A Motion Based Approach for Audio-Visual Automatic Speech Recognition

by

Nasir Ahmad

A doctoral thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Department of Electronic and Electrical Engineering

Loughborough University, United Kingdom

May 2011

To my parents and teachers

ABSTRACT

The research work presented in this thesis introduces novel approaches for both visual region of interest extraction and visual feature extraction for use in audio-visual automatic speech recognition. In particular, the speaker's movement that occurs during speech is used to isolate the mouth region in video sequences and motion-based features obtained from this region are used to provide new visual features for audio-visual automatic speech recognition. The mouth region extraction approach proposed in this work is shown to give superior performance compared with existing colour-based lip segmentation methods. The new features are obtained from three separate representations of motion in the region of interest, namely the difference in luminance between successive images, block matching based motion vectors and optical flow. The new visual features are found to improve visual-only and audio-visual speech recognition performance when compared with the commonly-used appearance feature-based methods.

In addition, a novel approach is proposed for visual feature extraction from either the discrete cosine transform or discrete wavelet transform representations of the mouth region of the speaker. In this work, the image transform is explored from a new viewpoint of data discrimination; in contrast to the more conventional data preservation viewpoint. The main findings of this work are that audio-visual automatic speech recognition systems using the new features extracted from the frequency bands selected according to their discriminatory abilities generally outperform those using features designed for data preservation.

To establish the noise robustness of the new features proposed in this work, their performance has been studied in presence of a range of different types of noise and at various signal-to-noise ratios. In these experiments, the audio-visual automatic speech recognition systems based on the new approaches were found to give superior performance both to audio-visual systems using appearance based features and to audio-only speech recognition systems.

LIST OF PUBLICATIONS

Ahmad, N., Mulvaney, D., Datta S., and Farooq, O. (2010), "Stream Weights Optimisation for Audio-Visual Automatic Speech Recognition", *Proceedings of National Symposium on Acoustics NSA-2010,* Pt. L. M. S. Govt P. G. College, Rishikesh, India, pp. 112-117.

Ahmad, N., Mulvaney, D., Datta S., and Farooq, O. (2009), "Dynamic Visual Features for Audio-Visual Automatic Speech Recognition", *Proceedings of National Symposium on Acoustics NSA-2009*, Research Centre Imarat, Andhra Pradesh, India, pp. 91-96.

Ahmad, N., Datta S., Mulvaney, D., and Farooq, O. (2008), "A comparison of visual features for audiovisual automatic speech recognition", *Proceedings of the 2nd joint conference of the Acoustical Society of America (ASA) and the European Acoustics Association (EAA), Acoustics'08*, Paris, pp. 6445-6449.

ACKNOWLEDGMENTS

I take this opportunity to formally thank those who have contributed to the accomplishment of this work.

First of all I am thankful to my supervisor Dr. David. J. Mulvaney for his thorough guidance, valuable suggestions and encouragement throughout this research. In particular his consistent assistance through weekly meetings has made it possible to achieve this goal. I am also thankful to my previous supervisor Dr. Sekharjit Datta, who has been participating in our weekly meeting voluntarily despite having no official responsibility for my research after his retirement from the University.

I would also like to offer my special gratitude to Dr. Omar Farooq who has been providing me with his valuable suggestions throughout this research.

I am also thankful to my parents, family members and especially my wife, Ayesha Tahir, for their moral support, encouragement and patience throughout this long journey. Particularly, my son Zeerak Nasir, who remained deprived of my presence in this early stage of his life.

GLOSSARY

AAM	Active Appearance Models
ADC	Analog to Digital Converter
ANN	Artificial Neural Network
ARPS	Adaptive Rood Pattern Search
ASM	Active Shape Models
ASR	Automatic Speech Recognition
AVASR	Audio-Visual Automatic Speech Recognition
CDHMM	Continuous Density HMM
CDI	Cumulative Difference Image
CWT	Continuous Wavelet Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DHMM	Discrete HMM
DI	Difference Image
DI	Direct Identification
DR	Dominant Recording
DTW	Dynamic Time Warping
DWT	Discrete Wavelet Transform
FFT	Fast Fourier Transform
FT	Fourier Transform
HCI	Human Computer Interaction
HMM	Hidden Markov Model
HTK	HMM Tool Kit
JPEG	Joint Photographic Expert Group

LDA	Linear Discriminant Analysis
LPC	Linear Predictive Coefficients
MAD	Minimum Absolute Difference
MFCC	Mel Frequency Cepstral Coefficient
MPEG	Moving Picture Expert Group
MR	Motor Recording
MRA	Multi Resolution Analysis
MSE	Mean Square Error
NIST	National Institute of Standards and Technology
PCA	Principal Component Analysis
PLP	Perceptual Linear Prediction
ROI	Region Of Interest
SI	Separate Identification
SMQT	Successive Mean Quantization Transform
SNR	Signal-To-Noise Ratio
SVM	Support Vector Machine
RASTA	RelAtive SpecTrA
STFT	Short Time Fourier Transform
TIMIT	Texas Instrument/ Massachusetts Institute of Technology
TSS	Three Step Search
VidTIMIT	Video TIMIT
WT	Wavelet Transform
WP	Wavelet Packets

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Computer usage has now become an integral part of our lives, having applications in commerce, industry and education, as well as in social and domestic spheres. However, many of our personal interfaces with computers are not natural, in the sense that they differ from how we interact with each other. Improving the ability of computers to interact using vision, touch and speech will provide more a natural communication with humans. This thesis makes contributions in the area of audio-visual automatic speech recognition (AVASR) and in particular in extracting the mouth region and applying new speech recognition approaches that make use of both visual and audio information.

This introductory chapter describes what is meant by automatic speech recognition (ASR) using machines, the purpose and scope of ASR and its relation to human communications. It also discusses the challenges of ASRs and the approaches that have been adopted by previous researchers in attempting to address these challenges. The use of the visual modality in human speech recognition and its potential for improving the performance of ASRs is also described. Also introduced are the series of operations that are carried out both by a typical ASR system and also by such a system when augmented by a visual modality. The objectives of the research presented in this thesis and the specific contributions of this research are also included.

1.1 AUTOMATIC SPEECH RECOGNITION

Speech, being the most effective way of directly communicating all but the simplest information between humans, is also frequently considered as a suitable candidate for human-computer interaction (HCI). With the ever-increasing interaction with computers for performing both business and personal tasks, developing machines that can speak and listen will make HCI more natural and increase productivity in many applications [1]. Speech recognition by machine, or ASR, is the process of translating speech signals into a set of words. The recognized speech transcript may be the

desired final output in applications such as data entry or word processing, while in other cases it could serve as an input for further processing by the machine, such as in multilingual communication, or the automatic ticketing of vehicles [2].

The practical implementation of ASR systems is made more difficult due to the differences in the speech signals that emanate from individual speakers, even when articulating the same word or phrase; this being known as inter-speaker variability. An example of this variability is the distinct pronunciation of words according to the speaker's geographical region of origin. Intra-speaker variability may also occur, in which the manner in which a word is pronounced is affected by conditions such as age or emotional state. A further major source of variation is co-articulation, where the pronunciation of a phoneme is influenced by the presence of neighboring phonemes [3].

Much of the ASR research attempts to imitate human speech recognition by machine [4], yet very little is known about the exact mechanisms we use [5]. Humans can often continue to recognize speech in challenging environments, such as in the presence of audible noise from the environmental including that generated by other speakers, known as 'cross talk' or, more colloquially, the 'cocktail party effect'. Human beings are able to isolate the wanted speech of one individual even though many others are speaking at the same time. Although many ASR solutions have performed well enough to be used in commercial products, none can yet achieve the level of performance of human speech recognition. Consequently, ASR has met success in relatively well-controlled environments, but to deliver acceptable performance in many real-world situations remains a considerable challenge [6].

1.2 CHALLENGES OF ASR

The goal of ASR research is to develop machines that have near human recognition capabilities in natural environments. In practice, ASR research in recent decades has improved its capabilities to near human performance, but only in noise-free environments. Our abilities to distinguish between speakers and to isolate speech have proved a challenging task to replicate by machine and even in controlled environments the performance of current ASR systems still lags behind that of humans [7]. In terms of developing an ASR solution, the difficulties that need to be

overcome include inter-speaker and intra-speaker variability, co-articulation, cross talk and more generally, susceptibility to environmental noise [1]. In the current work, the main contribution is to reduce the susceptibility of ASR to influences arising from background noise and so this issue provides the main focus for discussion in the remainder of the thesis.

1.3 APPROACHES TO SPEECH RECOGNITION UNDER NOISY CONDITIONS

The speech recognition performance of ASR systems that process only audio signals deteriorate severely in the presence of even moderate levels of background noise [8]. Speech recognition under noisy conditions is recognized as a major hurdle to the deployment of ASR systems in real-world situations and has attracted the attention of many research groups in last three decades. To overcome the problem of ASR performance degradation in the presence of noise approaches based on robust feature extraction, compensation techniques, noise reduction and audio-visual feature extraction have been proposed.

The techniques used in the extraction of the features inherently resistant to noise include RASTA (RelAtive SpecTrA) processing [9], one-sided auto-correlation [10] and auditory model processing of speech [11]. In the RASTA method, the features are extracted after filtering the components that represent both slow and rapid signal changes that lie beyond the normal speech range, thus attempting to mitigate the noise prior to feature extraction. In the one-sided auto-correlation approach, linear predictive coefficients (LPC) are extracted from the autocorrelation of the speech signal after noise filtering rather than being obtained directly from the original signal. As the autocorrelation of the noise component often remains constant, it can be removed in the correlation domain by high pass filtering of the resulting signal. These auditory model based methods extract features based explicitly on the knowledge of human auditory system and have generally shown to increase robustness to noise.

The compensation model attempts to recover the original speech from a corrupted version either in the feature parameter domain or at the pattern-matching stage. Suitable methods reported include the cepstral normalization [12], probabilistic optimum filtering [13] and parallel model combination [14].

To reduce noise content, spectral enhancement techniques such as spectral subtraction and Wiener filtering have been used and shown to improve recognition performance [15]. Other authors have developed speech signal 'denoising' processes based on the soft and hard thresholding of wavelet coefficients [16].

It is well documented that visual information from a speaker's face, referred as 'visual speech', is used by humans for speech recognition; particularly when audio noise is present and is widely used by people with hearing impairments [1]. The approach has been adopted to improve the performance of ASR systems in the presence of noise [17], [18], [19], [20] and the use of the video modality to augment the audio modality has become increasingly popular in recent years.

1.4 THE BIMODAL NATURE OF SPEECH

Human speech production and perception are bimodal in nature. The visual modality is particularly important in the understanding of speech for people with hearing impairments and for the deaf, who will use information obtained from observing lip movements and gestures [21]. The skill of using mouth shapes, and other visible articulators to estimate the underlying sound is known as lip-reading and can be refined through training [22]. For general communication between humans, facial expressions are a source of information about psychological state. Visual cues from the face of speakers and visible articulators such as lips, teeth, jaw and the tongue-tip of the talker are used as aids in human speech recognition. The accuracy and robustness of human speech recognition is generally improved when the complementary and supplementary information available from these multiple modalities is present [23].

Speech production is the result of the movement of articulators, vibrations of the vocal cavity and changes in the geometry of the vocal tract. Vowels or consonants are produced as a result of stable or transient configurations of the vocal tract, respectively. Phones, defined as the smallest segments of sound, are characterized by attributes such as open-closed, front-back, oral-nasal, and rounded-unrounded.

Audio and visual speech are mutually correlated and provide information whose complementary and supplementary natures have not yet been fully explored in the literature [1]. The two modalities provide continuous, independent streams of

information that contribute simultaneously to speech perception but whose integration is achieved in such a manner that the speech recognition performance is enhanced rather than compromised. Acoustically, easily confused sounds such as the unvoiced pair /k/ and /p/, the voiced pair /b/ and /d/, and the nasal /m/ and nasal alveolar /n/ can be potentially distinguished using information obtained from the place of articulation derived in the visual modality [23].

Summerfield [25] identified three main ways in which vision can aid speech perception. Firstly, it helps in localizing the audio source that provides the linguistic and paralinguistic information, so supporting the analysis of the signal and helping to distinguish it from noise. Secondly, it provides information not assessable in the audio stream, such as consonants of short duration that may be more easily masked by audible noise. Thirdly, it provides information about the place of articulation, such as labial, dental, palatal, alveolar or glottal.

The understanding that speech perception is bimodal has motivated interest in acquiring visual information to improve automatic speech recognition quality, and a field known as audio-visual automatic speech recognition (AVASR) has emerged [26].

1.5 OVERVIEW OF ASR AND AVASR SYSTEMS

A schematic diagram of a typical ASR system is depicted in Figure 1.1. ASR systems generally require a training stage, during which the signal processing front-end extracts features from speech utterances whose written interpretation are known and also presented to the system. During this stage, the training module configures the models of speech units, also known as acoustic models. In the decoding stage, unknown speech is applied and features are again extracted, but this time passed to the decoder for recognition purpose. The decoder normally classifies the unknown speech not only using the acoustic models developed at the training stage, but also using constraints imposed by the lexicon and language models [27]. The lexical model assesses the validity of alternative words proposed by the decoder and the language model generates probabilities according to how well candidate sequences of words match linguistic rules. Most modern ASR systems use statistical representations, the most popular being Hidden Markov Models (HMMs) [28].

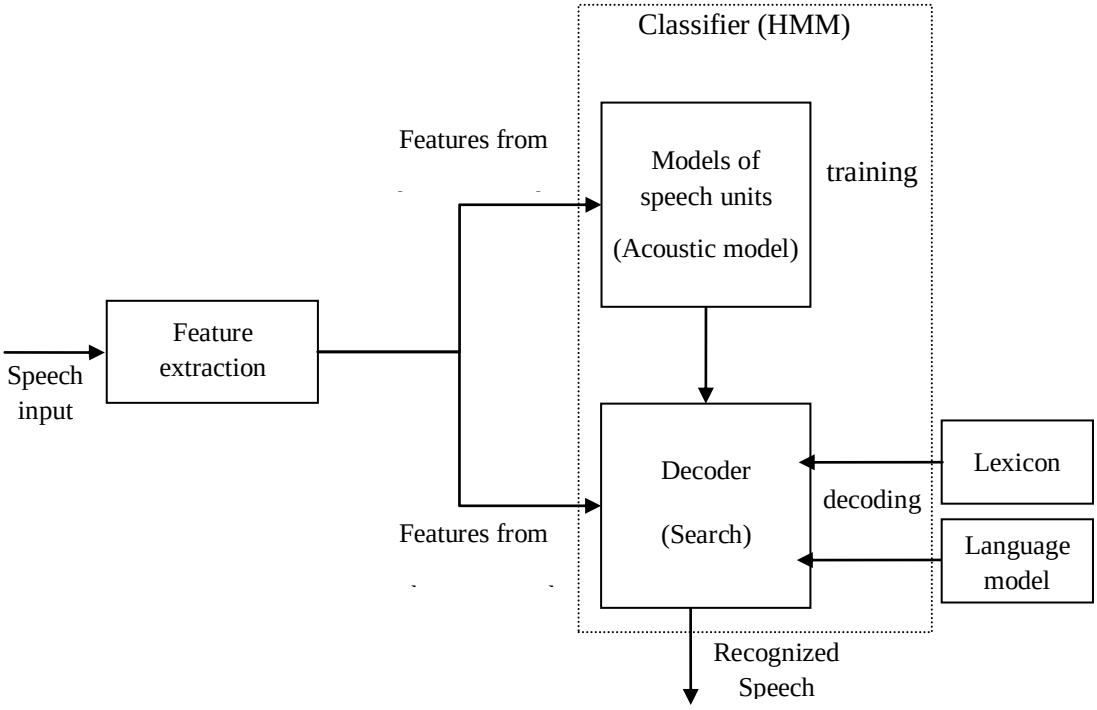


Figure 1.1 Schematic diagram of a typical ASR system [27]

The main difference between audio-only and audio-visual ASR lies in the design of the front-end, as two input streams (the audio stream and the video stream) are now available. Additionally, at some stage in the recognition process, the streams of information from the audio and visual modalities need to be fused and this can occur either after the front end by amalgamating features or, if separate recognizers are used for each modality, the results from the two separate decoders can be combined [29].

The reported research on AVASR systems has mainly focused on the design of the visual front-end and the effective integration of audio and visual modalities for improved speech recognition performance. The visual front-end consists of subtasks such as visual signal pre-processing, region of interest (ROI) extraction and feature extraction as shown in Figure 1.2. The signal pre-processing tasks may concern illumination, distance compensation and audio-visual synchronization. ROI extraction may include the detection of the face and mouth region of the speaker and either the extraction of specific geometric parameters of the speaker's lips such as width, height and curvature, or a region from the speaker's face deemed informative about the visual speech. Where geometric parameters are obtained, feature extraction could be

part of ROI identification, otherwise it could form a separate stage in which suitable transformation and dimensionality reduction techniques are applied to the ROI.

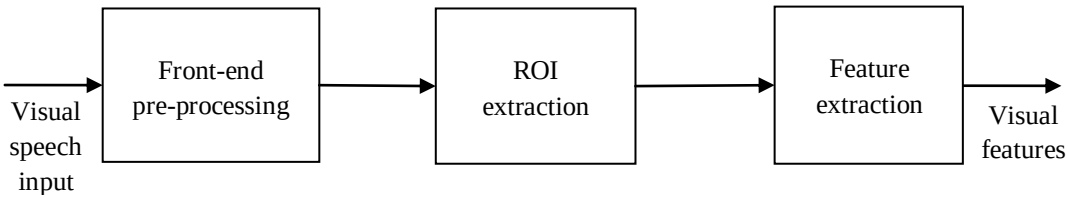


Figure 1.2 Visual front-end

A more detailed discussion of the AVASR components and the approaches applied in AVASR literature can be found in Chapter 2. The classification of AVASR systems and the speech units commonly used in AVASR research are discussed in following sub-sections.

1.5.1 Classification of AVASR systems

AVASR systems can be categorized based on parameters such as vocabulary size, mode of speaking, style of speech and speaker enrolment. These classifications are summarized in Table 1.1.

Table 1.1 Classification of AVASR systems

Parameter	Types
Vocabulary size	Small vocabulary
	Medium vocabulary
	Large vocabulary
Mode of speaking	Isolated word or digit
	Continuous speech
Style of speech	Spontaneous speech
	Read speech
Speaker enrolment	Speaker dependent
	Speaker-independent systems

Into which categories a particular AVASR systems falls is related to the complexity of the speech recognition performed. For example, a large vocabulary, speaker-independent continuous speech recognition system would be the most complex,

whereas isolated word recognition among a small vocabulary would pose a relatively simple task [2].

1.5.2 Speech units

The audio-visual feature vector obtained as the output of the front-end processing is used to train the acoustic models of the speech units. Isolated word and digit recognition systems generally use a whole word (or digit) as a speech unit and the models are trained at this level of granularity. In continuous speech recognition tasks, small components of audio and visual speech, known as phonemes and visemes respectively, are used as speech units [30]. Models are normally developed for each of the phonemes, as well as their context-dependent combinations, bi-phones and tri-phones, consisting of two and three phonemes respectively, and similarly for each of the visemes and their combinations, termed bi-visemes and tri-visemes.

Audio speech units (Phonemes)

Phonemes are the smallest segment of sound that conveys useful linguistic information. Phonologically, each language is made of these basic units and each language or dialect consists of a set of phonemes. For example, there are 45 phonemes in UK English, 46 in American English, 36 in Mandarin and 35 in French. These basic audio sounds are used in most speech recognition systems to provide a set of basic units for recognition; these can then be combined to form the words and sentences using the additional information stored in the lexicon and language models.

Visual speech units (Visemes)

Visemes are distinguishable segments obtained from videos of speakers. They represent particular oral or facial shapes, as well as the positions and movements adopted during speech utterances. They may coincide with the generation of one or more phonemes and are derived either manually by human observation of visual speech or automatically by the clustering of visual speech data. A number of phoneme-to-viseme mappings have been derived by researchers [29], but, unlike phonemes, there is no standard set of visemes for a given language.

1.6 RESEARCH MOTIVATION

One of the major challenges currently facing ASR researchers is to improve system robustness in the face of audible noise. As the visual modality is not directly affected by audio noise, its use can potentially make ASR systems more robust.

In the AVASR research reported in the literature, the feature extraction approaches taken are mainly borrowed from data compression and communication research in which the main aims are to achieve a compact representation with the aim to retain as much of the data as possible in a small number of dimensions. In these approaches, the focus is mainly to preserve the visual quality in the compressed domain with no attention to highlight the discriminative characteristics of the data classes. While low frequency coefficients in the discrete cosine transform (DCT) and discrete wavelet transform (DWT) representations of images of speakers may well capture the essentials of images for compression and communication purposes, this may not be the appropriate frequency band in which to find the bulk of the recorded information that relates to speech articulators and their movements. In this work, it is proposed that the extraction of such information be tackled from a pattern recognition viewpoint, and, in particular, a thorough investigation involving medium frequency coefficients is likely to prove fruitful.

Research in AVASR is mainly carried out by researchers with background experience in either image analysis or audio ASR. As such, the reported AVASR investigations have largely been carried out on individual images extracted from videos of speakers and dynamic information that could be obtained from sequences of frames has been largely ignored. The dominant AVASR approaches in the visual feature extraction paradigm are appearance-based and shape-based methods. Although speech dynamics are incorporated in these methods to a limited extent through the use of temporal derivatives of extracted static features, to best of the author's knowledge, explicit use of motion information has not been explored in AVASR. Further, ROI extraction is currently performed either manually or by applying techniques borrowed mainly from image analysis research and, as such, operate on individual images without exploiting motion information present in the video sequences. Motion detection and estimation have become common techniques in video processing and the isolation of motion information is important in both video compression and communication applications.

Speech is a dynamic phenomenon and the effect on ASR performance that results following the incorporation of motion information obtained from videos of speakers would intuitively appear to be worthy of further investigation. Again, to the best of author's knowledge, explicit use of motion detection and estimation in a visual front-end in an AVASR has not been previously reported.

1.7 RESEARCH AIM AND OBJECTIVES

The aim of the work in this thesis is to incorporate dynamic visual information in the front-end of an AVASR system in order to improve the overall speech recognition performance. The particular objectives of the work are as follows.

1. Investigate methods that use dynamic information obtained from sequences of images in order to automatically and robustly extract visual ROIs.

2. Provide additional speech-related information in the form of dynamic features obtained from the mouth region.

3. Where these features are frequency-based, determine which frequency regions produce information that is best able to improve AVASR performance.

1.8 ORIGINAL CONTRIBUTIONS

This research investigates AVASR from the perspective of visual speech dynamics. A new approach is adopted for the visual front-end based on the motion information from the video of speech for both ROI extraction and visual feature extraction. In addition, the limitations of traditional image transform methods are addressed from the new perspective of discriminative feature extraction instead of the data reduction point of view where the main goal is the preservation of maximum data variance.

This thesis reports novel contributions to the basic operations of AVASR, namely the automatic isolation of the visual ROI and the extraction of visual features for AVASR. The contributions of this work are as follows.

1. New features obtained from specific frequency bands are proposed from the DCT and DWT representations of visual speech images. These features are shown to

yield better performance compared to those extracted from low frequency coefficients.

2. A novel motion-based visual ROI extraction approach has been proposed for use with both appearance-based and shape-based feature extraction methods. The same ROI has also been used for feature extraction purposes, giving a completely automatic visual front-end.

3. New visual features based on a motion-based approach have been proposed. These features extract dynamic information from the video stream of speakers and were found to improve speech recognition performance compared to that obtained using static features extracted from individual frames.

1.9 ORGANISATION OF THE THESIS

This thesis consists of seven chapters. Chapter 1 has provided an introduction to the ASR and AVASR, the challenges of current ASRs and an overview of the objectives and contributions of this thesis.

Chapter 2 gives an overview of AVASR systems and reviews existing approaches described in the literature.

In chapter 3, the pattern recognition research relevant to AVASR systems is covered in detail. A description of the commonly-used AVASR data transformation and dimensionality reduction techniques is given.

Chapter 4 reviews the image transform based feature approach and its limitation in capturing speech information. New features derived from specific frequency bands of the discrete cosine transform (DCT) and discrete wavelet transform (DWT) representation of the images of speaker's mouth are proposed. The performance of the new features in their application to speech recognition is compared with existing low-frequency features, both for clean speech and in the presence of background noise.

Chapter 5 discusses the current approaches for automatic extraction of visual ROI, their limitations and the effect robust ROI extraction has on the overall performance of AVASR. A new method for ROI extraction based on motion detection in video frames is proposed and its performance assessed on a range of image sequences.

Chapter 6 reviews the use of motion compensation techniques in MPEG based video compression and discusses how the approach can be extended to the AVASR task. New motion-based features that use frame difference, block-matching and optical flow approaches are investigated and their performances are compared with the more commonly-used appearance based features. The recognition performance of the new features is studied in the presence of noise and compared with that of audio-only ASR.

Chapter 7 concludes the findings of this research and also proposes directions for future research.

1.10 REFERENCES

[1] Chibelushi, C. C., Deravi, F., Mason, J. S. D. (2002), "A Review of Speech-Based Bimodal Recognition", *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23-37.

[2] Zue, V., Cole, R., and Ward, W. (1996), "Speech Recognition", Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen A., Zue, V., Varile, G. B., and Zampolli, A. (Eds.), *Survey of the state of the art in human language technology*, Cambridge University Press, pp. 1-62.

[3] Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S. Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C. (2006), "Impact of variabilities on speech recognition", *Proceedings of 11th International Conference on 'Speech and Computer' SPECOM'2006*, Saint-Petersburg, Russia, pp. 3-16.

[4] Juang, B. H., and Rabiner, L. R. (2005), "Automatic speech recognition–a brief history of the technology", *Elsevier Encyclopaedia of Language and Linguistics*, Second Edition, Elsevier.

[5] Chu, S. M., Libal, V., Marcheret, E., Neti, C., Potamianos, G. (2004), "Multistage information fusion for audio-visual speech recognition", *Proceedings of the IEEE International Conference on Multimedia and Expo, 2004 (ICME'04)*, vol. 3, pp. 1651-1654.

[6] Nefian, A., Liang, L., Pi, X., Liu, X., Mao, C., and Murphy, K. (2002), "A coupled HMM for audio-visual speech recognition", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, Florida, USA, pp. 2013-2016.

[7] Potamianos, G. (2005), "Audio-Visual Automatic Speech Recognition Theory, Applications and Challenges", http://www.ee.columbia.edu/~stanchen/e6884/slides/lecture12.avsr.pdf/.

[8] Dupont, S., and Luettin, J. (2000), "Audio-visual speech modeling for continuous recognition", *IEEE Transaction on Multimedia*, vol. 2, no. 3, pp. 141-151.

[9] Hermansky, H., and Morgan, N. (1994), "RASTA processing of speech", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589.

[10] You, K. H., and Wang, H. C. (1999), "Robust features for noisy speech recognition based on temporal trajectory filtering of short-time autocorrelation sequences", *Speech Communication*, vol. 28, pp. 13-24.

[11] Kim, D. S., Lee, S. Y., and Kil, R. M. (1999), "Auditory processing of speech signals for robust speech recognition in real world noisy environment", *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 1, pp. 55-69.

[12] Acero, A., and Stern, R. M. (1990), "Environmental robustness in automatic speech recognition", *Proceedings of International Conference on Acoustic, Speech and Signal Processing*, pp. 849-852.

[13] Neumeyer, L., and Weintraub, M. (1994), "Probabilistic optimum filtering for robust speech recognition", *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, Adelaide, Australia, pp. 417-420.

[14] Gales, M. J. F., and Young, S. J. (1996), "Robust continuous speech recognition using parallel model combination", *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352-359.

[15] Gauvain, J., and Lamel, L. (2000), "Large vocabulary continuous speech recognition: Advances and applications", *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1181-1200.

[16] Young, S. (1996), "A review of large vocabulary continuous speech recognition", *IEEE Signal Processing Magazine*, pp. 45-57.

[17] Grant, K. W., and Seitz, P. F. (2000), "The use of visible speech cues for improving auditory detection of spoken sentences", *Proceedings of the Journal of the Acoustical Society of America*, vol. 108, pp. 1197-208.

[18] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J. (2000), "Audio-visual speech recognition", *Workshop final Report*, Centre of Language and Speech Processing, Johns Hopkins University, Baltimore, MD.

[19] Kaynak, M. N., Zhi, Q, Cheok, A. D., Sengupta, K., Jian, Z., and Chung K. C. (2004), "Analysis of Lip Geometric Features for Audio-Visual Speech Recognition", *Proceedings of the IEEE Transaction on Systems, Man and Cybernetics*, vol. 34, no. 4, pp. 564-570.

[20] Heckmann, M., Berthommier, F., and Kroschel, K. (2002), "Noise Adaptive Stream Weighting in Audio-Visual Speech Recognition", *Proceedings of EURASIP Journal on Applied Signal Processing*, pp. 1260-1272.

[21] Massaro, D. W. (1987), "Speech Perception by Ear and Eye", Campbell, R., and Dodd, B. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*. Hove, United Kingdom: Psychology Press Ltd. Publishers, pp. 53-83.

[22] Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoit, C., Guiard, M. T., Goff, B. L., Ribes, J. R., Adjoudani, A., Defee, I., Munch, S., Hartung, K. and Blauert, J. (1995), "A Taxonomy of Multimodal Interaction in the Human Information Processing System", *A Report of the ESPRIT Project 8579*, Mianmi.

[23] Benoit, C., Martin, J. C., Pelachaud, C., Schomaker, L., and Suhm, B. (2000), "Audio-visual and Multimodal Speech Systems", Handbook of Standards and Resources for spoken language Systems, Kluwer, *Multimedia Systems*, pp. 1-95.

[24] Massaro, D. W., and Stork, D. G. (1998), "Speech Recognition and Sensory Integration", *American Scientist*, vol. 86, no. 3, pp. 236-244.

[25] Summerfield, Q. (1987), "Some Preliminaries to a Comprehensive Account of Audio-visual Speech Perception", Campbell, R., and Dodd, B. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*, Hove, United Kingdom: Psychology Press Ltd. Publishers, pp. 3-51.

[26] Stork, D. G., and Hennecke, M. E. (1996), "Speechreading: An Overview of Image Processing, Feature Extraction, Sensory Integration and Pattern Recognition Techniques", *Proceedings of 2nd International Conference on Automatic Face and Gesture Recognition*, pp. XVI-XXVI.

[27] Rabinar, L. R. (1989), "A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286.

[28] Jain, A. K., Duin, P. W., and Mao, J. (2000), "Statistical pattern recognition: A review", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4-37.

[29] Hazen, T. (2006), "Visual model structures and synchrony constraints for audiovisual speech recognition", *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 1082-1089.

