were not present in the training data, although the structure of the unseen words can often potentially be constructed from the trained phones. For example, the word ‘same’ is not used in the training, but its structure can be derived from the words ‘system’ and ‘breakdown’ (as shown in Figure 7.20), both of which can be found in the training set. Recognition involves matching the audio and visual information in a test file against each HMM model, and that model which matches with the greatest probability is chosen, as depicted in Figure 7.21.

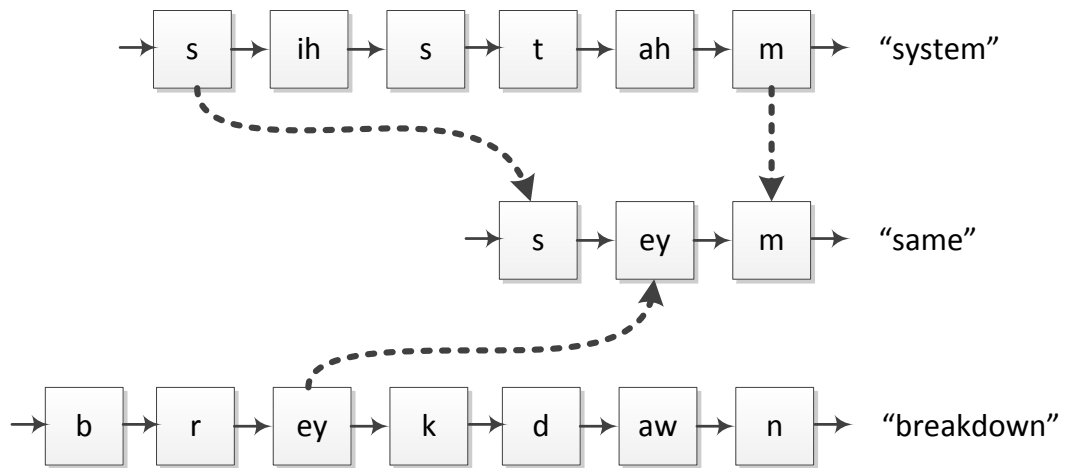


Figure 7.20 Construction of an untrained word using trained phones

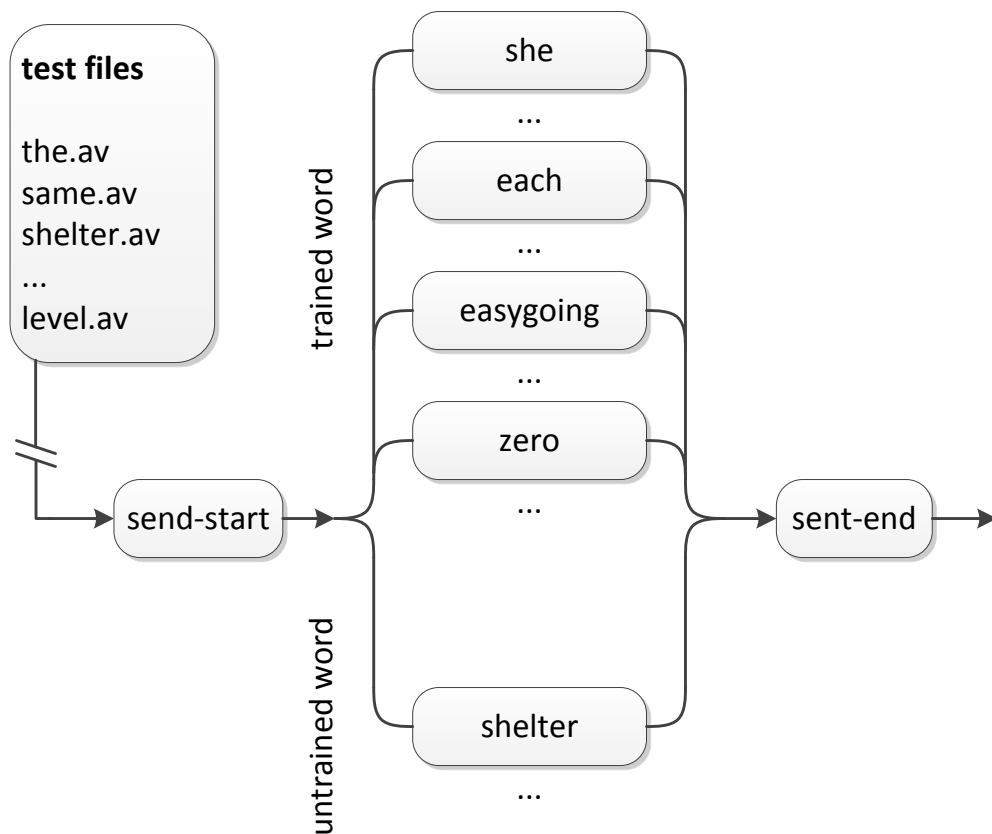


Figure 7.21 The recognition process involved the matching of words from the test file with models of trained words generated during training

7.3.3 Digit recognition results

This section presents the results obtained for an AVSR system that utilizes lip geometry information calculated from the visual information found in the LUNA-V data corpus, in order to determine whether the use of HD images improves the digit recognition rate in noisy environments. The experiments were conducted with ‘babble’, ‘factory1’, ‘factory2’ and ‘white’ noise obtained from the NOISEX-92 dataset [106] and added to speech signals such that specific SNRs are attained.

The experiments conducted here are similar to those carried out in Chapter 5, but now applied to the new LUNA-V data corpus. A visual-only digit recognition rate of 67.6% was achieved using five lip geometry features and the delta and delta-delta coefficients from the CUAVE corpus, whereas the corresponding recognition rate using the LUNA-V corpus was 92.5%. Figure 7.22 shows a comparison of the individual visual recognition results for all speakers using both data corpora. It can be seen that recognition performance for speakers in the CUAVE data corpus ranged from 30% to 95% with standard deviation of around 16%, while for the LUNA-V data corpus, the recognition performance varied from 80% to 100% with standard deviation of around 7%. The statistical results demonstrate that better visual word recognition performance can be achieved using HD visual information. However, attention should be drawn to the other differences between the data corpora, such as the speaker population, dialects and recording equipment, nevertheless the statistical improvement in recognition performance is clear.

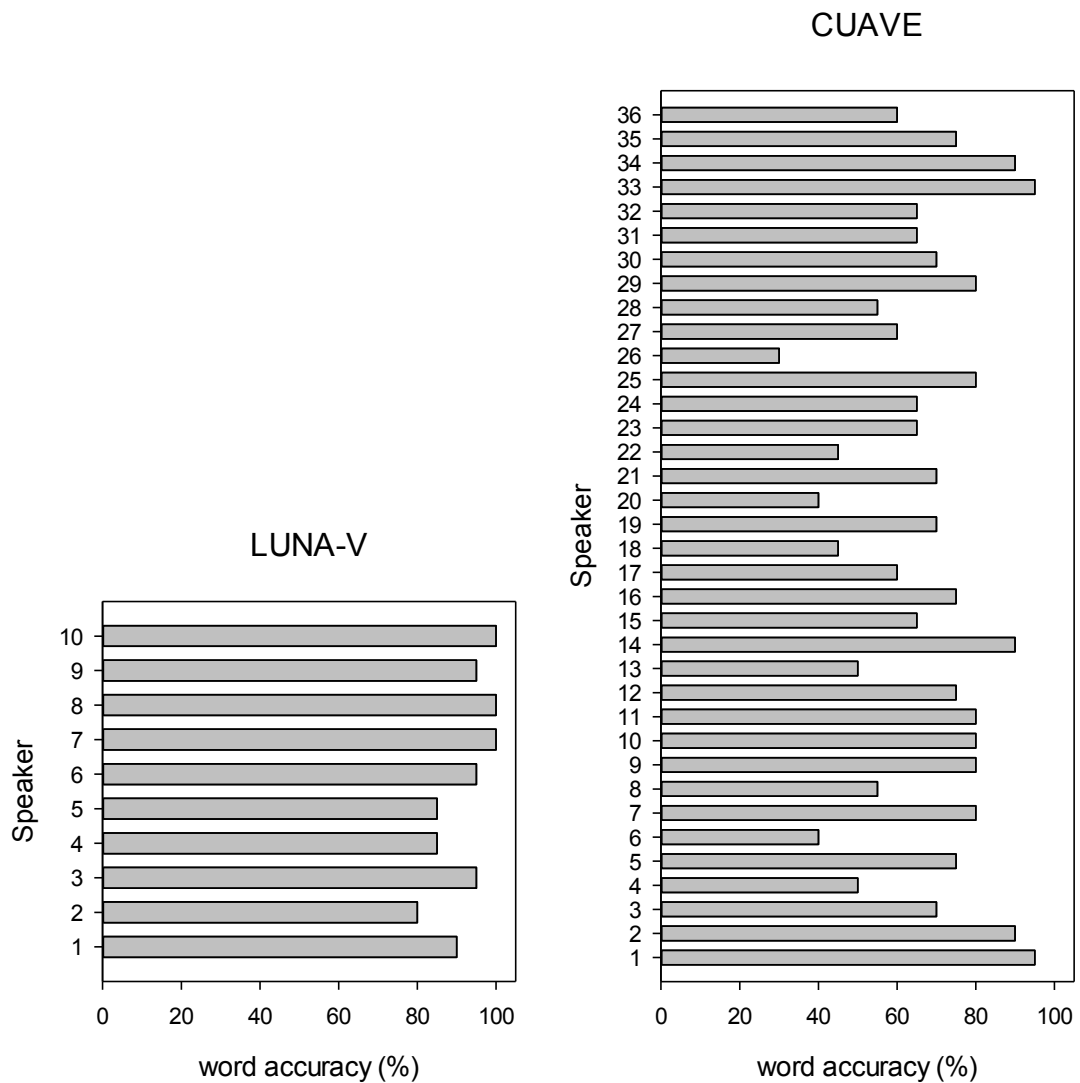


Figure 7.22 Visual speech recognition results for the individual subjects in the LUNA-V and CUAVE data corpora

Figure 7.23 shows how the word-accuracy performance of the geometrical-based AVSR system changed with SNR, where ‘babble noise’ from the NOISEX-92 database was added. It can clearly be seen that as the noise level is increased (and specifically for SNRs below 15dB), then using the combined information from the audio and visual modalities gives the best classification performance of the tests executed. For instance, at an SNR of 0dB, the improvement in performance is more than 40% compared with the corresponding audio-only figure.

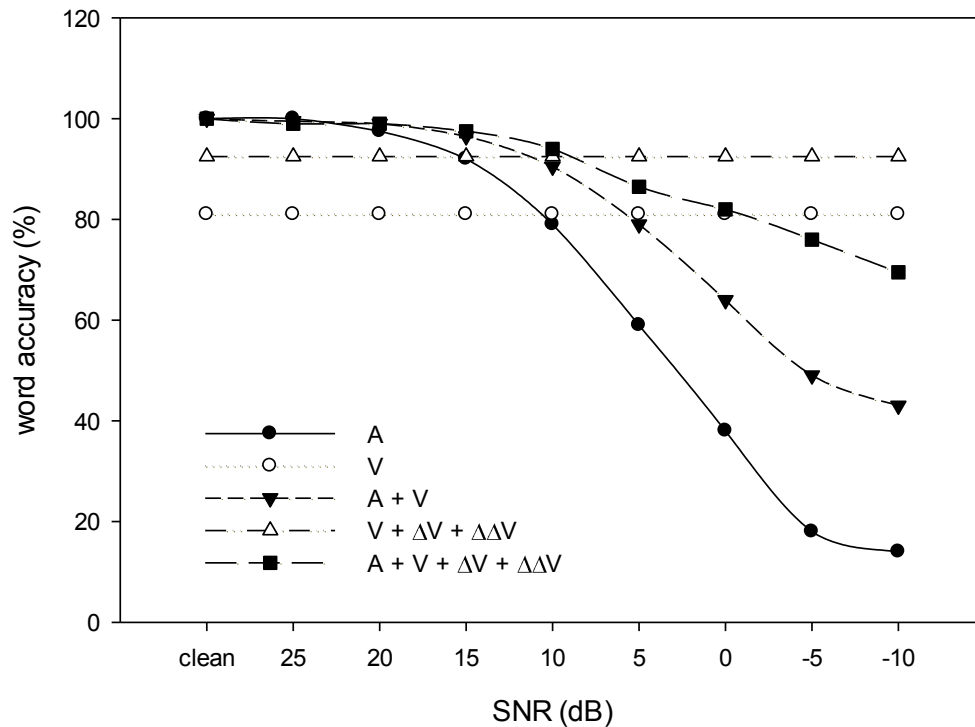


Figure 7.23 Performance of the geometrical-based AVSR system when ‘babble’ noise is applied and using the LUNA-V corpus. Shown are results when using audio-only data (A), visual-only data (V), dynamic visual information with delta and delta-delta features ($V + \Delta V + \Delta\Delta V$), a combination of audio and visual (A + V) features and a combination of audio, visual, visual delta and visual delta-delta features (A + V + $\Delta V + \Delta\Delta V$).

The classification performance of the visual modality was found to be improved if motion information is incorporated by the addition of the first-order (delta) and second-order difference (delta-delta) geometrical features. Such features are commonly used in audio speech recognition and Figure 7.23 shows the improvements in both audio and visual recognition that followed from the inclusion of the difference features; for example at an SNR of -10dB the performance can be seen to improve by more than 55% with respect to audio-only recognition.

Further tests of the geometrical-based AVSR system were carried out using ‘factory1’, ‘factory2’ and ‘white noise’ datasets from NOISEX-92 and the results can be seen in Figures 7.24, 7.25 and 7.26 respectively. The results obtained for these three types of noise showed similar improvements in word accuracy to those obtained using ‘babble’ noise, with the combined audio-visual performance at an SNR of -10dB improving by at least 40% in each case.

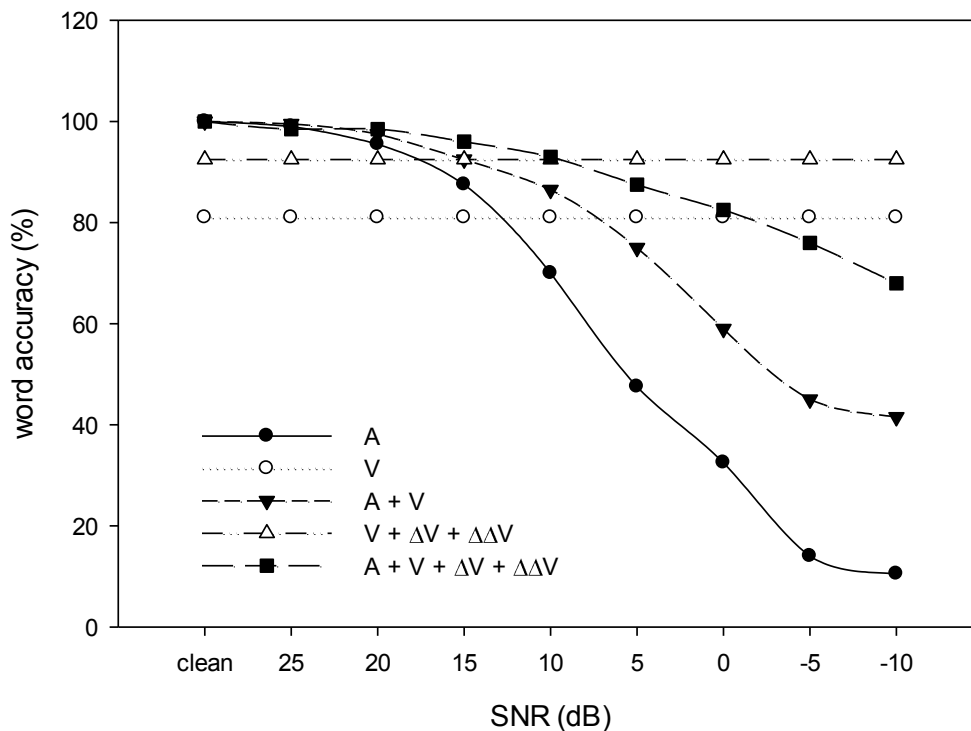


Figure 7.24 Performance of the geometrical-based AVSR system when ‘factory1’ noise is applied and using the LUNA-V corpus