[30] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev V., and Woodland, P., (2006), *The HTK Book V3.4.*

CHAPTER 2

AN OVERVIEW OF AUDIO-VISUAL AUTOMATIC SPEECH RECOGNITION SYSTEMS

2.1 INTRODUCTION

Chapter 1 discussed the challenges of traditional audio-only ASR systems and the various approaches taken in ASR research to make these systems robust in the presence of noise. In addition, it discussed how audio-visual automatic speech recognition (AVASR) systems can potentially improve the performance of audio-only ASRs when affected by audio noise by incorporating additional speech information from videos of speakers.

This chapter gives an overview of AVASR systems and its constituent parts. Following a general introduction to AVASR systems, the remainder of the chapter discusses these components in detail as well as reviewing the approaches adopted in the AVASR literature to achieve their implementation.

The use of visual speech information has introduced new challenges in the field of automatic speech recognition (ASR). These are robust face and mouth detection, extraction and tracking of a visual region of interest (ROI), extraction of informative visual features from the ROI, the integration of audio and visual modalities and the provision of suitable classifiers [1].

The AVASR process is depicted in Figure 2.1. In contrast to audio-only speech recognition systems, where only the audio stream of information is available, here there are two streams of speech information, namely the audio stream and the video stream. The process of AVASR system implementation consists of two stages: design of the front-end processing system and the training of the recognizer. Front-end processing transforms speech into a parameter vector suitable for subsequent processing and consists of the pre-processing of audio and video sources, followed by

feature extraction from each of the two sources [2]. The audio and visual features thus extracted can either be combined directly at the feature level (termed early integration) or could be used to train two separate audio and video recognizers and their results integrated at some later stage (known as decision integration). In Figure 2.1, the early integration method is represented by the solid lines and the alternative decision integration approach by the broken line. Between these two extremes there is a third possibility of intermediate integration, where the two modalities are combined at some point in the processing between the feature integration and decision integration [3]. Depending on the application, the recognition task could be as simple as isolated word or digit recognition or as complex as conversational speech recognition. The choice of classifier depends on the complexity of the recognition task and the integration strategy used. For example, for small vocabulary tasks involving isolated word or digit recognition, the features are normally passed directly to the classifier along with the corresponding features obtained from the unknown speech. The classifier performs the matching of the feature vector of unknown speech to those of the known speech units and assigns a symbol to the unknown speech corresponding to its matching pattern in the known speech. For large vocabulary continuous speech recognition tasks, acoustic models are developed for each phonetic symbol using these features. Features extracted from unknown speech are then passed to the recognizer which uses the acoustic model along with lexical and syntactic information to identify the unknown speech. Hidden Markov models (HMMs) are popular tools for such modelling [4], [5]. Depending on the application, the output from the recognizer is commonly either in the form of recognized text or is subjected to further processing.

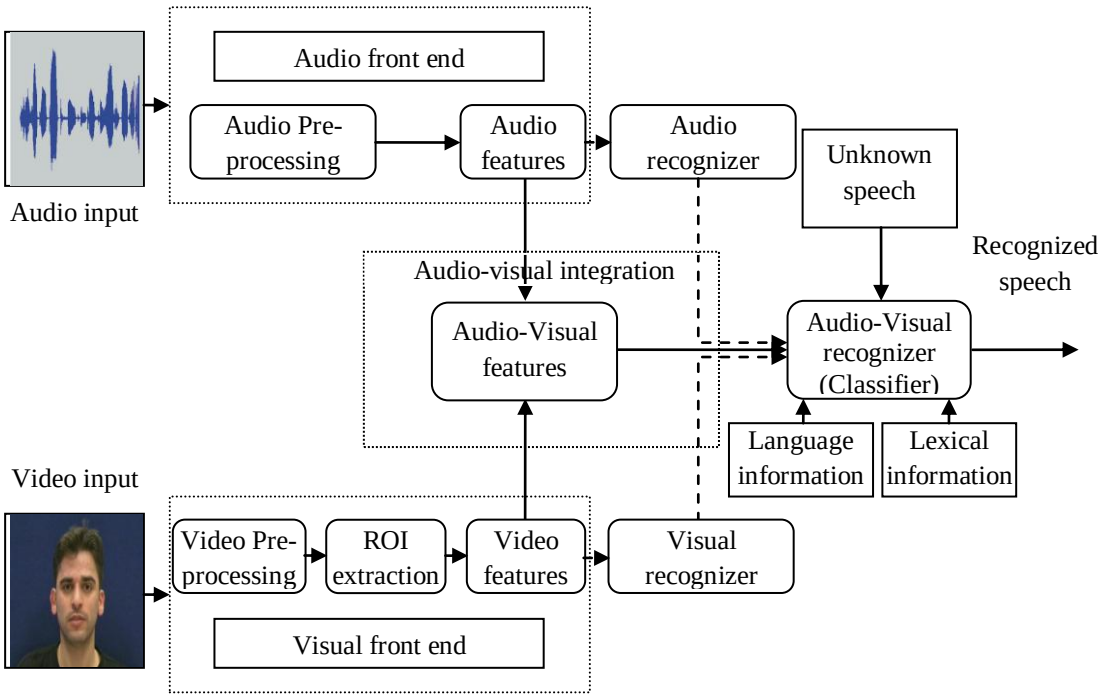


Figure 2.1 Block diagram of an AVASR system

The AVASR systems available to date are mainly the products of work carried out by individual researchers or small research groups in universities and vary greatly in the approach taken. Due to the unavailability of standard audio-visual databases, testing is normally carried out using databases developed by the researches themselves. As the feature extraction approaches are often unique to the research investigation, this may also dictate the requirement to develop task specific databases, particularly in those methods based on lip and mouth geometric parameters. Generally, the comparison of results between the research carried out by different authors is made difficult as there is no consistent use of speech or video databases, with each likely to involve different number of speakers and some containing only isolated words or digits [6], while others may contain large vocabularies of continuous speech [7].

This chapter is organised largely according to the processing stages identified in Figure 2.1. Section 2.2 provides an introduction to the front-end design while audio and visual front-ends are described in further detail in sections 2.3 and 2.4 respectively. Section 2.5 presents the approaches to audio-visual integration reported in literature and the range of classifiers that have been used for AVASR is outlined in section 2.6. Section 2.7 introduces the audio-visual databases that have been used in

AVASR research and the summary describes how the direction chosen for the current research has been influenced by the literature surveyed.

2.2 FRONT END DESIGN

Before being applied to the recognizer for training or recognition purposes, audio and visual streams need to be pre-processed to remove data irrelevant to speech and to enhance certain characteristics that help to improve speech recognition performance. These pre-processing stages of the audio and video data are known as the audio front-end and visual front-end respectively. 'Front-end' encompasses the pre-processing of the speech signal before the feature extraction phase, as well as the feature extraction itself. The design of the front-end, and particularly the feature extraction phase, plays an important role in maximizing the overall performance of a speech recognition system and is a core area of research in both audio-only and AVASR research. In the audio part of the front-end pre-processing, a number of techniques are available to enhance the speech signal and to reduce the effects of background and channel noise [7]. The design of video front-end is a rather more challenging task, as the video signal will contain substantial information about the speaker and background that are not relevant to the speech itself. This needs to be filtered out and a region of interest (ROI) around the mouth of the speaker defined and extracted [2], thereby greatly reducing both the dimensionality of the required feature vector and the computation cost of later processing. In comparison with the audio front-end, the visual front-end will also include the additional steps of speaker face and mouth detection and the extraction of a speech information region from the face of the speaker, collectively known as ROI extraction. The effects of variations in lighting conditions in both the spatial and temporal dimensions may also be addressed as part of the visual front-end, as well as distance and orientation normalization where relevant. Audio and visual front-end processing are performed separately on the two streams and the extracted feature vectors integrated to form a single feature vector or used to train two separate recognizers depending upon the modality fusion approach adopted.

As the original audio and visual speech signals have high dimensionality, then, to use them directly for training and recognition, the classifier will need computational time and resources that are not commonly found even in modern computer systems. Therefore, a more compact set of parameters representing the significant

characteristics of speech are extracted from both the audio and video signals. The compact sets of parameters extracted from the two streams are generally referred to as the audio and video features respectively. The performance of a speech recognition system is greatly dependent on the extraction of features which are robust, stable, and ideally retain all the speech information contained in the original source signal [8]. The main purpose of feature extraction is to capture speech information in a reasonably small number of dimensions with the aim of generating features with the following properties.

- A maximum variance between classes (here phonemes and visemes for the audio and visual modalities, respectively), while minimizing the variance between members of same class.
- Capture of the salient properties of speech in terms of both its spectral characteristics and its temporal variations.
- Robust against the effects of environmental changes in their respective streams, such as lighting conditions and image background in video and audible noise in audio.
- Independent of the speaker and of the speakers' displayed emotions. Note that the performance of a video recognizer may be affected more by certain factors than audio. For example, video feature extraction may well greatly be affected by a speaker having a beard, but this will have little or no effect on the audio modality.

The frequency at which these features are extracted depends upon the nature of recognition task. For digit and isolated word recognition task, usually the same number of features is extracted for each digit or word, irrespective of duration, while in the case of the continuous speech recognition task, features are generally extracted at a rate of 100 times a second. Features extracted from individual frames carrying the static speech characteristics are generally combined with first and second temporal derivatives (delta and delta-delta features) to include the dynamic characteristics [9]. The combined set of features is then used to train the acoustic models of individual phonemes or visemes and their context-dependent bi-phones/bi-visemes and tri-phones/tri-visemes [7]. Alternatively, the audio and visual features are combined to form a single audio-visual feature which is then used to train models for either phonemes or visemes.

2.3 AUDIO FRONT END

The two stages generally used for the audio front-end design are depicted in Figure 2.2. The first is the signal pre-processing stage that converts the sound pressure wave into a digital signal and enhances certain important spectral components. The second is the feature extraction stage that consists of the spectral analysis of the signal and the extraction of a set of parameters making the audio feature vector. These stages are now discussed in more detail.

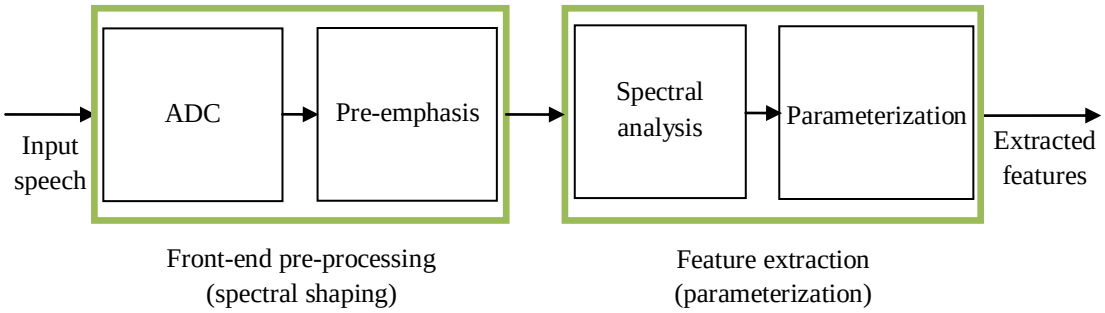


Figure 2.2 Stages of the audio front-end design

2.3.1 Front-end pre-processing (Spectral shaping)

Firstly, the audio front-end samples and quantizes the speech using an analog to digital converter (ADC). Sampling rates in the range 10 kHz to 16 kHz are commonly used for microphone inputs, while for telephone signals an 8 kHz sampling rate is more appropriate [10]. The analog to digital conversion process will introduce some noise to the signal in the form of quantization noise, non-ideal frequency response and fluctuating DC bias. In addition, the microphone used in the analog to digital conversion will also introduce non-linear distortion and both high and low frequency information loss. To minimize the effects of such noise, the digital speech is normally enhanced using a pre-emphasis filter to flatten its spectrum [8]. The advantage of this pre-emphasis is two-fold. Firstly it offsets the natural attenuation in voiced section of speech, caused by the physiological characteristics of speech production mechanism; secondly it assists the spectral analysis stage in modelling the perceptually important aspects of speech by emphasising the band above 1 kHz.

2.3.2 Audio feature extraction

Audio speech recognition has been an active field of research for more than five decades and considerable work has been carried out on audio feature extraction [8], [11], [12], [13]. The frequency domain representation of speech is generally useful for the extraction of the salient features for speech recognition as it reveals the spectral components present in the signal. Fourier transform (FT) and linear prediction techniques are the commonly used signal analysis techniques in the speech recognition literature [8].

Speech signals are generally non-stationary, implying that the amplitudes of spectral components present changes with time. To overcome this problem, the spectral analysis is carried out in a sequence of relatively short frames each of a duration in the range 10ms to 20ms, during which time the signal is assumed to be stationary. A separate set of audio features is extracted from each of these frames. Although the assumption of stationarity may not be strictly true for certain phonemes such as stop consonants, for most practical purposes the approach has been found to yield satisfactory results. The frames are extracted by applying a Hamming window of length 20ms to 30ms and non-overlapping frames of typical duration of 10ms to 20ms are formed for feature extraction. As the window length is normally longer than the frame period, a resulting overlap typically of about 50% occurs.

The most commonly used audio features are obtained from the Mel-frequency cepstral coefficients (MFCC), or less frequently linear predictive coefficients (LPC) or perceptual linear prediction (PLP) coefficients. More recently, features extracted from both the wavelet transform (WT) and wavelet packets (WP) have been found useful in addressing the limitations of MFCC based features in specific cases when the speech signal changes rapidly, such as stop consonants [14], [15], [16]. Although wavelet-based features have exhibited better performance when recognizing certain specific phonemes, MFCC coefficients remain the most commonly-used features in audio only ASR research. In AVASR research and in the work presented in this thesis, the primary focus is to explore the video modality for the extraction of additional features, whereas the MFCC features are used primarily for the audio modality. Nevertheless, as the AVASR work will involve integration with the audio

modality, it is appropriate here to provide a brief introduction to the most commonly used methods for audio feature extraction.

Linear predictive coefficients (LPC)

LPC is one of the most commonly used parametric modelling techniques in the speech recognition literature. In LPC analysis, it is assumed that the speech signal at any given time can be estimated from a linear combination of the speech samples in the past [10]. If *s*(*n*) is the current speech signal, it can be estimated from its previous values *s*(*n*-1), *s*(*n*-2), *s*(*n*-3),…, *s*(*n*-*p*) as

$$s(n) = \sum_{j=1}^{p} a(j)s(n-j) + e(n) \tag{2.1}$$

where *e*(*n*) is the error in the estimation of the current signal and the set of coefficients *a*(*j*) are the linear predictive coefficients. The number of predictive coefficients, *p*, is the number of previous samples used in the estimation. The predictive coefficients *a*(*j*) are computed by minimizing the mean-squared error between the predicted and the actual signal. The most frequently-used method to calculate the coefficients is autocorrelation, but covariance and lattice methods are also used [8].

Mel Frequency Cepstral Coefficients (MFCC)

It has been shown in psychophysical studies that humans do not perceive the variation in speech frequency on linear scale, but rather they are more sensitive to frequency variations below 500Hz. Above this, the same degree of variation in pitch is perceived by an unequal increase in frequency. The interval over which a certain level of change in pitch is observed becomes greater as the frequency increases on an ordinary hertz scale. The Mel scale representation is based on this non-linear response of the human ear to pitch perception. A more even distribution of coefficients according to pitch sensitivity is produced by mapping the pitch variations on hertz scale to the Mel scale [14]. The relation between the hertz scale and the Mel scale is given by

$$Mel(f_m) = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.2}$$

where f and f_m are the frequencies on hertz scale and Mel scale, respectively.

To obtain the MFCC values, the following procedure is normally followed [17]. The discrete Fourier transform (DFT) of the speech signal is taken over the frame duration and the power content of the resultant spectrum is mapped onto the Mel scale using triangular overlapping windows. The Fourier transform taken over a short duration such as the one above, is known as the short-time Fourier transform (STFT). The MFCC are calculated by taking the discrete cosine transform (DCT) of the logarithm of the power mapped on the Mel frequencies. The steps for calculating MFCC coefficients are depicted in Figure 2.3.

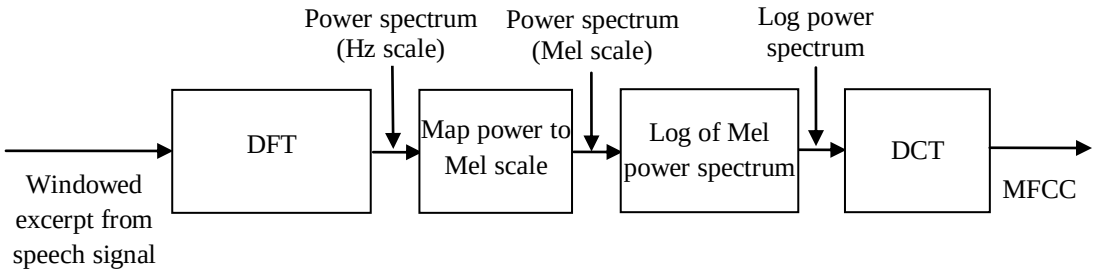


Figure 2.3 General stages for calculating the MFCC

Limitations of the STFT and wavelet transform

STFT is the most commonly-used technique for spectral analysis of speech signals. In the application of the STFT, it is assumed that the speech signal remains stationary for the duration of a frame. However, this is not strictly true and particularly for stop consonants where the spectral transition occurs rapidly. It has been found that for certain rapidly-changing consonant sounds, replacing the STFT by the wavelet transform gives better recognition performance [14], [15], [16]. A detailed discussion on the capabilities of the wavelet transform in addressing the limitations of the STFT can be found in section 3.2.

2.4 VISUAL FRONT END

The visual front-end encompasses the detection of the speaker's face and mouth regions, the extraction of a visual ROI, extraction of visual features and the matching and synchronization of video and audio streams [9]. The general stages of the visual front-end process are shown in Figure 2.4.

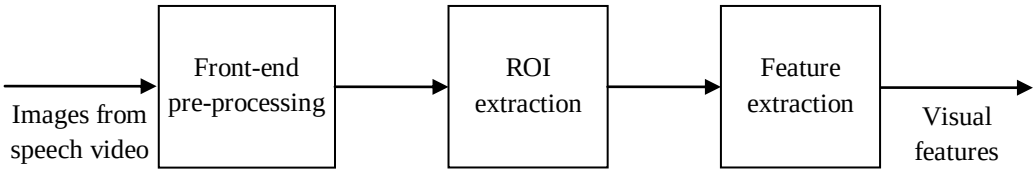


Figure 2.4 Visual front-end processes

Almost all visual feature extraction approaches obtain features from individual frames of speaker videos, requiring the frames to be isolated and stored as separate images. Most of the audio-visual databases used for AVASR research are also available as a sequence of separate images. As the videos of speakers contain information not related to the speech itself, such as the identity and background, the visual front-end needs to remove this superfluous information leaving only that related to speech. The mouth region of the speakers is identified and a region of interest (ROI) is isolated prior to the extraction of visual speech features. A number of feature extraction methods are described in the literature, but the general aim on normally to identify features which contain sufficient information for discriminating between speech classes (visemes) and which are stable and robust to changes in the environment [18].

The two tasks in visual front end design, namely ROI extraction and feature extraction, are greatly interdependent. The required accuracy of ROI identification depends on the feature extraction approach adopted. The linear transform based feature extraction techniques, such as the discrete cosine transform (DCT) [1] and discrete wavelet transform (DWT) [2], require only relatively crude mouth region detection. Conversely, both shape-based feature approaches [19] (that require parameters such as mouth length, width, area, perimeter and eccentricity), and face contour based approaches [20], [21], (such as the active appearance model) require more accurately defined face and lip region identification.

2.4.1 Front-end pre-processing

Although most of the databases are generated in constrained environmental conditions with constant lighting and a relatively static head and shoulders, to further improve invariance and simplify the recognition task, normalization is often performed. Typical of such operations are illumination and distance normalization, as well as head rotation compensation [1]. As features are typically extracted from audio speech

at a rate of hundred times a second whereas the video data are usually recorded at a rate of 25 or 30 frames per second, then, to allow a combined analysis, video data are normally up-sampled to the audio frame rate using some suitable interpolation technique. Up-sampling can be performed either by interpolation between the video frames before the feature extraction phase, or later by interpolation between the extracted features [9].

2.4.2 ROI identification

The area containing the mouth of the speaker is generally considered as the most informative for visual speech information [7]. In shape-based AVASR, the ROI is normally defined around the mouth of the speaker from which the geometric parameters are extracted as visual features, but the entire face or the lower half of face containing the mouth and other articulators may be used in appearance-based AVASR. A brief survey of the ROI extraction approaches used in the AVASR literature is included in this section, while a more detailed discussion of ROI extraction can be found in chapter 5 of this thesis, along with a novel motion based approach for ROI extraction.

An important area of research in AVASR is focused on extracting the most informative visual features for robust speech recognition. One approach is to colour or apply markings to the speaker's lips in order to simplify detection [22], [23]. Further, in some corpora, the mouth region is extracted manually [6], [24]. To realise a real-time and general purpose AVASR system, it is essential to be able to detect, track and extract the face and mouth in video frames automatically without applying any pre-defined marking. In recent years various techniques have been used for automatic extraction of the ROI, including statistical approaches [25], as well as traditional image processing based techniques such as colour segmentation [26], combining colour and edge detection techniques [27], template matching [28], symmetry detection [29], deformable templates [30]. To enhance performance these techniques are usually used in combination with simple image analysis and morphological operations.

Depending on the feature extraction approach adopted, ROI extraction may require only the detection and tracking of the face and mouth regions, but in other

applications a rather more accurate estimation of lip contour is required. These operations are discussed in further detail in the following subsections.

Face and mouth detection, and tracking

As the mouth region contains very few features to detect it directly, most of the ROI extraction approaches first detect the face of the speaker followed by the identification of the mouth region, from which the required ROI is then extracted [2]. A typical procedure is the one described in Steifelhagen *et al*. [31], in which the face of the speaker was first detected using a skin colour statistical model and the mouth position was subsequently identified by detecting the locations of the eyes before using geometric relationships between the eyes and the mouth.

Face and mouth detection for audio-visual speech recognition is normally performed using techniques similar to those found in image analysis and recognition literature. These include, skin colour based segmentation [32], [33], region based approaches [34] and methods based on knowledge of the geometric relations between facial features [35]. As the corpora in use for audio-visual speech recognition are usually face centred and variations in orientation and lighting conditions are restricted, these techniques generally yield outputs sufficient for the needs of AVASR systems. For appearance-based feature extraction, the detected face or mouth of the speaker is the desired ROI while for shape or model-based approaches, further processing is carried out to estimate lip and face contours in order to extract either their geometric parameters or the parameters of model.

Face and mouth detection could be performed for every frame of the video sequence, but when the head movement is limited, it is usually more computationally economical to track the face and mouth region between consecutive frames.

For appearance-based approaches, features are often extracted from a mouth region that is not carefully defined and, the coordinates of ROI are determined once in the first frame of video, and the coordinates so identified are then used for ROI extraction in the remainder of the frames. As shape-based feature extraction require a more accurate estimation of the lip contour, a tracking approach constrains the search area, thereby reducing the computation time.

Lip contour estimation

To extract the geometric parameters of the mouth a number of algorithms suitable for lip detection and contour estimation have been proposed in literature. In [25], a lip detection algorithm based on normalized RGB pixel values and refined by using neighbourhood-based processing is reported. Chandramohan *et al*. [30] have proposed a deformable template approach where an initial estimate of the lip contour is provided by comparing the image with pre-defined templates whose points iteratively converge to the lip contour, minimizing a penalty function. Other techniques used for lip contour extraction include edge tracking [27], active contour models [28], active shape and appearance models [20] and snakes [36].

2.4.3 Visual feature extraction

Research in AVASR has been being carried out for over two decades, but, unlike audio speech recognition where MFCC have emerged as *de facto* standard audio features, no agreed standard for visual feature extraction yet exists. The types of visual features found in the AVASR literature can be broadly grouped into three categories: (a) appearance-based (or low-level) features; (b) shape-based (or high-level) features; and (c) hybrid features obtained by using a combination of appearance and shape based features [2].

Although speech production is a dynamic activity, in AVASR research it is generally assumed that the individual static frames from the videos of speakers can provide important information to aid the recognition of speech. Nearly all of the visual feature extraction approaches found in the literature are based on information obtained from individual frames. To provide dynamic visual information, the first and second derivatives are taken of features extracted from consecutive frames and used to supplement the static features [37].

Appearance-based features

In the appearance-based feature extraction approaches, pixels from the speaker's mouth region are used as source of visual speech information for AVASR. Appearance-based approaches do not need sophisticated algorithms for feature extraction but are generally more sensitive to lighting conditions and pose than are shape-based features. The ROI used is typically either a rectangular or circular region that includes the speaker's mouth. A vector is then obtained either directly using the

colour or greyscale values of the pixels in the ROI or some suitable transformation of the pixel values is obtained, such as the DCT [6] or the DWT [2].

The dimensionality of this vector is generally too high to be used directly for statistical modelling of speech classes and one of a number of available dimensionality-reduction techniques is normally applied to render the information suitable for recognition purposes while retaining as much of the original speech information as possible. The two most commonly used techniques for dimensionality reduction are principal component analysis (PCA) [38], [39] and linear discriminant analysis (LDA) [1], [40]. PCA transforms data in such a way that the most of the variance in the data is contained to a small number of parameters called principal components. LDA transforms data so as to maximize the discrimination between different classes.

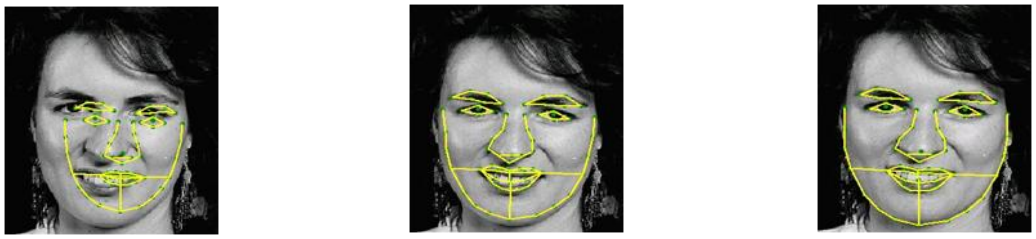
Shape-based features

Shape-based features are inherently of low dimensionality and are less affected by the lighting conditions and face orientation. However, compared with the appearance-based features, they are difficult to extract robustly and are computationally expensive.

In these approaches, the shape of speaker's lips or the face contour itself is used to generate the speech related information for speech recognition. One approach is to obtain geometric features such as the length, width, area and perimeter of the inner or outer parts of the lips [23], [41]. Also, statistical models have been developed to describe the shape of lips or the face. For example, Luettin and Thacker, [21] described active shape models (ASM) as deformable templates that can be iteratively adjusted to match the outlines of objects in an image. The parameters of the model are then used as visual features for recognition. Active appearance models (AAM) extend ASMs to include grey-level information in performing the template match [20]. Kass *et al.* [36] describe a method termed 'snakes' that track edges in image sequence and are used to extract the lip contour.

Initial pose | After 5 iterations | Convergence

Figure 2.5 An example of Active Shape Models (ASM) Cootes *et al.* [28]



Initial pose | After 2 iterations | Convergence

Figure 2.6 An example of Active Appearance Models (AAM) Cootes *et al.* [20]

Hybrid Features

The use of appearance and shape features each has their own strengths and limitations. In an attempt to harness the advantages of both, appearance and shape based features have been combined to make a third class of features known as hybrid features, normally by using a simple concatenation of the two types. In [42], the PCA projection of pixels from the mouth region was combined with lip geometric features. In another investigation, the combination of ASM-based features with PCA to produce a set of visual features was reported [43].

2.5 AUDIO VISUAL INTEGRATION

Although audio and visual streams have been used independently to design audio-only and video-only ASRs, the literature shows that recognizers combining information from both audio and visual modalities can outperform those using a single modality [2]. Consequently, the effective integration of audio and video streams of data is likely to be a fruitful area for research activities that are attempting

to improve ASR performance. There are different levels at which the two modalities could be integrated, namely feature level, state level, phoneme and word level, or even combining the recognition scores at a sentence level [44]. A number of AVASR design approaches have also attempted to integrate the information from audio and visual modalities at a number of levels, using methods that are as near as possible to those used by humans [2]. This is, however, proving a difficult task, as it is not really known how humans integrate audio and visual speech modalities. Cognitive studies have suggested that there may be four different architectures for modality integration [45] and integration strategies used in AVASR literature for auditory-visual fusion usually follow one of these architectures. These architectures are as follows.

1. In the Direct Identification (DI) model, the data from both the audio and video modalities are provided as direct inputs to a bimodal classifier. The classifier chooses the prototype from its vocabulary which is nearest to the input in some statistical sense.

2. The Separate Identification (SI) model employs a separate classifier for each modality and the results from the unimodel classifiers are fused for final decision making based on probabilistic values.

3. In the Dominant Recording (DR) model, the audio modality is taken as a dominant modality, with the video modality incorporated in the audio representation, such as the estimation of the vocal tract transfer function from both the audio and visual data. These estimates are integrated for final classification purposes.

4. In the Motor Recording (MR) model, both audio and visual inputs are projected into a common space before being passed to a classifier.

These models are graphically depicted in Figure 2.7.

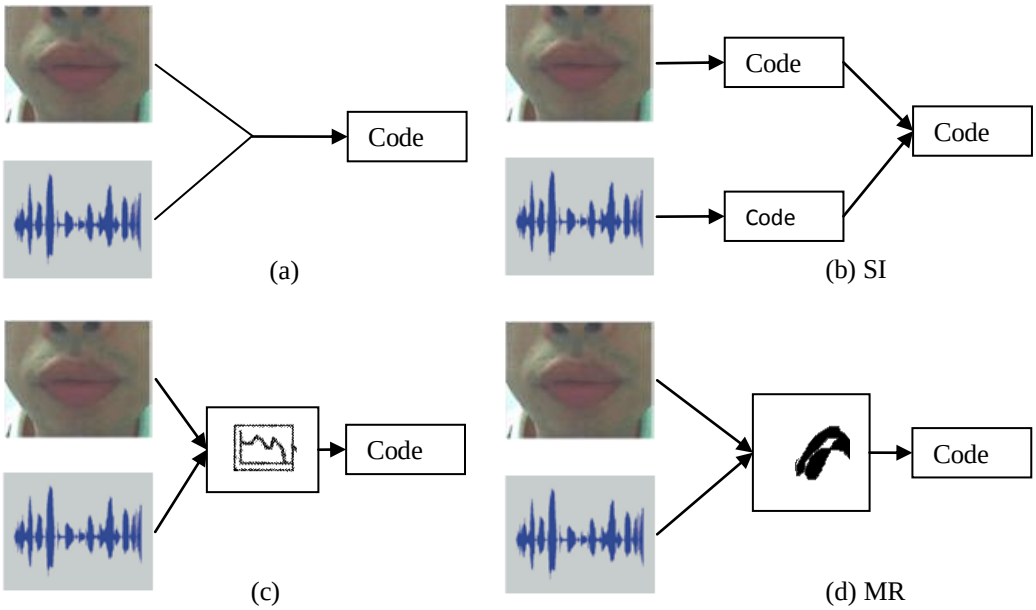


Figure 2.7 Models of audio-visual integration [45]

Although there is no general agreement among psychologists regarding which model most closely resembles the human speech perception process, empirical evidence favours the MR architecture [45].