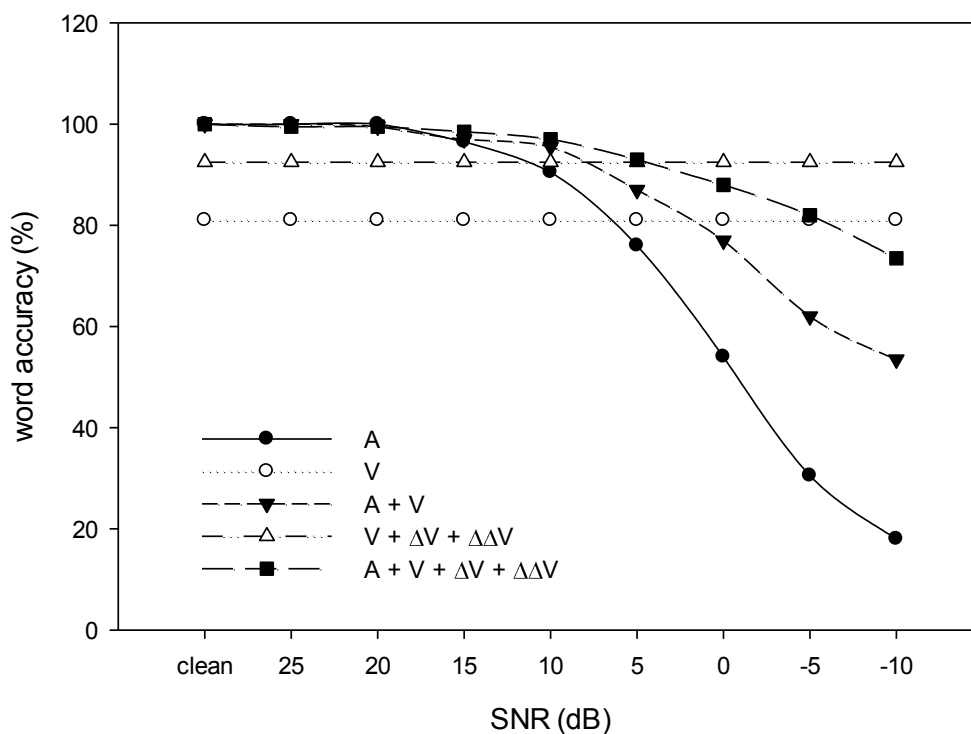


Figure 7.25 Performance of the geometrical-based AVSR system when ‘factory2’ noise is applied and using the LUNA-V corpus

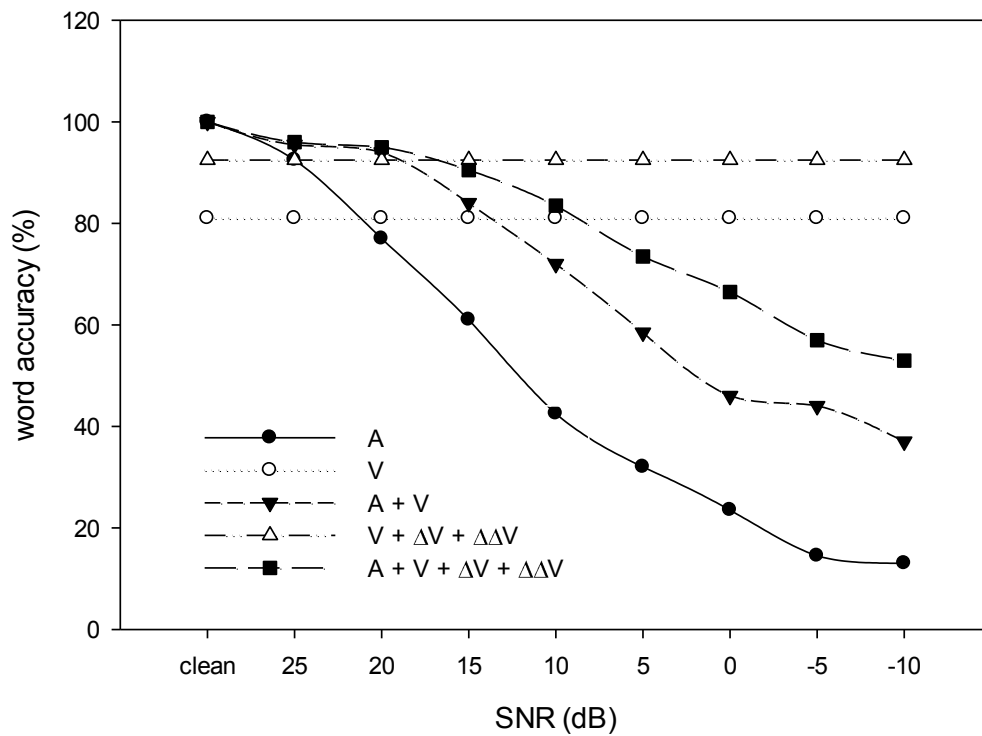


Figure 7.26 Performance of the geometrical-based AVSR system when ‘white’ noise is applied and using the LUNA-V corpus

Figure 7.27 shows the direct comparison between LUNA-V and CUAVE data corpus when simulated under ‘babble’ noise. When comparing the results from LUNA-V with CUAVE data corpus, there is very important finding related to the effect of the visual information on the word recognition performance. The performance of AVSR system can be improved significantly by increasing the accuracy of the visual-only recognition, as has occurred by using LUNA-V in place of CUAVE.

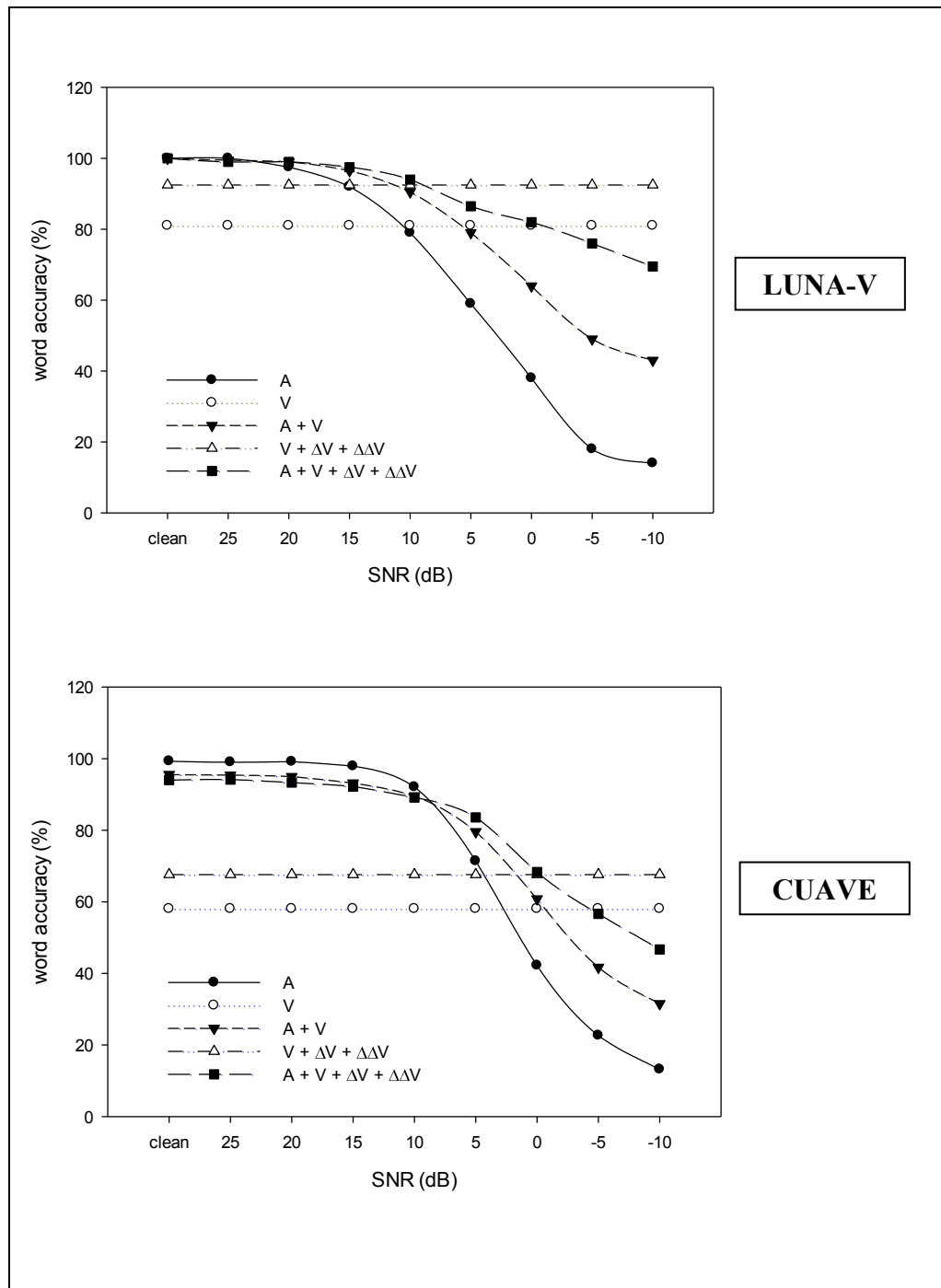


Figure 7.27 Direct comparison of the word accuracy AVSR system performance using the LUNA-V and CUAVE data corpora when ‘babble’ noise is applied

7.3.4 Speech recognition results

This section presents the results obtained for an AVSR system that utilizes lip geometry information, with the aim of improving speech recognition performance in noisy environments by using high quality images obtained from the LUNA-V data corpus. In a manner similar to that used for digit recognition in the previous sub-section, the experiments were conducted under the influence of ‘babble’, ‘factory1’, ‘factory2’ and ‘white’ noise from the NOISEX-92 dataset and which were added to the speech signals such that specific SNRs were attained.

Figures 7.28 and 7.29 show the word accuracy performance of the geometrical-based AVSR system for a range of SNR values when ‘babble’ and ‘white’ noise were added. For visual speech recognition, the AVSR system achieved 30% word accuracy when using lip geometry features and the recognition improved to 44% by including the delta and delta-delta coefficients. Compared to digit recognition, this word recognition task is more difficult as the system now has a broader selection of 51 different words, compared to the 10 used in the digit recognition experiments.

When no noise is added, it can clearly be seen that using a combination of audio and visual information improves classification performance by around 13% compared to audio-only recognition. As the noise level increases, the audio and audio-visual word accuracy performances reduce and follow a similar trend. It is interesting to see that at those SNRs when the visual-only performance is better than audio-only performance, the combined audio and visual modalities show a proportional greater improvement with respect to the audio curve. For instance, at an SNR of 0dB (both ‘babble’ and ‘white’ noise) the improvement in performance is more than 20% compared with the corresponding audio-only figure. Further tests of the geometrical-based AVSR system were carried out using ‘factory1’ and ‘factory2’ noise as shown in Figure 7.30 and 7.31. The trends of the results obtained for these two types of noise were similar to those found for ‘babble’ noise.

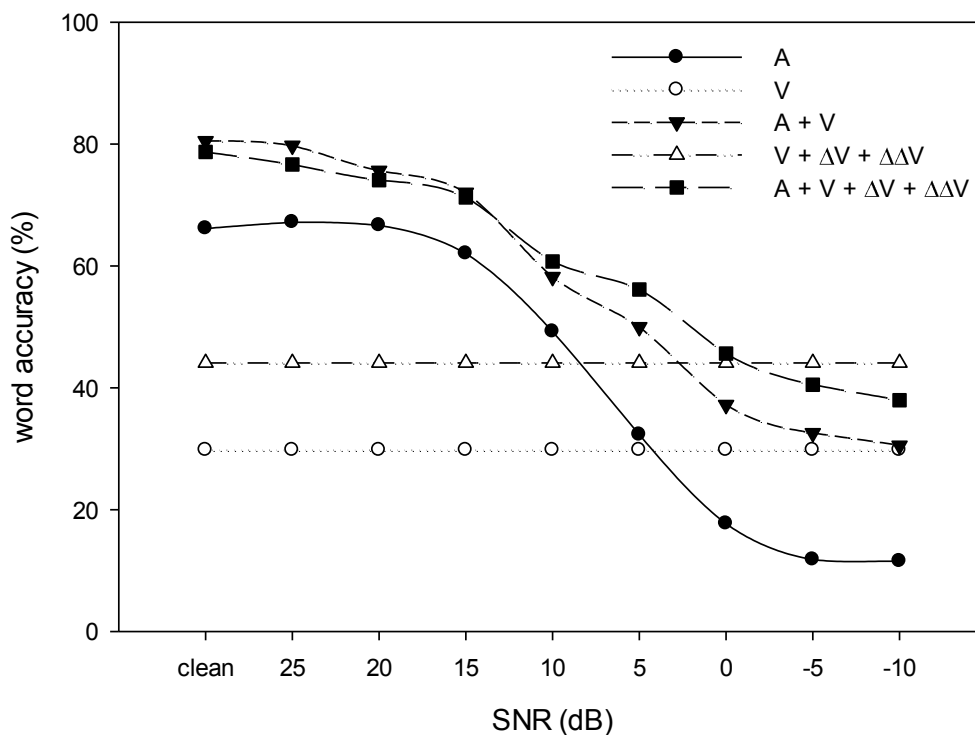


Figure 7.28 AVSR system performance using geometrical features when ‘babble’ noise was applied

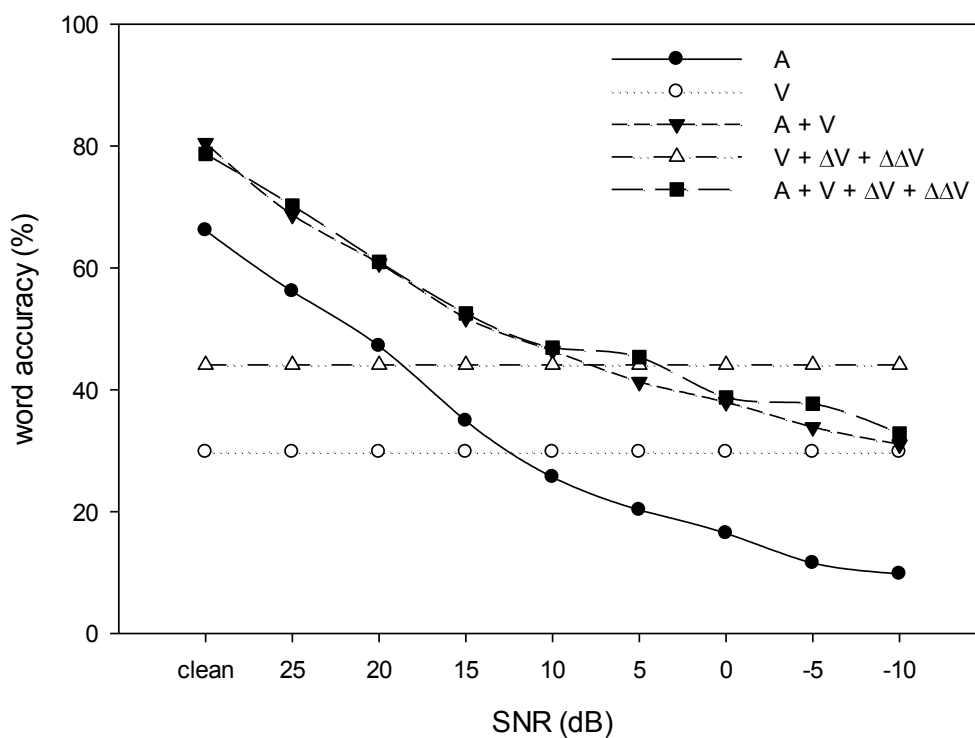


Figure 7.29 AVSR system performance using geometrical features when ‘white’ noise was applied

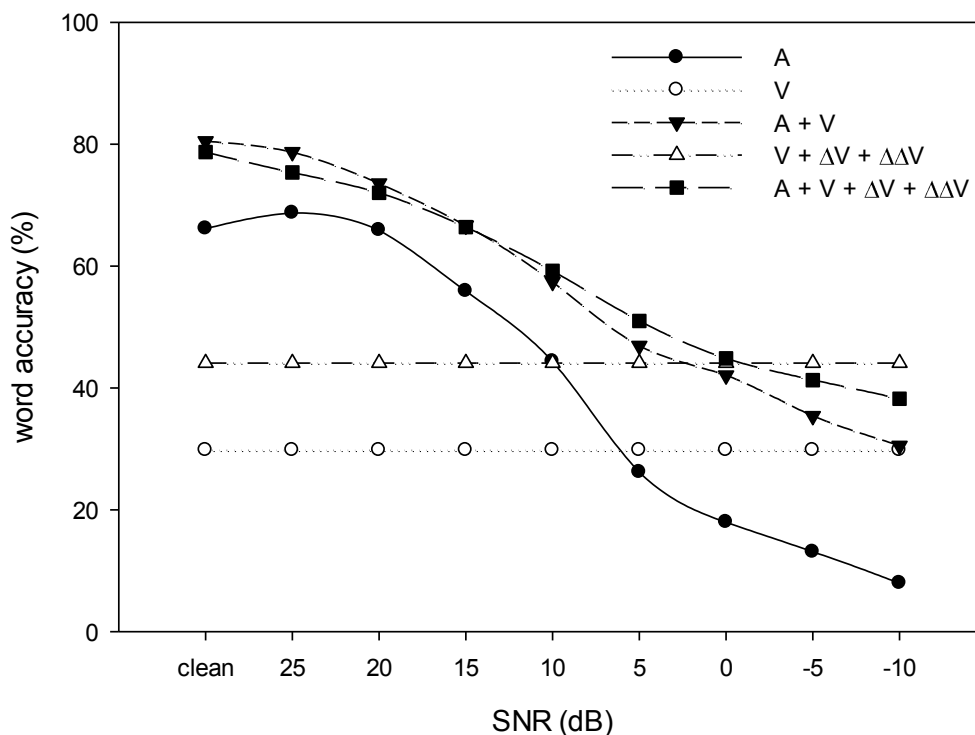


Figure 7.30 AVSR system performance using geometrical features when 'factory1' noise was applied

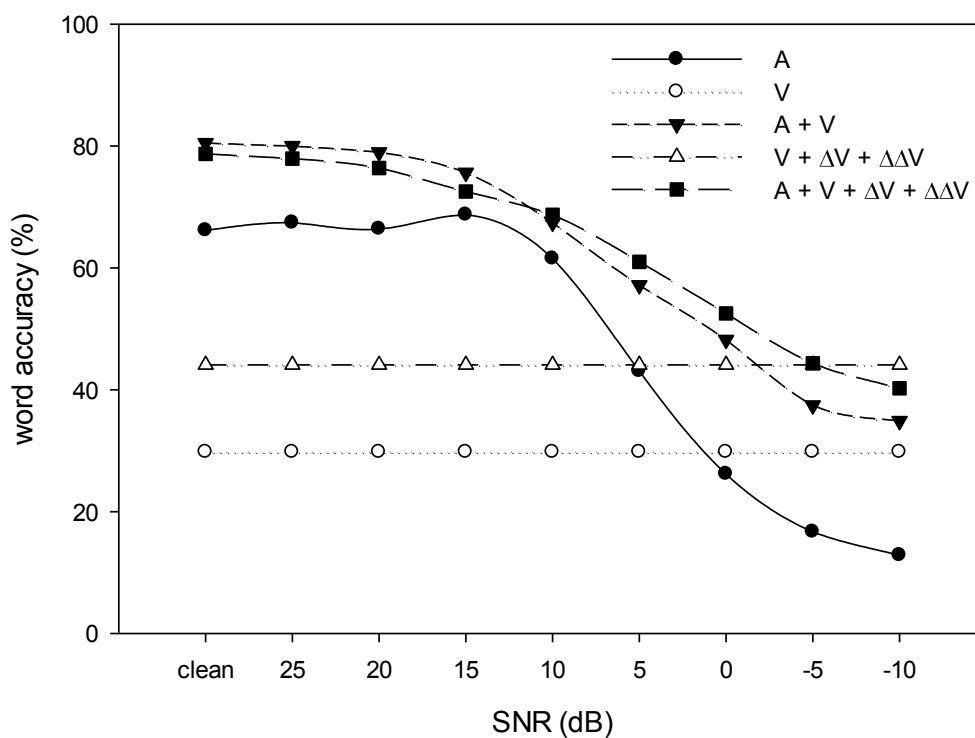


Figure 7.31 AVSR system performance using geometrical features when 'factory2' noise was applied.

To understand further how the performance of the AVSR system is affected by noise, part of the confusion matrix showing the actual and predicted classifications of 13 words is shown Table 7.8. Only the results of ‘babble’ and ‘white’ noise presented here as the trends of the results obtained for other two types of noise (factory1 and factory2) were similar to those found for ‘babble’ noise. The full confusion matrix can be found in Appendix F.

When performing visual-only recognition, the word ‘same’ is rarely confused and this is probably because substantial lip movements need to be made in its articulation. In contrast, the word ‘or’ is the most frequently confused during recognition, probably because its generation requires only minimal lip movements. By integrating good visual information into the audio data when they are affected by noise, the performance of the AVSR system can be significantly improved. For example, in the recognition of the word ‘same’, the performance of the AVSR improved by 43.3% and 20% when simulated under ‘babble’ and ‘white’ noise (SNR = 0dB) respectively. In contrast, under the same conditions, no improvement was apparent for the word ‘or’. Some participants pronounced the word ‘or’ as ‘o’ perhaps out of habit or due to dialect, making automatic classification difficult and adversely affecting performance.

When no noise was present then for certain words the visual recognition was better than for audio; examples being the words ‘be’ and ‘an’, where the improvement in performance was more than 50%. Consequently, when no noise is present, the combination of audio and visual information has the potential to yield a better performance compared to an audio-only system, providing a suitable fusion approach is taken. This also explained why the performance of the AVSR system especially in clean environment as shown in Figure 7.28 to 7.31 improved a lot, probably due to the performance contribution by word ‘be’ and ‘an’.

Under ‘babble’ and ‘white’ noise conditions at an SNR of 0dB, the recognition of a number of words improved significantly (by more than 30%) when video information is added and these cases have been highlighted in bold in Table 7.8. It can be seen that the recognition of all words tested improves when visual information is added to the audio system, except for the word ‘or’ as explained earlier.

Table 7.8 Summary of word recognition for each modality under different environmental noise condition

Word tested	Environmental noise						
	V	None		Babble ^a		White ^a	
		A	AV	A	AV	A	AV
the	46.7	90.0	93.3	0.0	46.7	0.0	6.7
same	83.3	93.3	100.0	56.7	100.0	50.0	70.0
shelter	50.0	100.0	93.3	0.0	10.0	0.0	10.0
could	10.0	16.7	40.0	3.3	20.0	0.0	3.3
be	46.7	33.3	93.3	0.0	23.3	33.3	46.7
built	66.7	93.3	100.0	3.3	66.7	36.7	93.3
into	43.3	90.0	86.7	0.0	3.3	0.0	0.0
an	40.0	40.0	93.3	66.7	96.7	63.3	100.0
embankment	73.3	96.7	66.7	3.3	20.0	0.0	6.7
or	10.0	10.0	36.7	0.0	0.0	0.0	0.0
below	50.0	80.0	70.0	30.0	86.7	6.7	56.7
ground	26.7	60.0	73.3	10.0	50.0	0.0	30.0
level	26.7	56.7	76.7	56.7	70.0	23.3	80.0

Note 1. ^a SNR = 0 dB

Note 2. A-audio, V-visual , AV-combination of audio and visual

Note 3. **Bold** figure indicates the AV recognition was 30% or better than A

Note 4. Percentages have been pooled over all participants and word contexts

7.4 Summary

This chapter has described a new data corpus named LUNA-V that has been collected at Loughborough University with the purpose of validating the AVSR methods proposed in the thesis. The data corpus consists of a total of 170 sentences containing 1820 words available in 10 separate video files (one for each speaker). The corpus has been made freely available to the community for research use.

The lip extraction process developed for the CUAVE data corpus has been successfully applied to the LUNA-V corpus and its effectiveness demonstrated using both qualitative and quantitative assessments.

Performance evaluations that have used the AVSR system to recognize the English digits ‘zero’ to ‘nine’, demonstrated that the higher resolution images of the LUNA-V data corpus compared to CUAVE could lead to a significant improvement in visual-only word recognition performance. In the speech recognition experiments, it was shown that the accuracy of word recognition improves when visual information is added to the audio system. Although the number of words in the vocabulary tested is currently small, it is sufficient to provide a good indication that substantial improvements in audio-only speech recognition can be achieved by the appropriate integration of information from visual features.

In addition, the speech recognition results presented also accord with those of Chițu *et al.* [26], in which the authors achieved similar trend of improvement in term of word accuracy when combining visual information in speech recognition system using DUTAVSC data corpus. This is an important result, since although the DUTAVSC data corpus in Dutch language, the LUNA-V data corpus presented here was able to perform at the same level with existing data corpus available in AVSR field.

CHAPTER 8

CONCLUSION AND FUTURE WORKS

This thesis has presented a number of new approaches relevant to audio-visual speech recognition, principally for feature extraction and classification, including integration in the presence of audio noise. For feature extraction, the work has established a novel lip geometry approach and demonstrated its robustness to head rotation and brightness changes. For classification, a template probability multi-dimension dynamic time warping method has been developed that allows improved modelling of feature changes during the utterance of a word. By using a small number of features, the new lip geometry approach also reduces the effect of the ‘curse of dimensionality’ in feature-fusion audio-visual speech recognition systems. For audio-visual integration, a novel adaptive fusion AVSR system based on skewness and kurtosis analysis is proposed which enhances the robustness of the system in noisy environments by using an appropriate combination of modalities where the selection threshold to choose modality was trained using white noise. Finally, a new data corpus namely LUNA-V that has been collected at Loughborough University that seeks to validate the methods proposed in the thesis that previously using CUAVE database.

8.1 Summary of the work in the thesis

This section summaries the main contributions of the thesis.

8.1.1 Lip geometry feature extraction

A new process has been established to extract lip geometry information from single images of video sequences on which classification can be performed that is able to identify visual speech based on dynamic lip movements. Extraction of lip

geometry features was carried out using a combination of a skin colour filter, a border following algorithm and a convex hull approach. The proposed method was compared with the popular ‘snake’ technique and was found to significantly improve lip shape extraction performance for the database studied. The lip geometry features obtained, including height, width, ratio, area, perimeter and various combinations of these features were evaluated in their classification using three separate methods, namely optical flow, dynamic time warping (DTW) and a new approach termed multi-dimensional DTW. Experiments using the English digits 0 to 9 as provided in the video sequences available in the CUAVE database, show that using only lip height, lip width and their ratio, the proposed system is capable of a recognition performance of 68%. As these results compare very favourably with previous results found in the literature, the approach appears to have the potential to be incorporated in a multimodal speech recognition system for use in noisy environments.

8.1.2 Robustness of lip geometrical features to head rotation and brightness changes in lip reading system

This work has provided a new automatic lip-reading system that uses geometrical information extracted from video sequences in the classification of dynamic lip movements and which has been implemented in four variants of hidden Markov models. In the recognition of the English digits ‘zero’ to ‘nine’ as spoken by the subjects in the CUAVE database, the proposed system was able to produce a word recognition performance better than that obtained using a conventional appearance-based discrete cosine transform technique. The two approaches were also compared when operating under simulated changes in environment conditions that arise from head movements and variations in image illumination. The performance of the appearance-based approach was adversely affected by such rotational and brightness changes, yet the performance of the geometrical-based method remained consistent, demonstrating its robustness for use as part of a multimodal speech recognition system.

8.1.3 Geometrical based lip reading system using TP-MDTW

Using the approach to generate high-quality geometrical information described in section 8.1.1, a complete lip-reading system has been implemented. There are three novel aspects to the work. Firstly, the application of a border following technique and the construction of a convex hull was able to provide lip extraction of an accuracy that improved considerably on previously reported results. Secondly, the features extracted from the lip geometry are determined over the duration of the video sequences and pattern matching with respect to stored templates is performed according to the dynamic variations characteristic of individual word utterances. Thirdly, the dynamic features are classified using a novel Template Probabilistic MDTW approach that is able to adapt to the dynamic differences in the way words are uttered by speakers. In the experiments, the results obtained from the new approach compared favourably to those of existing lip reading approaches, achieving a word recognition accuracy of up to 76%.

8.1.4 Lip geometry approach to reduce the ‘curse of dimensionality’ in feature-fusion based AVSR

A range of experiments was carried out to demonstrate that the geometrical features established in this work have information content that is highly descriptive for the recognition task, while also suffering little from the ‘curse of dimensionality’ (or scalability) issue that is often affects performance in feature-fusion based AVSR systems. The results were compared to those obtained using conventional appearance-based methods, namely the discrete cosine transform and principal component analysis techniques, when operating under a range of different signal to noise ratio conditions. Experimental results show that the implementations using the new geometrical-based feature approach outperformed those using appearance-based features in terms of recognition accuracy, even though a significantly reduced number of features was used.

8.1.5 Novel adaptive fusion AVSR using skewness and kurtosis analysis

In the decision fusion approach, recognition is performed separately for each of the modalities, with the partial results from each sub-process being combined to produce the final classification. As the models may often deliver different partial classification decision outcomes, the decision-fusion approach must provide a suitable method for their ranking and collation. A range of experiments were carried out to identify a suitable method to estimate a suitable means of selecting between the audio and video modalities based on the quality of the audio received. The selected method used an assessment of the audio quality based on skewness and kurtosis measurements and the decision boundary was established by training on white noise examples. The results generated were very promising and the new system achieved a recognition improvement by up to 67% with respect to audio-only recognition at an SNR of -10dB when tested using ‘white’, ‘babble’, ‘factory1’ and ‘factory2’ noise.

8.1.6 Development of the LUNA-V data corpus

The current work has demonstrated that AVSR systems are able to perform better than audio-only speech recognizers in noisy environments. In order to develop reliable AVSR systems and to demonstrate their utility, appropriate audio-visual databases are needed. Although a small number of such databases have been collected and made generally available, none has the high-definition quality that can be achieved using modern equipment. Consequently, a new data corpus, LUNA-V, was developed that contains audio-visual data of 10 speakers each uttering 10 isolated digits and five sentences. The sentence design was adopted from the CUAVE and TIMIT corpora, both of which have been widely used for audio-speech recognition experiments. Experimental results have shown that the lip geometry features extracted from the LUNA-V corpus are able to provide results that significantly improve on those achievable using the same feature set extracted from the CUAVE database, both in terms of the accuracy of lip segmentation produced and also in terms of the subsequent AVSR process.

8.2 Future work

As is normally the case in research of this nature, there are many areas that can be identified that warrant further investigation as well as new areas that have yet to be explored. In this section, an overview of the opportunities for future work are discussed.

8.2.1 Scale invariant features

One issue in using lip geometric features in AVSR systems is that the dimensions of the lips in the scene depend on the distance of the mouth relative to the camera lens, thereby affecting the values of geometrical feature acquired. This effect can be mitigated and the AVSR system made more robust by using features that are invariant to scale. As the subjects in the CUAVE and LUNA-V data corpora were not allowed translational movement in the direction of the camera view, simple measurements rather than scale invariant parameters could be used. However, such limited translational movement cannot be guaranteed for video sequences captured in real-world environments and scale invariant parameters would be required. One possible scale invariant feature is the lip angle [124] and Figure 8.1 shows examples of five lip angle parameters that can be derived directly from the lip height and width. A further advantage of using lip invariant features is that speaker independent experiments can be performed more easily as the features are effectively normalized automatically for all the speakers in the database. Using the robust approach to lip extraction that has been developed in the current work, an interesting further study would be to assess AVSR performance using scale invariant features obtained from a database of subjects who are allowed greater movement within the camera's frame of view.