The approaches used in the AVASR literature for the integration of the audio and visual streams of information can be grouped into three categories: feature fusion, decision fusion and hybrid fusion. These integration strategies are graphically depicted in Figure 2.8, and discussed in more detail in the following subsections.

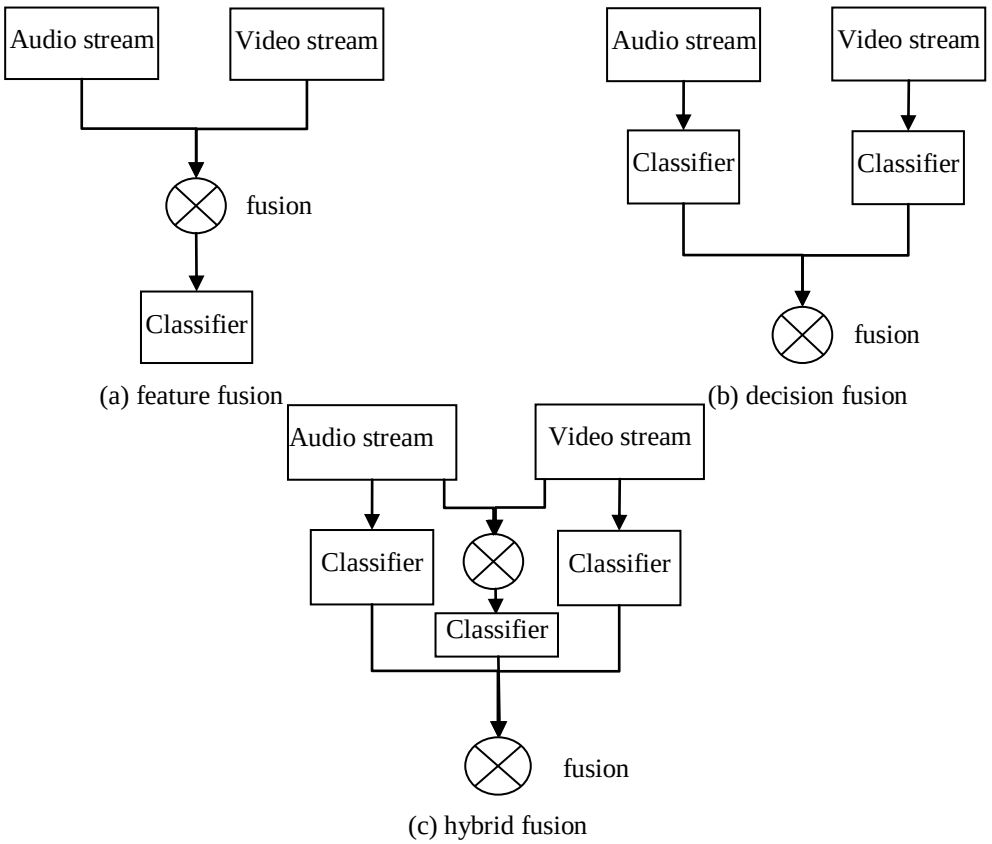


Figure 2.8 Depiction of the alternative types of Audio-visual integration

2.5.1 Feature fusion

Feature fusion is the most straight forward form of modality integration. In this method, a single classifier is trained on data obtained by a simple concatenation of audio and video data or their transformation [2], [7]. There are two common implementation approaches. In the first approach, data from both the audio and visual modalities are combined directly and, using suitable dimensionality reduction tools, mapped to a common lower dimensional space with little correlation and a small number of dimensions. In the second approach, instead of combining raw audio and visual data, features extracted from the two speech modalities are concatenated to form the audio-visual feature vector. Although simple to implement, this type of integration suffers from a number of limitations, such as the audio and visual features having different dynamic range, the time offset often found to exist between audio and video signals and the absence of a one-to-one mapping from phoneme to viseme set [46]. As the data or features are combined directly, the asynchrony between the

two streams can't be adequately modelled. Also training a single set of units (phonemes or visemes) will favour one modality over the other.

2.5.2 Decision fusion

In the decision fusion approach, audio and video streams are used to train two separate classifiers, one for each modality [3], [46]. The most popular scheme used to combine the recognition results from the two modalities uses the classification value obtained from the class conditional probabilities of the individual modality classifiers based on stream reliability scores. As the audio and visual streams have different speech information content and speech discriminative performance, and also the two streams are affected by different types of noise, the use of reliability measures provide better control of the contribution of each modality in calculating the final likelihood score. In decision fusion, both audio and video channel noise and the reliability of the visual ROI extraction can all be modelled by using appropriate stream weights for each modality.

2.5.3 Hybrid fusion

In hybrid fusion, the audio and visual modalities are integrated at a stage intermediate between the two extremes of feature and decision fusion. Although there is a range of possible levels at which integration could take place, most commonly hybrid fusion occurs at state level due to its simplicity of implementation in a multi-stream HMM framework [47]. Hybrid fusion thus attempts to exploit the individual advantages of both feature and decision fusion, in particular capturing the mutual dependencies of the audio and visual modalities while at the same time giving a better control of modality reliability compared to feature fusion [2].

2.6 TYPES OF CLASSIFIER

The overall performance of any speech recognition system is greatly dependent on the classifier adopted. In AVASR systems research, the classification tools used tend to be the same as those found in the audio-only speech recognition literature. The dynamic time warping (DTW) [48] algorithm has been historically used for audio-only speech recognition. It tracks the similarities between two time series that differ in speed or time and adopts a dynamic programming approach to optimize the match

between two time series within certain constraints. Another popular tool is linear discriminant analysis (LDA) [6], [49], a statistical pattern recognition technique that classifies objects on the basis of a set of features representing these objects. LDA falls into the category of supervised classification, as the output of the classifier is one of the set of pre-defined classes. LDA-based classifiers do not make use of a language model and classification is based solely on the basis of acoustic evidence. This makes LDA classifiers unsuitable for complex tasks like continuous speech recognition, where the use of lexicon and language models are usually helpful to guide the recognition process and greatly improve the recognition performance. Apart from its use as a classifier, LDA is most commonly used in AVASR literature for dimensionality reduction [1], [7], [11], as it maximizes the variance between different classes while minimizing the variance between members of same class. At present, the most popular classifiers for speech recognition are artificial neural networks (ANNs) and hidden Markov models (HMMs) and their variants. Of the two approaches, HMMs are the more commonly used due to their simplicity of implementation, ease of training and computational efficiency.

2.6.1 Artificial neural networks (ANNs)

ANNs are models that imitate the human brain activity and consist of a set of interconnected ‘neurons’, whose outputs are formed by taking the product of a weighted sum of its inputs before applying either a linear or non-linear activation function. Neurons in a network can be connected in a number of alternative ways often using a layered architecture. The most popular architecture is the feed-forward architecture with a single hidden layer, as shown in Figure 2.9, where a neuron is connected to each neuron in the previous layer and each connection is associated with a weight that can be adjusted during the training process.

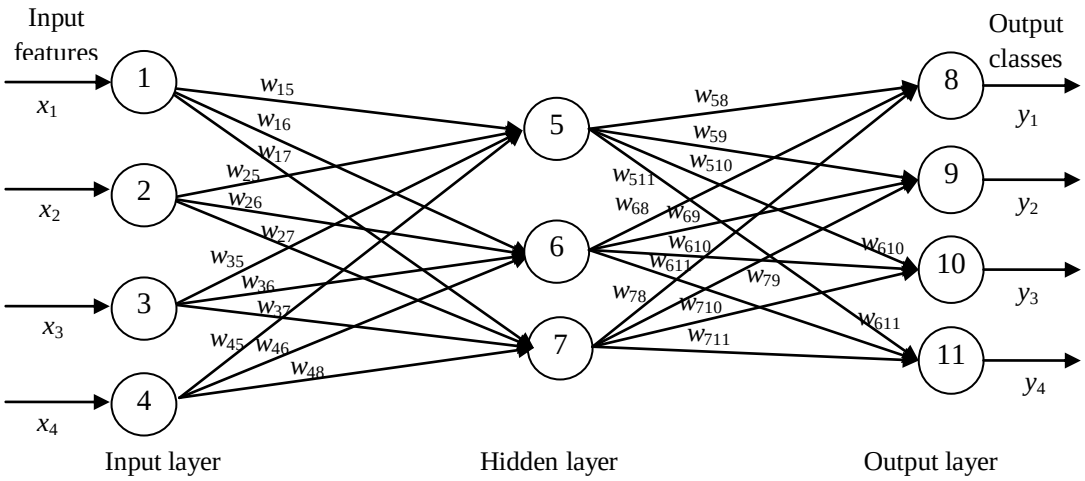


Figure 2.9 Feed-forward artificial neural network

ANNs are found to perform well when larger quantities of training data are available and in applications that require only a limited speech vocabulary. ANNs outperform HMMs on phoneme recognition and small vocabulary tasks, but ANNs perform less well for large vocabularies and on continuous speech recognition applications due to the more effective language modelling capabilities of HMMs [50]. Hybrid approaches that make use of both ANN and HMM classification have also been reported to give better performance than individual ANN or HMM classifiers [23].

2.6.2 Hidden Markov models (HMMs)

HMMs are statistical models suitable for performing pattern recognition of sequential data [11] and are the most commonly-used classifiers in audio-only and audio-visual speech recognition [4], [31]. HMMs are able to model any time series using two stochastic variables, where the first variable models the state transition probability between hidden states while the second models the probability of state output observation. An example of a four state left-right HMM is shown in Figure 2.10.

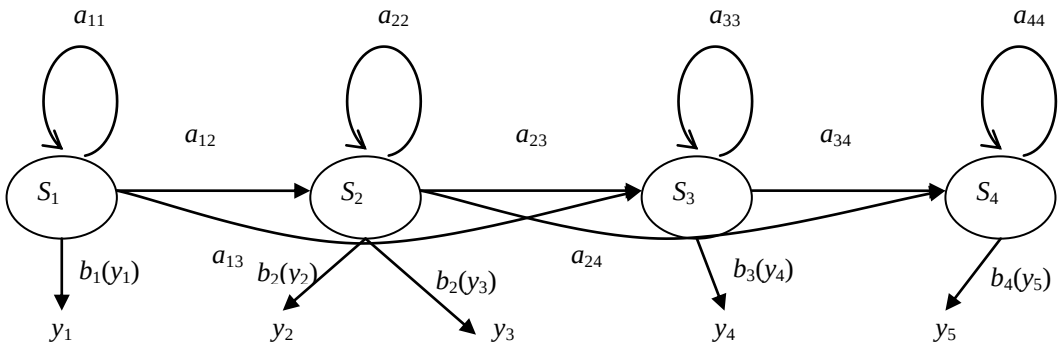


Figure 2.10 Four-state left-right HMM

In HMMs, each state transition from state S_i to S_j is associated with a transition probability a_{ij} whereas each state j generates an output probability distribution $b_j(y_k)$. In their application to speech recognition, HMMs are used to produce both an acoustic model based on the features extracted from the speech signal and a language model based on the language grammar. A model is normally produced for each of the speech units (phoneme/viseme) and these are concatenated to form an HMM for a word or a sequence of words. Variants of HMMs, such as state synchronous multi-stream HMMs and product HMMs, have been used to incorporate the concepts of early and intermediate level integration [2], [6]. A brief introduction to the HMM theory and its applications to speech recognition, and the HMM based speech recognition toolkit (HTK) [51] is provided in chapter 3, while a more detailed discussion on HMM can be found in [4] and [51].

2.7 AUDIO-VISUAL DATABASES

The research in audio-only speech recognition is relatively mature compared to AVASR research and while databases for audio-only speech recognition are abundantly available, there are few databases available for AVASR research. Most of these audio-visual databases have been developed by individual researchers or small research groups and suffer from a number of limitations, such as only showing sequences from a small number of speakers, being of short duration, audio and video stream asynchrony, limited phonetic coverage, having been designed for specific recognition tasks such as isolated word recognition, or having only a small vocabulary. These issues give rise to practical problems in their use such as the models produces being undertrained or lacking generality [2], [3], [7]. In addition,

AVASR database capture often requires expensive hardware to achieve high-quality image capture, additional data storage capacity, synchrony of audio and video streams, as well as needing to satisfy privacy issues related to the use of video information.

A number of databases have been developed in various languages and for a range of applications and the principal databases that have been used in AVASR research are given in Table 2.1.

Table 2.1 Popularly-used audio-visual databases

Database	Language	Task	Number of speakers	Reference
XM2VTS	UK English	Isolated digits	295	[52]
University of Sheffield	UK English	Isolated letters	34	[53]
Tulips1	US English	Isolated digits	12	[54]
IBM	US English	Continuous digits	50	[55]
AV-ViaVoice	US English	Continuous speech	290	[2]
AV-TIMIT	US English	Continuous speech	1	[56]
VidTIMIT	Australian English	Continuous speech	43	[57]
ICP	French	Vowels	1	[58]
M2VTS	French	Isolated digits	37	[59]
ATR	Japanese	Isolated words	1	[60]

AV-ViaVoice and VidTimit are suitable for large vocabulary continuous speech recognition and VidTIMIT is used in the research in this thesis. A detailed discussion

of VidTimit database is given in chapter 3 and the use of various component parts of VidTIMIT used in the experiments in this work are discussed in their respective chapters.

2.8 SUMMARY

In this chapter, the architecture of AVASR systems has been discussed. Current approaches reported in the AVASR literature and their relative advantages and disadvantages have been identified. After this general overview of AVASR systems, chapter 3 discusses some important concepts used in AVASR and also in this research, in further details.

Some of component parts of AVASR systems, such as classifier methods and modality fusion are multidisciplinary research areas while the audio front-end design and ROI detection and extraction are performed by approaches borrowed from other research areas such as audio-only ASR and image analysis research. The main focus of AVASR is the extraction of speech informative visual features to complement and supplement the audio steam, particularly when the audio channel is noisy. As the quality of extracted visual feature values are greatly dependent on the accurate extraction of the mouth ROI, it is potentially beneficial to view the ROI extraction task from an AVASR perspective, in particular to exploit the information available in video sequences, in contrast to image analysis approaches where the image segmentation is based only on information available in individual images. In this thesis a novel motion based approach is used for both accurate ROI extraction and for the generation of high quality visual features. In addition, a new frequency-band based approach to the extraction of appearance-based visual feature is also investigated. A suitable database has been identified for use in the experimental work performed in this thesis.

2.9 REFERENCES

[1] Connell, J. H., Haas, N., Marcheret, E., Neti, C., Potamianos, G., and Velipasalar, S. (2003), "A Real-Time Prototype for Small-Vocabulary Audio-Visual ASR", *Proceedings of the International Conference on Multimedia and Expo (ICME)*, Baltimore, Maryland, vol. II, pp. 469-472.

[2] Potamianos, G., Neti, C., Luettin J., and Matthews, I. (2004), "Audiovisual automatic speech recognition: An overview", Bailly, G., Bateson, E. V., and Perrier, P. (Eds.), *Issues in Visual and Audio-Visual Speech Processing*, MIT Press.

[3] Stork, D. G., and Hennecke, M. E. (1996), "Speechreading: An overview of image processing, feature extraction, sensory integration and pattern recognition techniques", *Proceedings of 2nd International Conference on Automatic Face and Gesture Recognition*, Killington, VT , USA, pp. XVI–XXVI.

[4] Rabinar, L.R. (1989), "A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286.

[5] Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. (2003), "Recent advances in the automatic recognition of audiovisual speech", *Proceeding of IEEE*, vol. 91, no. 9, pp. 1306-1326.

[6] Nefian, A. V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., and Murphy, K. (2002), "A Coupled HMM for Audio-Visual Speech Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, Florida, vol. 2, pp. 2013-2016

[7] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J. (2000), "Audio-Visual Speech Recognition", *Workshop 2000 Final Report*, Centre for Language and Speech Processing, JHU, Baltimore, MD.

[8] Picone, J. W., (1993), "Signal modelling techniques in speech recognition", *Proceedings of the IEEE*, vol. 81, no. 9, pp. 1215-1247.

[9] Shiell, D. J., Terry, L. H., Aleksic, P., and Katsaggelos, A. K. (2009), "Audio-Visual and Visual-only Speech and Speaker Recognition- Issues about theory, system design, and implementation", Liew, A., and Wang, S. (Eds.), *Visual Speech Recognition: Lip Segmentation and Mapping*, Hershey, PA, IGI Global, pp. 1-38.

[10] Chou W., and Juang B. W., (Eds.), (2003), "*Pattern Recognition in Speech and Language Processing*", CRC Press.

[11] Gauvain, J., and Lamel, L. (2000), "Large vocabulary continuous speech recognition: Advances and applications", *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1181-1200.

[12] Davis, S. B., and Mermelstein, P. (1980), "Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357-366.

[13] Young, S. (1996), "A review of large vocabulary continuous speech recognition", *IEEE Signal Processing Magazine*, pp. 45-57.

[14] Farooq, O., and Datta, S. (2002), "Speech recognition with emphasis on wavelet based feature extraction", *IETE Journal of Research*, vol. 48, no. 1, pp. 3-13.

[15] Tan, B. T., Fu, M., Spray, A., and Dermody, P. (1996), "The use of wavelet transform in phoneme recognition", *Proceedings of 4th International Conference on Spoken Language Processing*, Philadelphia, PA, USA, vol. 4, pp. 2431-2434.

[16] Long, C. J., and Datta, S. (1996), "Wavelet Based Feature Extraction for Phoneme Recognition", *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, PA, USA, vol. 1, pp. 264-267.

[17] Zheng, F., Zhang, G., and Song, Z. (2001), "Comparison of Different Implementations of MFCC", *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582-589.

[18] Chibelushi, C. C., Deravi, F., and Mason, J. S. D. (2002), "A Review of Speech-Based Bimodal Recognition", *IEEE transaction on multimedia*, vol. 4, no. 1, pp. 23-37.

[19] Li, X., and Kwan, C. (2005), "Geometric Feature Extraction for Robust Speech Recognition", *Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, pp. 558-562.

[20] Cootes, T. F., Taylor, C. J., and Edward, G. J. (1998), "Active Appearance Models", Lecture Notes in Computer Science, Springer, vol. 1407, pp. 484-498.

[21] Luettin, J., and Thacker, N. A. (1997), "Speechreading using probabilistic models", *Journal of Computer Vision and Image Understanding*, vol. 65, no. 2, pp. 163-178.

[22] Stork, D. G., and Hennecke, M. E., (Eds.) (1996), "*Speechreading by Humans and Machines*", Berlin, Germany: Springer.

[23] Heckmann, M., Berthommier, F., and Kroschel, K. (2001), "A Hybrid ANN/HMM Audio-Visual Speech Recognition System", *Proceedings of International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 190-195.

[24] Petajan, E., Bischoff, B., and Bodoff, D. (1988), "An Improved Automatic Lipreading System to Enhance Speech Recognition", *Proceedings of SIGCHI conference on Human factors in computing systems***,** Washington, D. C., United States, pp. 19-25.

[25] Jian, Z., Kaynak, M. N., Vheok, A. D., and Chung, K. C. (2001), "Real-Time Lip Tracking for Virtual Lip Implementation in Virtual Environments and Computer Games", *proceedings of 10th IEEE International conference on Fuzzy systems*, Melbourne, Australia, pp. 1359-1362.

[26] Chiou, G. I., and Hwang, J. -N. (1997), "Lipreading from color video", *IEEE Transaction on Image Processing*, vol. 6, no. 8, pp. 1192-1195.

[27] Zhang, X., Mersereau, R. M. (2000), "Lip Feature Extraction towards an Automatic Speechreading System", *Proceedings of International Conference on Image Processing*, Vancouver, BC, Canada, vol. 3, pp. 226-229.

[28] Cristinacce, D., and Cootes, T. F. (2006), "Facial Feature Detection and Tracking with Automatic Template Selection", *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, pp. 429-434.

[29] Reisfeld, D., and Yeshurun, Y. (1992), "Robust Detection of Facial Features by Generalised Symmetry", *Proceedings of 11th International Conference on Pattern Recognition*, The Hague , Netherlands, vol. 1, pp. 117-120.

[30] Chandramohan, D., and Silsbee, P. L. (1996), "A Multiple Deformable Template Approach for Visual Speech Recognition", *Proceedings of Fourth International Conference on Spoken Languages*, Philadelphia, PA , USA, vol. 1, pp. 50-53.

[31] Steifelhagen, R., Yang, J., and Meier, U. (1997), "Real Time Lip Tracking for Lipreading", *Proceedings of Eurospeech '97*, pp. 2007-2010.

[32] Yang, J., and Waibel, A. (1996), "A Real-Time Face Tracker", *Proceedings of 3rd IEEE Workshop on Applications of Computer Vision*, Sarasota, FL , USA, pp. 142-147.

[33] Hsu, R. -L., Abdel-Mottaleb, M., and Jain, A. K. (2002), "Face Detection in Color Images", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696-706.

[34] Nilsson, M., Nordberg, J., and Claesson, I. (2007), "Face Detection Using Local SMQT Features and Split up SNoW Classifier", *proceedings of IEEE international conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 589-592.

[35] Bourel, F., Chibelushi, C. C., and Low, A. A. (2000), "Robust Facial Feature Tracking", *Proceedings of 11th British Machine Vision Conference*, Bristol, England, pp. 232-241.

[36] Kass, M., Witkin, A., and Terzopoulos, D. (1988), "Snakes: Active Contour Models", *International Journal of Computer Vision*, vol.1, no. 4, pp. 321-331.

[37] Drugman, T., Gurban, M., and Thiran, J. (2007), "Relevant feature selection for audio–visual speech recognition", *proceedings of the IEEE 9th Workshop on Multimedia Signal Processing (MMSP)*, Crete, Greece, pp. 179-182.

[38] Arsic, I., and Thiran, J. P. (2006), "Mutual Information Eigenlips for Audio-Visual Speech Recognition", *14th European Signal Processing Conference (EUSIPCO)*, Lecture Notes in Computer Science.

[39] Petar, S. A., and Aggelos, K. K. (2004), "Comparison of Low and High-level Visual Features for Audio-Visual Continuous Automatic Speech Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 917-920.

[40] Katz, M., Meier, H. G., Dolfing, H., Klakow, D. (2002), "Robustness of Linear Discriminant Analysis in Automatic Speech Recognition", *Proceedings of 16th International Conference on Pattern Recognition*, Quebec, Canada, vol. 3, pp. 30371-30374.

[41] Petajan, E. D. (1984), "Automatic Lipreading to Enhance Speech Recognition", *Proceedings of the IEEE Communication Society Global Telecommunications Conference,* Atlanta, Georgia, pp. 265-272.

[42] Chan, M. T. (2001), "HMM-Based Audio-Visual Speech Recognition, Integrating Geometric and Appearance-Based Visual Features", *Proceedings of Fourth Workshop on Multimedia Signal Processing*, pp. 9-14.

[43] Mok, L. L., Lau, W. H., Leung, S. H., Wang, S. L., and Yan, H. (2004), "Person Authentication Using ASM Based Lip Shape And Intensity Information", *International Conference on Image Processing*, vol. 1, pp. 561-564.

[44] Saenko, K., Darrell, T., Glass, J. R. (2004), "Articulatory features for robust visual speech recognition", *Proceedings of the 6th international conference on Multimodal interfaces - ICMI '04*, New York, pp. 152-158.

[45] Schwartz, J. L., Ribes, J. R., and Escudier, P. (1998), "Ten Years after Summerfield: A Taxanomy of Model for Audio-Visual Fusion in Speech Recognition", Campbell, R., Dodd, B., and Burnham, D. (Eds.), *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers, pp. 85-108.

[46] Verma, A., Faruquie, T., Neti, C., Basu, S., and Senior, A. (1999), "Late Integration in Audio-Visual Continuous Speech recognition", *Proceedings of Automatic Speech Recognition and Understanding Workshop*, vol. 1, pp. 71-74.

[47] Chu, S.M., Marcheret, V. L. E., Neti, C., and Potamianos, G. (2004), "Multistage Information Fusion for Audio-Visual Speech Recognition", *Proceeding of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 3, pp. 1651-1654.

[48] Liu, Y., Lee, Y. C., Chen, H. H., Sun, G. Z. (1992), "Speech Recognition Using Dynamic Time Warping with Neural Network Trained Templates", *International Joint Conference on Neural Networks*, vol. 2, pp. 326-331.

[49] Balakrishnama, S., Ganapathiraju, A., and Picone, J. (1999), "Linear discriminant analysis for signal processing problems", *Proceedings of the IEEE Southeastcon*, pp. 78-81.

[50] Zue, V., Cole, R., and Ward, W. (1996), "Speech Recognition", Cole, R. A., Mariani, J., Uszkoreit, H., Zaenen A., Zue, V., Varile, G. B., and Zampolli, A. (Eds.), *Survey of the state of the art in human language technology*, Cambridge University Press, pp. 1-62.

[51] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev V., and Woodland, P. (2006), *The HTK Book V3.4.*

[52] Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G. (1999), "XM2VTS: The extended M2VTS database", *Proceedings of International Conference on Audio video-based Biometric Person Authentication*, Washington, DC, USA, pp. 72-76.

[53] Cooke, M., Barker, J., Cunningham, S., and Shao, X. (2006), "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition", *Journal of Acoustical Society of America*, vol. 120, no. 5, pp. 2421-2424.

[54] Movellan, J. R. (1995) "Visual Speech Recognition with Stochastic Networks", Tesauro, G., Touretzky, D., and Leen, T. (Eds.), *Advances in Neural Information Processing Systems 7*, MIT Press Cambridge.

[55] Potamianos, G., and Neti, C. (2001), "Automatic speechreading of impaired speech", *Proceedings of international Conference on Audio-Visual Speech Processing*, Aalborg, Denmark, pp. 177-182.

[56] Hazen, T. J., Saenko, K., La, C. H., and Glass, J. (2004), "A segment based audio-visual speech recognizer: Data collection, development, and initial experiments", *Proceedings of ICMI*, 2004, pp. 235-242.

[57] Sanderson, C., and Paliwal, K. K. (2004), "Identity Verification Using Speech and Face Information", *Digital Signal Processing*, vol. 14, no. 5, pp. 449-480.

[58] Heckmann, M., Kroschel, K., Berthommier, F., and Savariaux, C., (2002), "DCT-based video features for audio-visual speech recognition", *Proceedings of International Conference on Spoken Language Processing*, Denver, USA, pp. 1925-1928.

[59] Pigeon, S., and Vandendorpe, L. (1999), "The M2VTS multimodal face database (release 1.00)", *Proceedings of the First International Conference on Audio and Video-based Biometric Person Authentication*, Crans-Montana, Switzerland, pp. 403-409.

[60] Nakamura, S. (2001), "Fusion of audio-visual information for integrated speech processing", Bigun, J., and Smeraldi, F. (Eds.), *Audio-and Video-Based Biometric Person Authentication*, Berlin, Germany: Springer, pp. 127-143.

CHAPTER 3

AN OVERVIEW OF IMPORTANT CONCEPTS IN AVASR

3.1 INTRODUCTION

Chapter 2 introduced the structure of a typical AVASR system and its components parts, and the approaches taken in the literature to achieve their implementation. This chapter concentrates on developing the background to the methods adopted in the work presented in this thesis.

AVASR systems contain two streams of information, namely audio and visual rather than the one audio stream used in audio-only ASR. Since ASR generally has its basis in pattern recognition research, this approach has also been adopted in AVASR work. The chapter is organized as follows. Section 3.2 discusses the concept of image transformations and their use in AVASR, with emphasis on the discrete cosine transform (DCT) and discrete wavelet transform (DWT). In order to make the quality of data to be processed more manageable, dimensionality reduction is normally applied in AVASR implementations. Section 3.3 discusses the concept of dimensionality reduction and describes the operation of the two most popular dimensionality reduction techniques, principal component analysis (PCA) and linear discriminant analysis (LDA). As audio and video streams have different speech units, namely phonemes and visemes respectively, a form of mapping between the two is needed for their use in AVASR; the existing approaches are introduced in section 3.4. A Hidden Markov Model (HMM) based classifier is used in this work and Cambridge University's toolkit, HTK, is used for training, recognition and analysis of the results. Section 3.5 describes the use of HMM for speech recognition and section 3.6 provides a brief introduction to HTK. Lastly, VidTIMIT, the audio-visual database used in this research is described in section 3.7.

3.2 IMAGE TRANSFORMATION

Transforms are mathematical operations that map data from one domain to another. Transformations may be carried out for the purposes of data compression, to reduce computational complexity or to view certain hidden patterns in the data. For example, compressed images (such as JPEG) [1] require less storage capacity than the same image in an uncompressed format (such as bit map pattern). Similarly, Laplace transforms can make handling of differential equations easier as the algebraic operation can be applied in transform domain rather than in the original time domain. Transformations may also aid visualization, for example by using the logarithm of intensity in a plot of luminance readings whose values cover a wide dynamic range. Important considerations when selecting a data transformation are its generality, compactness and computational feasibility.

In signal analysis research, transforms are used for accessing specific aspects of a signal. For example, the temporal variations in a signal are more readily available in the time domain, while the different frequency components that make up a signal can more easily be assessed when viewed in the frequency domain. The frequency domain representation is usually preferred for signal analysis purposes as this provides information best able to characterize the data [1].

Audio-only ASR solutions, perhaps due to their relative maturity, have generally settled on the use of a single set of feature types, namely the Mel-frequency cepstral coefficients (MFCC) derived from the short time Fourier transform (STFT) of the audio speech signal. Conversely, in video feature extraction, a range of transformation techniques, such as discrete cosine transform (DCT) [2], discrete wavelet transform (DWT) [3], [4], principal components analysis (PCA) [5] and linear discreminent analysis (LDA) [6], [7] are still being described in the literature. These techniques are largely inherited from techniques described in the data compression literature. In AVASR research, the DCT and DWT are the most commonly-used image transformation methods, while PCA and LDA are more popular in dimensionality reduction applications. Consequently, visual feature extraction usually takes the form of either DCT or DWT based representations followed by the application of PCA or LDA to reduce the dimensionality of final feature vector [8], [9].

3.2.1 Discrete cosine transform

The DCT is one of the most popular tools used in image analysis research. It describes an image in terms of its frequency components and is widely used in image reconstruction, filtering and image compression applications. The use of the DCT in pattern recognition research is well established with the majority of AVASR systems using the DCT transformation as a first stage of the visual front-end [10], [11]. The DCT transformation is lossless and an inverse transform can be performed to reconstruct the original image from the DCT coefficients. The DCT is often used to exploit the inter-pixel and inter-frame redundancies present in images and in video data for compression. A detailed discussion on DCT theory and its applications to image and video analysis can be found in [12].

The number of frequency components generated in the DCT transform domain corresponds to the dimensionality of the input signal. Thus the output of the DCT transform on a sequence of length *N* will be a sequence of same length *N*. For a two dimensional image signal of dimensionality *M*x*N*, the output of the DCT transform is a matrix of the same order *M*x*N*. As the DCT is a separable transform, the two-dimensional DCT of an image can be performed by applying a one-dimensional DCT in one dimension followed by a second one-dimensional DCT performed in the second.

A one-dimensional DCT $y[f]$ of a sequence $x[n]$ of length N can be performed as follows

$$y[f] = r[f] \sum_{n=0}^{N-1} x[n] \cos\left[\frac{\pi(2n+1)f}{2N}\right] \quad f = 0,1,2,\dots,N-1 \qquad (3.1)$$

where the coefficient $r[f]$ is defined as

$$r[f] = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } f = 0 \\ \sqrt{\frac{2}{N}} & \text{for } f = 1,2,\dots N-1 \end{cases} \qquad (3.2)$$

The first coefficient f [0] in the DCT domain represents the mean value (or energy) of the sequence known as DC component.

A plot of the term $\cos\left[\frac{\pi(2n+1)f}{2N}\right]$ for N=8 is shown in Figure 3.1. Each one of these plots represents the waveforms associated with one value of f. These are called the one-dimensional cosine basis function. The DCT performs the matching between the input signal and these basis functions. The output values of the DCT transform represent what proportion each of the basis functions contributes in forming the signal.

Figure 3.1 One-dimensional DCT basis functions

The two-dimensional DCT $y[u,v]$ of a matrix $x[m,n]$ of dimension MxN is given by

$$y[u,v] = r[u]r[v]\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x[m,n]\cos\left[\frac{\pi(2m+1)u}{2M}\right]\cos\left[\frac{\pi(2n+1)v}{2N}\right] \tag{3.3}$$

$$u = 0,1,\dots,M-1$$

$$v = 0,1,\dots,N-1$$

The coefficients $r[u]$ and $r[v]$ are defined as

$$r[u] = \begin{cases} \sqrt{\frac{1}{M}} & \text{for } u = 0 \\ \sqrt{\frac{2}{M}} & \text{for } u = 1,2,\dots,M-1 \end{cases} \tag{3.4}$$

and

$$r[v] = \begin{cases} \sqrt{\frac{1}{N}} & \text{for } v = 0 \\ \sqrt{\frac{2}{N}} & \text{for } v = 1,2,\dots,N-1 \end{cases} \tag{3.5}$$

The two-dimensional basis functions for M=N=8 are shown in Figure 3.2.

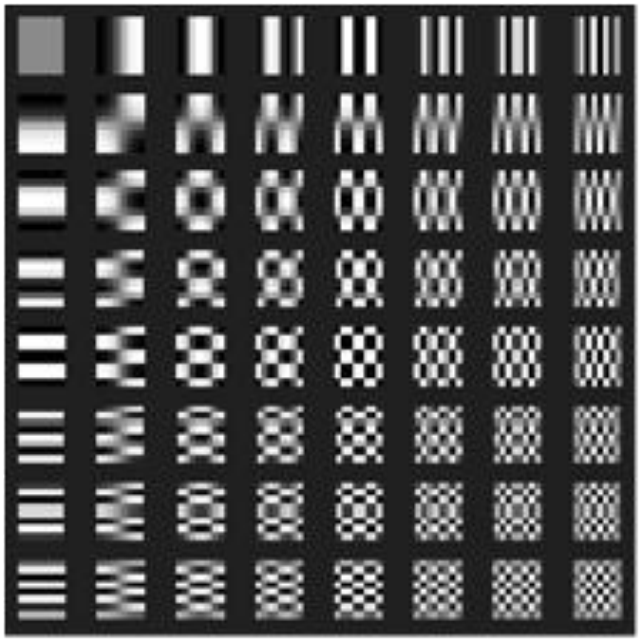


Figure 3.2 two-dimensional DCT Basis function

As the neighbouring pixels in an image are generally highly correlated, the DCT transform coefficients of the high frequency components are usually small and, as they contribute little to the perceived image, are often discarded; this being known as *lossy* compression. In this process, although some information contained in the original image is lost, frequencies containing important information are retained, thus resulting in little or no effect on the perceived visual quality of the image [13]. Such inter-pixel redundancy has also been applied in DCT based AVASR systems in an attempt to achieve a compact representation of speech related information from the speaker's mouth region (ROI).

3.2.2 Discrete wavelet transform

Wavelets are commonly used to represent both the time and frequency information in a signal and can provide a more efficient representation of signals than do the discrete Fourier transform (DFT) or DCT [1], by providing information relating not only the spectral components present in a signal, but also to the time at which these spectral components exist.

Although the word wavelet was first coined by Haar in 1909 [14], the use of wavelets in science and engineering became established only after the introduction of the Fast

Fourier (FFT) algorithm in 1965 [1]. Since the 1980s wavelets have extensively been used as a signal analysis tool in audio and visual processing [15].

Background

Signals can be considered to be made up of a collection of sinusoidal waves of different frequencies. Signals with constant amplitude are considered to have zero or no frequency, whereas other signals will be composed of one or more frequency components. The Fourier transform (FT) extracts the amplitudes of the component frequencies of a signal. Mathematically it can be written as

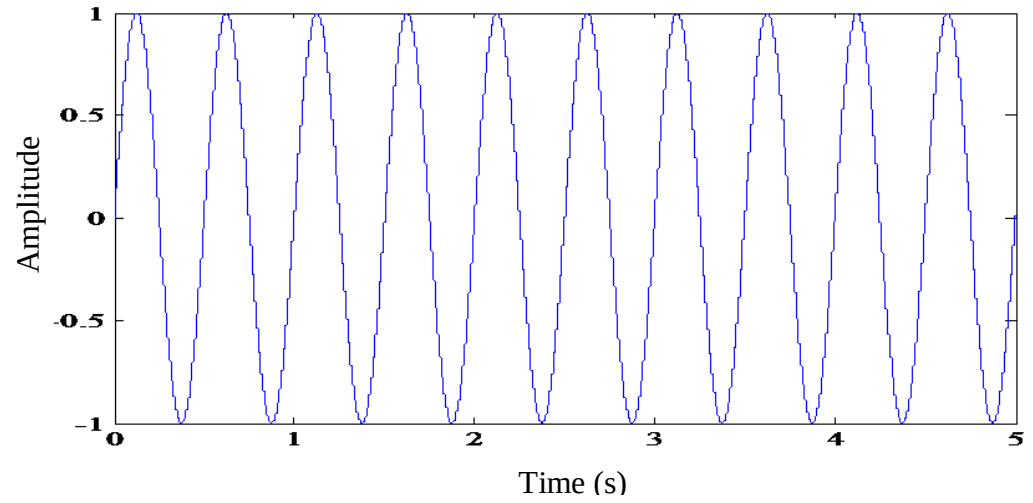
$$X(f)= \int_{-\infty}^{\infty} x(t)e^{-2\pi ft}\,dt \tag{3.6}$$

where t and f are the time and frequency; and the notation x and X are used to represent the input time domain signal and the resulting frequency domain representation respectively.

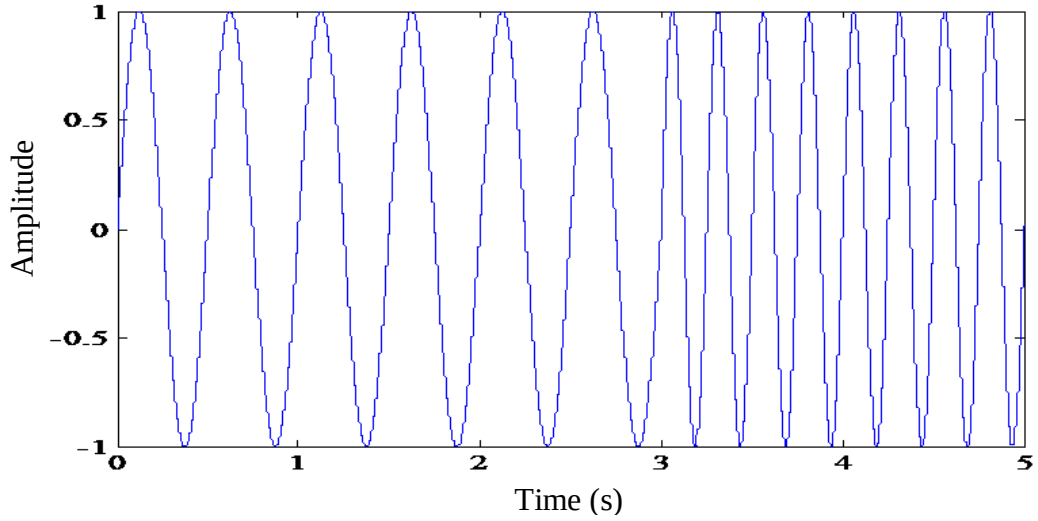
The Fourier transform is a reversible transform and the time domain signal can be recovered from the frequency domain representation using equation

$$x(t)= \int_{-\infty}^{\infty} X(f)e^{2\pi ft}\,df \tag{3.7}$$

A signal is said to be stationary if its spectral characteristics remain constant over time. Otherwise it is non-stationary. In Figure 3.3, (a) is stationary signal with a constant frequency of 4Hz while (b) is non-stationary as its frequency changes from 2Hz to 4Hz after 3 seconds.



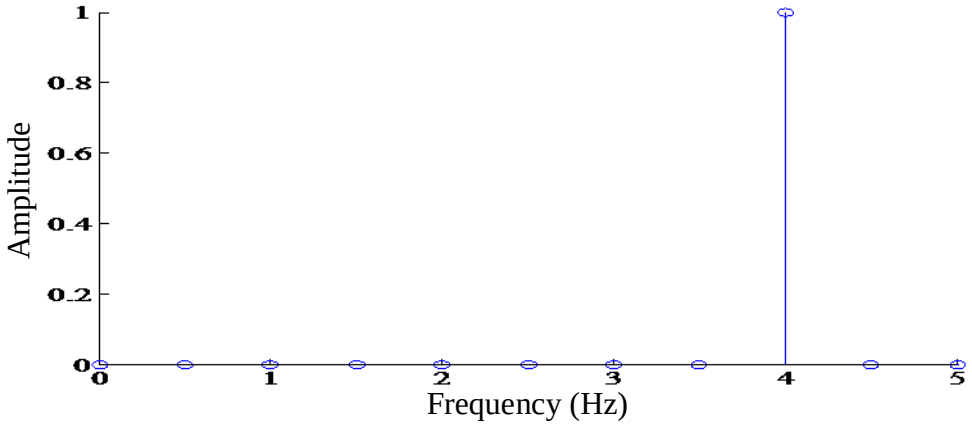
(a) Stationary signal



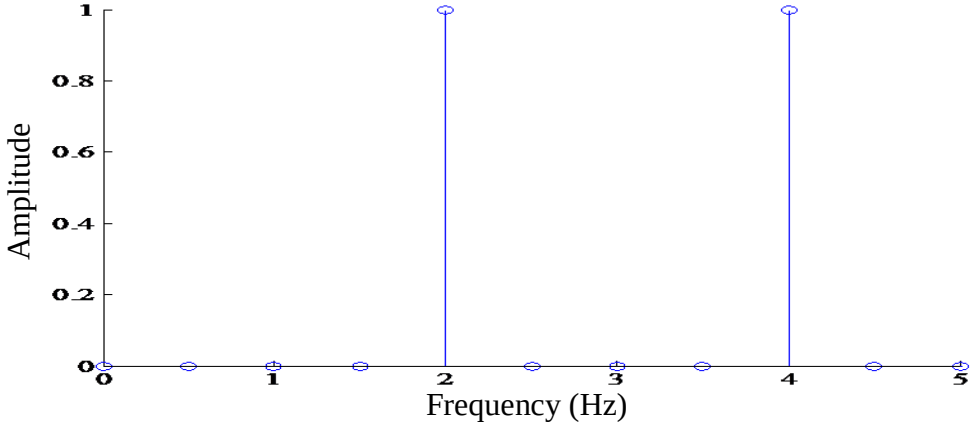
(a) Non stationary signal

Figure 3.3 Examples of stationary and non-stationary signals

Using the Fourier transform, the frequency information of the signals in Figure 3.3 can be obtained. As shown by the power spectra in Figure 3.4, in case (a) there will be a single peak on the frequency axis at 4 Hz while in case (b) there will be principally two peaks, one at 2 Hz and the second at 4 Hz (the additional frequency components caused due to the transition in frequency are ignored). This information is sufficient to determine which frequency components are present in the signal, but it does not reveal the times at which these frequency components are found in the original signal.



(a) Power spectrum of signal in Figure 3.3 (a)



(a) Power spectrum of signal in Figure 3.3 (b)

Figure 3.4 Power spectra of signals in Figure 3.3

To localize in time the frequency information in the signal, either the short time Fourier transform (STFT) or the wavelet transform (WT) can be used. The two methods are briefly described below.

Short time Fourier transform (STFT)

The limitation of the Fourier transform in not being able to represent changes in the frequency content of signals over time is addressed in the STFT in which the signal is divided into short time segments during which the signal can be considered to be stationary. A 'window' of width equal to the segment length is used to extract samples at a number of positions along the duration of the signal. The Fourier transform is applied to each individual windowed section of the signal generating a series of frequency responses. A number of window types exist, including rectangular, Hamming and Blackman windows; but in speech recognition applications, the

Hamming window is most commonly used. The window length and the frame duration are selected in pairs that result in smoothly varying estimates of the spectral components while avoiding over smoothing. A window of length about double the frame duration is commonly used, resulting in an overlap of around 50%. The resulting STFT can be written as

$$STFT_X^{(\omega)}(t',f) = \int [x(t)\cdot\omega^*(t-t')]\cdot e^{-j2\pi ft}dt \qquad (3.8)$$

where $\omega(t)$ is the window function, $x(t)$ is the input time domain signal to be transformed and t' is the time shift. The STFT is thus the FT of the product of $x(t)$ and the shifted version of window function $\omega(t)$.

The STFT so obtained includes not only the frequency components present in the signal but also the time at which these components exist. In the STFT, spectral amplitudes are plotted against both time and frequency and the tiling of time-frequency is shown in Figure 3.5, where Δt and Δf are the time and frequency resolutions respectively.

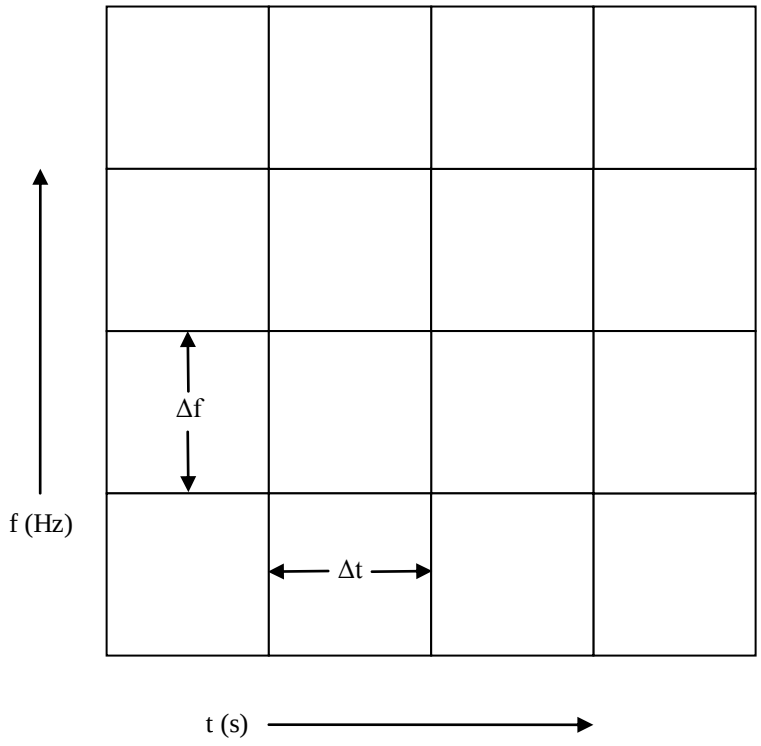


Figure 3.5 STFT based time-frequency tiling

Although STFT gives better time localization of frequency components, it suffers from the *resolution* problem. Time resolution means how well separated are the spectral values in time, while the frequency resolution indicates how well separated are the frequency components. The application of a window of finite length causes degradation in the frequency resolution as it gives a band of frequencies rather than individual frequencies. Increasing the window duration improves the frequency resolution but results in a reduction in time resolution. This time-frequency resolution conflict is related to Heisenberg's uncertainty principle [16], meaning that for a given window size high resolution can be attained either in time or frequency but not both. This time-frequency relation is mathematically given by

$$\Delta t * \Delta f \leq \frac{1}{4\pi} \tag{3.9}$$

This implies that an increase in resolution of either time or frequency will result in a decrease in resolution of the other.

Multi resolution analysis

High frequency components have better time resolution as they last for a shorter duration while low frequency components have poorer time resolution as they last for a longer duration. In the wavelet transform (WT) this time-frequency resolution problem is addressed by using variable size window, instead of the fixed size window as used in STFT. This scheme of analyzing signal at multiple resolutions is known as Multi Resolution Analysis (MRA). The time-frequency tiling for MRA is shown in Figure 3.6.

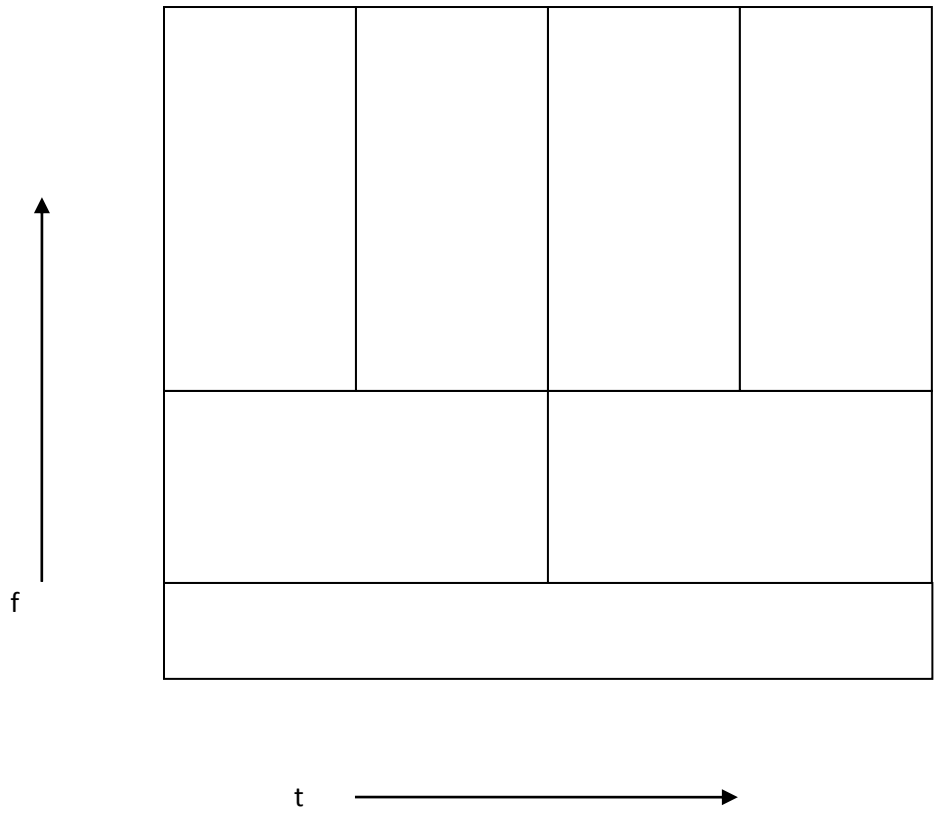


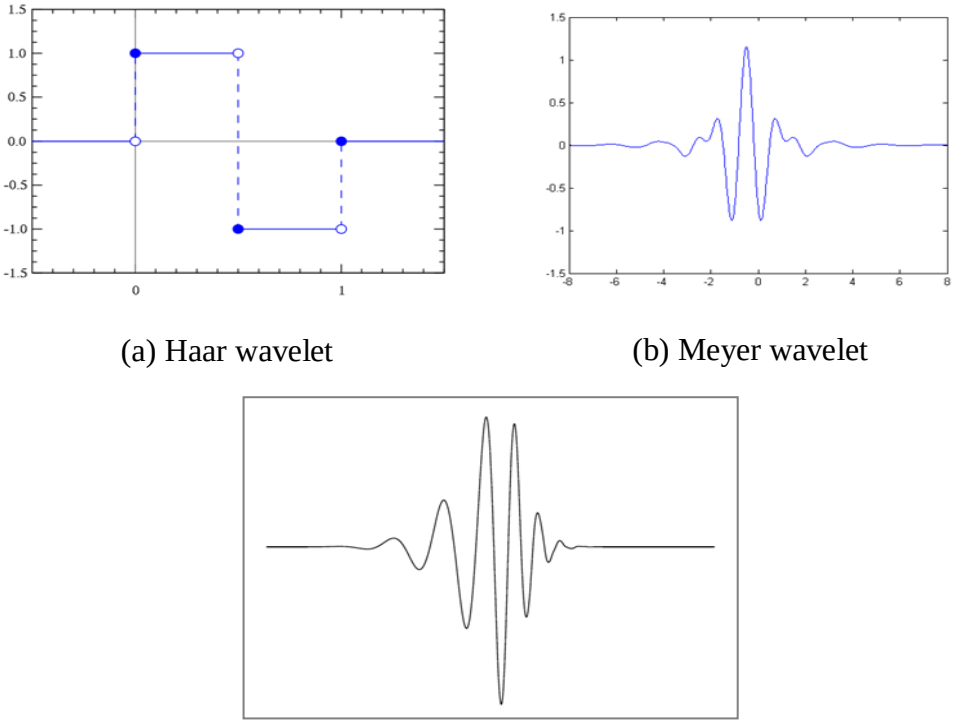
Figure 3.6 WT based time-frequency tiling (MRA)

The continuous wavelet transform (CWT) of signal $x(t)$ is given by

$$\mathrm{CWT}_{\chi}^{\varphi}(\tau,s)= \psi_{\chi}^{\varphi}(\tau,s)= \frac{1}{\sqrt{|s|}} \int x(t)\varphi^{*}\left(\frac{t-\tau}{s}\right)\mathrm{dt} \qquad (3.10)$$

Where the transformed signal depends on the two variables τ and s, termed the translation and scale variables respectively. The function φ is called the mother wavelet and can generate small waves (window) by varying τ and s, effectively determining the similarity between the waves and $x(t)$ at different scales and times. One major difference between the FT and the WT is that the FT has uses only sine and cosine as basis functions, while the WT has available an infinite set of basis functions. Examples of commonly used mother wavelets functions are Haar, Meyer and Daubechies wavelet, Figure 3.7. The discrete wavelet transform (DWT) is performed by using a dyadic scheme, where the translation and scale values are repeatedly increased by a factor of two, giving a high pass and a low pass version of

the signal. A detailed discussion of the DWT and its filtering scheme can be found in [16] and [17].



(a) Haar wavelet

(b) Meyer wavelet

(c) Daubechies wavelet of order 2

Figure 3.7 Examples of mother wavelet functions [17]

3.3 DIMENSIONALITY REDUCTION

In pattern recognition, an observation is commonly represented by a set of parameters known as features. These features are extracted from the input observation or as a result of the application of a suitable transformation in order to emphasise distinguishing characteristics. In general, increasing the number of dimensions provides additional input information and hence improves the performance of the pattern recognition system. This, however, may not always be the case, since adding features that contain no additional information not only increase the storage requirement but could also increase computation time. Furthermore, using high dimensional data for pattern recognition also increases the quantity of data required for training the recognizer and, if training data is limited, this can reduce the effectiveness of the training process and thus worsen the performance of the recognition system [18]. The source dimensionality of both audio and visual data is

high and some form of reduction is required to convert the inputs into more compact representations in which the number of dimensions is reduced to that intrinsic in the data [19]. However, a general transformation that leads to such a compact representation for a wide range of applications has not yet been identified in the literature. For data compression applications, the intrinsic dimensions are those carrying most of the information present in the original data, while for pattern recognition application they are those maximizing the discrimination between the elements of the different classes present in the source.

Dimensionality reduction methods can be grouped into two categories, namely feature selection and feature extraction techniques. In feature selection approaches, features are selected from the original data based on scores assigned using a recognition criterion. Discrimination criteria that have been used in the literature include discriminative features analysis, F-ratio and recognition rate [20], with the resulting feature set obtained depending greatly on the selection approach adopted. The main limitation of feature selection methods is that they do not consider the correlation between selected features, meaning that although the selected feature may have high discriminative power, it could be highly correlated with one or more other features and thus add little or no additional information [18]. In contrast, feature extraction methods transform data to orthogonal dimensions to reduce the correlation between the original feature set. The aim is to produce a new set of features that contain all the information present in the original set, but with a different representation that minimizes the correlation between the features. Two popular methods used for feature extraction are principal component analysis (PCA) and linear discriminant analysis (LDA) [21]. PCA transforms the original data in a manner such that the feature with the maximum data variance lies along the first dimension, the one with the second highest variance lies along the second dimension, and so on. LDA transforms the data in such a manner so as to maximize the discrimination between members of different classes while minimizing the discrimination between members of same class. A detailed survey on different types of dimensionality reduction techniques can be found in [22]. The operations of PCA and LDA and their application to data reduction are discussed briefly in the following subsections.

3.3.1 Principal component analysis

Principal component analysis (PCA) is one of the most popular linear dimensionality reduction techniques and is widely used in pattern recognition applications [23]. In PCA, the data are transformed into a transform space whose dimensions are ordered according to decreasing variance. A certain number of these dimensions, called the principal components, are then indentified as containing sufficient information to represent the original data [24]. These dimensions are considered to capture useful information to provide a distinction between the classes contained in the data and so reveal a hidden underlying pattern in the data which would be difficult to extract in the original data space. A detailed discussion on the theory, calculation and various applications of PCA can be found in [25].

For a given set of data of N dimensions, PCA finds a new space of D orthogonal dimensions ($D < N$) such that the data points mainly lie along these D dimensions. Let M observations of an N dimensional data vector $\boldsymbol{x}$ be represented by a matrix $\mathbf{X}$ of order NxM such that each column of $\mathbf{X}$ represent one observation of the data vector $\boldsymbol{x}$. Let the D principal axes be denoted by T_1, T_2, …, T_D. These principal axis could be given by the eigenvectors of the covariance matrix $\boldsymbol{S}$, such that

$$\boldsymbol{S}T_i=\lambda_i T_i \quad i=1, 2, 3….D \tag{3.11}$$

where λ_i is the i^{th} largest eigenvalue of $\boldsymbol{S}$ and

$$\boldsymbol{S} = \frac{1}{M}\sum_{j=1}^{M}(\boldsymbol{x}_j - \boldsymbol{\mu})^T(\boldsymbol{x}_j - \boldsymbol{\mu}) \tag{3.12}$$

Where $\boldsymbol{\mu}$ is the mean of the observation vectors and $\boldsymbol{x_j}$ is the j^{th} observation vector.

As the larger is the value of λ, then the larger is the variance and so the maximum data variance can be found by selecting the first few components in the projected space. A measure for representing the portion of data is the percentage variance. The projected D dimension matrix is given by

$$\mathbf{Y}=[\mathrm{Y}_1, \mathrm{Y}_2,…, \mathrm{Y}_D]=\left[T_1^T\mathbf{X}, T_2^T\mathbf{X}, T_3^T\mathbf{X}, …,T_D^T\mathbf{X}\right]=\mathbf{T}^T\mathbf{X} \tag{3.13}$$

where **T** is the transformation matrix whose columns are made of the principal axis T_i.

The *D*x*M* dimensional matrix **Y** thus obtained contains the desired principal components of input matrix **X** of dimensionality *N*x*M*. Although the features extracted using PCA have a minimum correlation along the direction of the principal axis, the approach does not guarantee the separation of classes among data as no class information is used in the PCA calculation. PCA also has a limitation of scale sensitivity implying that the principal components may be affected by the relative scaling of variables in original data.

3.3.2 Linear discriminant analysis

The transformation performed by linear discriminant analysis (LDA) is able to separate the elements of different classes while at the same time minimizing the distance between elements of same class [26]. This approach comes under the domain of supervised dimensionality reduction methods, meaning that prior knowledge of the classes present in the data is used in performing the transformation.

Let the data matrix **X** contain observation vectors from *k* classes, x_1, x_2, x_3 …x_k, each having *N* dimensions. If the j^{th} observation of class *i* is represented by x_{ij} such that i = 1, 2, …, k and j = 1, 2, …, M_i, where the M_i are the number of observations in class *i*. The mean of observations in class *i* is then given by

$$\mu_i = \frac{1}{M_i}\sum_{i=1}^{M_i} x_{ij} \tag{3.14}$$

and the covariance matrix for class *i* is given by

$$\boldsymbol{S}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} (x_{ij}-\mu_i)(x_{ij}-\mu_i)^T \tag{3.15}$$

For k classes, the within-class variance S_w is given by

$$\boldsymbol{S}_w = \sum_{i=1}^{k} \boldsymbol{S}_i \tag{3.16}$$

and the between-class variance S_b is

$$\boldsymbol{S}_b = \sum_{i=1}^{k} M_i(\mu_i - \mu)(\mu_i - \mu)^T \tag{3.17}$$

where μ is the mean of all the data given by

$$\mu = \frac{1}{M} \sum_{j=1}^{k} \sum_{i=1}^{M_i} x_{ij} \tag{3.18}$$

and M is the total number of data vectors such that $M = \Sigma M_i$ for $i = 1, 2, \ldots, k$.

The transformation from N-dimensional space to a lower D-dimensional space is performed by

$$\mathbf{Y} = \mathbf{W}^T\mathbf{X} \tag{3.19}$$

where $\mathbf{W}$ is the transformation matrix. The greatest separation between classes can be achieved by maximizing the Fisher Linear Discriminant operator

$$\mathbf{J(W)} = \frac{\mathbf{W}^{\mathrm{T}}\boldsymbol{S}_b\mathbf{W}}{\mathbf{W}^{\mathrm{T}}\boldsymbol{S}_w W} \tag{3.20}$$

The optimum $\mathbf{W}$ consists of the D largest eigenvectors, where D is the desired dimensionality of the transformed space.