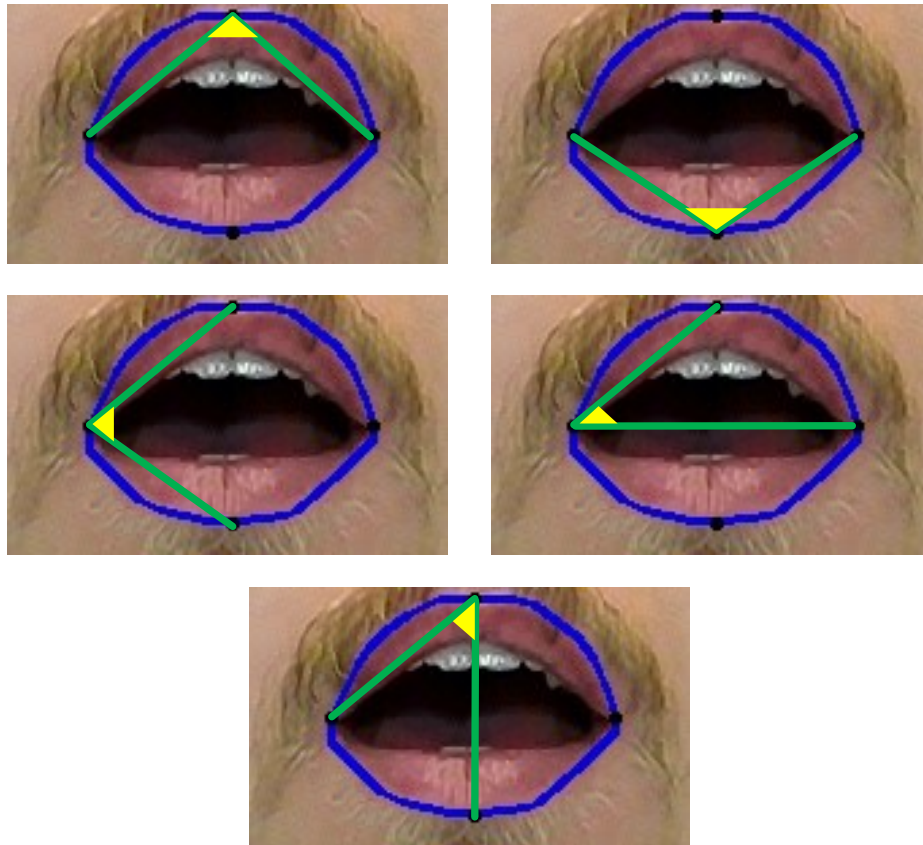


Figure 8.1 Five possible scale invariance lip angle features for robust AVSR in environments where subjects are allowed some freedom of movement

8.2.2 Extending the LUNA-V data corpus

The current LUNA-V data corpus was sufficient to make initial explorations of the behaviour and trends of the audio and visual parameters. Although the image quality obtained was better than that available in other audio-visual speech data corpora, the information available can be extended in a number of ways. At present, the corpus contains 10 speakers and more should be included to improve the statistical reliability of results obtained. In addition, only one female speaker is currently in the database and more need to be added in order to allow the detection of characteristics that may differ between the two sexes and so potentially consider the development of separate models in order to improve overall performance. Initial indications are that there may be differences in the audio speech parameters, such as in the fundamental frequency and the first three formant frequency [125], but verification requires more speakers. The LUNA-V data corpus could also be

extended to include sentences that cover additional parts of the phone space that involve different phone contexts.

8.2.3 Using the Microsoft Kinect as the hardware interface

Recent advances in 3D camera sensor technology have created many new opportunities for human computer interaction (HCI). The Microsoft Kinect [126] is one such 3D device and its sensor is able to access depth information in an environment, potentially allowing more accurate models to be defined and so simplifying the recognition task [127]. The active appearance model (AAM) [13] and the active shape model (ASM) [87] have been investigated by a number of researchers, but their practical realization requires considerable computational resources and are generally not able to provide adequate accuracy and robustness for real-time implementation. One solution to this computational problem has been developed in [128], in which the authors successfully performed 3D tracking based on an AAM model constrained by depth data provided by Kinect sensor. This gives many potential opportunities for new AVSR work, particularly since the Kinect has available a microphone array with excellent sound quality and Microsoft has released a software development kit containing image and signal processing algorithms optimized for the sensor.

8.2.4 Using lip gestures for HCI

There is an increasing demand for the development of new HCIs for situations when it is difficult, ineffective or not possible to use the traditional mouse and keyboard for input. As an example, the television HCI has evolved from pressing buttons on the set, to remote controls and now to the detection of hand gestures by 3D camera sensors [129]. Lip gestures, where the mouth movements are made but sound is not necessarily emitted, have been tested as a potential HCIs by paralyzed and severely disabled people [130]. As this thesis has provided a robust method to extract lip geometry, an interesting investigation would be to assess its ability to recognize lip gestures as part of an HCI system.

REFERENCES

- [1] J. Aron, “How innovative is Apple’s new voice assistant, Siri?,” *New Sci.*, vol. 212, no. 2836, p. 24, 2011.
- [2] W. Kim and J. H. L. Hansen, “Feature Compensation Employing Variational Model Composition for Robust Speech Recognition in In-Vehicle Environment,” in *Digital Signal Processing for In-Vehicle Systems and Safety*, J. H. L. Hansen, P. Boyraz, K. Takeda, and H. Abut, Eds. Springer US, 2012, pp. 175–185.
- [3] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.
- [4] BBC, “Horizon: Is Seeing Believing?,” United Kingdom, 2010.
- [5] G. Potamianos and C. Neti, “Improved ROI and within frame discriminant features for lipreading,” *Int. Conf. Image Process.*, vol. III, pp. 250–253, 2001.
- [6] X. Zhang, C. C. Broun, R. M. Mersereau, and M. a. Clements, “Automatic Speechreading with Applications to Human-Computer Interfaces,” *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1228–1247, 2002.
- [7] E. Benhaim, H. Sahbi, and G. Vitte, “Designing relevant features for visual speech recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 2420–2424.
- [8] Q. Summerfield, “Use of visual information for phonetic perception.,” *Phonetica*, vol. 36, no. 4–5, pp. 314–331, 1979.
- [9] G. Potamianos, C. Neti, J. Luetin, and I. Matthews, “Audio-visual automatic speech recognition: An overview,” *Issues in Visual and Audio-Visual Speech Processing*. MIT Press, 2004.
- [10] C. Neti, G. Potamianos, J. Luetin, I. Matthews, H. Glotin, and D. Vergyri, “Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins Summer 2000 Workshop,” in *Proc. Works. Multimedia Signal Processing*, 2001, pp. 619–624.
- [11] G. Potamianos, H. P. Graf, and E. Cosatto, “An image transform approach for HMM based automatic lipreading,” in *International Conference on Image Processing. ICIP98*, 1998, vol. 3, pp. 173–177.

-
- [12] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimed.*, vol. 2, no. 3, pp. 141–151, 2000.
- [13] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Proc. Eur. Conf. Comput. Vis.*, pp. 484–498, 1998.
- [14] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Processing*. Prentice Hall Signal Processing Series, 1993.
- [15] P. Aleksic and G. Potamianos, "Exploiting visual information in automatic speech processing," *Image Video Process.*, 2005.
- [16] I. S. Pandzic and R. Forchheimer, *MPEG-4 facial animation: the standard, implementation and applications*, vol. 13, no. 5. John Wiley and Sons, 2002, p. 299.
- [17] C. Xu and J. L. Prince, "Gradient Vector Flow: A New external force for Snakes," *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, 1997 (CVPR 1997)*, pp. 66–71, 1997.
- [18] J. Canny, "A Computational Approach to Edge Detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679 – 698, 1986.
- [19] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, IEEE, pp. 9–21, 2001.
- [20] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "Moving-Talker, Speaker-Independent Feature Study, and Baseline Results Using the CUAVE Multimodal Speech Corpus," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1189–1201, Jan. 2002.
- [21] S. Theodoridis and K. Koutroumbas, *Pattern recognition 2nd Ed*, vol. 8, no. 3. Academic Press, 2003.
- [22] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001 (ICASSP'01)*, 2001, vol. 1, pp. 177–180.
- [23] W. Yau, D. Kumar, and S. Arjunan, "Voiceless speech recognition using dynamic visual speech features," in *HCSNet Workshop on the Use of Vision in HCI*, 2006.
- [24] A. F. Bobick and J. W. Davis, "The Recognition of Human Movement Using Temporal Templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.
- [25] A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemin, and J. Gubbi, "Lip Reading using Optical Flow and Support Vector Machines," in *3rd*

- International Congress on Image and Signal Processing (CISP)*, 2010, pp. 327–330.
- [26] A. G. Chițu, L. J. M. Rothkrantz, P. Wiggers, and J. C. Wojdel, “Comparison between different feature extraction techniques for audio-visual speech recognition,” *J. Multimodal User Interfaces*, vol. 1, no. 1, pp. 7–20, Mar. 2007.
- [27] S. Lucey, T. Chen, S. Sridharan, and V. Chandran, “Integration strategies for audio-visual speech processing: applied to text-dependent speaker recognition,” *IEEE Trans. Multimed.*, vol. 7, no. 3, pp. 495–506, Jun. 2005.
- [28] F. Lavagetto, “Converting speech into lip movements: a multimedia telephone for hard of hearing people,” *IEEE Trans. Rehabil. Eng.*, vol. 3, no. 1, pp. 90–102, Mar. 1995.
- [29] D. Crystal, *A Dictionary of linguistics and phonetics*, 6th ed. Wiley-Blackwell, 2008, p. 560.
- [30] M. Z. Ibrahim and D. J. Mulvaney, “Geometry based lip reading system using Multi Dimension Dynamic Time Warping,” in *IEEE Visual Communications and Image Processing (VCIP 2012)*, 2012, pp. 1–6.
- [31] M. Z. Ibrahim and D. J. Mulvaney, “Robust geometrical-based lip-reading using Hidden Markov models,” in *IEEE Eurocon*, 2013, pp. 2011–2016.
- [32] M. Z. Ibrahim and D. J. Mulvaney, “A lip geometry approach for feature-fusion based audio-visual speech recognition,” in *IEEE 6th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, 2014, pp. 644–647.
- [33] K. O. Jian-Tong Wu, Shinichi Tamura, Hiroshi Mitsumoto, Hideo Kawai, Kenji Kurosu, “Neural network vowel recognition jointly using voice features and mouth shape image,” *Pattern Recognit.*, vol. 24, no. 10, pp. 921–927, 1991.
- [34] T. Chen and R. R. Rao, “Audio-visual integration in multimodal communication,” *Proc. IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.
- [35] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, “A review of speech-based bimodal recognition,” *Multimedia, IEEE Trans.*, vol. 4, no. 1, pp. 23–37, Mar. 2002.
- [36] A. Adjoudani and C. Benoit, “Audio-visual speech recognition compared across two architectures,” in *European Conf. Speech Communication and Technology (Eurospeech’95)*, 1995, pp. 1563–1566.
- [37] S. Cox, I. Matthews, and J. A. Bangham, “Combining noise compensation with visual information in speech recognition,” in *Auditory-Visual Speech Processing (AVSP’97)*, 1997.

- [38] M. McGrath and Q. Summerfield, "Intermodal timing relations and audio-visual speech recognition," *J. Acoust. Soc. Amer.*, vol. 77, no. 2, pp. 678–685, 1985.
- [39] J. Luettin, G. Potamianos, and C. Neti, "Asynchronous stream modeling for large vocabulary audio-visual speech recognition," *Int. Conf. Acoust. Speech Signal Process.*, vol. 1, pp. 169–172, 2001.
- [40] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, 1998, vol. 6, no. 2, pp. 3733–3736.
- [41] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust.*, vol. 28, no. 4, 1980.
- [42] B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, vol. 50, pp. 637–655, 1971.
- [43] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," *IEEE Trans. Acoust.*, vol. ASSP-34, no. 1, pp. 52–59, 1986.
- [44] J. G. Wilpon, C.-H. Lee, and L. R. Rabiner, "Improvements in connected digit recognition using higher order spectral and energy features," [*Proceedings*] *ICASSP 91 1991 Int. Conf. Acoust. Speech, Signal Process.*, 1991.
- [45] H. P. Graf, E. Cosatto, and M. Potamianos, "Robust recognition of faces and facial features with a multi-modal system," *1997 IEEE Int. Conf. Syst. Man, Cybern. Comput. Cybern. Simul.*, vol. 3, pp. 2034–2039, 1997.
- [46] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2001, vol. 1, pp. 511–518.
- [47] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. International Conference on Multimedia and Expo*, 2001, no. 2, pp. 2–5.
- [48] G. Pomianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.
- [49] C. Bregler and Y. Konig, "'Eigenlips' for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994.*, 1994, vol. 2, pp. 669–672.

- [50] P. Duchnowski, U. Meier, and A. Waibel, "See Me, Hear Me: Integrating Automatic Speech Recognition And Lip-Reading," in *Int. Conf. Spoken Lang. Process*, 1994, pp. 547–550.
- [51] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audio-visual speech recognition," *Int. Conf. Spok. Lang. Process.*, vol. 3, pp. 20–23, 2000.
- [52] A. V. Nefian, L. Liang, X. Pi, X. Liu, and K. Murphy, "Dynamic Bayesian Networks for Audio-Visual Speech Recognition," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1274–1288, 2002.
- [53] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction," in *International Conference on Multimedia and Expo. ICME '03.*, 2003, pp. 9–12.
- [54] G. Potamianos, C. Neti, J. Huang, J. H. Connell, S. Chu, V. Libal, E. Marcheret, N. Haas, and J. Jiang, "Towards practical deployment of audio-visual speech recognition," *IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 3, pp. 777–780, 2004.
- [55] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition," *EURASIP J. Adv. Signal Process.*, vol. 2002, no. 11, pp. 1260–1273, 2002.
- [56] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.
- [57] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *Int. J. Comput. Vis.*, vol. 8, no. 2, pp. 99–111, Aug. 1992.
- [58] X. Huang, S. Z. Li, and Y. Wang, "Statistical learning of evaluation function for ASM/AAM image alignment," in *ECCV Workshop BioAW*, 2004, pp. 45–56.
- [59] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Macmillan Publishing Company, 1993.
- [60] D. G. Stork and M. E. Hennecke, *Speechreading by humans and machines*. Berlin, Germany: Springer, 1996.
- [61] R. Campbell, B. Dodd, and D. Burnham, *Hearing by Eye II: Advances in the psychology of speechreading and audio-visual speech*, vol. 115, no. 4. Psychology Press, 1998, p. 338.

- [62] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *Am. Sci.*, vol. 86, no. 3, pp. 236–244, 1998.
- [63] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Rationale for phoneme-viseme mapping and feature selection in visual speech recognition," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, Ed. Berlin, Germany: SpringerVerlag, 1996, pp. 505–515.
- [64] A. Rogozan, "Discriminative Learning of Visual Data for Audiovisual Speech Recognition," *Int. J. Artificial Intell. Tools*, vol. 8, no. 1, pp. 43–52, 1999.
- [65] C. Neti, G. Potamianos, J. Luetttin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Center for Language and Speech Processing, g, The Johns Hopkins University, Baltimore, MD, Final Workshop 2000 Report, 2000.
- [66] A. ChiÑu and L. Rothkrantz, "Building a Data Corpus for Audio-Visual Speech Recognition," vol. 1, no. Movellan 1995, 2007.
- [67] J. R. Movellan, "Visual Speech Recognition with Stochastic Networks," in *Advances in Neural Information Processing Systems 7*, G. Tesauro, D. S. Touretzky, and T. K. Leen, Eds. MIT Press, 1995, pp. 851–858.
- [68] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, 2002.
- [69] J. B. Millar and R. Goecke, "The audio-video australian English speech data corpus AVOZES.," in *INTERSPEECH*, 2004.
- [70] C. Sanderson and K. K. Paliwal, "Fast features for face authentication under illumination direction changes," *Pattern Recognit. Lett.*, vol. 24, no. 14, pp. 2409–2419, 2003.
- [71] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. Int. Conf. Spoken Lang. Process*, 2004, pp. 2489–2492.
- [72] J. C. Wojdel, P. Wiggers, and L. J. M. Rothkrantz, "An audio-visual corpus for multimodal speech recognition in Dutch language," in *the International Conference on Spoken Language Processing (ICSLP) 2002*, 2002.
- [73] G. Richard, Y. Mengay, I. Guis, N. Suaudeau, J. Boudy, P. Lockwood, C. Fernandez, F. Fernandez, C. Kotropoulos, A. Tefas, P. Pitas, R. Heimgartner, P. Ryser, C. Beumier, P. Verlinde, S. Pigeon, G. Matas, J. Kittler, J. Biglin, Y. Abdeljaoued, E. Meurville, L. Besacier, M. Ansorge, G. Maitre, J. Luetttin, S. Ben-Yacoub, B. Ruiz, K. Aldama, and J. Cortes, "Multi modal verification for teleservices and security applications (M2VTS)," *Proc. IEEE Int. Conf. Multimed. Comput. Syst.*, vol. 2, 1999.