3.4 PHONEME AND VISEME MAPPING

Audio and visual speech units are termed phonemes and visemes respectively [27]. A phoneme is the smallest segment of audio speech that conveys linguistic information, whereas a viseme is the smallest visually distinguishable segment of speech that may represent one or more phonemes. The difference between the lengths of the segments arises due to the fact that in practice not the entire vocal tract is visible making it impossible to identify each phoneme visually [28]. The phoneme grouping that corresponds to a viseme is determined either manually [29] or by using statistical clustering techniques [30]. While several studies investigating suitable phoneme

clustering techniques have been reported in literature [27], [28], [31], there is no general agreement among the researchers as which phoneme set actually corresponds to a specific viseme. As a result there are a number of distinct phoneme-to-viseme mappings being used in AVASR research, containing a number of visemes ranging from 12 to 20. Three of the most commonly used phoneme-to-viseme mappings are shown in Table 3.1.

Table 3.1 Examples of the phone-viseme mapping

Viseme	Phonemes	Viseme	Phoneme
Hazen *et al*. [29]			
(1)	/ax/, /ih/, /iy/, /dx/	(8)	/b/, /p/
(2)	/ah/, /aa/	(9)	/bcl/, /pcl/, /m/, /em/
(3)	/ae/, /eh/, /ay/, /ey/, /hh/	(10)	/ch/, /jh/, /sh/, /sz/
(4)	/el/, /l/	(11)	/t/, /d/, /th/, /dh/, /g/, /k/
(5)	/er/, /axr/, /r/	(12)	/gcl/, /kcl/, /ng/
(6)	/y/	(13)	/f/, /v/
(7)	/s/, /z/, /epi/, /tcl/, /dcl/, /n/, /en/	(14)	/aw/, /uh/, /uw/, /ow/, /ao/, /w/, /oy/
Lewis *et al*. [32] (consonants)			
(1)	/p/, /b/, /m/	(6)	/s/, /z/
(2)	/f/, /v/	(7)	/l/
(3)	/th/, /dh/	(8)	/r/
(4)	/sh/, /zh/	(9)	/d/, /t/, /n/, /g/, /k/, /ng/, /h/
(5)	/w/		
Yau *et al*. [33] (consonants)			

(1)	/p/, /b/, /m/	(6)	/sh/, /j/, /ch/
(2)	/f/, /v/	(7)	/s/, /z/
(3)	/th/, /dh/	(8)	/n/, /l/
(4)	/t/, /d/	(9)	/r/
(5)	/k/, /g/		

Example of mouth shapes and their corresponding phonemes are given in appendix I.

3.4.1 Phoneme and viseme based AVASR

Although audio and visual speech have different sets of units (phoneme and viseme), in AVASR research the recognition is generally performed using phonemes only [34]. Both audio and video streams are used to train models for a set of phonemes and their context-dependent bi-phonemes and tri-phonemes. In early integration, as the data or features are combined before passing into the recognizer for training or testing, it is not possible to have two different sets of units for audio and visual streams. However, in a late integration approach, the use of separate phoneme and viseme models for audio and visual streams has been studied showing no significant improvement in performance [28]. An inspiration for the use of separate audio and visual units is the inherent asynchrony in audio and visual streams due to the inertia of articulators due to which video speech lags slightly behind the audio speech. The asynchronous modelling of audio and video streams is reported by Hazen *et al*. in [29], who showed that there is no performance gain in the approach with respect to synchronous modelling. In this work, for all audio, video and audio-visual ASR tasks, the speech units used are phonemes and their bi-phonemes and tri-phonemes.

3.5 HIDDEN MARKOV MODEL (HMM)

Hidden Markov Models (HMMs) are statistical models suitable for modelling and recognition of sequential data and are most commonly used technique in speech recognition applications today [35]. HMMs drive two stochastic processes: one models the transition between Markov chain of hidden states while the second models the output observation omitted for being in a specific state. The transition from state *i* to

state j is governed by the state transition probability a_{ij} while the output observations k from state j is given by probability distribution $b_j(k)$.

Consider a system with N number of distinct states $S = \{S_1, S_2, S_3, ..., S_N\}$ such that the system changes its state at regular time intervals t = 1, 2...T. In a Markov chain it is assumed that the state of the system at any time t depends only on previous state, and is independent of all the states before the previous. If S_t represents the state of the system at any time t, and $S_t=j$, $S_{t-1}=i$ and $S_{t-2}=k$ then the Markov process can be probabilistically described as

$$\begin{aligned} & P[S_t = j \mid S_{t-1} = i, S_{t-2} = k, \dots] \\ & = P[S_t = j \mid S_{t-1} = i] \end{aligned} \tag{3.21}$$

If the right-hand side of equation (3.21) is independent of time then it leads to state transition probability

$$\begin{aligned} & = P[S_t = j \mid S_{t-1} = i] \\ & = a_{ij} \qquad\qquad 1 \leq \mathrm{i,j} \leq \mathrm{N} \end{aligned} \tag{3.22}$$

In systems where any state can be reached from any other state in a single step, $a_{ij} > 0$ for all i and j, while for others $a_{ij} = 0$ for one or more pairs of i, j. The state transition probabilities obey the following general rule

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{3.23}$$

The observation symbol O_k is emitted in state j according to the output probability distribution $b_j(k)$ such that

$$b_j(k) = P[O_k \mid S_t = j] \tag{3.24}$$

Depending on the nature of observation probability distributions, the HMMs can be discrete HMM (DHMM) or continuous density HMM (CDHMM).

If the probability of being in state i at the beginning of HMM chain, that is, at t=1, is given by

$$\pi_i = P[S_1 = i] \tag{3.25}$$

The sets of probabilities given by equations (3.22), (3.24) and (3.25), can be used both to compute the probability of generating an observation $O = O_1O_2,...,O_T$, and to find a most likely state sequence $S = S_1S_2,...,S_T$, given the observation O.

HMM models can thus be completely specified by the number of states *N*, the number of output observations per state *M* and three sets: (a) the set of state transition probabilities *A*; (b) the set of observation probabilities *B*; and (c) and the probabilities for the states initializing the HMM chain π, commonly referred as the components of HMM. The HMM model is compactly described as

$$\Delta = (A, B, \pi) \tag{3.26}$$

The operation of HMMs is governed by the solution of three fundamental problems, these are.

(1) To compute the probability of occurrence of a specific observation given a model $\Delta = (A, B, \pi)$ that is to find $P(O/\Delta)$. This is an evaluation problem or how well a model matches a certain observation sequence.

(2) Given an HMM model $\Delta=(A, B, \pi)$ and observation $O = O_1\ O_2\ O_3\ \dots\ O_T$, how to choose a sequence of state transition $S = S_1\ S_2\ S_3...S_T$, so that to maximises the joint probability of *O* and *S*, that is $P(O,S/\Delta)$. This is a decoding problem.

(3) Given an observation $O = O_1\ O_2\ O_3\ \dots\ O_T$, adjust the parameters of the HMM, that is $\Delta = (A, B, \pi)$ so that $P(O/\Delta)$ is maximized. This is a training problem.

The solutions of these problems are provided in Appendix II.

3.5.1 Speech recognition using HMM

Continuous speech recognition by HMM is performed by connecting HMMs of individual speech units in sequence [36][37]. An example of a five state left-right HMM with three emitting states is shown in Figure 3.8. This topology is the most commonly used in speech recognition applications. The two non-emitting states S_1 and S_5 have the purpose of providing an interface between individual HHMs.

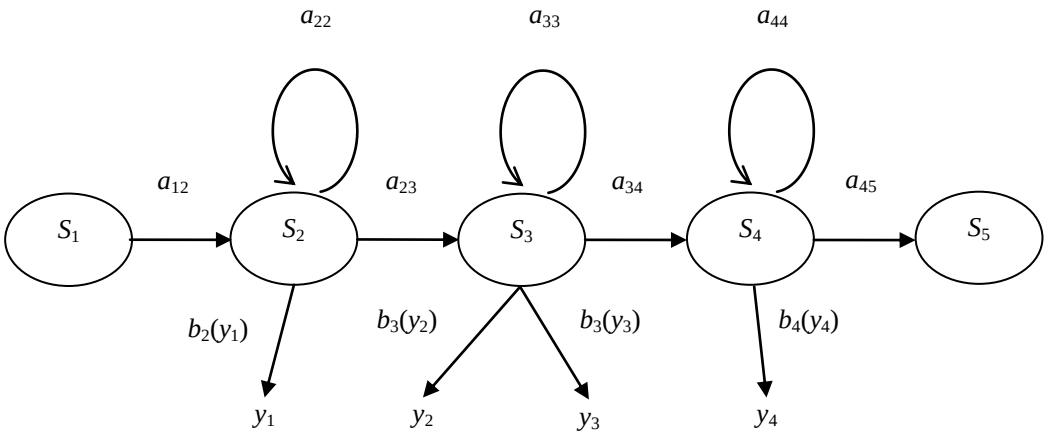


Figure 3.8 Five state left-right HMM with three emitting states

In this section, the simpler case of isolated word recognition is first considered and then extended to the more complex case of continuous speech recognition.

Let the utterance of a word *w* produce an observed speech signal *S* and *Y* contain a sequence of parameter vectors y_r extracted from *S* at regular intervals such that

$$Y = \{y_1, y_2, y_3, \dots, y_T\} \tag{3.27}$$

The word recognition problem can be stated as solving

$$\arg\,(\max\{P(w_i|Y)\}) \tag{3.28}$$

Where $P(w_i|Y)$ is the probability of word w_i being identified given observation *Y*, and can be expressed according to Bayes' Rule as

$$P(w_i|Y) = \frac{P(Y|w_i)P(w_i)}{P(Y)} \tag{3.29}$$

In practice, however, it is not practically feasible to compute the conditional probability of *Y* due to the high dimensionality of the observation vector. In HMMs, this complex problem is replaced by estimating the parameters of the Markov model. For the Markov model shown in Figure 3.8, the above problem can be stated as the joint probabilities of state transitions and observations, as

$$\mathrm{P}(Y, X|M) = a_{22}b_2(y_1)a_{23}b_3(y_2)a_{33}b_3(y_3) \ldots \quad (3.30)$$

where X is the sequence of hidden states and M is the model. $P(Y|M)$ can be computed by summing equation (3.30) over all allowed state sequences, although in practice, the summation is replaced by the maximum operator.

Equation (3.29), and hence equation (3.28), can be solved by assuming that

$$P(Y|w_i) = P(Y|M) \quad (3.31)$$

For isolated speech recognition, a HMM is built for every word in the vocabulary.

For continuous speech recognition it is not computationally practical to build an HMM for each word, so instead HMMs are developed for the speech units (phonemes) and their context-dependent bi-phones and tri-phones. The initial and final states of HMMs are non-emitting so that they can be combined together to form a composite model. More detailed discussion of HMM concatenation and development of composite HMM models can be found in [36] and [37].

3.6 HMM TOOLKIT (HTK)

An extensively used HMM based speech recognition package is the Cambridge University HTK toolkit [36]. HTK is a general purpose HMM toolkit that provides specific library modules for the range of operations needed for speech recognition research, such as speech recoding, parameterization and the formation of lexicon and language model.

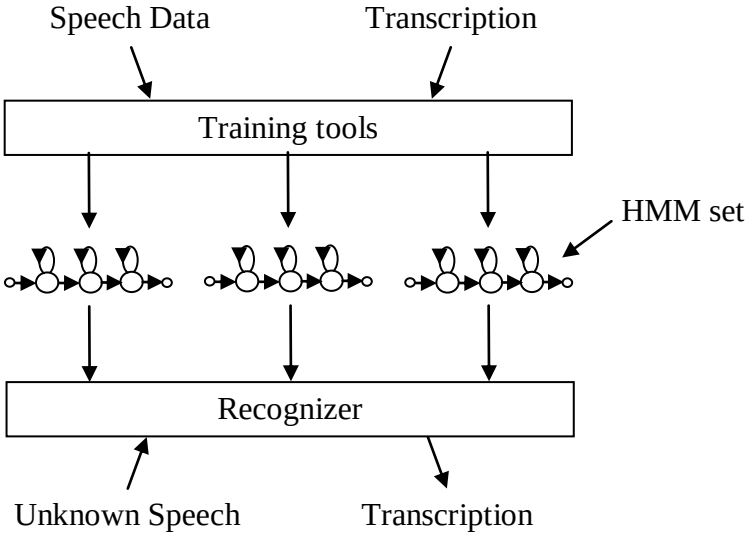


Figure 3.9 HTK speech recognition mechanism [36]

The mechanism of speech recognition by HTK is depicted in Figure 3.9. It consists of two major processing stages. In the first stage, HMMs are developed for each of the speech units based on the features extracted from the known speech samples and their associated transcriptions using HTK training tools. In the recognition stage, test data are transcribed based on the acoustic models and both lexical and language constraints using the HTK recognition tools. In addition to standard parameterization and embedded training tools, HTK also provides specialized tools for HMM adaptation and a number of linear transformations.

The speech recognition process and hence HTK tools are sub-divided into the following four stages discussed below.

Data preparation tools

To build HMMs, training speech data and its association transcription are needed. The speech data is usually obtained from available databases, however HTK also provide tools for audio recording and manual annotation of the recorded speech. As HTK was originally developed for audio speech recognition, it does not have tools for video recording and hence the video must be obtained from a database archive or recorded offline. Tools are also provided for converting transcriptions into the form accepted by HTK. A number of choices for the extraction of audio feature are available by using HTK tools; however the features from video data need to be extracted separately. The HTK tools are mostly designed for audio and therefore the data

preparation stage for video streams needs to be performed offline except for the labelling of the transcript that can be used for both audio and video speech recognition.

Training tools

Unlike the data preparation tools, the training tools operate in the same manner for both audio and video data. First, a topology for the HMM is defined by using a prototype HMM. Although HTK has provision for a number of common topologies to be generated automatically, user-generated topologies can be specified using a simple text editor. An initial set of models can be created with known phoneme/viseme boundaries, known as *bootstrap data*, or all the HMMs can be initialised with same mean and variance, known as *flat start*. Once the HMMs are created, they are refined incrementally using the embedded re-estimation tool HERESET. Context dependent bi-phone and tri-phone HMM are created and refined in a similar fashion. Tools for parameter tying (to address the issue of limited data) and speaker adaptation are also provided among the training tools.

Recognition tools

The HVITE tools perform recognition using acoustic and language models. For audio speech the recognition can be performed on stored audio as well as direct audio input, however, for video speech, recognition can be performed only for already-prepared test data as HTK does not have the capability of extracting video features.

Analysis tools

The performance of the developed recognizer can be assessed using test data for which the transcription is known. The HRESULT tools are able to compare known and recognized transcriptions in a number of different aspects, such as word, phoneme, speaker-by-speaker, and confusion matrix. The results are produced in a format compatible with that specified by National Institute of Standards and Technology (NIST) [36].

3.7 VIDTIMIT DATABASE

The VidTIMIT database used in this work contains recordings of audio and their corresponding videos of continuous speech [38]. It consists of 43 speakers (24 male and 19 female) uttering short sentences taken from the test section of the TIMIT database [39]. The database is recorded in three different sessions with a gap of seven days between the first two sessions and six days between the last two. Each speaker utters ten sentences in front of a camera centred on the face of the speaker resulting in a total of 430 sentences. The ten sentences are distributed among the three sessions so that six sentences are uttered in the first session and two sentences in each of the remaining two sessions. Two out of the ten sentences are common among all the speakers and the remaining eight are generally different for any two speakers. On average, a single sentence has a duration of 2.4 seconds with 106 frames per utterance. The audio is recorded at a sampling rate of 32 kHz and 16 bits quantization; the video is recorded at a rate of 25 frames per second and resolution of 512x384 pixels with 24 bits per pixels and available in JPEG format. Office paper was placed between the lamps (fluorescent and tungsten) and the speaker to reduce glare from the head and face of the speakers. The database was recorded in an office environment with background noise emanating from a computer fan.

3.8 SUMMARY

This chapter discussed in detail those approaches found in the AVASR literature that have been adopted in this thesis. The image transformation techniques discussed in this chapter are used in chapter 4 and 6 of this thesis in order to analyse the images from the videos of the speakers and to achieve a compact representation of speech information present in the images. The dimensionality reduction tools, PCA and LDA are used to reduce the dimensionality of observation vectors obtained from the image transformations into a small number of dimensions, suitable for use in the recogniser. In the work reported in this thesis, phonemes are used as common speech units for both audio and video streams. The extracted audio and visual features are combined using an early integration strategy and the audio-visual features thus obtained are used to develop HMM models for a set of phonemes and their context dependent bi-phonemes and tri-phonemes. The MFCC features from speech audio are extracted

utilizing the HTK built-in tools while the video features are extracted separately. The HTK recognition toolkit is used for training, recognition and analysis of results in the experiments reported in this thesis. The recognition is performed using only acoustic models and without the aid of a language model, so as to yield a direct comparison of the performance of the visual features proposed in this work with those of baseline systems. The results are produced based on percentage of words recognized correctly. Various subsets of the VidTIMIT database have been used in different experiments and these are divided into training and test sets such that all the phonemes present in test set are also available in training set.

3.9 REFERENCES

[1] Polikar, R. (2001), "The Engineer's Ultimate Guide to Wavelet Analysis -The Wavelet Tutorial", available online on **http://users.rowan.edu/~polikar/WAVELETS/WTtutorial.html**.

[2] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., and Vergyri, D. (2001), "Large-Vocabulary Audio-Visual Speech Recognition: A Summary of The John Hopkins Summer 2000 Workshop", *IEEE 4th Workshop on Multimedia Signal Processing*, Cannes, France, pp. 619-624.

[3] Matthews, I., Potamianos, G., Neti, C., and Luettin, J. (2001), "A Comparison of Model and Transform-Based Visual Features for Audio-Visual LVCSR", *Proceedings of International Conference on Multimedia and Expo. (ICME '01)*, Tokyo, Japan, pp. 825-828.

[4] Potamianos, G., Graf, H. P., and Cosatto, E. (1998), "An Image Transform Approach for HMM Based Automatic Lipreading", *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 173-177.

[5] Hazen, T. J. (2005), "Visual model structures and synchrony constraints for audio-visual speech recognition", *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 1082-1089.

[6] Potamianos, G., and Graf, H. P. (1998), "Linear Discriminant Analysis for Speechreading", *Proceedings of IEEE Second Workshop on Multimedia Signal Processing*, Redondo Beach, CA, USA, pp. 221-226.

[7] Liu, P., and Wang, Z. (2003), "Visual Information Assisted Mandarin Large Vocabulary Continuous Speech Recognition", *Proceedings of International Conference on Natural language Processing and Knowledge Engineering*, pp. 72-77.

[8] Arsic, I. and Thiran, J. P. (2006), "Mutual Information Eigenlips for Audio-Visual Speech Recognition", *Proceedings of 14th European Signal Processing Conference (EUSIPCO)*, Lecture Notes in Computer Science.

[9] Potamianos, G., Neti, C., Huang, J., Connell, J. H., Chu, S., Libal, V., Marcheret, E., Haas, N., and Jiang, J. (2004), "Towards Practical Deployment of Audio-Visual Speech Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. III-777-780.

[10] Potamianos, G., Verma, A., Neti, C., Iyengar, G., Basu, S. (2000), "A cascade image transform for speaker independent automatic speechreading" *Proceedings of the IEEE International Conference on Multimedia and Expo*, New York, vol. II, pp. 1097-1100.

[11] Zhong, D., Defee, I. (2004), "Pattern Recognition by Grouping Areas in DCT Compressed Images", *Proceedings of the 6th Nordic Signal Processing Symposium* – (*NORSIG 2004*), Espoo, Finland, pp. 312-315.

[12] Khayam, S. A. (2003), "The discrete cosine transform (DCT): Theory and application", Technical Report: WAVES-TR-ECE802.602, Michigan State University.

[13] Cabeen, K., and Gent, P., "Image Compression and the Discrete Cosine Transform", Math 45, College of the Redwoods.

[14] Graps, A. (1995), "An Introduction to Wavelets", *IEEE transaction on Computational Science and Engineering*, vol. 2, no. 2, pp. 50-61.

[15] Meyer Y. (1993), "Wavelets: Algorithms and Applications", Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

[16] Farooq, O. (2002), "Wavelet-Based Techniques for Speech Recognition", *PhD thesis*, Loughborough University.

[17] Sarkar, T. K., Su, C., Adve, R., Salazar-Palma, M., Garcia-Castillo, L., and Boix, R. R. (1998), "A tutorial on wavelets from electrical engineering perspective, part1: Discrete wavelet techniques", *IEEE Antennas and Propagation Magazine*, vol. 40, no. 5, pp. 49-68.

[18] Wang, X., and Paliwal, K. K. (2002), "A Modified Minimum Classification Error (MCE) Training Algorithm for Dimensionality Reduction", *Journal of VLSI Signal Processing*, vol. 32, no. 1, pp. 19-28.

[19] Maaten, L. J. P. van-der., Postma, E. O., and Herik, H. J. van-den. (2007) "Dimensionality Reduction: A Comparative Review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Preprint.

[20] Nicholson, S., Milner, B., and Cox, S. (1997), "Evaluating feature set performance using f-ratio and j-measures", *proceedings of EUROSPEECH 97*, Rhodes, Greece, pp. 413-416.

[21] Wang, X., and Paliwal, K. K. (2003), "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition", *Pattern Recognition*, vol. 36, no. 10, pp. 2429-2439.

[22] Fodor, I. K. (2002), "A survey of dimension reduction techniques", LLNL Technical report: UCRL-ID-148494.

[23] Ma, Z., and Leijon A. (2008), "A Probabilistic Principal Component Analysis Based Hidden Markov Model for Audio-Visual Speech Recognition", *Proceedings of 42nd Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, CA, USA, pp. 2170-2173.

[24] Koren, Y., and Carmel, L. (2003), "Robust linear dimensionality reduction", *IEEE Transactions on Visualization and Computer Graphics*, vol. 10, no. 4, pp. 459-470.

[25] Shlens, J. (2005), "A Tutorial on Principle Component Analysis", Systems Neurobiology Laboratory, University of California at San Diego, CA, pp. 1-13.

[26] Yu, H., and Yang, J. (2001), "A direct LDA algorithm for high-dimensional data with application to face recognition", *Pattern Recognition*, vol. 34, pp. 2067-2070.

[27] Chen, T., and Rao, R. R. (1998), "Audio-Visual Integration in Multimodal Communication", Proceedings of IEEE, vol. 86, no. 5, pp. 837-852.

[28] Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J. (2000), "Audio-Visual Speech Recognition", *Workshop 2000 Final Report*, Baltimore, MD: Centre for Language and Speech Processing, JHU, Baltimore, MD.

[29] Hazen, T. J., Saenko, K., La, C. H., and Glass, J. R. (2004), "A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments", *Proceedings of 6th International Conference on Multimodal Interfaces, (ICMI '04)*, PA, pp. 235-242.

[30] Rogozan, A. (1999), "Discriminative learning of visual data for audiovisual speech recognition", *International Journal of Artificial Intelligence Tools*, vol. 8, no. 1, pp. 43-52.

[31] Silsbee, P. (1994), "Sensory integration in audiovisual automatic speech recognition", *Proceedings of 28th Annual Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, vol. 1, pp. 561-565.

[32] Lewis, T. W., and Power, D. M. W. (2002), "Lip Feature Extraction Using Red Exclusion and Neural Networks", *proceedings of 25th Australasian Conference on Computer Science*, Melbourne, Australia, vol. 4, pp. 149-156.

[33] Yau, W. C., Kumar, D. K., and Arjunan, S. P. (2006), "Voiceless Speech Recognition Using Dynamic Visual Speech Features"**,** *Proceedings Of HCSNet workshop on Use of vision in human-computer interaction*, Canberra, Australia, vol. 56, pp. 93-101.

[34] Potamianos, G., Neti, C., Gravier, G., Garg, A. and Senior, A. W. (2003), "Recent Advances in Automatic Recognition of Audio-Visual Speech", *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306-1326.

[35] Juang, B., and Rabiner, L. R. (2005), "Automatic speech recognition—a brief history of the technology", *Elsevier Encyclopaedia of Language and Linguistics. 2nd edition,* Elsevier.

[36] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2006), *The HTK Book V3.4.*

[37] Rabinar, L.R. (1989), "A Tutorial on Hidden Markov Model and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286.

[38] Sanderson, C., and Paliwal, K. K. (2002), "Polynomial Features for Robust Face Authentication", *Proceedings of IEEE International Conference on Image Processing*, vol. 3, pp. 997-1000.

[39] Fisher, W. M., Doddington, G. R., and Kathleen, M. G.-M. (1986), "The DARPA Speech Recognition Research Database: Specifications and Status", *Proceedings of DARPA Workshop on Speech Recognition*, pp. 93-99.

CHAPTER 4

FREQUENCY-BAND BASED VISUAL FEATURES FOR AVASR

This chapter presents a novel frequency-band based approach to visual feature extraction for AVASR using both DCT and DWT domain representations of the images obtained from the videos of speakers. The new visual features use a novel discriminative approach to the DCT and DWT transformation, in contrast to the more commonly-used data reduction viewpoint. Where the work presented in this chapter lies within the general AVASR system of Figure 2.1 is shown in Figure 4.1.

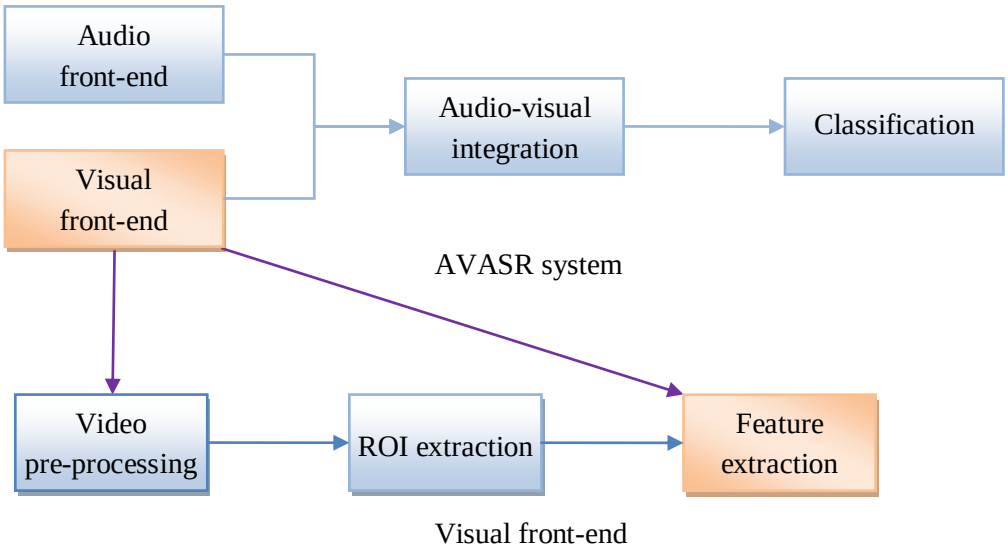


Figure 4.1 Location of the feature extraction process in the general AVASR system

The chapter is organized as follows. Section 4.1 provides an introduction to visual feature extraction for AVASR and discusses the visual feature extraction approaches currently used in AVASR research. Section 4.2 describes the feature extraction approach adopted in this work and the rationale of using the frequency-band based DCT and DWT for AVASR. The experimental setup used in this work, including the data used for training and testing of the recognizer, extraction of the mouth ROI, extraction of visual features, integration of audio and video modality and training of the classifier, are presented in Section 4.3. Section 4.4 presents the results obtained

from a series of experiments conducted for both visual-only and audio-visual ASR. Section 4.5 concludes the outcomes of this investigation and highlights the specific findings of the study.

4.1 VISUAL FEATURE EXTRACTION FOR AVASR

Visual feature extraction is a core area of research in AVASR. The purpose of feature extraction is to retain as much speech related information as possible from the original images of the speaker in a reasonably small number of parameters. Feature extraction techniques aim to develop models based on the knowledge of human speech production and perception mechanisms [1]. Visual features could be used to develop visual-only speech recognition systems, but, in most cases, they are combined with features extracted from the audio stream to form an AVASR system. Consequently, these visual features should be robust, but also supplement and complement the audio features so as to be able to improve on the performance of ASR systems under certain challenging conditions [2]. Three types of features namely shape-based features, appearance-based features and hybrid features, which are a combination of the first two types, have been used in literature. Shape (sometimes termed geometry or model-based) features may represent various aspects of the speaker's mouth region, such as length, width, curvature or eccentricity. Alternatively, the shape of mouth is fitted to a statistical model, whose parameters are then used as visual features for ASR [3]. In the extraction of appearance (or transform based) features, the assumption is that the whole mouth region provides useful information about the speech. Features are extracted directly from the mouth pixel values or following some suitable transformation of the mouth region [4]. Shape-based feature extraction techniques require robust face and mouth tracking and mouth contour extraction, while appearance-based techniques require an approximate mouth region for their implementation [5]. In the last two decades, a number of AVASR systems both on shape-based [6], [7], [8] and appearance-based [9], [10], [11] features have been reported and have been demonstrated to yield improved recognition performance compared with audio-only ASRs in the presence of noise. It has been claimed that appearance-based methods, when applied to AVASR, outperform approaches based on shape-based features [12], [13].

The visual feature extraction approach presented in this chapter falls into the category of appearance-based features extraction techniques. The most commonly used transforms in appearance-based feature extraction approaches for AVASR research are the DCT and the DWT. In AVASR, coefficients from the low frequency region of the DCT and DWT transforms matrices have been used as visual features, or alternatively as observation vectors for PCA and LDA based feature extraction. Appearance-based visual feature extraction approaches are adopted principally from the data reduction literature, where the main goal is to achieve a compact representation of images or video for reducing the memory capacity required for storage. Retaining just a few of the low frequency DCT and DWT coefficients is generally sufficient for restoring an image whose subjective quality is adequate for many practical imaging purposes [14]. However, this approach does not guarantee that these coefficients also contain the most discriminating information for speech recognition; thus being the main concern here and this is in contrast with data compression applications, where the aim is to present image data in a compact set containing a small number of dimensions [15], [16]. In the work presented in this chapter, visual features for AVASR are extracted from either the DCT or DWT coefficients based on a novel pattern recognition approach. PCA and LDA techniques are used to reduce the dimensionality of extracted features so as to render the final feature vector suitable for use in a classifier.

4.2 RESEARCH RATIONALE

The DCT and DWT feature extraction approaches used in AVASR research are taken mainly from the image compression literature, in which the primary goal is to accomplish a compact representation of image and video information while maintaining high visual quality. The concept of psychovisual redundancy forms the basis of image and video compression research and takes advantage of the fact that human eye is less sensitive to high frequency information in video and so an acceptable video quality can be achieved by retaining only low-frequency components [14]. However, it has not been established that psychovisual redundancy will not remove visual information that may be useful for speech recognition. In particular, while the low-frequency coefficients represent the gross features in an image, visual

speech information is at least partly contained in the higher-frequency, finer details of the image, such as the curvature of the mouth.