-
- [74] J. Sharp, *Microsoft Visual C# 2010 Step by Step*. Redmond, Washington: Microsoft Press, 2010.
- [75] G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.
- [76] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *International Conference on Image Processing 2002 (ICIP2002)*, 2002, pp. 900–903.
- [77] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.
- [78] C. Zhang and Z. Zhang, "A Survey of Recent Advances in Face Detection," Microsoft Research, 2010.
- [79] A. Albiol, L. Torres, and E. J. Delp, "Optimum color spaces for skin detection," in *International Conference on Image Processing 2001*, 2001, vol. 1, no. xL, pp. 122–124.
- [80] P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A survey of skin-color modeling and detection methods," *Pattern Recognit.*, vol. 40, no. 3, pp. 1106–1122, Mar. 2007.
- [81] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on Pixel-Based Skin Colour Detection Techniques," in *Proceedings of Graphicon 2003*, 2003.
- [82] S. Suzuki and K. Be, "Topological structural analysis of digitized binary images by border following," *Comput. Vis. Graph. Image Process.*, vol. 30, no. 1, pp. 32–46, 1985.
- [83] M. de Berg, M., Cheong, O., van Kreveld, M., Overmars, *Computational Geometry: Algorithms and Applications*, Third Edit. Springer-Verlag, 2008, p. 386.
- [84] N. Eveno, A. Caplier, and P. Coulon, "Accurate and Quasi-Automatic Lip Tracking," *IEEE Trans. Circuits Syst.*, vol. 14, no. 5, pp. 706–715, 2004.
- [85] P. Kuo, P. Hillman, and J. Hannah, "Improved lip fitting and tracking for model-based multimedia and coding," in *IEE International Conference on Visual Information Engineering (VIE 2005)*, 2005, pp. 251–258.
- [86] Y. Yokogawa, N. Funabiki, T. Higashino, M. Oda, and Y. Mori, "A proposal of improved lip contour extraction method using deformable template matching and its application to dental treatment," *Syst. Comput. Japan*, vol. 38, no. 5, pp. 80–89, May 2007.

- [87] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active Shape Models - Their Training and Application," *Comput. Vis. image ...*, vol. 61, no. 1, pp. 38–59, 1995.
- [88] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The subspace Gaussian mixture model—A structured model for speech recognition," *Comput. Speech Lang.*, vol. 25, no. 2, pp. 404–439, 2011.
- [89] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK Book (for HTK Version 3.4)," *Cambridge University*, vol. 2, no. 2. Cambridge University Engineering Department, 2006.
- [90] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1988, pp. 19–25.
- [91] C. Bregler, H. Hild, S. Manke, and A. Waibel, "Improving connected letter recognition by lipreading," in *Proceedings of the 1993 IEEE international conference on Acoustics, speech, and signal processing*, 1993, pp. 557–560.
- [92] D. G. Stork, G. Wolff, and E. Levine, "Neural network lipreading system for improved speech recognition," in *International Joint Conference on Neural Networks (IJCNN 1992)*, 1992, pp. 289 – 295.
- [93] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1248–1259, Jan. 2002.
- [94] P. Yin, I. Essa, and J. M. Rehg, "Asymmetrically boosted hmm for speech reading," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2004 (CVPR 2004)*, 2004, vol. 2, pp. 755–761.
- [95] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.
- [96] H. Liu, "Study on Lipreading Recognition Based on Computer Vision," in *2nd International Conference on Information Engineering and Computer Science*, 2010, vol. 2, no. 1, pp. 1–4.
- [97] M. Müller, *Information Retrieval for Music and Motion*. Springer, 2007, p. 318.

- [98] H. Li and M. Greenspan, "Model-based segmentation and recognition of dynamic gestures in continuous video streams," *Pattern Recognit.*, vol. 44, no. 8, pp. 1614–1628, Aug. 2011.
- [99] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *ICASSP 1989. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989, pp. 532–535.
- [100] IBM, "IBM SPSS Statistics," 2014. [Online]. Available: <http://www-01.ibm.com/software/analytics/spss/products/statistics/index.html>.
- [101] M. Gurban and J.-P. Thiran, "Information theoretic feature extraction for audio-visual speech recognition," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4765–4776, 2009.
- [102] S. W. Chin, K. P. Seng, and L.-M. Ang, "Audio-Visual Speech Processing for Human Computer Interaction," in *Advances in Robotics and Virtual Reality*, vol. 26, Springer Berlin Heidelberg, 2012, pp. 135–165.
- [103] R. E. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton Univ. Press, 1961.
- [104] C. Chatfield and A. J. Collins, *Introduction to Multivariate Analysis*. London, United Kingdom: Chapman and Hall, 1991.
- [105] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [106] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
- [107] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 14, no. 6, 2006.
- [108] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995, vol. 1, pp. 153–156.
- [109] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, 2001.
- [110] E. Nemer, R. Goubran, and S. Mahmoud, "SNR estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Process. Lett.*, vol. 6, no. 7, pp. 171–174, 1999.

- [111] T. Pyzdek, *The Six Sigma handbook: a complete guide for green belts, black belts, and managers at all levels*. McGraw-Hill, 2003.
- [112] W. L. Winston, *Microsoft Excel 2010 Data Analysis and Business Modeling*. Microsoft Press, 2011, p. 700.
- [113] Donald J. Wheeler, *Advanced Topics in Statistical Process Control*. SPC Press, 2004.
- [114] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust.*, vol. 37, no. 11, 1989.
- [115] Sony, *AVCHD Introduction Handbook*, 2nd Editio. Japan: Sony Corporation, 2009, p. 18.
- [116] "Power Director 12," 2014. [Online]. Available: http://www.cyberlink.com/products/powerdirector-ultra/features_en_GB.html.
- [117] D. Mazzoni and R. Dannenberg, "Audacity: Free audio editor and recorder," 2014. [Online]. Available: <http://audacity.sourceforge.net>.
- [118] Sony Corporation, "HXR-MC2000E." [Online]. Available: <http://www.sony.co.uk/pro/product/broadcast-products-camcorders-nxcam-avchd/hxr-mc2000e/features/#features>. [Accessed: 14-Oct-2014].
- [119] R. Lippmann, E. Martin, and D. Paul, "Multi-style training for robust isolated-word speech recognition," *ICASSP '87. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 12, 1987.
- [120] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition using hidden Markov models," *IEEE Trans. Acoust.*, vol. 37, no. 8, 1989.
- [121] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. Acoust.*, vol. 38, no. 4, pp. 599–609, 1990.
- [122] L. Bahl, R. Bakis, P. Cohen, A. Cole, F. Jelinek, B. Lewis, and R. Mercer, "Further results on the recognition of a continuously read natural corpus," *ICASSP '80. IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 5, 1980.
- [123] R. Schwartz, Y. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved hidden Markov modeling of phonemes for continuous speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1984, vol. 9, pp. 21–24.
- [124] ABC, "ABC," *IET Image Process.*

-
- [125] M. Iseli, Y.-L. S. Y.-L. Shue, and A. Alwan, "Age-and Gender-Dependent Analysis of Voice Source Characteristics," *2006 IEEE Int. Conf. Acoust. Speech Signal Process. Proc.*, vol. 1, 2006.
- [126] Microsoft, "Kinect for Windows," 2014. [Online]. Available: <http://www.microsoft.com/en-us/kinectforwindows/>. [Accessed: 15-Oct-2014].
- [127] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimed.*, vol. 19, no. 2, pp. 4–10, 2012.
- [128] N. Smolyanskiy, C. Huitema, L. Liang, and S. E. Anderson, "Real-time 3D face tracking based on active appearance model constrained by depth data," *Image Vis. Comput.*, vol. 32, no. 11, pp. 860–869, Nov. 2014.
- [129] S. H. Lee, M. K. Sohn, D. J. Kim, B. Kim, and H. Kim, "Smart TV interaction system using face and hand gesture recognition," in *Digest of Technical Papers - IEEE International Conference on Consumer Electronics*, 2013, pp. 173–174.
- [130] P. Dalka and A. Czyzewski, "Lip movement and gesture recognition for a multimodal human-computer interface," in *Proceedings of the International Multiconference on Computer Science and Information Technology, IMCSIT '09*, 2009, vol. 4, pp. 451–455.

Appendix A – Statistical analysis based on McNemar test

In this section, the details analysis of statistical results based on McNemar's test will be described. The results presented here were generated using IBM SPSS Statistics version 22.

A.1 McNemar's test between TP-MDTW(Model 4) and DTW

```
*Nonparametric Tests: Related Samples.
      NPTESTS
/RELATED TEST(DTW TP_MDTW_Model4) MCNEMAR(SUCCESS=FIRST)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.001 CILEVEL=99.9.
```

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of different values across DTW and TP_MDTW_Model4 are equally likely.	Related-Samples McNemar Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .00.

A.2 McNemar's test between TP-MDTW(Model 4) and HMM

```
*Nonparametric Tests: Related Samples.
      NPTESTS
/RELATED TEST(HMM TP_MDTW_Model4) MCNEMAR(SUCCESS=FIRST)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.001 CILEVEL=99.9.
```

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of different values across HMM and TP_MDTW_Model4 are equally likely.	Related-Samples McNemar Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .00.

A.3 McNemar's test between TP-MDTW(Model 4) and DCT

```
*Nonparametric Tests: Related Samples.
      NPTESTS
/RELATED TEST(DCT TP_MDTW_Model4) MCNEMAR(SUCCESS=FIRST)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.001 CILEVEL=99.9.
```

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of different values across DCT and TP_MDTW_Model4 are equally likely.	Related-Samples McNemar Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .00.

A.4 McNemar's test between TP-MDTW(Model 4) and OF

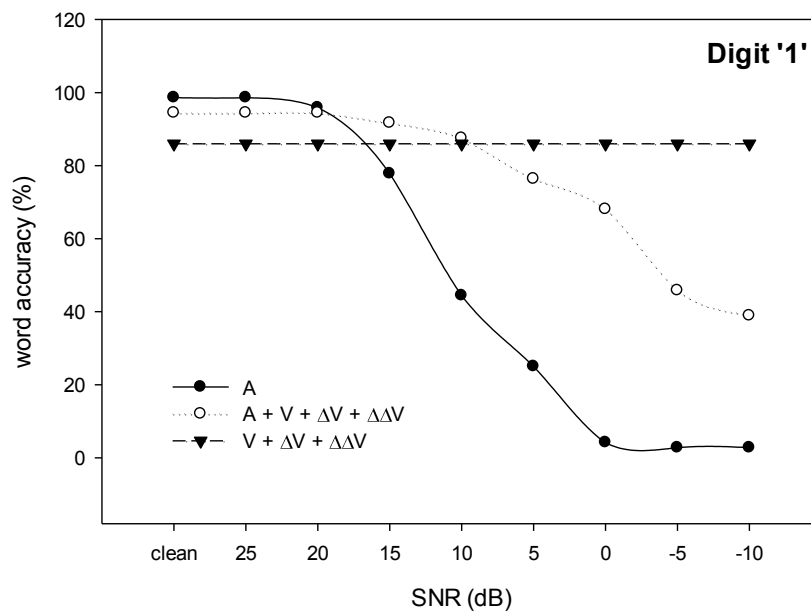
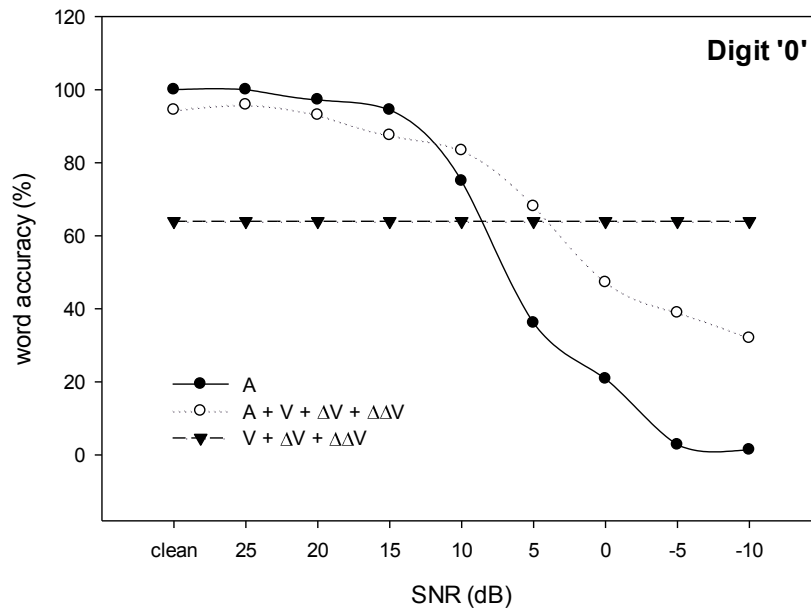
```
*Nonparametric Tests: Related Samples.
  NPTESTS
/RELATED TEST(OF TP_MDTW_Model4) MCNEMAR(SUCCESS=FIRST)
/MISSING SCOPE=ANALYSIS USERMISSING=EXCLUDE
/CRITERIA ALPHA=0.001 CILEVEL=99.9.
```

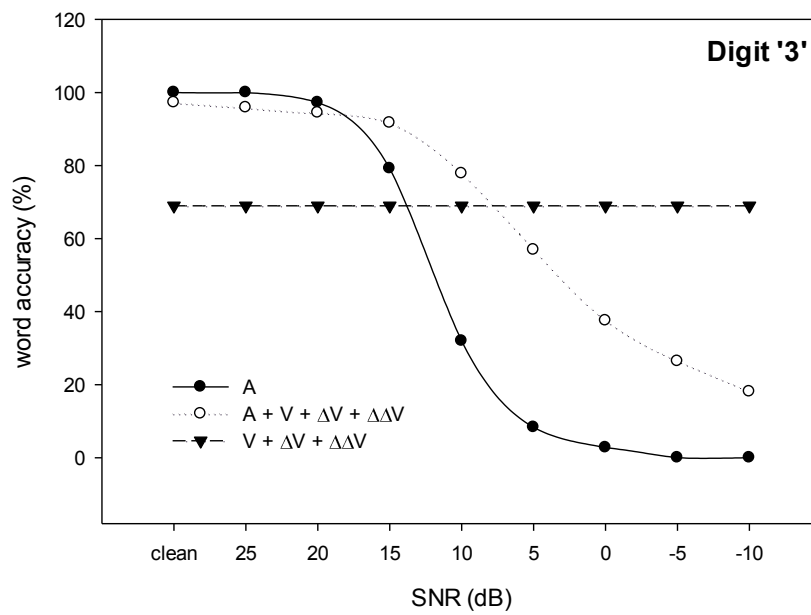
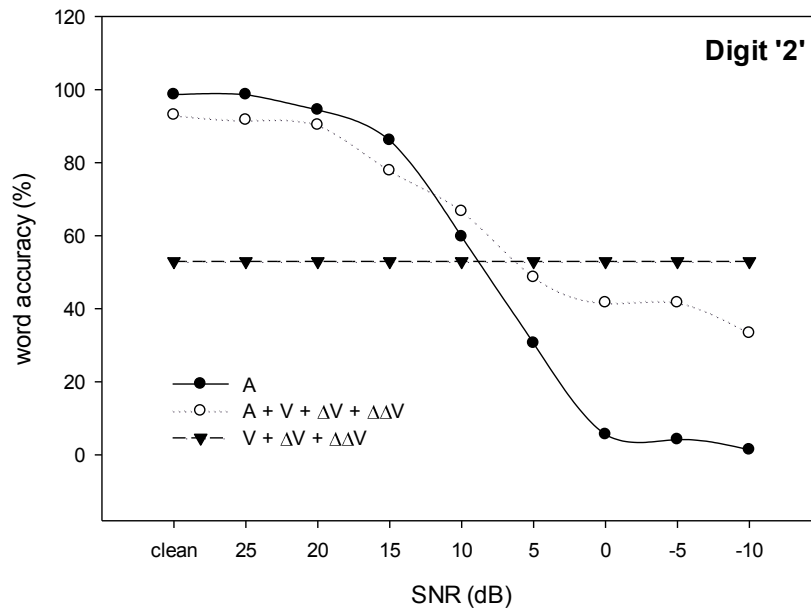
Hypothesis Test Summary