The DCT transform of an image $I(x_i,y_i)$ of size MxN (where $1 \leq i \leq M$ and $1 \leq j \leq N$) is a matrix $D(u_p,v_q)$ of the same dimensions MxN, where the coefficients $d(u_p,v_q)$ of transform matrix $D(u_p,v_q)$ represent the p^{th} and q^{th} frequency component in the vertical and horizontal direction in the image, respectively. The DCT transform places the frequency information in the image in such a way that the low frequency coefficients lie towards the upper left corner while the high frequency coefficients lie towards the bottom right of the transform matrix, as shown in Figure 4.2.

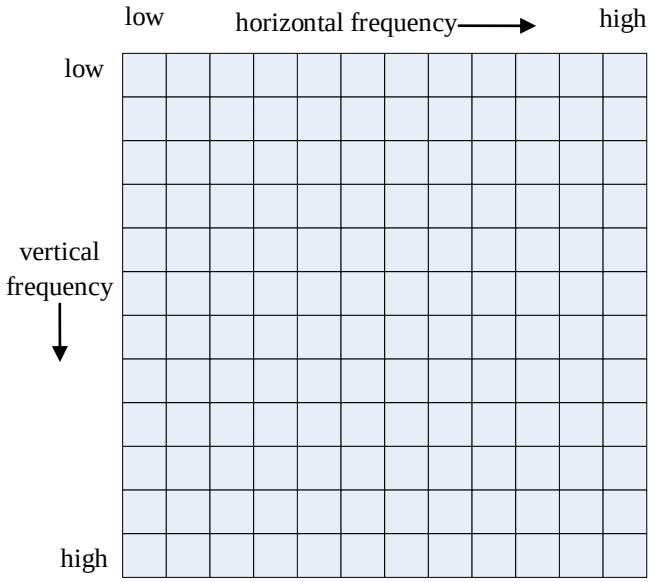


Figure 4.2 Frequency coefficients distribution by DCT

Similarly, the DWT transform decomposes the input image into a low-frequency sub-band (known as the approximate image) and high-frequency sub-bands (known as detailed images), as shown in Figure 4.3. The LL region of the DWT transform in Figure 4.3, contain the low frequency contents of the image, the HL region contains the high-frequency horizontal details, LH the high-frequency vertical details and HH the high-frequency details for both the horizontal and vertical direction. The application of the DWT to an image results in high-pass and low-pass filtering of the image. Further refined details of an image can be extracted by applying higher levels of decomposition. This is achieved by the application of DWT to the sub-images obtained in the lower level, starting from the original input image. First-level decomposition means the DWT of the original image; second-level decomposition

means the DWT of sub-images obtained in first level and so on, whereas the low-frequency components are known as approximate coefficients while the high-frequency components are known as detailed coefficients.

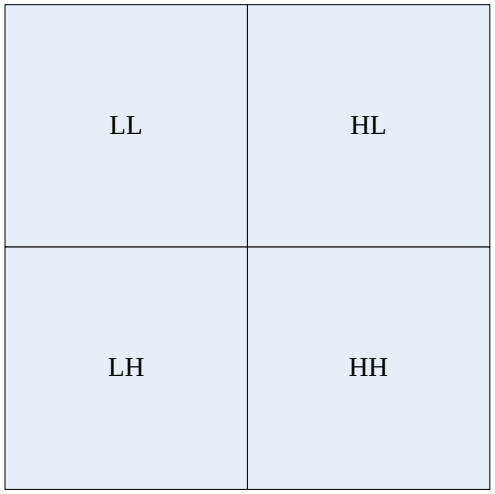


Figure 4.3 Single level DWT decomposition of an image

The use of frequencies other than low frequencies from the DCT and DWT transform space has been reported in pattern recognition applications. In [17], for text capture application, the DCT coefficient matrix is partitioned into three regions, low, medium and high frequency, as shown in Figure 4.4. It was shown that medium frequency components performed better in this application as compared with using only the low frequencies coefficients. A similar partition of DCT coefficients has been reported in [18] for face recognition applications, where it was found that the salient features for face recognition are contained in medium frequency components and that a weighted combination of all frequencies outperformed a solution using only low frequency coefficients. As the low frequency components are more sensitive to illumination variations, in [19] an illumination invariant face recognition system was proposed that truncated the low frequency coefficients in DCT transformed space. In this work, low frequency components in the DCT transform matrix were set to zero and it was found that the features extracted from the resulting matrix, containing only medium and high frequency coefficients, were more robust to illumination variations. Similarly the medium frequency coefficients from the DWT decomposition of fingerprint images has been used for fingerprint recognition purposes [20]. In Wong *et al*. [21], the features for face recognition were extracted from a multiple sub-band decomposition

based on the DWT transform and certain frequency bands were identified with giving better recognition performance.

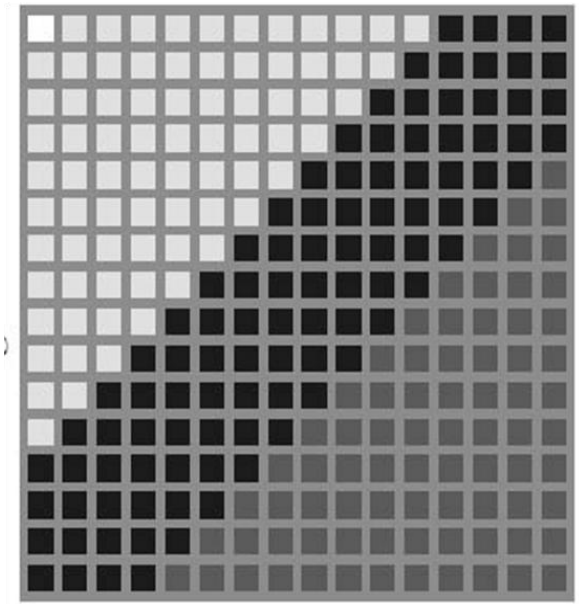


Figure 4.4 Partitioning of the DCT coefficients matrix in [17]

Figure 4.5 and Figure 4.6 show the reconstruction of the mouth region of speakers from low and high frequency coefficients of the DWT and DCT domains respectively. In both figures, images (a) and (b) are reconstruction from low frequency coefficients, while (c) and (d) are reconstructions from high frequency coefficients. In Figure 4.5, images (a) and (b) are reconstructions from the 2^{nd} level and 3^{rd} level approximate coefficients of DWT decomposition, while (c) and (d) are reconstructions from the remaining detailed coefficients. The corresponding coefficients from the DCT transform are used for the DCT-based reconstructions of Figure 4.6. These image reconstructions suggest that while the overall subjective appearance of the image is well retained in low frequency coefficients, the edges of the mouth are better preserved in detailed coefficients, and hence the use of these coefficients could potentially be useful for AVASR purposes.

(a) approximate coefficients at level 2

(a) approximate coefficients at level 3

(c) detailed coefficients at level 2

(d) detailed coefficients at level 3

Figure 4.5 Image reconstructions from DWT coefficients

(a) approximate coefficients at level 2

(a) approximate coefficients at level 3

(c) detailed coefficients at level 2

(d) detailed coefficients at level 3

Figure 4.6 Image reconstructions from DCT coefficients

Although the use of discriminative information extracted from the medium and higher frequency coefficients of the DCT and DWT transform domains have been reported in various pattern recognition applications, to the best of author's knowledge it has not been applied to AVASR. Assuming that the visual speech information is contained in the motions of the lips and other visible articulators, the motion information is likely to be found in the edges and texture of this region. As edge information is captured in

mid and high frequency coefficients, in this work the visual features for AVASR purpose are extracted from regions in the DCT and DWT transform space that include these frequency bands. In particular, a detailed investigation of visual feature extraction for speech recognition purposes has been carried out that includes specific frequency bands of DCT and DWT transform and the results compared with those obtained using features from only the low frequency coefficients, as is commonly used in the literature. It was found that the speech recognition performance of the visual modality can be improved by the inclusion of certain intermediate and higher frequency coefficients. Furthermore, the visual features from the frequency bands giving the best visual-only recognition performance were combined with MFCC based audio features to form an audio-visual feature set that can be used for AVASR. This system was tested in the presence of acoustic noise at a range of signal-to-noise ratios and the results obtained are compared with audio-only ASR. The results of this study are presented in the next section, and show that while the performance of audio-only speech recognition system degrades drastically in presence of noise, the AVASR system remains relatively robust under such conditions.

4.3 EXPERIMENTAL SETUP

This study has investigated the use of a range of different frequency bands in the generation of visual features in AVASR system design. The audio-visual database, ROI identification, feature extraction, and audio-visual integration techniques used in this work are presented in this section. Using this new method, the experimental work involves investigation of visual-only, audio-only and audio-visual features for their use in speech recognition.

4.3.1 Audio-visual database

The VidTIMIT [22] database used in this thesis is discussed in detail in section 3.7. A subset of VidTIMIT database having 32 speakers (16 male and 16 female speakers) was used in the experiments presented in this chapter. To avoid over training of specific phonemes, the two sentences common to all speakers in VidTIMIT are not used in these experiments. The data thus obtained has each speaker uttering eight different sentences in front of a camera centered on the face of the speaker. The sentences are all examples of continuous speech taken from the TIMIT database and

contain a total of 256 utterances and a vocabulary of 925 words. Of these, 216 utterances are used for training and the remaining 40 are used for testing such that all the phonemes in the test set are also available in the training set. To make the comparison between different visual feature fair, these training and test sets were used in all the experiments on visual feature extraction reported in section 4.4 of this chapter and sections 6.4 and 6.5 of chapter 6. The video is provided at a rate of 25 frames per second with a resolution of 384x512, while the audio stream has a sample rate of 32 kHz and 16 bits depth. As, in this work, features are extracted from the audio stream at a rate of 100 times a second, so, to match this rate, video frames were up-sampled to the rate of 100 frames per second using linear interpolation.

4.3.2 Face detection and mouth ROI extraction

Local successive mean quantization transform (SMQT) features [23], were used to locate the face of the speaker in the first frame of the utterance. A bounding box of size 72x96 around the center of the lower half of the face is extracted as the mouth ROI. As the training and recognition by HMMs requires that all the observation vectors have the same dimensionality, the dimensions of mouth bounding box needs to be the same for all images in the training and test data sets. The size of the bounding box was adopted following manual observation of the mouth region across all the utterances in the test and training sets. To reduce the computation time, the coordinates found for the mouth region extracted from first frame of utterance are used for ROI extraction in the remaining frames of that utterance. As the mouth movement in these utterances is limited, the above approach was found to work well and greatly reduced the time required to extract the mouth region in each image of the video sequence at a rate of 100 frames per second, resulting in 92732 and 46740 frames for the training and test data respectively. One such mouth region extracted in this manner is shown in Figure 4.7 (a). In a small number of cases where the face region was not found correctly, the coordinates of the mouth centre were provided manually. Figure 4.7 (b) shows one such case where the face region is not found correctly and therefore the mouth region not properly located, while Figure 4.7 (c) shows the manually corrected mouth ROI for this frame.

(a) Accurately extracted ROI

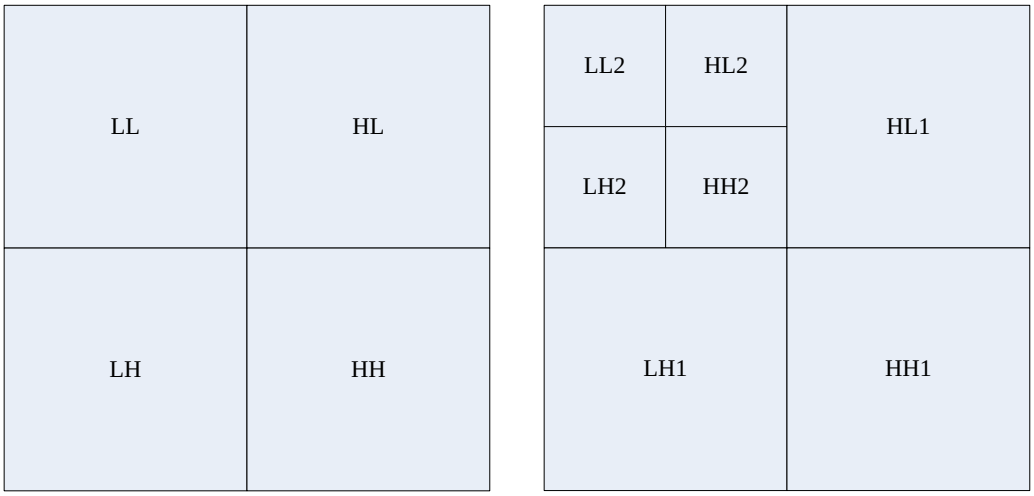
(b) Missed ROI

(c) Manually corrected ROI

Figure 4.7 Region of interest (ROI) extraction

4.3.3 Feature extraction

The two dimensional DCT and DWT display the spatial-frequency information contained in images in the transformed space. The wavelet transform decomposes an image into sub-images at a range of resolutions, corresponding to different frequency bands. The image decomposition by the DWT transform is shown in Figure 4.8, where (a) shows a single level of decomposition while (b) shows two levels of decomposition. The four sub-images in Figure 4.8(a), namely LL, HL, LH and HH are known as approximate, vertical detail, horizontal detail, and diagonal detail coefficients, respectively, and contain information about the spatial frequencies present in the image in the horizontal and vertical dimensions. Figure 4.8(b) shows a two-level decomposition where the symbols L and H represent low and high frequency coefficients respectively, while the numbers 1 and 2 here represent the level of decomposition. Further frequency sub-bands for the DWT transform can be obtained by applying higher levels of decomposition.



(a) Single-level decomposition (a) Two-level decomposition

Figure 4.8 Image decomposition by DWT transform

Similarly, the DCT transform of an image, shown in Figure 4.9, places the spatial frequency information in the image in ascending order of frequency. Thus regions R1 to R4 in Figure 4.9 contain the horizontal and vertical components in order of increasing frequency. Additional sub-bands can be obtained for the DCT by further subdivision of the regions R1 to R4. In this work, both four and eight frequency bands are used and this was achieved in the DWT by taking two and three levels of decomposition and while for the DCT the appropriate regions from the transform space were selected.

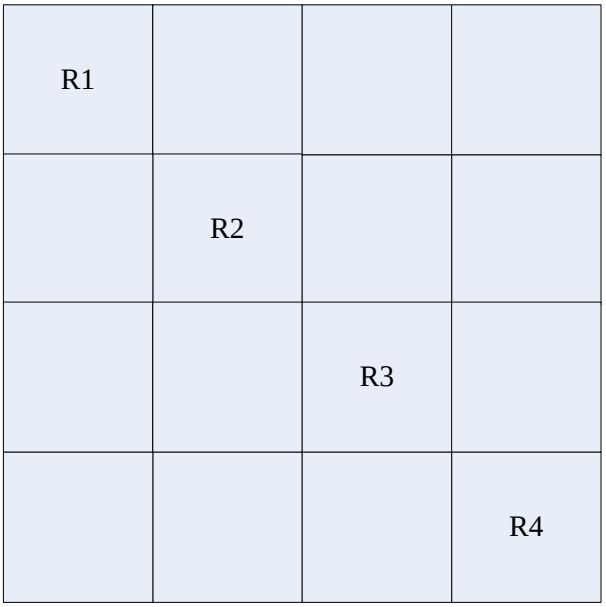


Figure 4.9 DCT based frequency regions

The two-dimensional DCT and DWT of the mouth region of interest (ROI) were taken and separated into a number of frequency bands, named R1, R2, R3, …, RN. (where N is the number of bands used). These frequency bands were used as input observations for the extraction of visual features for AVASR. The coefficients from these regions were re-arranged to form an observation vector. As the dimensionality of the observation vectors obtained from these regions is too high to be used directly for training the recognizer, PCA or LDA are applied to reduce the dimensionality to a common 30 dimensional feature vector.

4.3.4 Audio-visual integration and HMM modelling

In the visual-only experiments, the extracted 30 static visual features were appended with their first and second derivatives so as to incorporate dynamic information and resulting in a total of 90 features. Similarly, for the audio-only experiments, the 13 MFCC coefficients were extracted from the speech audio and appended with their first and second derivatives to form a feature vector of 39 dimensions. In the audio-visual experiments, an early integration strategy was adopted in which the 13 MFCC coefficients were combined with the 30 visual features to form a 43 dimensional audio-visual feature vector. Similarly, the audio-visual vectors were appended with their derivatives resulting in a 129 dimension feature vector. Using the HTK Toolkit

[36], the three state HMM model shown in Figure 4.10 was developed for each of the 46 phonemes used in this work, along with their context-dependent bi-phones and tri-phones. The recognition is performed on the basis of phoneme models for all audio, video and audio-visual recognition and only the acoustic model was deployed without the aid of language information.

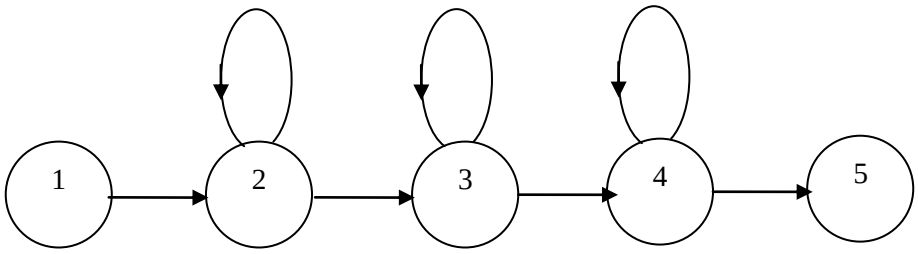


Figure 4.10 HMM with three emitting states

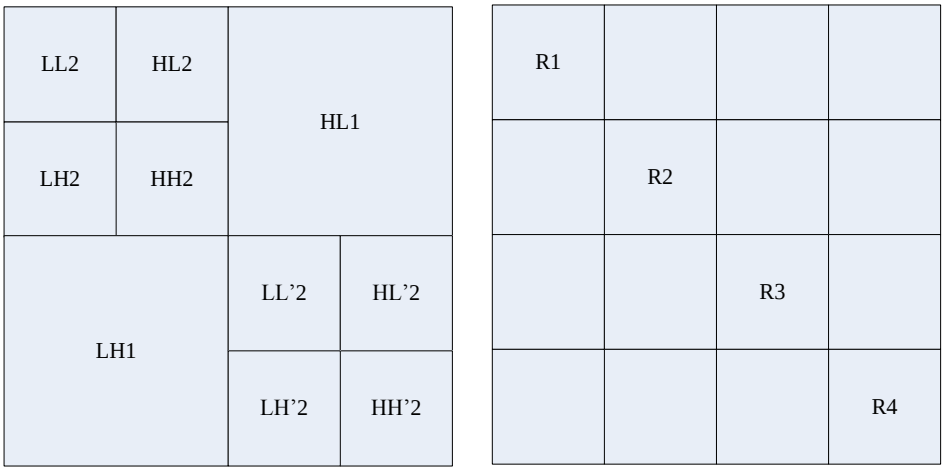
4.4 EXPERIMENTS AND RESULTS

The experiments on the extraction of visual features from the frequency bands were carried out in two stages. In the early experiments related to this work and reported in [25], the DCT and DWT transform coefficients were divided into four frequency bands and experiments were conducted to assess the effect on performance of including the different bands, the types of features included and the choice of transformation technique. To investigate the influence of the different frequency bands in finer detail, later experiments extended the number of frequency bands to eight and two additional factors were investigated, namely the choice of dimensionality reduction technique and performance under noise.

4.4.1 Experiments using four frequency bands

In these experiments, the DCT and DWT transform spaces were each divided into four frequency bands, as shown in Figure 4.11. Here, the Haar mother wavelet is applied at level 1 to perform the first level DWT decomposition of input image into four sub-images, LL1, HL1, LH1 and HH1. Further single level decompositions were carried out of both the low-frequency sub-image LL1 (LL1 is not visible in Figure 4.11) to obtain LL2, HL2, LH2 and HH2 and of the high-frequency sub-image HH1 to obtain LL'2, HL'2, LH'2 and HH'2. The four sub-images along the diagonal

containing the horizontal and vertical frequency details, namely LL2, HH2, LL'2 and HH'2 were then used as the input frequency bands for visual feature extraction. A similar operation was performed for DCT, where the output of the transform was divided into the four regions, R1, R2, R3 and R4 (in order of increasing horizontal and vertical frequency), as shown in Figure 4.11(b). To simplify the comparison between DCT and DWT transform features, in the investigations that follow, the frequency bands from the DWT are referred to as R1, R2, R3 and R4, rather than LL2, HH2, LL'2 and HH'2. Note that as the ROI is of dimension 72x96, each of these regions is of dimension 18x24.



(a) DWT image decomposition (b) DCT image decomposition

Figure 4.11 Image decompositions in the transform domain

Visual features obtained from the low frequency coefficients of the DCT transform (top-left corner of transform matrix) of the mouth ROI have been reported by Jun and Hua [26]. In addition, in Matthews *et al.* [27], a number of regions from the low frequency region of the DCT and DWT transform matrices of the mouth ROI have been used for AVASR visual feature extraction. In Huang *et al.* [28], the visual ROI was reduced in size and visual features extracted by applying the LDA to the whole transform matrix. In Gagnon *et al.* [29], high energy coefficients from the DCT transformed space have been selected and reduced to lower dimensions using LDA. In order to compare the performance of the newly proposed frequency band-based features with these approaches, the coefficients from the four regions of the DCT transform matrix were re-shaped to form an observation vector of 432 dimensions. In one set of experiments, the entire observation vector was passed to LDA to reduce the

dimensionality of final feature vector to a total of 30 values. In a second set of experiments, 100 high energy coefficients from the DCT transform matrix were retained and reduced to 30 values using LDA. As LDA is a supervised dimensionality reduction technique requiring that the class membership of input observations are provided, these were obtained from an audio-only HMM developed earlier, using forced alignment. In addition, to compare the frequency band-based method with the ROI resizing approach, the ROI was reduced to 18x24 pixels (the same as that of the frequency bands) using nearest neighbour interpolation and features obtained from the DCT transform of the resized ROI in the same manner as used in the method for determining the frequency-based features. The resulting 30 dimensional feature vectors for the four frequency bands and the resized ROI were then used to train a video-only recognizer using the training set described in section 4.3, and the performance evaluated using the corresponding test set. The results obtained using DCT are shown in Figure 4.12.

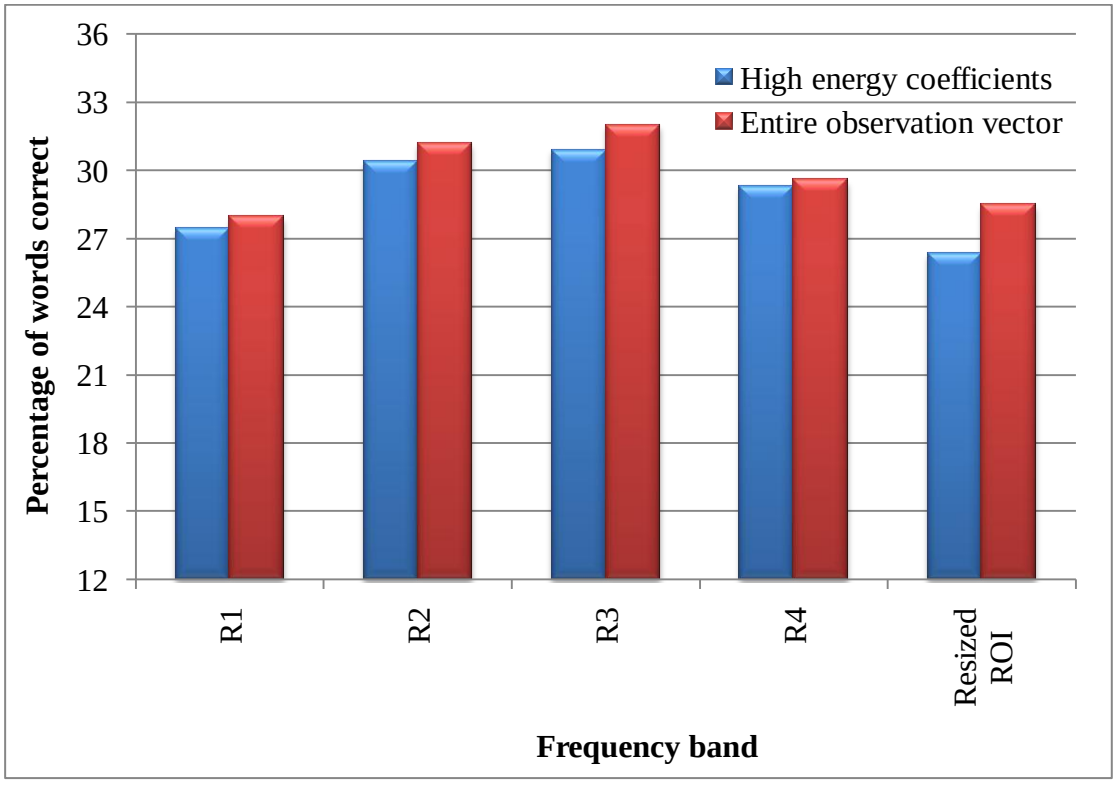


Figure 4.12 Recognition performance of DCT based frequency-band features

To determine the performance of these frequency bands on the DWT transform, a 30 dimension feature vector was extracted from the DWT transform of the resized ROI and each of the four regions of the DWT transform of the original ROI, in a way

similar to that described for the DCT transform above, and their performances evaluated using the same training and test sets used for the DCT. The results obtained from the DWT transform feature are shown in Figure 4.13.

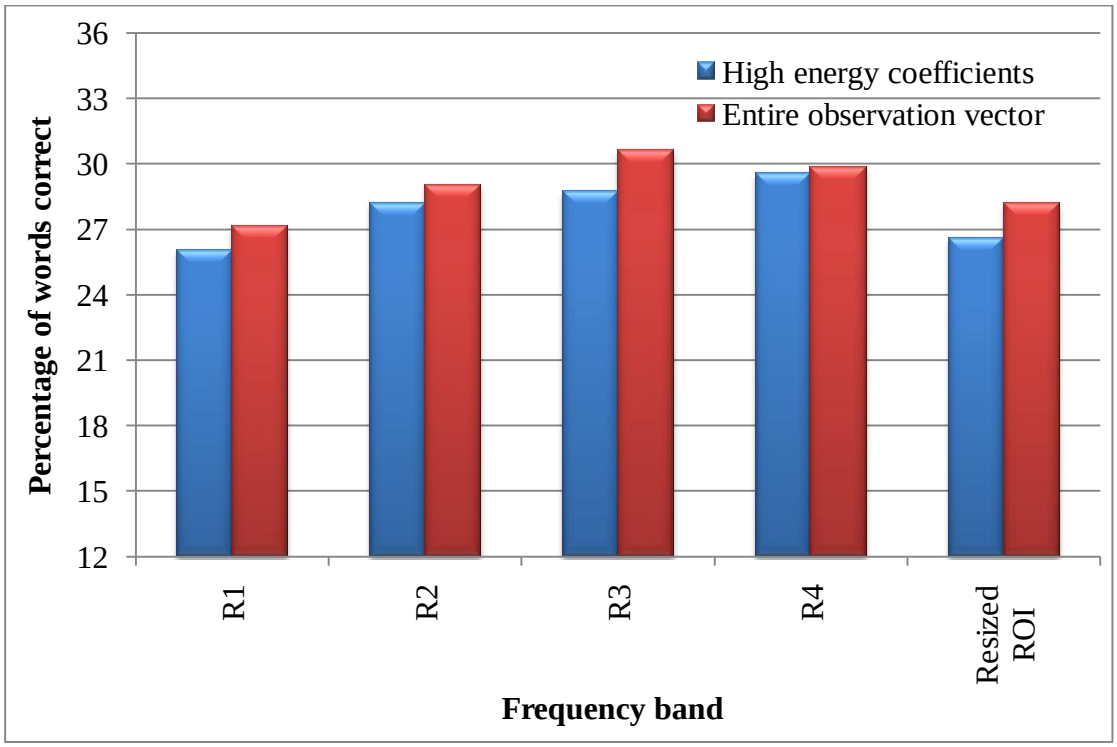


Figure 4.13 Recognition performance of DWT based frequency band features

The results for both the DCT and DWT frequency bands show that the features extracted from mid-frequencies bands (R2 and R3) gave better recognition performance than using only the low frequency band and that the features obtained by using the entire observation vector from the frequency bands as input for LDA outperformed those obtained from using only the 100 highest energy coefficients. This is probably because the highest energy coefficients do not necessarily represent the same spectral component among the sequence of video images and therefore result in an improper comparison. Also, the resizing of the ROI to smaller dimensions adversely affected the recognition performance, perhaps due to the loss of the visual speech information present around the lip edges and texture of the original mouth ROI. This suggests that the visual speech information is retained better in the mid-frequency bands rather than at the low frequencies.

To compare the performance of the DCT transform with that of DWT, the results for the DCT features obtained using the entire frequency-band as input observation vector were compared with those of DWT, as shown in Figure 4.14.

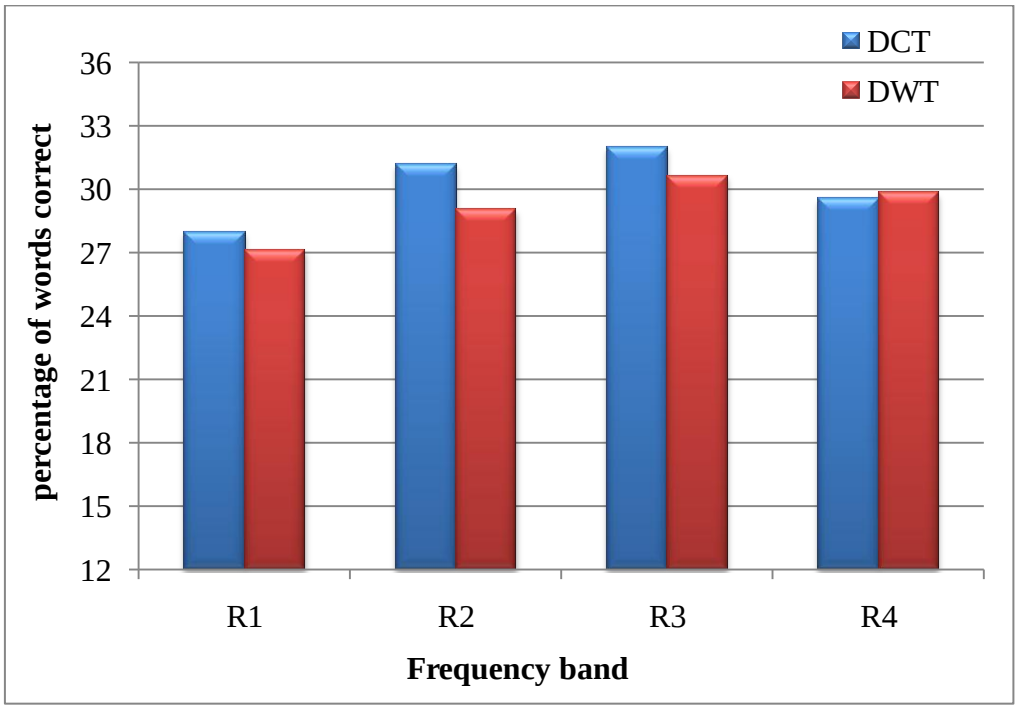


Figure 4.14 Comparison of DCT and DWT based features

Figure 4.14 shows that DCT based features in general gave better performance compared to DWT features. The reason for this may be that the DWT can better represent certain specific phonemes, but may be less effective in representing others. DWT is thus suitable for recognising certain specific phonemes but, for overall speech recognition, the DCT performs better than the DWT.

The results obtained from the four frequency bands above have shown that the use of the mid-frequency coefficients of the DCT and DWT transforms of the mouth ROI give improved recognition performance compared to the low frequency coefficients that are commonly used in the AVASR literature. In next section the frequency-band based visual feature are explored in finer detail by using eight frequency bands.

4.4.2 Experiments using eight regions

To investigate the performance of the frequency band based features in greater detail, the DCT and DWT transform spaces were further divided into a total of eight frequency bands. This was achieved by applying an additional single level of

decomposition on the DWT regions obtained in earlier experiments. Corresponding regions in the DCT transform were also subdivided to give a total of eight frequency bands R1, R2, R3...R8, each of dimensions 9x12, as shown in Figure 4.15.

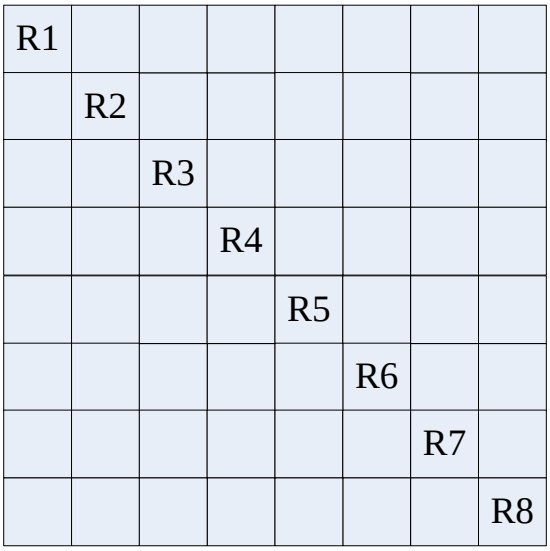


Figure 4.15 DCT and DWT frequency bands for eight regions

Consequently, 108 values from each of the eight regions of DCT and DWT transform domain were reshaped into a vector, followed by the application of LDA, in order to reduce the number of dimensions in the final visual feature vector to 30. These vectors were then used to train a video-only recognizer for each region and the performance evaluated using the same test set as used in earlier experiments on four frequency bands. The recognition performance achieved in each of the eight frequency bands is shown in Figure 4.16.

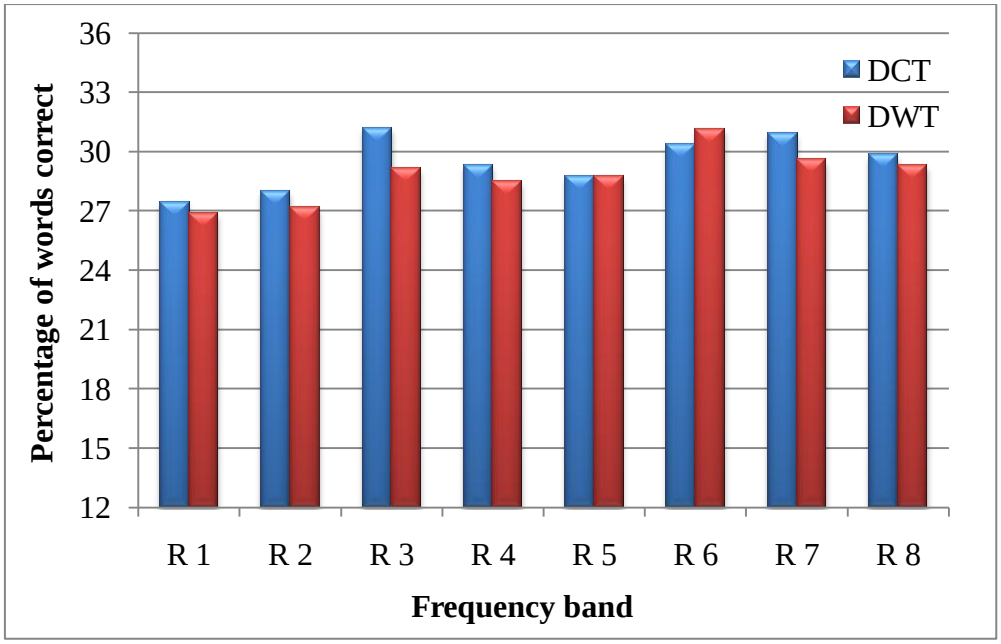


Figure 4.16 Recognition performances of DCT and DWT transform coefficients for eight frequency bands features using LDA for dimensionality reduction

Figure 4.16 shows that the DCT based features in general perform better than the DWT based features. This further verifies the results shown in Figure 4.14 for four frequency bands, namely that for speech recognition applications, containing both vowel and consonant phonemes as used in these experiments, the DCT performs better than the DWT. In addition, comparing the speech recognition results of eight frequency-bands in Figure 4.16 with those obtained for four bands in Figure 4.14 shows that there is no significant difference in the performance of mid-frequency bands for speech recognition purposes following the increase in the number of bands. This implies that mid-frequency components, although containing useful visual speech information, contain some form of redundancy and therefore the addition of extra components in the mid frequencies adds little to the information content.

PCA is another commonly used dimensionality reduction technique [30]. To compare the performance of the LDA based approach with that of the PCA, the experiments for eight frequency bands discussed above were repeated using PCA to reduce the dimensionality of observation vector. The results obtained are shown in Figure 4.17.

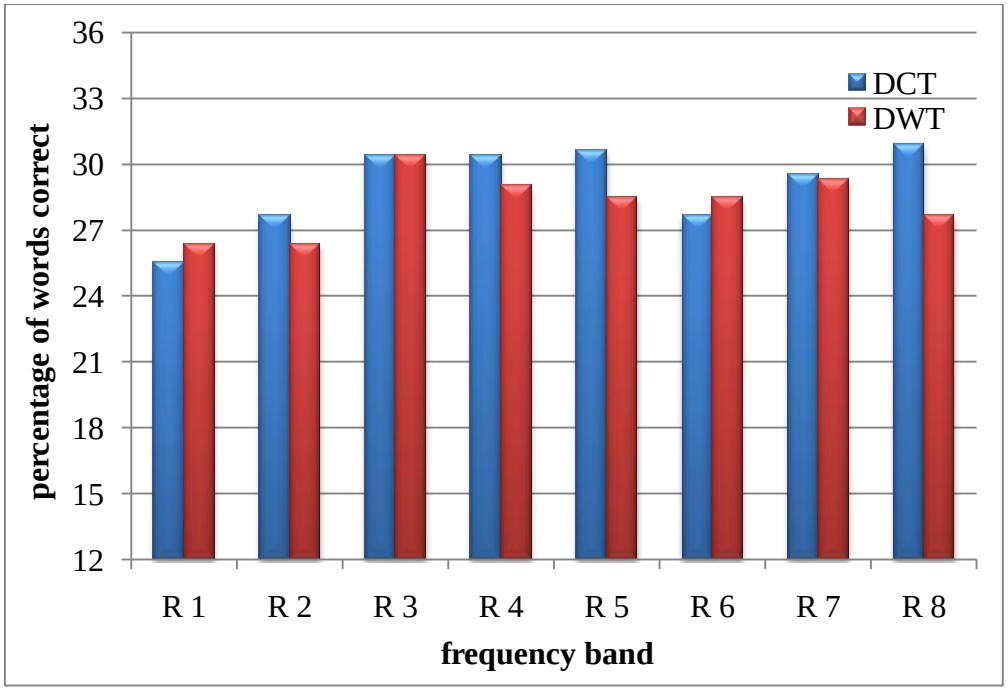


Figure 4.17 Recognition performance of PCA based frequency bands features from DCT and DWT transform coefficients

From Figures 4.16 and Figures 4.17, it can be seen that LDA generally gives better results than PCA. In PCA, the data in the transform space are arranged in order of decreasing variance, so that retaining a few principal components represent most of the variance in the original data, but does not guarantee to separate the different classes present, while LDA transforms the input data such that the separation between the classes present in the data is maximized. This demonstrates that LDA is a better option for dimensionality reduction in pattern recognition applications. Although PCA reveals certain patterns in the input data, it appears better suited to data compression than speech recognition.

An important aspect in the selection of features for speech recognition is their robustness to changes in the environmental conditions. For the video modality, the most common challenge to the performance of visual features is changes in the illumination. As the available databases for AVASR and also the VidTIMIT database used in this research do not have video sequences allowing such investigations, the intensity of the images from the videos was artificially altered to provide the test data for assessing ASR performance under different illumination conditions. This was achieved by altering the intensity values of each pixel of the images in the test data

and then determining the DCT-based frequency-band features extracted from each of the eight regions and using LDA for dimensionality reduction. The speech recognition performance of the visual features for the original test data and those obtained after the intensity change are shown in Figure 4.18.

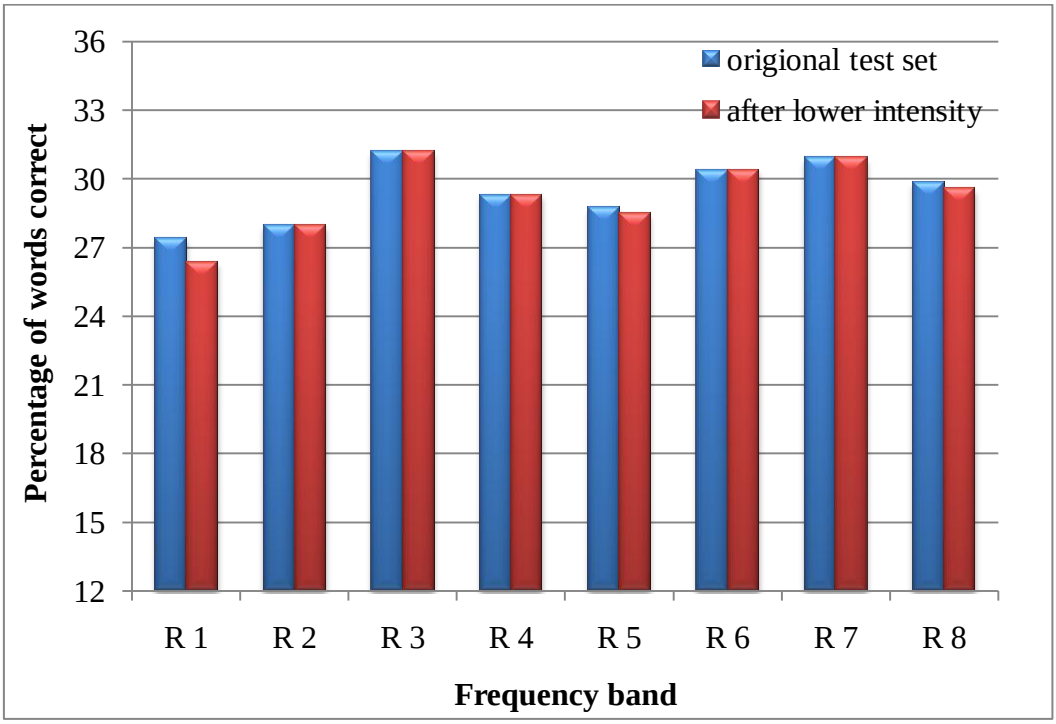


Figure 4.18 Speech recognition performance of frequency-band based features after lowering the illumination

The results in Figure 4.18 show that, while the speech recognition performance of the low frequency features (R1) is affected more by the intensity changes, the features from medium and high frequency bands remain quite robust to these changes. This is due to the changes in intensity affecting mainly the coefficients of the lower frequencies in the image that contain the overall appearance information in the image, while the mid and high frequency coefficients containing information about edges are largely unaffected.

These experiments show that the features obtained from the mid-frequency bands consistently performed better than those obtained from the low frequency bands, irrespective of the transformation and dimensionality reduction method used. This shows that intermediate frequencies coefficients are probably more informative about visual speech than low frequency coefficients.

In the above experiments, the performance of the new frequency-band based features was evaluated on visual-only recognition task and compared with the visual features reported in literature. To investigate the performance of the new visual features for the AVASR task, features from the regions that give the best performance for visual-only speech recognition in each of the combinations of transform and dimensionality reduction techniques were combined with 13 MFCC features from the audio modality to form a 43 dimension audio-visual feature vector. An AVASR system was thus developed for each of the combinations DCT-PCA, DCT-LDA, DWT-PCA and DWT-LDA using the training set described in section 4.3. For comparison purposes, an audio-only speech recognition system was implemented using MFCC features obtained from the same training set. All of the recognition systems described above were tested both on clean speech and noisy speech at a range of signal to noise ratios (SNR). The test data for clean speech was provided by the test set described in section 4.3, whereas the noisy speech signal at different signal-to-noise ratios was obtained by weighted summation of the test set with the speech noise obtained from the NOISEX database [31]. Figure 4.19 show the results for these audio-only and audio-visual recognition experiments.

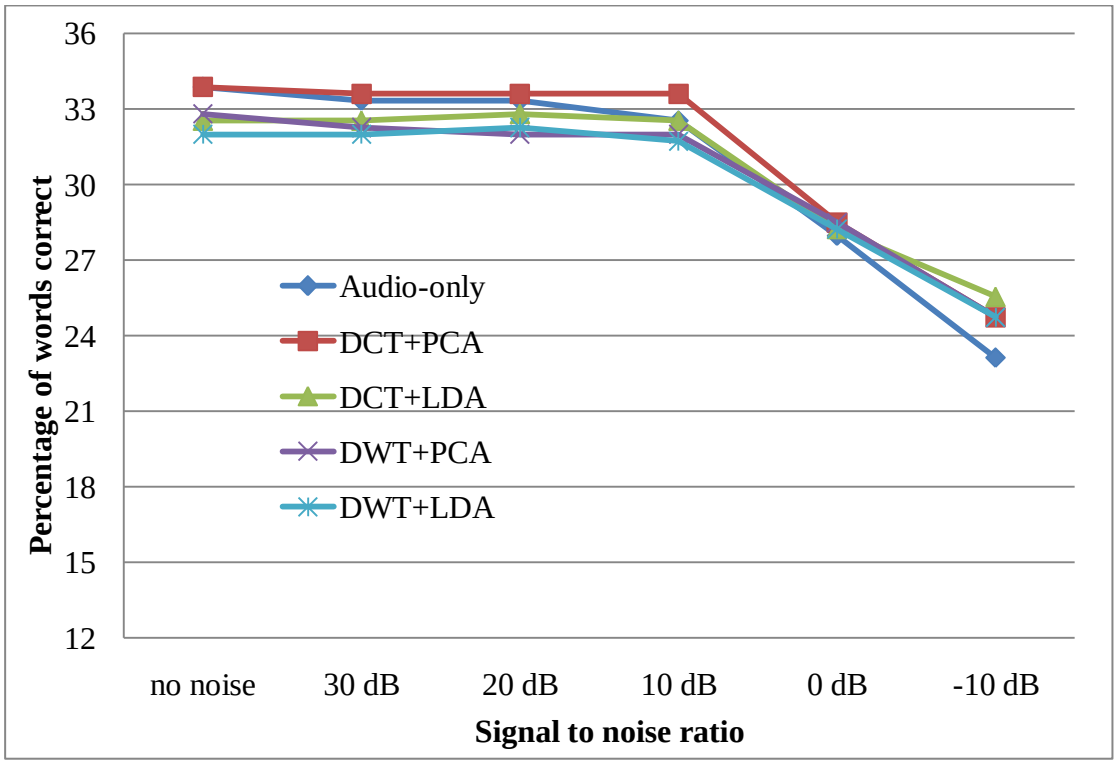


Figure 4.19 Performance of audio-only and audio-visual ASRs under noise

Figure 4.19 shows that, for clean speech, the performances of both the audio-only and AVASR approaches using DCT and PCA, are very similar. With the increase in audio noise, the performance of the audio-only recogniser degrades rapidly; the performance of AVASR is affected to a lesser extent as the video modality is unaffected by the audio noise and thus gave better recognition compared to audio-only ASR.

For clean speech, the information content in the audio stream is superior to that of video stream. Below a signal-to-noise ratio of 0 dB, the performance of the audio modality solution was severely affected while the video stream is unaffected by this noise. However, due to the equal contribution of both modalities to the features used, the AVASR implementation was affected by a degradation in the performance of audio modality, as can be seen from the AVASR results of Figure 4.18. A remedy for this problem could be to introduce an appropriate weighing of the two modalities in accordance with the modality reliability. In this work, experiments were carried out to adjust the weights for the two modalities in order to obtain the best performance under a variety of noise conditions. This was achieved by using a multi-stream HMM where the audio and video streams were assigned weights α and β respectively, such that $\alpha+\beta=1$. The value of α was varied from 1 to 0 in steps of 0.1, thus effectively providing the flexibility to alter the recognizer from being audio-only ($\alpha=1$) through a combination of audio and video, to video-only ($\alpha=0$). The results from the AVASR with stream weights determined to provide the best performance, are shown in Figure 4.20.

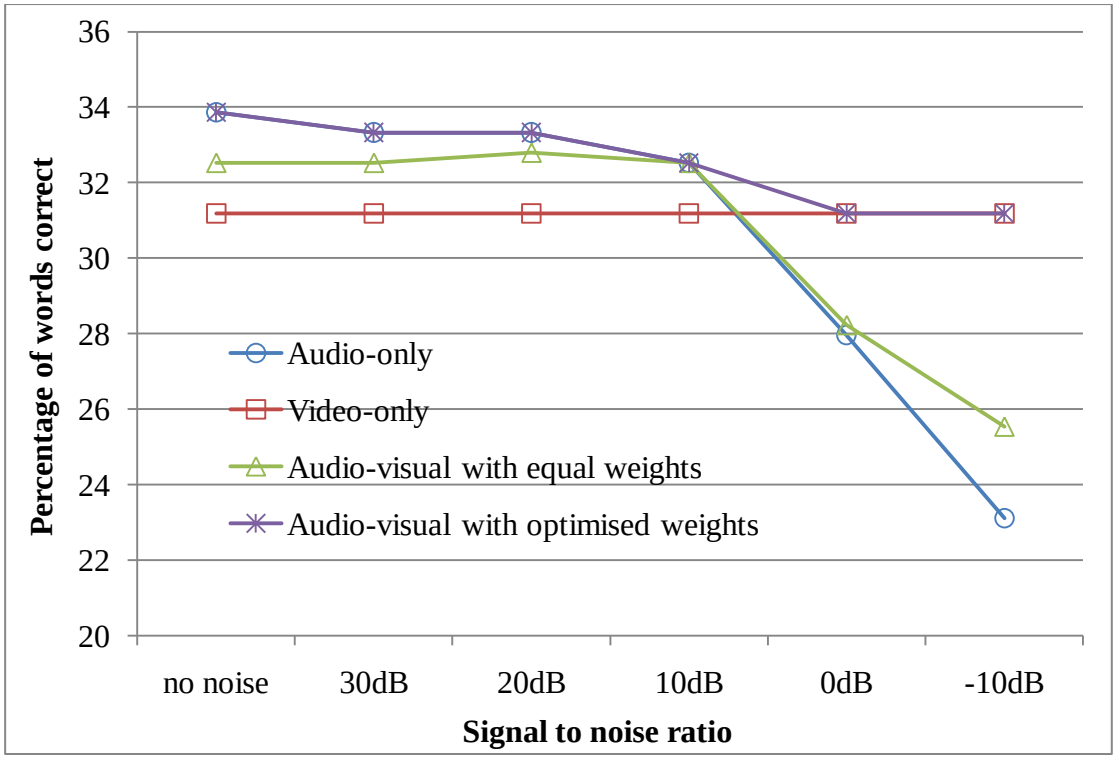


Figure 4.20 AVASR performance with streams optimised according to noise level

As can be seen from Figure 4.20, the AVASR system with tuned audio and video weights gave the best recognition results for all signal-to-noise ratios tested, and this is because the approach is able to exploit information content present in the two modalities at many different levels of noise. In traditional AVASR systems with equal weights for audio and video streams, then, for clean speech, the classification may become confused by the one-to-many mapping of viseme to phonemes and the relatively low speech information content found in video streams. Where tuned weights are used, these problems are overcome by rectified resorting to an audio-only mode. When audio noise is present, the audio stream is corrupted while the video stream remains unaffected. Consequently applying higher weightings to video streams in noisy conditions helps to avoid the ASR misclassifications that result from the presence of audio noise. This ability to select different sensors according to environmental conditions is somewhat akin to the approach taken by humans performing speech recognition.

4.5 DISCUSSION AND CONCLUSION

The work in this chapter has investigated how a range of different types of video features affects automatic speech recognition performance. Because there is no

common agreement on which benchmark database AVASR community should adopt, and as face and mouth extraction techniques are not yet mature, it is not possible to directly compare the results obtained here with previous work. To overcome this, the results obtained in this chapter have been obtained by re-implementing the techniques reported in the AVASR literature and using a single continuous speech recognition database obtained for a large number of subjects. The proposed new region-based features are compared with commonly-used low frequency features and the results are reported on both visual only and audio-visual speech recognition, implemented without using a language model in order to provide a direct comparison of the methods. The results show that mid-frequencies in both the DCT and DWT transform were able to give better speech recognition performance than the commonly-used low-frequency coefficients, irrespective of the dimensionality reduction method applied. This is probably because the intermediate level features contain information at frequencies similar to those exhibited by the lip moments. LDA is able to separate the speech classes and was shown to provide a superior dimensionality reduction technique for ASRs when compared to PCA. The results also demonstrate that, in general, the DCT-based features give better performance compared to wavelet transform based features. The visual modality inherently contains less information about speech than the audio modality, mainly because the audio modality is richer in information content, but also due to the total or partial occlusion of various articulators such as the tongue, the teeth and larynx. In addition, the mapping between visemes and phonemes is one-to-many, implying that not all phonemes are visually distinguishable. Due to these limitations of visual speech recognition, adding visual features may have no benefit or even a degrading effect as it may cause confusion during phoneme classification. The real benefit of the video modality occurs in the presence of audio noise where the performance of the audio speech recognition worsens, but the visual speech information remains unaffected. The results of experiments on changing the individual stream weights according to noise level demonstrated that a noise adaptive scheme could make good use of the visual modality by controlling the contribution of the two modalities in accordance with the noise level in the environment of application.

In this chapter one of the core areas of AVASR, namely visual feature extraction, has been investigated. The feature extraction approach presented in this chapter falls into the category of appearance-based methods. The next chapter presents a novel approach

to the automatic extraction of the mouth ROI, another important area in AVASR research, while chapter 6 presents a new approach to visual feature extraction based on the motion information obtained from videos of speaker.

4.6 REFERENCES

[1] Chibelushi, C. C., Deravi, F., and Mason, J. S. D. (2002), "A Review of Speech-Based Bimodal Recognition", *IEEE transaction on multimedia*, vol. 4, no. 1, pp. 23-33.

[2] Zhang, X., Mersereau, R. M., Clements, M., and Broun, C. C. (2002), "Visual speech feature extraction for improved speech recognition", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1993-1996.

[3] Aleksic, P. S., and Katsaggelos, A. K. (2004), "Comparison of Low and High-Level Visual Features for Audio-Visual Continuous Automatic Speech Recognition", *Proceedings of the ICASSP 2004*, pp. 917-920.

[4] Potamianos, G., Verma, A., Neti, C., Iyengar, G., and Basu, S. (2000), "A Cascade Image Transform for Speaker Independent Automatic Speechreading", *Proceedings of the IEEE International Conference on Multimedia and Expo*, New York, vol. II, pp. 1097-1100.

[5] Zhao, G., Pietikainen, M., and Hadid, A. (2007), "Local spatiotemporal descriptors for visual recognition of spoken phrases", *Proceedings of 2nd international workshop on Human-Centred Multimedia (HCM 2007)*, pp. 57-65.

[6] Jang, K. S. (2007), "Lip Contour Extraction Based on Active Shape Model and Snakes", *International Journal of Computer Science and Network Security*, vol. 7, no. 10, pp. 148-153.

[7] Gurbuz, S., Tufekci, Z., Patterson, E., and Gowdy, J. N. (2001), "Application of Affine-invariant Fourier Descriptors to Lipreading for Audio-Visual Speech Recognition", *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, pp. 177-180.

[8] Kaynak, M. N., Zhi, Q., Cheok, A. D., Sengupta, K., Jian, Z., and Chung, K. C. (2004), "Analysis of Lip Geometric Features for Audio-Visual Speech Recognition", *IEEE Transaction on System, Man and Cybernetics-Part A: System and Humans*, vol. 34, no. 4, pp. 564-570.

[9] Potamianos, G., Graf, H. P., and Cosatto, E. (1998), "An Image Transform Approach for HMM Based Automatic Lipreading", *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 173-177.

[10] Arsic, I., and Thiran, J. P. (2006), "Mutual Information Eigenlips for Audio-Visual Speech Recognition", *Proceedings of 14th European Signal Processing Conference*, Lecture Notes in Computer Science.

[11] Potamianos, G., Neti, C., Huang, J., Connell, J. H., Chu, S., Libal, V., Marcheret, E., Haas, N., and Jiang, J. (2004), "Towards Practical Deployment of Audio-Visual Speech Recognition", *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, pp. III-777-780.

[12] Connell, J. H., Haas, N., Marcheret, E., Neti, C., Potamianos, G., and Velipasalar, S. (2003), "A Real-Time Prototype for Small-Vocabulary Audio-Visual ASR", *Proceedings of the International Conference on Multimedia and Expo*, Baltimore, Maryland, vol. II, pp. 469-472.

[13] Reilly, R., and Scanlon, P. (2001), "Feature Analysis for Automatic Speechreading", *Proceedings of Workshop on Multimedia Signal Processing*, pp. 625–630.

[14] Khayam, S. A. (2003), "The discrete cosine transform (DCT): Theory and application", Technical Report WAVES-TR-ECE802.602, Michigan State University.

[15] Strang, G. (1999), "The discrete cosine transform", *SIAM Review*, vol. 41, no. 1, pp. 135–147.

[16] Walker, J. S., and Nguyen, T. Q. (2000), "Wavelet-based image compression", Yip, P. C., and Rao, K. R., (Eds.), *The Transform and Data Compression Handbook*, vol. 1, ch. 6, pp. 267-312, CRC Press, Boca Raton, 2000.

[17] Shiratori, H., Goto, H., and Kobayashi, H. (2006), "An efficient text capture method for moving robots using DCT feature and text tracking", *Proceedings of the 18th international Conference on Pattern Recognition (ICPR)*, pp. 1050-1053.

[18] Dabbaghchian, S., Aghagolzadeh, A., and Moin, M. S. (2008), "Reducing the effects of small sample size in DCT domain for face recognition", *proceedings of the IEEE International Symposium on Telecommunication*, pp. 634-638.

[19] Chen, W., Er, M. J., and Wu, S. (2006), "Illumination Compensation and Normalization For Robust Face Recognition Using Discrete Cosine Transform in Logarithm Domain", *IEEE Trans on Systems, Man and Cybernetics – Part B*, vol. 36, no. 2, pp. 458-466.

[20] Tico, M., Kuosmanen, P., and Saarinen, J. (2001), "Wavelet domain features for fingerprint recognition", *IEEE Electronic Letters*, vol. 37, no. 1, pp. 21-22.

[21] Wong, Y. W., Seng, K. P., and Ang, L. M. (2008), "M-Band Wavelet Transform in Face Recognition System", *Proceedings of ECTI-CON*. pp. 453-456.

[22] Sanderson, C., and Paliwal, K. K. (2002), "Polynomial Features for Robust Face Authentication", *Proceedings of IEEE International Conference on Image Processing*, Vol. 3, pp. 997-1000.

[23] Nilsson, M., Nordberg, J., and Claesson, I. (2007), "Face Detection Using Local SMQT Features and Split up SNoW Classifier", *ICASSP*, pp.589-592.

[24] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2006), *The HTK Book V3.4.*

[25] Ahmad, N., Datta S., Mulvaney, D., and Farooq, O. (2008), "A comparison of visual features for audiovisual automatic speech recognition", *Proceedings of the 2nd joint conference of the Acoustical Society of America (ASA) and the European Acoustics Association (EAA), Acoustics'08,* Paris, pp. 6445-6449.

[26] Jun, H., and Hua, Z. (2009), "Research on Visual Speech Feature Extraction", *proceedings of IEEE 2009 international Conference on Computer Engineering and Technology*, pp. 499-502.

[27] Matthews, I., Potamianos, G., Neti, C., and Luettin, J. (2001), "A comparison of model and transform-based visual features for audio-visual LVCSR", *Proceedings of International Conference on Multimedia and Expo*, pp. 22-25.

[28] Huang, J., Marcheret, E., and Visweswariah, K. (2005), "Rapid feature space speaker adaptation for multi-stream hmm-based audio-visual speech recognition", *proceedings of IEEE International Conference on Multimedia and Expo*, pp. 338-341.

[29] Gagnon, L., Foucher, S., Laliberte, F., Boulianne, G. (2008), "A simplified audiovisual fusion model with application to large-vocabulary recognition of French Canadian speech", *Canadian Journal of Electrical and Computer Engineering*, vol. 33, no. 2, pp. 109-119.

[30] Potamianos, G., Neti, C., Luettin J., and Matthews, I. (2004), "Audiovisual automatic speech recognition: An overview", Bailly, G., Bateson, V. V., and Perrier, P. (Eds.), *Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.

[31] Varga, A. P., Steenekan, H. J. M., Tomlinson, M., and Jones, D. (1992), "The noisex-92 study on the effect of additive noise on automatic speech recognition", *Technical Report*, DRA Speech Research Unit.

CHAPTER 5

VISUAL REGION OF INTEREST EXTRACTION FOR AVASR

This chapter presents a novel motion based approach for visual region of interest (ROI) extraction for AVASR purposes. The movements of the speakers in videos of speech are used to identify the mouth region, which is further processed to isolate a ROI from which visual features for AVASR can be generated. The work presented in this chapter depicted as part of the general AVASR system of Figure 2.1 is shown in Figure 5.1.