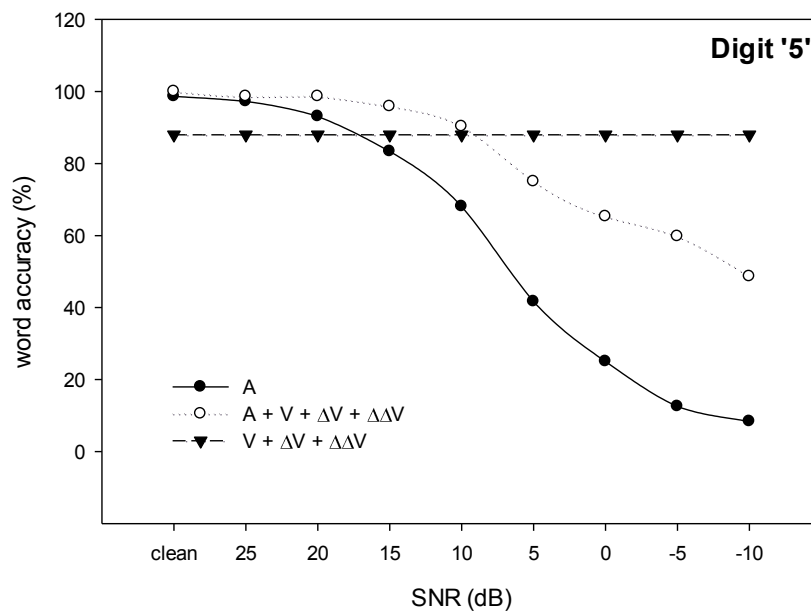
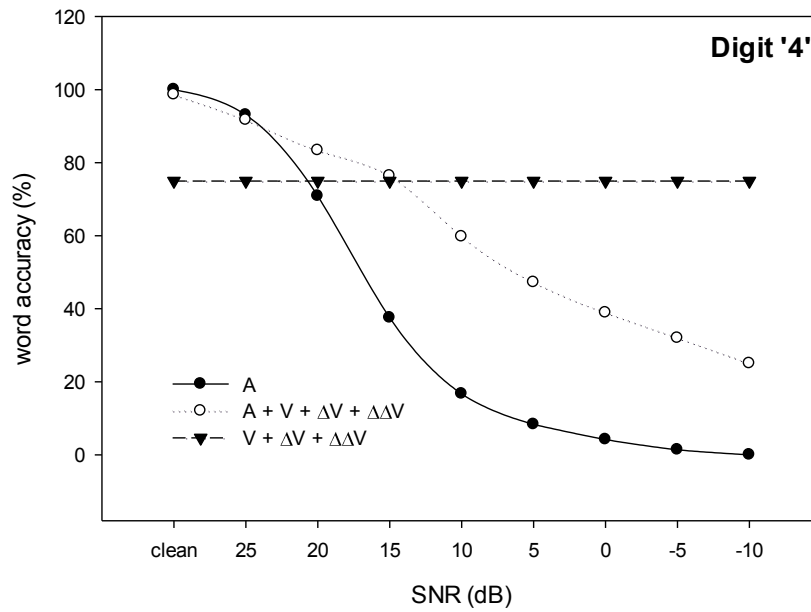
	Null Hypothesis	Test	Sig.	Decision
1	The distributions of different values across OF and TP_MDTW_Model4 are equally likely.	Related-Samples McNemar Test	.000	Reject the null hypothesis.

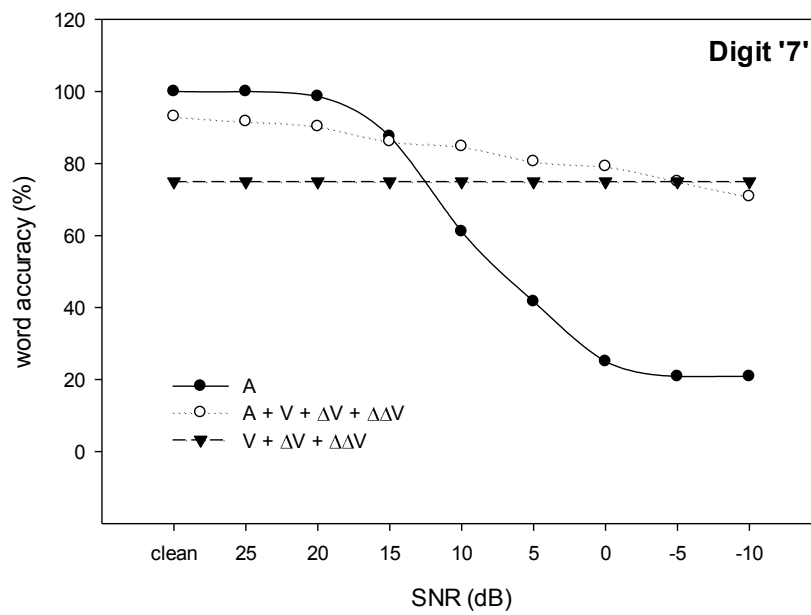
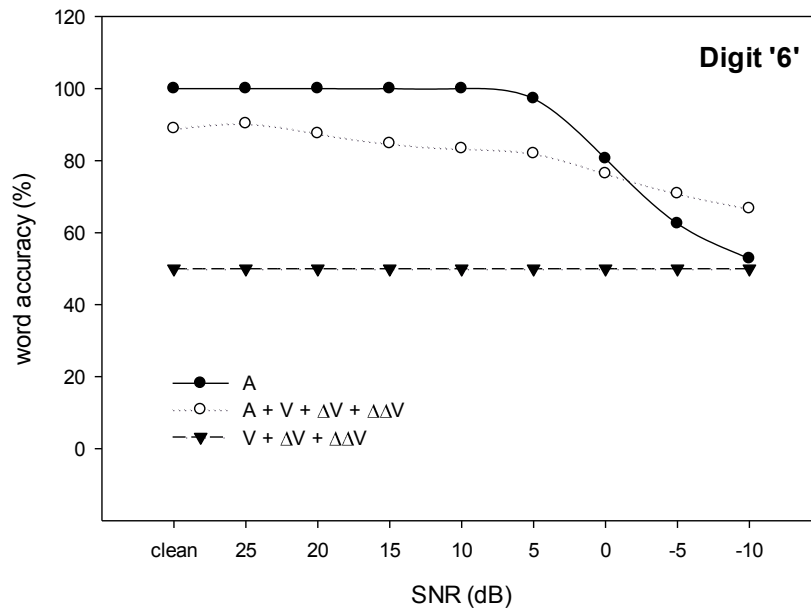
Asymptotic significances are displayed. The significance level is .00.

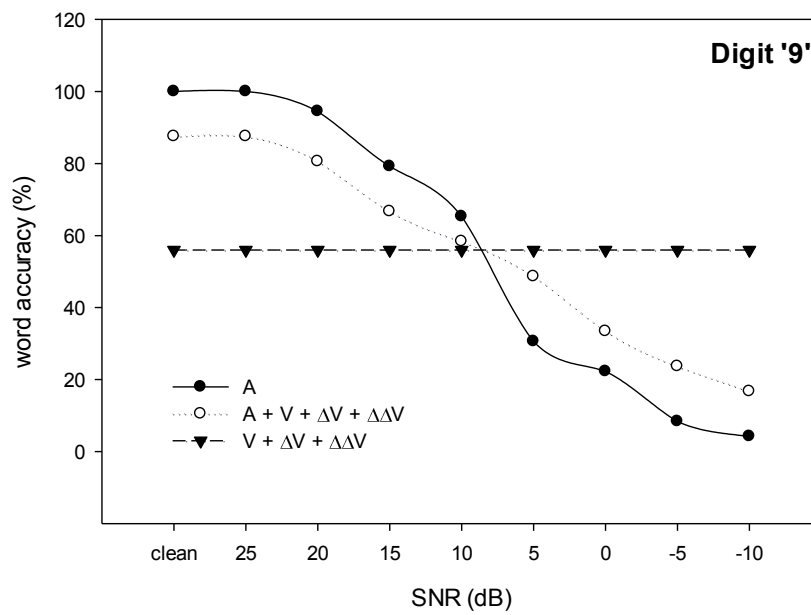
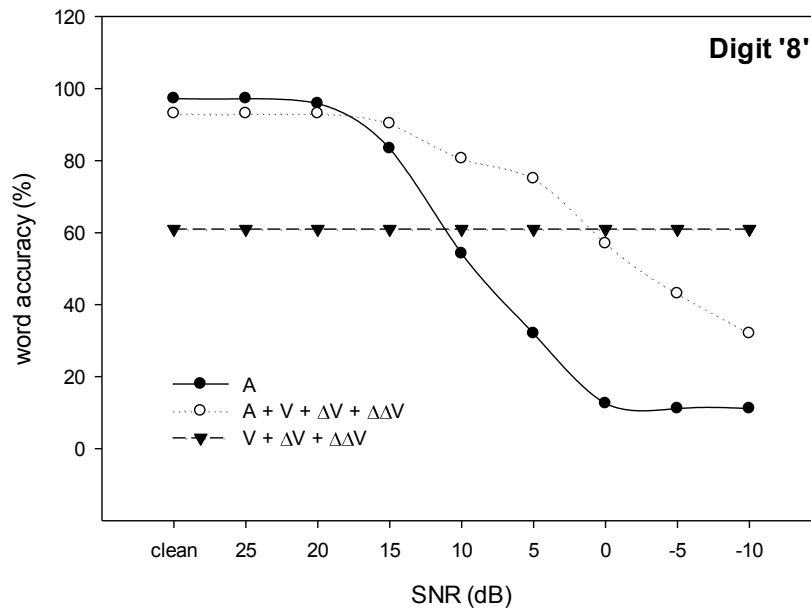
Appendix B – Digit performance in early integration simulated with white noise



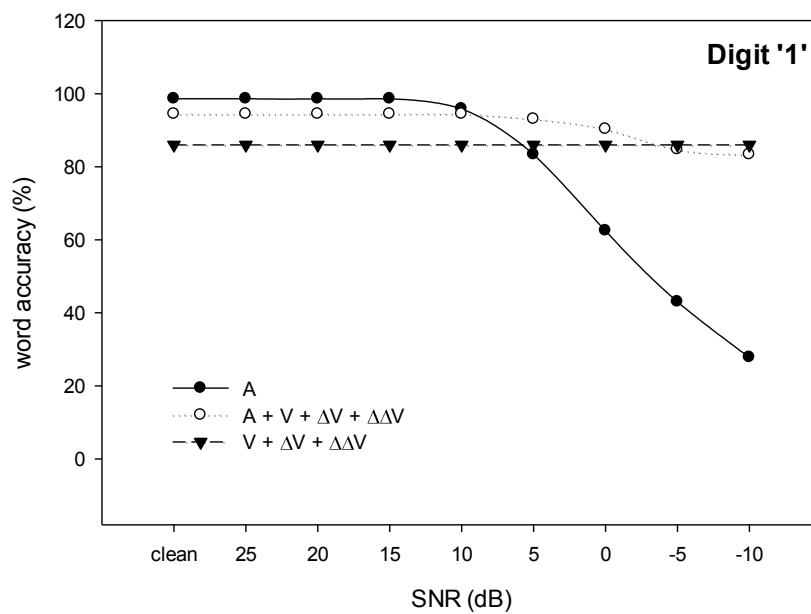
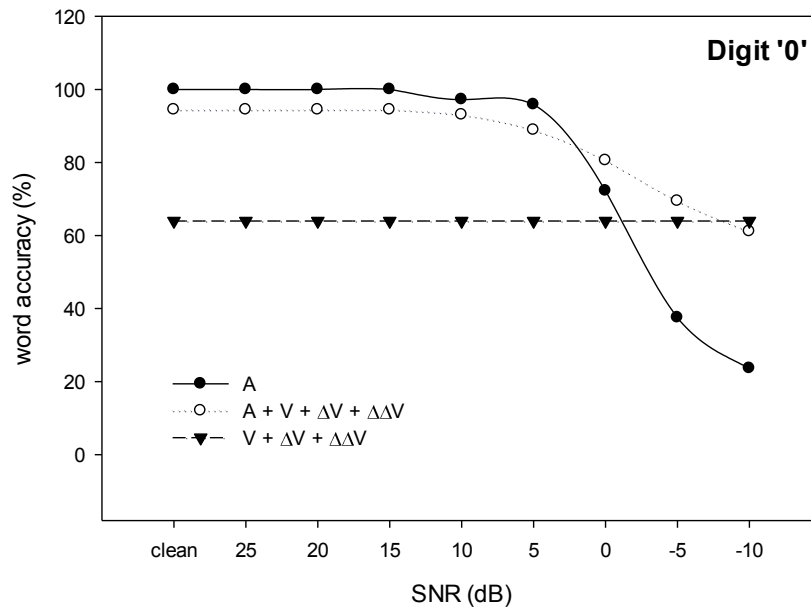


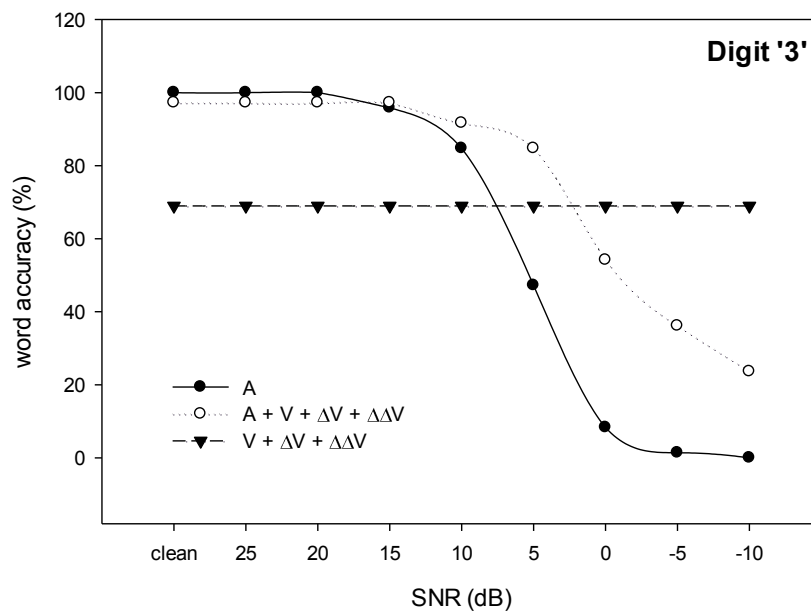
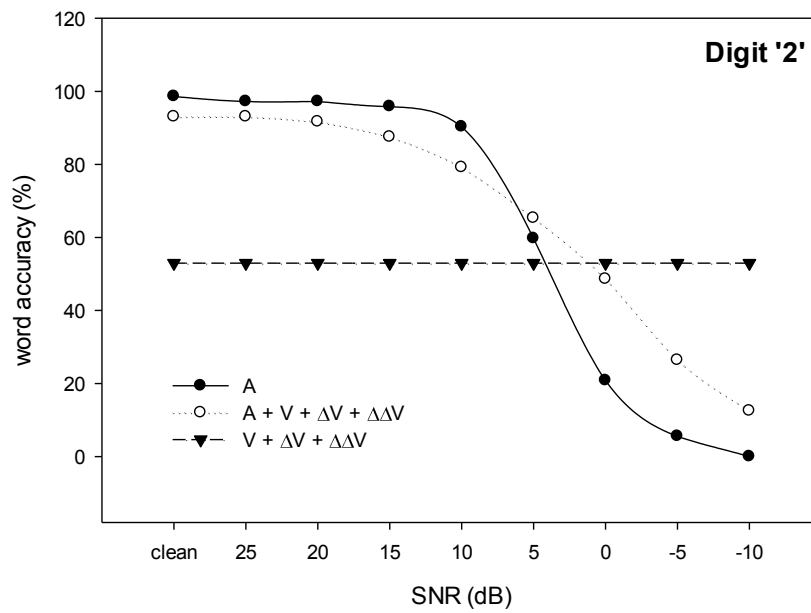


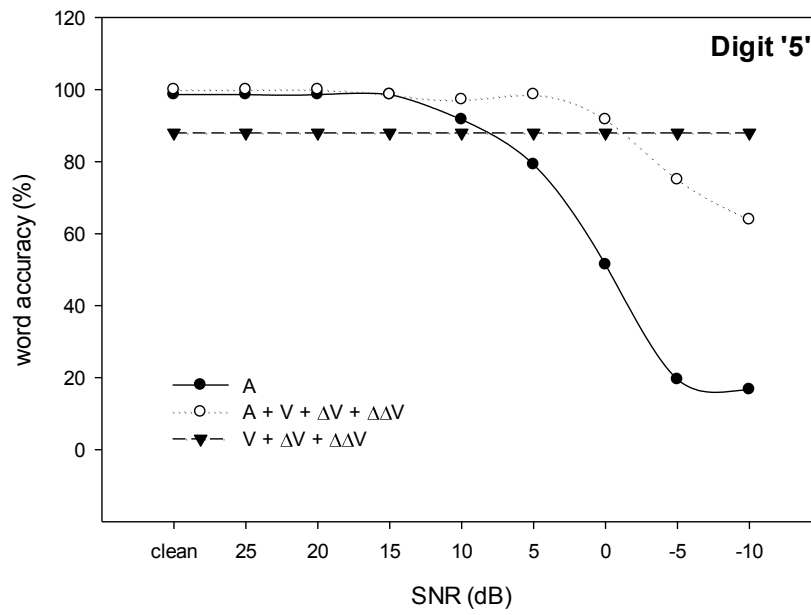
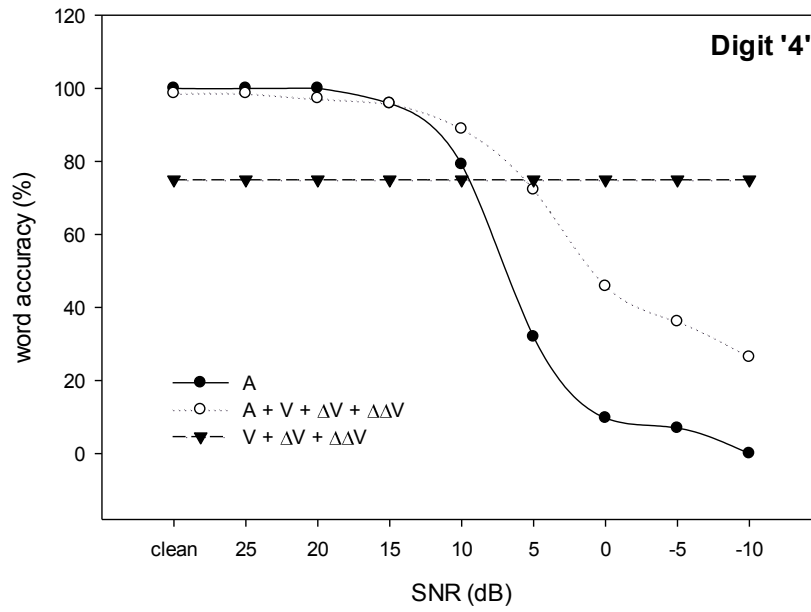


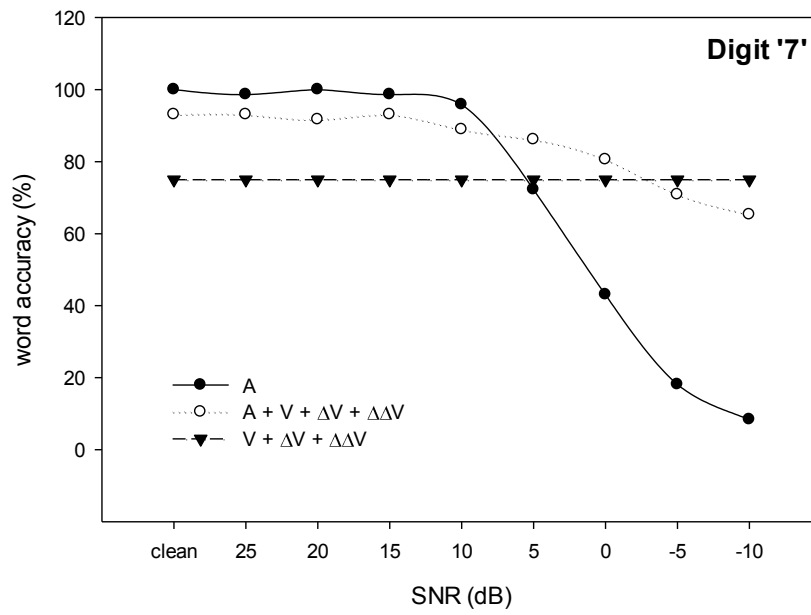
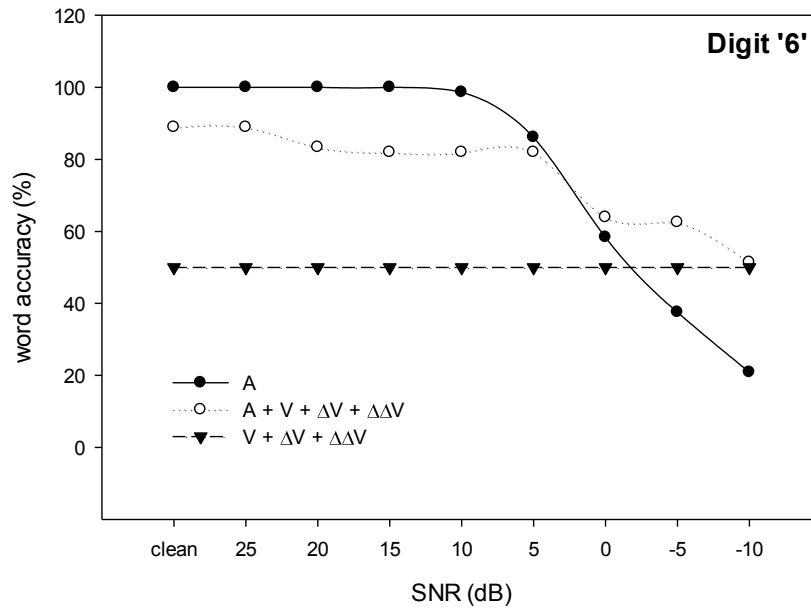


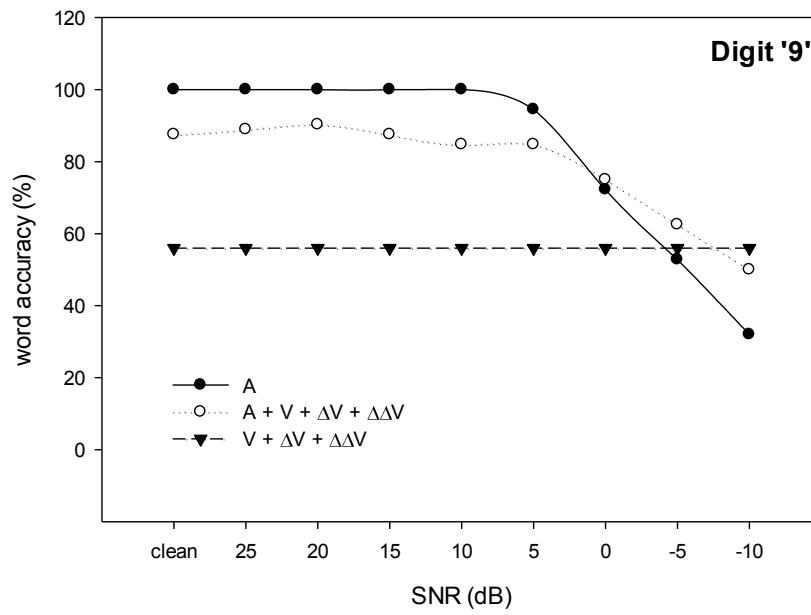
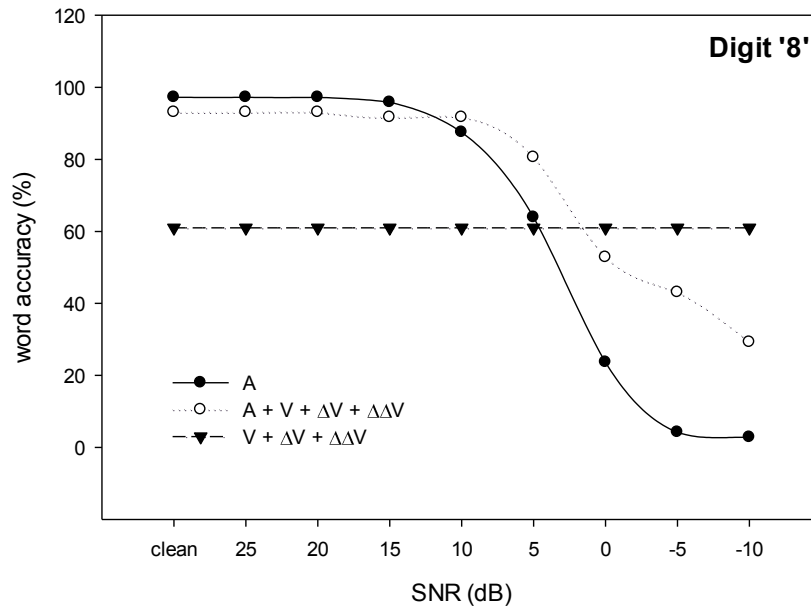
Appendix C – Digit performance in early integration simulated with babble noise












Appendix D – Human participant documents

<p>Ethics Approvals (Human Participants) Sub-Committee</p>	
<p>Research Proposal for Studies Involving Human Participants</p>	
<p>Project Details</p>	
<p>1. Project Title: Building a Data Corpus for Audio-Visual Speech Recognition</p>	
<p>2. Aims and objectives of the study The aim is to improve speech recognition in noisy environments by incorporating lip movement information, with the following objectives.</p> <ul style="list-style-type: none">- Investigate the extent to which speech recognition can be performed using visual information- Establish relationships between lip movements and audible speech signals- Determine whether a combination of visual and audio information can improve speech recognition	
<p>3. Lay summary of the study The study will record the audio and video of participants while they are speaking English sentences (e.g. "She had your dark suit in greasy wash water all year"). The recordings of audio and video will be used to form a database for each sentence which will then be used test the performance of new automatic audio-visual speech recognition systems. There is currently a shortage of such training databases as the field is relatively immature. The work has potential applications in all aspects of human-computer interaction as well as security applications.</p>	
<p>4. Start date of study: 13th May 2013</p>	
<p>5. End date of study: 30th June 2014</p>	
<p>6. Duration of the study: 13 months</p>	
<p>7. Start date for data-collection: 13th May 2013</p>	
<p>Note: Data collection should not commence before final ethical approval is confirmed.</p>	
<p>8. Location of the study: School of Electronic, Electrical and Systems Engineering, Loughborough University.</p>	
<p>9. Reasons for undertaking the study (e.g. contract, student research): Student Research (PhD)</p>	

10. Do any of the researchers stand to gain from a particular conclusion of the research study?	Yes	No
If Yes , how do the researchers stand to gain?		
Applicant Details		
11. Name of Researcher (applicant): Mohd Zamri bin Ibrahim		
12. Status: Postgraduate Research Student		
13. Email address: m.z.ibrahim@lboro.ac.uk		
14. Contact address: Room W.269, School of Electronic, Electrical and Systems Engineering, Loughborough University, LE11 3TU, UK		
15. Telephone number: 01509 227050		
All other researchers (including supervisors if applicant is a student)		
16. Name(s): Dr David Mulvaney		
17. Status(es): Supervisor		
18. Email address(es): d.j.mulvaney@lboro.ac.uk		
19a. Contact Address(es): Room W.268, School of Electronic, Electrical and Systems Engineering, Loughborough University, LE11 3TU, UK		
19b. Telephone number(s): 01509 227042		
20. Experience of all investigators in the methods to be used in this study		
Mohd Zamri – Execution of research and experiments using data gathered by third parties, principally audio speech recordings used for automatic speaker recognition.		
David Mulvaney – Collection and analysis of data from video, image and ultrasonic sensors for the purposes of object identification and avoidance in robotic applications.		
Note: Please ensure the experience of all investigators is included in this section.		
Participant Information		
21. Number of participants to be recruited: 20 - 40 people		
22. Details of participants (age, gender, special interests etc.): 18 years old and above, any gender and no special interests.		
23. How will participants be selected?		
- Inclusion criteria: 18 years old and above.		
- Exclusion criteria: There are no additional exclusion criteria.		

<p>24. How will participants be recruited and approached?</p> <ul style="list-style-type: none"> - Primary: Word of mouth. - Secondary: Emails to students containing 'participant information sheet' and cover note. <p>25. Please state the demand on participants' time:</p> <ul style="list-style-type: none"> - Approx. 30 minutes <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: center;">Time (minutes)</th> <th style="text-align: center;">Task</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">5</td> <td>Prior briefing and signing consent forms</td> </tr> <tr> <td style="text-align: center;">15</td> <td>Recording audio and video of participant when pronouncing sentences</td> </tr> <tr> <td style="text-align: center;">10</td> <td>De-briefing and signing release forms</td> </tr> </tbody> </table>			Time (minutes)	Task	5	Prior briefing and signing consent forms	15	Recording audio and video of participant when pronouncing sentences	10	De-briefing and signing release forms
Time (minutes)	Task									
5	Prior briefing and signing consent forms									
15	Recording audio and video of participant when pronouncing sentences									
10	De-briefing and signing release forms									
26. Will control participants be used?		Yes	No							
<p>If No, please go to Question 30. If Yes, please answer the questions below.</p>										
<p>27. How will control participants be selected? Note: Include the inclusion/exclusion criteria to be used.</p> <p>28. How will control participants be recruited and approached? Note: If an advertisement or forum post is to be used, please include this in your application to the Sub-Committee.</p> <p>29. Please state the demand on control participants' time: Note: Where possible, include a breakdown of how long each part of the study will take, as well as a total time demand.</p> <p>30. Please provide procedures for the chaperoning and supervision of the participants during the study. The primary researcher will chaperone participants into the audio visual lab and remain within the lab during the recording.</p> <p>31. Possible risks, discomforts and/or distress to participants. There is no risk or discomfort to the participant.</p> <p>32. Please provide details of any payments to be made to the participants. Light refreshment may be provided during or after the study.</p> <p>Researcher Safety</p>										
33. Are there any potential risks to the researchers in this study?		Yes	No							
<p>If Yes, please answer the following questions:</p> <p>34. What are the potential risks to the researchers?</p> <p>35. What measures have been put in place to address these risks?</p>										

Study details		
36. Brief outline of study design and methodology: The participants will first be briefed on the requirements of the experiment and be asked to sign the consent form. Participants will be required to sit in the audio-visual room (layout attached – Appendix 2) and view a monitor on which will be displayed a series of sentences that the participant will be asked to read out loud. After the recording, participants will be debriefed on the aims and objectives of the study and asked to sign the release form.		
37. Measurements to be taken: Audio and video recordings will be taken when the participant speaks the sentences. The video coverage will be from the shoulder to the top of the head of the participant (examples are shown in Appendix 3).		
38. Please indicate whether the proposed study:		
38a. Involves taking bodily samples	Yes	No
38b. Involves procedures which are physically invasive (including the collection of body secretions by physically invasive methods)	Yes	No
38c. Is designed to be challenging:		
Physically	Yes	No
Psychologically	Yes	No
38d. Involves procedures that are likely to cause:		
Physical distress to participants	Yes	No
Psychological distress to participants	Yes	No
Social distress to participants	Yes	No
Emotional distress to participants	Yes	No
38e. Involves intake of compounds additional to daily diet, or other dietary manipulation/supplementation	Yes	No
38f. Involves pharmaceutical drugs (Please refer to published guidelines)	Yes	No
38g. Involves testing new equipment	Yes	No
38h. Involves procedures which may cause embarrassment to participants	Yes	No
38i. Involves collection of personal and/or potentially sensitive data	Yes	No
38j. Involves use of radiation (Please refer to published guidelines and contact the University's Radiological Protection Officer before beginning any study which exposes participants to ionising radiation)	Yes	No
38k. Involves use of hazardous materials (Please refer to published guidelines)	Yes	No
38l. Assists/alters the process of conception in any way	Yes	No
38m. Involves methods of contraception	Yes	No
38n. Involves genetic engineering	Yes	No
If Yes, please give specific details of each of the procedures to be used and arrangements to deal with adverse effects: Consent		
39. Is written consent to be obtained from participants?	Yes	No
If yes , please attach a copy of the consent form to be used. If no , please justify.		

40. Will any of the participants be from one of the following vulnerable groups?		
40a. Children under 18 years of age	Yes	No
40b. Persons incapable of making an informed decision for themselves	Yes	No
40c. Prisoners/other detained persons	Yes	No
40d. Other vulnerable groups Please specify:	Yes	No
<p>If Yes, to any of the above, please answer the following questions:</p> <p>41. What special arrangements have been made to deal with the issues of consent?</p> <p>42. Have the researchers obtained necessary police registration/clearance? Note: Please provide details or indicate why this is not applicable to your study.</p> <p>Withdrawal</p> <p>43. How will participants be informed of their right to withdraw from the study? Via the 'participant information sheet', the consent form, the release form and reiterated verbally.</p> <p>44. How will participants be informed of the issues with withdrawing their data once this has been aggregated in the study? Via the 'participant information sheet', the consent form, the release form and reiterated verbally.</p> <p>Storage and Security of Data</p>		
45. Will the investigation include the use of any of the following:		
Observation of participants	Yes	No
Audio recording	Yes	No
Video recording	Yes	No
<p>If Yes, to any, please provide details of how the recording will be stored, where specifically the recording will be stored, when the recordings will be destroyed and how confidentiality will be ensured?</p> <p>Recordings will be stored in their original form on a secure hard-drive on a computer (with no remote access) within the audio-visual lab (W2.69). The computer will be accessible only to the 'primary investigator' and 'supervisor'. A back-up of the data will be made and encrypted on a removable hard drive to be stored in the office of the supervisor.</p> <p>If the primary investigator or the supervisor were to leave the university to continue research elsewhere, a copy of the collected data may be made and transferred to another institution provided the supervisor is satisfied they will continue to be used for research purposes. It is the intention that the primary investigator and the supervisor will continue to have access to the data in their future research activities, even if this is longer than 10 years, but will destroy the data on termination of their activities in the research area.</p> <p>Subject to a suitable memorandum of understanding (MOU), the data will be made available to colleagues at other research institutions. The MOU will stipulate that the data be held at the receiving research institutions in circumstances similar to those pertaining at Loughborough.</p>		

<p>46. What steps will be taken to safeguard anonymity of participants/confidentiality of personal data? Each participant will be allocated a unique identification number immediately following the experiment to which all personal information will be associated and independently securely stored.</p> <p>47. Please give details of where the data collected will be stored, and how the collection and storage of the data complies with the Data Protection Act 1998? - Only relevant/necessary data will be collected. - Data will be anonymous and stored securely in original form, prior to processing (accessible by only investigator and supervisor). - Participants will be advised of their right to withdraw from the study, even if they wish to do after publication. - Irrelevant data will be disposed of as soon as possible.</p> <p>48. If human tissue samples are to be taken, please give details of, and the timeframe for, the disposal of the tissue. Note: Please also ensure that this information is included on the Participant Information Sheet.</p> <p>Sponsorship and Insurance Cover</p>		
49a. Is the study being sponsored?	Yes	No
If Yes , please state source of funds including a contact name and address for the sponsor:		
If No , please go to question c.		
49b. Is the study to be covered by the sponsors insurance?	Yes	No
If No , please confirm who will be insuring the study:		
49c. Is the study to be covered by the University's insurance?	Yes	No
If No , please confirm who will be insuring the study:		
<p>Insurance Cover</p> <p>Note: It is the responsibility of investigators to ensure that there is appropriate insurance cover for the procedure/technique.</p> <p>The University maintains in force a Public Liability Policy, which indemnifies it against its legal liability for accidental injury to persons (other than its employees) and for accidental damage to the property of others. Any unavoidable injury or damage therefore falls outside the scope of the policy.</p>		
50. Will any part of the study result in unavoidable injury or damage to participants or property?	Yes	No
If Yes , please detail the alternative or additional insurance cover arrangements and include the supporting documentation in this application.		
The University Insurance relates to claims arising out of all normal activities of the University, but Insurers require to be notified of anything of an unusual nature.		

51. Is the study classed as normal activity?	Yes	No
<p>If No, please check with the University Insurance Officer that the policy will cover the activity. If the activity falls outside the scope of the policy, please detail the alternative or additional insurance cover arrangements and include the supporting documentation in this application.</p>		
<p>Declaration</p>		
<p>I have read the University's Code of Practice on Investigations on Human Participants and have completed this application. I confirm that the above named investigation complies with published codes of conduct, ethical principles and guidelines of professional bodies associated with my research discipline.</p>		
<p>I agree to provide the Ethics Approvals (Human Participants) Sub-Committee with appropriate feedback upon completion of my study.</p>		
<p>Signature of applicant:</p>		
<p>Signature of Supervisor (if applicable):</p>		
<p>Signature of Head of School/Department:</p>		
<p>Date:</p>		
<p>Note: Please check to ensure you have attached all necessary documents to your application.</p>		



Building a Data Corpus for Audio-Visual Speech Recognition

INFORMED CONSENT FORM
(to be completed after Participant Information Sheet has been read)

The purpose and details of this study have been explained to me. I understand that this study is designed to further scientific knowledge and that all procedures have been approved by the Loughborough University Ethics Approvals (Human Participants) Sub-Committee.

I have read and understood the information sheet and this consent form.

I have had an opportunity to ask questions about my participation.

I understand that I am under no obligation to take part in the study.

I understand that I have the right to withdraw from this study at any stage for any reason, and that I will not be required to explain my reasons for withdrawing.

I understand that all the information and data I provide will be treated in strict confidence and will be kept anonymous and confidential to researchers at Loughborough and collaborating institutions, unless (under the statutory obligations of the agencies which the researchers are working with), it is judged that confidentiality will have to be breached for the safety of the participant or others.

I agree to participate in this study.

Your name _____

Your signature _____

Signature of investigator _____

Date _____



Building a Data Corpus for Audio-Visual Speech Recognition

RELEASE FORM

I confirm that the true purpose and details of this study have been explained to me and can be summarized as

‘The collection of video and audio recordings of me while speaking specific sentences designed for the purpose of developing a data corpus for audio-visual speech recognition.’

I understand that this study is designed to further scientific knowledge and that all procedures have been approved by the Loughborough University Ethics Approvals (Human Participants) Sub-Committee.

I have read and understood the information sheet and this release form.

I have had the opportunity to ask questions about my participation.

I understand that I am under no obligation to take part in the study.

I understand that I have the right to withdraw from this study at any stage for any reason, and that I will not be required to explain my reasons for withdrawing.

I understand that all the information I provide will be treated in strict confidence and will remain confidential to the researchers unless (under the statutory obligations of the agencies with which the researchers are working), it is judged that confidentiality will have to be breached for the safety of the participant or others.

I agree that the video and audio data collected from me will be kept indefinitely in a database for the sole purpose of academic research.

I agree that individual images of me from the database may be published in academic literature.

Your name _____

Your signature _____

Signature of investigator _____

Date _____

Appendix E – LUNA-V sentence selection

In this section, the coverage of the phones and visemes in LUNA-V data corpus will be described. Table D.1 shows the list of the sentences found in the data corpus that contains several sentences from the well-known TIMIT data corpus and the English digits ‘zero’ to ‘nine’. The phone coverage for these sentences can be seen in Table D.2. Please note that the 39 phones shown in Table D.2. are commonly used in speech processing field [114]. It can be seen that all the phones were chosen except phone ‘dx’.

Table E.1 The sentences collected for the LUNA-V corpus

sentence	content
S1	She had your dark suit in greasy wash water all year
S2	Each untimely income loss coincided with the breakdown of a heating system part
S3	The easygoing zoologist relaxed throughout the voyage
S4	The same shelter could be built into an embankment or below ground level
S5	zero, one, two, three, four, five, six, seven, eight, nine

Table E.2 Phone coverage

	Phone		sum	Sentence				
	S1	S2		S3	S4	S5		
1	iy		10	3	3	2	1	1
2	ih	ix	17	2	6	4	3	2
3	eh		3	0	0	0	2	1
4	ey		3	0	1	0	1	1
5	ae		4	1	0	1	2	0
6	aa	ao	9	4	2	1	1	1
7	aw		3	0	1	1	1	0
8	ay		4	0	2	0	0	2
9	ah	ax,ax-h	14	0	6	3	3	2
10	oy		1	0	0	1	0	0
11	ow		5	0	1	2	1	1
12	uh		2	1	0	0	1	0
13	uw	ux	4	1	0	1	1	1
14	er	axr	2	1	0	0	1	0
15	jh		2	0	0	2	0	0
16	ch		1	0	1	0	0	0
17	b		5	0	1	0	4	0
18	d		7	2	3	0	2	0
19	g		3	1	0	1	1	0
20	p		1	0	1	0	0	0
21	t		15	2	4	3	4	2
22	k		8	1	3	1	2	1
23	dx		0	0	0	0	0	0
24	s		12	2	4	2	1	3
25	sh	zh	3	2	0	0	1	0
26	z		3	0	0	2	0	1
27	f		2	0	0	0	0	2
28	th		2	0	0	1	0	1
29	v		5	0	1	1	1	2
30	dh		5	0	2	2	1	0
31	m	em	6	0	3	0	3	0
32	n	en,nx	13	1	4	0	4	4
33	ng	eng	3	0	1	1	1	0
34	l	el	10	1	2	2	5	0
35	r		13	4	2	2	2	3
36	w		4	2	1	0	0	1
37	y		2	2	0	0	0	0
38	hh	hv	2	1	1	0	0	0
39	silence							

The visemes coverage for sentences found in LUNA-V data corpus can be seen in Table D.3. This viseme map was derived by [65] which composed of 12 classes and silent class. It can be seen that the sentences in LUNA-V data corpus covers all the viseme class.

Table E.3 Viseme coverage

Code	Shape	List	sum
V1	Lip-rounding	/ao/ /ah/ /aa/ /er/ /oy/ /aw/ /hh/	31
V2		/uw/ /uh/ /ow/	11
V3		/ae/ /eh/ /ey/ /ay/	14
V4		/ih/ /iy/ /ax/	41
A	Alveolar-semivowels	/l/ /el/ /r/ /y/	25
B	Alveolar-fricatives	/s/ /z/	15
C		/t/ /d/ /n/ /en/	35
D		/sh/ /zh/ /ch/ /jh/	6
E		/p/ /b/ /m/	12
F		/th/ /dh/	7
G		/f/ /v/	7
H		/ng/ /k/ /g/ /w/	18
S	Silence		

Appendix F – Confusion matrix for speech recognition

Table F.1 Confusion matrix for visual-only recognition

Actual	Predicted												
	the	same	shelter	could	be	built	into	an	embankment	or	below	ground	level
the	46.7			3.3			3.3						
same		83.3											
shelter		3.3	50.0			3.3		3.3				6.7	10.0
could	16.7		10.0	10.0				10.0		6.7	0.0	3.3	
be					46.7	3.3					10.0		3.3
built						66.7			3.3				
into				3.3			43.3	13.3					10.0
an	6.7						3.3	40.0					
embankment		6.7							73.3				20.0
or				16.7						10.0	3.3		
below						26.7				3.3	50.0		
ground												26.7	10.0
level		26.7										10.0	26.7

Note. Percentages of correct predicted have been pooled over participants and word contexts.

Table F.2 Confusion matrix for audio at clean environment

Actual	Predicted												
	the	same	shelter	could	be	built	into	an	embankment	or	below	ground	level
the	90.0				6.7								
same		93.3											
shelter			100										
could		6.7		16.7	3.3	73.3							
be	10.0	6.7			33.3								
built						93.3						6.7	
into		10.0					90.0						
an		40.0						40.0			13.3	3.3	
embankment						3.3			96.7				
or				20.0						10.0			
below		10.0				10.0					80.0		
ground						30.0					10.0	60.0	
level	6.7			3.3		3.3				3.3	26.7		56.7

Note. Percentages of correct predicted have been pooled over participants and word contexts.

Table F.3 Confusion matrix for combination of audio and visual at clean environment

Actual	Predicted												
	the	same	shelter	could	be	built	into	an	embankment	or	below	ground	level
the	93.3										3.3		
same		100											
shelter		3.3	93.3										3.3
could		3.3		40.0	50.0		3.3		3.3				
be					93.3	3.3				3.3			
built						100							
into						10.0	86.7						
an		3.3						93.3		3.3			
embankment		6.7				3.3			66.7				23.3
or				30.0						36.7			3.3
below						23.3					70.0		6.7
ground						23.3					3.3	73.3	
level		23.3											76.7

Note. Percentages of correct predicted have been pooled over participants and word contexts.

Table F.4 Confusion matrix for audio at 0 dB SNR using babble noise

Actual	Predicted												
	the	same	shelter	could	be	built	into	an	embankment	or	below	ground	level
the	0.0	10.0						26.7			23.3		30.0
same		56.7											43.3
shelter		3.3	0.0			6.7	3.3				26.7		60.0
could		13.3		3.3		6.7	10.0				26.7		40.0
be		46.7		3.3	0.0		6.7				23.3		13.3
built				3.3		3.3	3.3				26.7		63.3
into		13.3				13.3	0.0	13.3			13.3		46.7
an		3.3						66.7			10.0		13.3
embankment		3.3		6.7				26.7	3.3		13.3	3.3	43.3
or		3.3		6.7				13.3		0.0	13.3		63.3
below		13.3		3.3				10.0			30.0		40.0
ground				3.3		3.3		40.0			20.0	10.0	23.3
level		3.3						30.0			10.0		56.7

Note. Percentages of correct predicted have been pooled over participants and word contexts.

Table F.5 Confusion matrix for combination of audio and visual at 0 dB SNR using babble noise

Actual	Predicted												
	the	same	shelter	could	be	built	into	an	embankment	or	below	ground	level
the	46.7	3.3		6.7				13.3			3.3		23.3
same		100											
shelter		6.7	10.0	10.0		10.0		10.0			6.7	16.7	30.0
could	6.7			20.0		6.7		26.7			20.0	3.3	16.7
be					23.3	33.3					36.7		6.7
built						66.7			3.3		10.0		20.0
into		3.3		10.0		10.0	3.3	46.7			3.3		23.3
an		3.3						96.7					
embankment		23.3							20.0				56.7
or	3.3			3.3						0.0	40.0		53.3
below						6.7					86.7		6.7
ground						6.7						50.0	43.3
level		30.0											70.0

Note. Percentages of correct predicted have been pooled over participants and word contexts.

Table F.6 Confusion matrix for audio at 0 dB SNR using white noise

Actual	Predicted												
	the	same	shelter	could	be	built	into	an	embankment	or	below	ground	level
the	0.0	23.3			6.7	16.7		20.0			3.3		30.0
same		50.0				13.3		20.0					16.7
shelter		30.0	0.0			20.0		10.0					40.0
could		33.3		0.0		20.0		23.3					23.3
be		6.7			33.3	10.0		30.0			3.3		16.7
built		23.3				36.7		6.7					33.3
into		20.0			6.7	36.7	0.0	16.7			3.3		16.7
an		13.3				13.3		63.3					10.0
embankment		26.7				30.0		20.0	0.0			3.3	20.0
or		30.0				16.7		26.7		0.0	3.3		23.3
below		16.7				33.3		10.0			6.7		33.3
ground		20.0				30.0		30.0			3.3	0.0	16.7
level		20.0				36.7		20.0					23.3

Note. Percentages of correct predicted have been pooled over participants and word contexts.

Table F.7 Confusion matrix for combination of audio and visual at 0 dB SNR using white noise

Actual	Predicted												
	the	same	shelter	could	be	built	into	an	embankment	or	below	ground	level
the	6.7	30.0		3.3	10.0	10.0		23.3			6.7		10.0
same		70.0											30.0
shelter		23.3	10.0	3.3		33.3		6.7				6.7	16.7
could		33.3	3.3	3.3		26.7		30.0					3.3
be					46.7	43.3					6.7		3.3
built						93.3					3.3		3.3
into		30.0			3.3	26.7	0.0	33.3					6.7
an								100.0					
embankment		20.0							6.7				73.3
or		6.7				30.0				0.0	26.7		36.7
below						40.0					56.7		3.3
ground				3.3		10.0						30.0	56.7
level		13.3				3.3		3.3					80.0

Note. Percentages of correct predicted have been pooled over participants and word contexts.