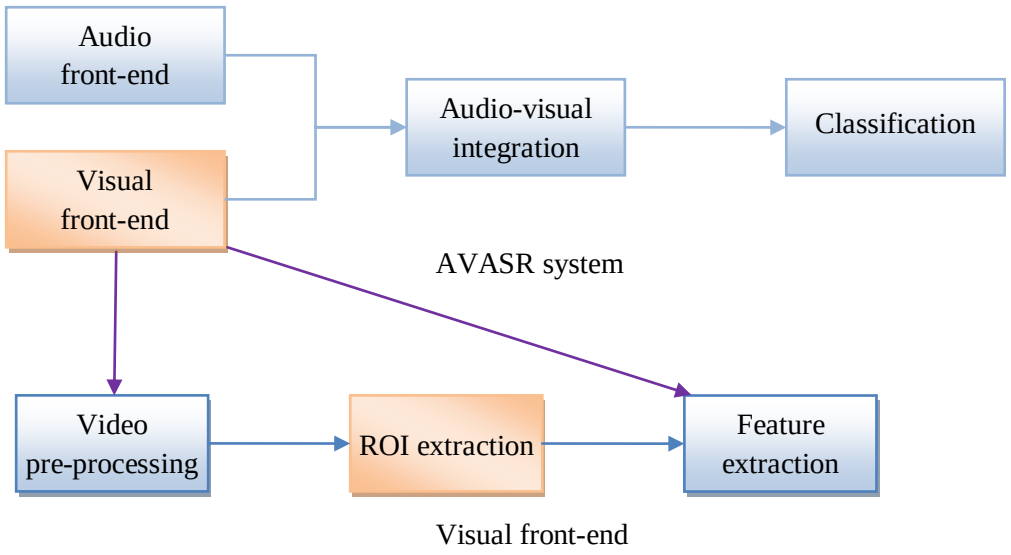


Figure 5.1 Location of the ROI extraction process in the general AVASR system

The chapter is organized as follows. Section 5.1 provides an introduction to the extraction of visual ROI for AVASR and discusses as how different feature extraction approaches affect the required approach to ROI extraction. It also outlines the impact of robust ROI extraction on the extraction of informative visual features and its role in the overall performance of the AVASR system. Section 5.2 discusses currently used ROI extraction approaches and provides a detailed discussion of the concepts behind these approaches. Section 5.3 provides an introduction to motion estimation in video and also discusses the most commonly used motion estimation approaches, namely

the intensity-based and feature-based methods. Section 5.4 discusses motion based approaches for ROI extraction for application in AVASR. In section 5.5, the proposed intensity-based ROI extraction approach introduced in this thesis is described and results obtained by practical application of the method are given. In section 5.6, a new feature-based ROI extraction method is described and initial results given. Section 5.7 concludes the findings of this study and provides a commentary on the performances of the proposed methods.

5.1 VISUAL REGION OF INTEREST (ROI) FOR AVASR

The visual front-end identifies the portion of the speaker's face that contains the most speech information and extracts that information in a parametric form suitable for processing by the recognizer. Front-end design can be divided into two sub-tasks, region of interest (ROI) extraction and feature extraction [1]. Though often considered separately, the two tasks are largely interdependent. The ROI provides the raw input data for visual feature extraction and thus the overall performance of an AVASR system is greatly influenced by the accurate extraction of ROI [2].

In appearance-based feature approaches the whole mouth region is considered as a source of speech information. Some researchers argue that the jaw and chin moments also provide useful information about the speech and therefore need to be included as part of the ROI [3], reducing the ROI identification task to one of detection of the lower half of the face containing the mouth along with other articulators. This crude initial estimate is, in practice further refined by filtering out un-required parts such as the nostrils and the background on either side of the chin. The shape-based feature approaches extract information regarding the lip geometry and compared to appearance-based approaches, require a more robust lip contour estimation.

The required ROI thus depends upon the feature extraction approach used and the use of a wide range of different ROIs have been reported in literature, ranging from entire face of speaker to the lower half of face and mouth region only [4]. For appearance-based feature approaches, the desired ROI is obtained by extracting a bounding box around the detected mouth/lips region containing the mouth region and perhaps other articulators, from which the visual features are then extracted by applying a suitable transformation to the ROI such as DCT [9] or DWT [1]. For shape-based features, the

required ROI is the region around the detected mouth or lips from which the geometric parameters of lips can then be extracted by employing a suitable algorithm. In some cases the boundary points of the detected lip are used to provide initial estimates of the lip model and the model then determines the exact lip contour by iteratively refining its parameters. Various approaches such as edge tracking [6], template matching [7] active shape and appearance model [8] and snakes [9] have been used for lip contour estimation.

ROI detection and extraction is fundamentally an image analysis task and development in image analysis literatures leads to robust ROI extraction and thus to the effective extraction of informative visual features for AVASR. For example, the approaches that involve certain pre-defined lip models have not included other visible articulators such as the tongue and teeth, probably due to the difficulty in modelling these articulators. In [10], Saenko argues that the use of lip shape features alone cannot differentiate visemes in different contexts, and that a multi-articulator based approach is more useful for such classification. In this work, the articulatory features are used implicitly by developing HMM models as outputs of multiple underlying articulators rather than extracting features from individual articulators separately. Extracting features from visible articulators other than lips and their use along with lip contour features could potentially improve the performance of current AVASR systems. Developments reported in image analysis research have lead to the more accurate mouth/lip detection and lips parameter estimation and thus to new approaches to feature extraction.

Early research on AVASR system design focused both on the analysis of the visual modality for the extraction of informative visual speech features and on the integration of audio and video streams, while the ROI extraction task was generally ignored [1]. The AVASR tasks reported in these works is commonly limited to frontal face AVASR in a controlled environment. The ROI in those studies was extracted either manually or otherwise the extraction task was simplified by applying visible markers to the lips of the speaker. Also the corpora used are face-centred with limited variation in orientation and lighting. In some corpora, the mouth region coordinates are determined manually for use in a series of subsequent research studies [11], [12]. However, to achieve a real-time and general purpose speech recognition system, it is essential to detect and track the face and mouth automatically with any pose and in

unrestricted environmental conditions without any artificial marking. Due to this realization and tremendous impact of accurate ROI extraction on overall performance of AVASR systems, research on visual ROI has attracted the interest of many researchers. In recent research, a number of automatic ROI extraction methods have been proposed [13], [14].

5.2 AN OVERVIEW OF VISUAL ROI EXTRACTION

The identification of the ROI is made more difficult due to the high deformation of lip shape, as well as the variation in the content of the mouth region due to the presence or absence of tongue, teeth, and opening and closing of mouth during speech. Mouth/lip detection approaches are also often influenced by variations in lighting conditions and changes in the pose and orientation of the speakers. The presence or absence of a beard or moustache also presents a possible source of confusion that reduces the effectiveness of generic ROI extraction algorithms.

The general steps of the typical ROI extraction task are depicted in Figure 5.2. Although attempts have been made to detect directly the mouth or lips of the speaker [13], [14], they have met limited success. This is because mouth/lips exhibit few easily distinguishable features, the face and lip colours are largely correlated and the mouth contour becomes deformed during speech. To help define an initial estimate of the mouth ROI, most ROI extraction methods use face detection as a first step.

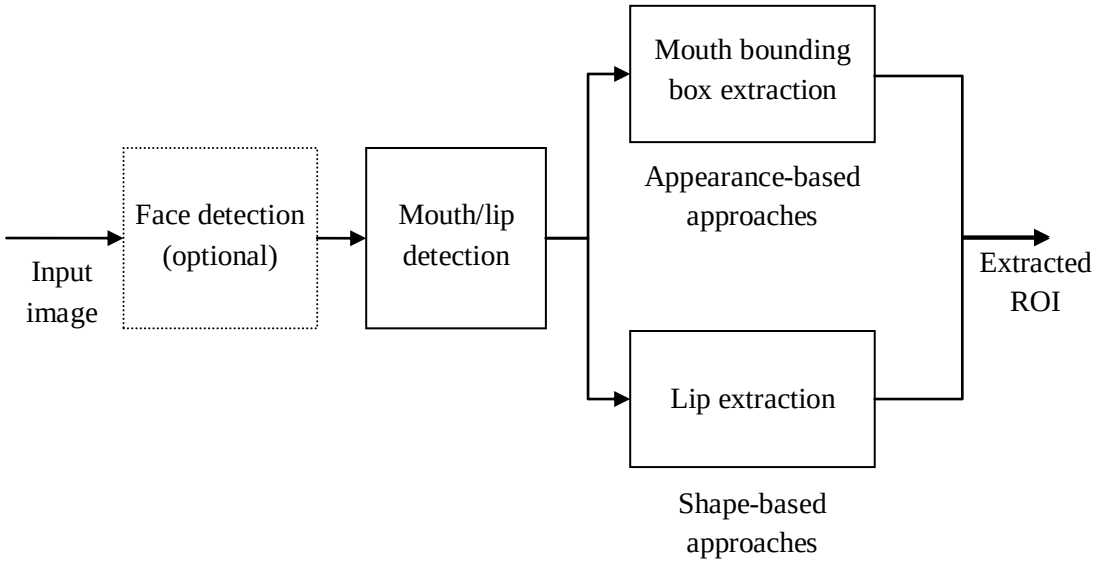


Figure 5.2 Block diagram of visual ROI extraction

First the face of the speaker is identified and isolated from the background. Various face detection techniques based on skin colour, geometry of face and facial features have been reported in literature [15], [16], [17] and [18]. Face detection is followed by mouth/lip region detection. The lower half of the face region is assumed to contain the mouth and other visible articulators and is generally used as a starting point for a further refined search for the detection of the mouth/lips region. Techniques similar to face detection have also been used for mouth/lip detection. The approaches used for mouth/lip detection include traditional image processing techniques such as colour segmentation and edge detection techniques [19], template matching [20], deformable templates [21], symmetry based methods [17] and statistical techniques such as [22]. As the shape of the mouth and lips goes through considerable deformation during speech and contains very few distinguishable features, the detection is often facilitated by referring to other facial features. For instance, a region aligned with the centre of eyes may be regarded as the best candidate for an initial reference in frontal view AVASR tasks. Detecting the ROI in each frame of video separately would be a time consuming task, and hence a ROI tracking approach is often adopted instead. The desired ROI is commonly identified in first frame of utterance and tracked in the remaining frames of video. As the mouth movement in consecutive frames is small, the tracking process is typically easier and less time consuming.

Face and mouth detection, extraction of ROI and tracking of ROI along the frames of video are discussed below in further detail.

5.2.1 Face and mouth detection

Face detection is used in many different branches of research such as surveillance systems, expression recognition, emotional/cognitive state recognition and audio-visual speech recognition. Typical challenges in face detection are orientation, presence of beard and moustache, facial expression, non rigidity, lighting conditions, size, partial occlusion and noise. Profile view face detection (which is of interest in profile view AVASR), has also received attraction recently but that is still a challenging task. Face detection is a special case of image segmentation and object detection. Face detection approaches could be further extended to mouth and lip detection by applying additional constraints. Different approaches used for face and mouth/lip detection can be broadly grouped into one of the four categories [23].

Knowledge-based methods

Knowledge based methods for face detection describe the face in terms of parameters based on human knowledge. The rules are defined on the basis of constituent parts of face and their mutual relationship. These rules are then utilized to guide the search for face in the target image. The regions that fit the rules are identified as faces [24]. Yang and Huang in [25], proposed a three-level face detection method. At the first level, the image is scanned at the lowest resolution and all the regions in the image with uniform intensity are extracted as face candidates. At the second level, the number of candidate faces is further refined by scanning the regions identified in the first level and its neighbouring pixels at higher resolutions. At the highest level, the face candidates are searched for the structural components of the face and either verified or rejected as being part of a face based on geometric relationships and intensity information. The method described in [24] uses a symmetry operator to locate the line of symmetry in the image. Facial features, such as the eyes and mouth, are then located with reference to this symmetry line. This method can be used to locate a single face in the image and claims to locate faces for a wide range of rotation, scale and lighting variations. It is simple to devise some basic rules but as human can’t completely transform its knowledge to definite rules; the model cannot completely reflect the human knowledge. Pose variations and cluttered images are

challenging issues in defining a common rule. In the case of mouth detection, defining these rules become further challenging problem due to its non rigidity and presence and absence of different parts such as tongue and teeth.

Feature invariant approaches

Feature invariant methods use a bottom-up approach for face detection and localization [26]. They search for features that are present in faces irrespective of the pose, size and variation in lighting conditions. As human can detect facial features even at different poses and illumination, it is assumed that detecting individual features could lead to complete face detection. Facial features such as eyes, nose, mouth and chin are detected using shape, colour, texture or edge detection techniques. In [27], a number of regions of connected edges are detected in the image with the assumption that they form facial landmarks. The eyes and eyebrows are then detected by identifying landmark pairs of horizontal orientation and the centre point defined by these features are used to determine the nose and mouth locations. The landmarks thus obtained are verified by the use of facial geometry. Bevilacqua *et al.* in [28] detected the eyes using template matching and support vector machine (SVM), with their positions used as initial reference locations to determine the nose and mouth positions. The detected features are then combined to fit a model of the entire face. Although facial features are substantially invariant to pose and location, they are difficult to detect in presence of noise, occlusion and in case of complex background.

Template matching

In these methods, predefined face patterns or templates are stored and the input image is searched for these templates during the detection process. The correlation between the stored templates and the searched regions in the image itself are used to produce the similarity measures. In [29] a frontal-view face template was generated by taking the mean of 36 frontal face vectors. Regions in the image of uniform skin colour and containing at least one hole are selected as candidate faces. The regions are correlated with the face template, and those having correlation values above 0.6 are classified as detected faces. These methods are simple to implement, but as the templates are based on specific poses and orientations, they cannot detect faces with different poses and sizes and their generality is limited. There has been some research to make the templates more flexible to fit into cases of varying poses and size. Deformable

templates, sub-templates and multi-scale templates have been used to address the issue of non-rigidity, occlusion and size of faces and facial features. Chandramohan and Silsbee [22], proposed a multiple deformable template model. In this work they argue that a single template, even if deformable, cannot describe all the possible two-dimensional projections of an object that may occur in practice. The search for target lip detection took place in two stages. In the first stage, a rough scan of the image produced a coarse selection of template and initial parameters, while the second stage involved varying parameter values so that the penalty function converged to a local minimum. The use of mouth templates and its deformable variants have extensively been reported in AVASR research both to detect the mouth region and to extract the mouth parameters that are then used in as a visual feature vector for shape-based AVASRs.

Appearance-based methods

Unlike the template based methods where the patterns are pre-defined, appearance-based methods learn the face patterns from the training data [30], [31] often for use in identifying whether an object is a face [32] or for face recognition purposes [33]. These methods use statistical analysis techniques to classify objects or regions into either face or non-face classes based on a probabilistic framework. The images are represented as variable x associated with class conditional probabilities, P(x|face) and P(x|non-face). As the dimensionality of variable x is usually high, then, to compute these probabilities directly, they are transformed to a lower dimensional space using suitable dimensionality reduction techniques, such that

$$y = Wx \tag{5.1}$$

where y is the output lower dimension vector and W is the transformation matrix. The dimensionality of y is substantially lower as compared to x, suitable for calculating the class conditional probabilities. Bayesian classifiers, artificial neural networks, the Fisher linear discriminant or other suitable classifier can then be used to classify the transformed variable as a face or a non-face class. These methods have been widely used in AVASR research to reduce the high dimensional video data to a reasonably small number of dimensions [5].

5.2.2 ROI Extraction

After identification of speakers mouth/lips region the next stage is the extraction of the ROI. For appearance based feature approaches a bounding box around the lower half of the face containing the mouth region is extracted as desired ROI. As same size of bounding box need to be used for the application of transformation, the size of the bounding box is selected according to the maximum mouth size along all utterances. For shape-based feature approaches, the region containing mouth or lips is extracted as desired ROI, which is then processed further to extract lips and determine the lip parameters. Typical approaches used for lip contour extraction are edge detection [35] and colour based segmentation techniques [18].

5.2.3 ROI tracking

As visual features are extracted from each frame of video, the ROI needs to be extracted from every video frame. This could be done either by detecting a ROI in each frame of video independently or alternatively the coordinates of ROI are found in one frame and tracked along the remaining frames of video. The latter approach is commonly preferred due to much reduced computation time. In restricted conditions where head movement is small, the tracking task can be omitted for appearance based features by selecting a larger spatial window around the detected mouth region. The coordinates of bounding box are selected in such a way that it contains the desired ROI in all the frames of utterance.

For shape features, a more sophisticated mechanism for ROI tracking is implied. In this case the tracking mechanism instead of extracting the ROI helps to reduce the search area for lip boundary detection and also improves the performance of lip boundary estimation algorithm by avoiding the occurrences of false positives in the background and other parts of the face such as eyes and face wrinkles.

5.3 MOTION ESTIMATION IN VIDEO

The work in this thesis reports a novel motion-based approach to visual ROI detection and extraction for AVASR systems. Before discussing the proposed motion-based approach for ROI detection purposes, the background of motion detection in video

(sequence of frames) is discussed as this has been the most fertile application area for these approaches.

Sensing and estimating motion in video is of great interest in many fields of research and has many practical areas of use ranging from defence, security and surveillance to medical applications. Humans have the ability to discern objects, sense their state of motion or rest, and to comprehend their motion in three-dimensional space. Computer vision research attempts to replicate the human ability in a machine. However, this is rather a difficult task as it is not known how exactly the human motion sensing mechanism work. Neuropsychologists and psychophysicists are aiming to understand human vision systems while computer scientists and engineers conduct research on developing machine vision systems to detect objects in images for identification and motion tracking applications. The findings in the two areas of research have a cross impact on each other [36].

Motion in a sequence of image frames can come from the motion of the camera or of the objects in a scene. The problem is ill posed as the three dimensional motion of objects and camera zoom have an impact than the actual motion perceived. Two distinct approaches, tracking objects feature and the change in brightness level have been adopted in the literature for motion estimation in video. Object feature approaches detect certain distinct features of objects such as vertices, edges and curves in the image and track these features in a sequence of frames to estimate the speed and direction of motion. Three dimensional motion is calculated from the two-dimension motion in frames based on 2-D to 3-D motion conversion models. An alternative method, referred to as the optical flow approach, determines motion from the rate of temporal variation in the intensity values of pixels. Motion estimation has also been used for communication applications and video compression and due to its relevance to the work presented in this thesis, a discussion on the use of motion compensation approach in MPEG based video compression is provided in chapter 6.

Feature-based motion estimation

Feature-based (sometimes termed region-based) motion estimation extracts a set of features from a region in a frame of video and searches for the same features in subsequent frames and identifies the region in the frame that provide the best match. The matching criteria is defined on the basis of a similarity measure, such as

maximizing the cross correlation between the region in the first frame and the corresponding region (and neighbouring regions) in later frames. Alternately an error minimization criteria could be used such as the minimum absolute difference (MAD) or the mean square error (MSE). In practice, the region-based motion estimation is implemented by dividing the image into a number of macro-blocks of size 8x8 or 16x16. For a block B of N x N pixels, the MAD is given as

$$MAD(\Delta m,\Delta n) = \frac{1}{N^2} \sum_{(m,n)\in B} |y(m+\Delta m, n+\Delta n, t+\Delta t) - y(m,n,t)| \qquad (5.2)$$

where $y(m,n,t)$ is the value of a pixel in B in the reference frame at time t and $y(m+\Delta m, n+\Delta n, t+\Delta t)$ is its value at time $t+\Delta$t, assuming a displacement of Δm and Δn in the horizontal and vertical dimensions respectively.

The MSE can be written as

$$MSE(\Delta m,\Delta n) = \frac{1}{N^2} \sum_{(m,n)\in B} [y(m+\Delta m, n+\Delta n, t+\Delta t) - y(m,n,t)]^2 \qquad (5.3)$$

Although the method primarily estimates the translational motion between frames, it could be extended to rotational and scaling matching by piecewise translation of regions. The search window may be of fixed size where the matching is assessed for all the points in the region, or alternatively the search is terminated based on a match, where measure of correlation or minimum error exceeds a given threshold. A number of search patterns for block-matching have been reported in the literature for the fast and efficient implementation of the feature-based motion estimation approach. Two of these methods, namely the 'three step search' (TSS) and 'adaptive rood pattern search' (ARPS) were use in this work. In the TSS method, the best match to a block in the current frame is searched in the subsequent frame by iteratively updating the location of the centre and altering the size of the search window. The initial starting point for the search is to use the coordinates of the block in the current frame and a search window of size 8x8. The ARPS method utilises the fact that the macro-blocks in the neighbouring location often have similar motion patterns and therefore the search direction and the step size is determined statistically from the motion pattern in

the neighbouring blocks. A detailed description of these methods can be found in [41].

Feature-based motion estimation performs well for rigid bodies with sharp features such as edges and corners, but its performance is affected adversely by occlusion, detection of false features and deformation of non-rigid objects.

Intensity-based motion estimation (Optical-flow method)

The relative motion of the objects in a scene with respect to the image sensor gives rise to in brightness changes in the objective plane. Intensity (or optical-flow) based motion estimation approaches determine motion by calculating the instantaneous variation in the intensity pattern in a sequence of video images. A velocity map known as optical flow is obtained by analyzing the changes in the brightness values at each pixel position. The motion and structure of entire object is then recovered by optical flow based clustering of regions in the image.

If $f(x,y,t)$ is the intensity of a point p(x,y) in the image at time t and point p$(x+\Delta x, y+\Delta y)$ has the same intensity at time $t+\Delta t$ then

$$f(x+\Delta x, y+\Delta y, t+\Delta t) = f(x,y,t) \tag{5.4}$$

where Δx, Δy and Δt are small changes in the horizontal, vertical and temporal dimensions respectively. Expanding the left-hand side of equation 5.4 using the Taylor series gives

$$\begin{aligned} & f(x,y,t) + f_x \Delta x + f_y \Delta y + f_t \Delta t + \text{higher order terms} \\ & = f(x,y,t) \end{aligned} \tag{5.5}$$

where f_x, f_y and f_t are the partial derivatives in the x, y and t dimensions. Ignoring higher order terms, then

$$f_x \Delta x + f_y \Delta y + f_t \Delta t = 0 \tag{5.6}$$

or

$$f_x u + f_y v + f_t = 0 \quad (5.7)$$

where $u = \Delta x/\Delta t$ and $v = \Delta y/\Delta t$ are the desired velocity components along horizontal and vertical direction of motion.

Equation 5.7 can be solved by making additional assumptions, such as the optical flow is constant for all the points of same object. A variety of methods are available for solving equation 5.7, and these are discussed in detail in [36]. A detailed discussion of different approaches have been taken in the implementation of intensity-based motion estimation can be found in [37]. For motion detection applications, the approach is often implemented using the difference in the intensities of the corresponding pixels of consecutive frames of videos [38].

The performance of optical flow methods can be adversely affected by the presence of noise and brightness variations due to other sources such as changes in lighting conditions as well as occlusion which violates the continuity assumption.

5.4 MOTION BASED APPROACH FOR ROI EXTRACTION IN AVASR

In most current AVASR research, the ROI, if not extracted manually, uses approaches developed in image analysis research, such as differences in skin and lip colour, facial features and their spatial relationships. These methods commonly operate on individual images, and so use only a small portion of the available information, well are affected by the high correlation between the skin and lip colours or fail due to miss-identification of features in the mouth region or the non-rigidity of lips.

Although speech is a dynamic phenomenon and the advantage of motion information for both audio and visual speech recognition is well proven, an explicit use of motion information is rarely reported in AVASR research. In the previous approaches taken that include visual speech dynamics have involved taking the temporal derivatives of static features extracted from individual frames or by concatenating image frames before feature extraction. However, to the best of author's knowledge the explicit use of motion information for ROI detection purpose has not been reported. This may be because the research on AVASR has mostly focussed on discriminative feature extraction and modality integration with little attention paid to the ROI extraction

task. Also, the ROI extraction methods used in AVASR research are mainly borrowed from the image analysis literature where the inputs are individual images rather than sequences. The work presented in this thesis takes a new approach to visual ROI extraction based on the motion between frames of video of speakers. Motion-based approaches have a significant advantage over the appearance-based approaches due to their greater tolerance to changes in lighting conditions. In the following sections, two new motion-based ROI methods are introduced. Firstly, an intensity-based method in described and compared with baseline colour based approaches, this being a popular method for lip region detection found in literature [18]. Secondly, the implementation of a feature-based method is described.

Database

The database used in this work was a subset of the VidTimit audio-visual database [57], taken from its video part and consist of 16 speakers (8 male and 8 female) each uttering 10 sentences. The data obtained is composed of 160 utterances with a total number of 16510 images at a resolution of 512x384 pixels and 24 bit depth. The video is recorded at 25 frames per second.

5.5 INTENSITY-BASED ROI EXTRACTION

The intensity-based ROI extraction method introduced in this section is depicted in Figure 5.3. The process of ROI extraction is performed in two stages. The first stage utilizes the relative motion of objects in the image sequence to detect the mouth of the speaker, while the second stage extracts the desired mouth ROI for AVASR purposes. As the required ROI depends on the feature extraction approach adopted; in this work one such bounding box containing the mouth region is extracted, suitable for appearances-based feature approaches while for shape based approaches, the lip region is extracted from which the geometric parameters can be computed or alternatively the model parameters can be estimated.

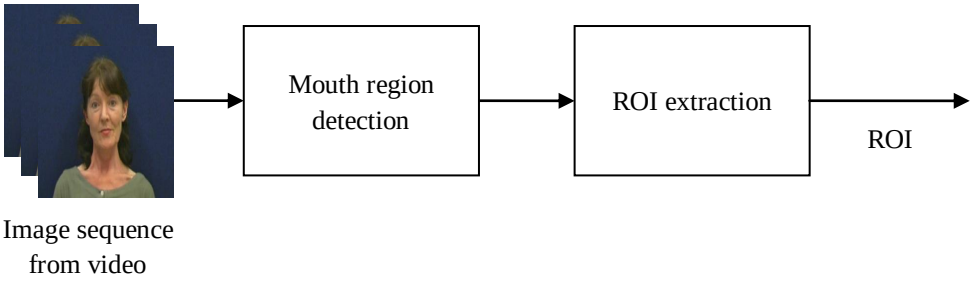


Figure 5.3 Block diagram of the proposed motion based ROI extraction

5.5.1 Mouth region detection

Motion in different parts of the face conveys different expressions such as happiness, sadness, fear and surprise [40]. Humans perceive these expressions by observing the motion of some facial parts relative to others. During speech the lips undergoes through the highest amount of motion compared to other parts of the face and background. Eyes are other such parts that undergoes through significant motion. The higher relative motion in the mouth region during the speech could be used to automatically detect the mouth region of the speakers. The mouth region detection approach is shown in Figure 5.4.

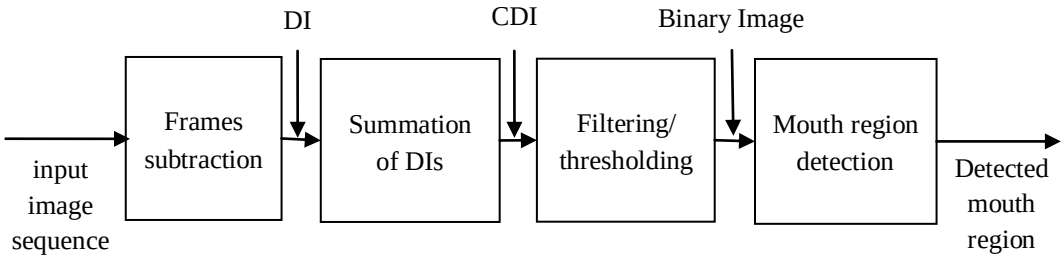


Figure 5.4 Mouth region detection process

Motion calculation

The motion between consecutive frames of video is represented by changes in intensity values. Here the change is first determined by simply finding the difference in values of corresponding pixels between frames. The resultant image is referred to as difference image (DI). An alternative approach that was considered is a macro-block based motion vector method that was adopted for estimating the motion of lips and other articulators and is further discussed in section 5.6. The DI is given by

$$\mathrm{DI}_i(x,y) = I_{i+1}(x,y) - I_i(x,y) \qquad 1 \le i \le T-1 \tag{5.10}$$

where $I_i(x,y)$ and $I_{i+1}(x,y)$ are i^{th} and $(i+1)^{\text{th}}$ image in the sequence respectively. Typical examples of DIs obtained are shown in Figure 5.5.

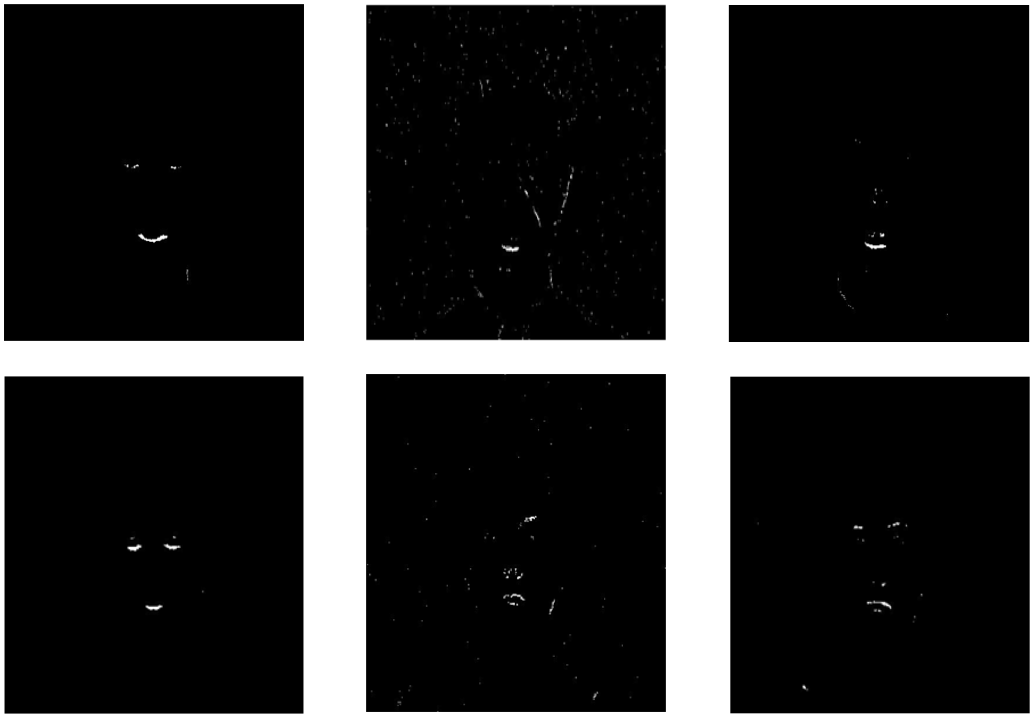


Figure 5.5 Examples of difference images

The motion in the scene between successive frames of video is usually small and therefore difficult to detect robustly. To improve the reliability of the motion information obtained, difference information gathered from a number of consecutive video frames can be used. This can be achieved by adding together the DIs for several frames, resulting in a cumulative difference image (CDI). For N consecutive difference images, the CDI is calculated by

$$\mathrm{CDI} = \sum_{i=1}^{N} DI_i(x,y) \tag{5.11}$$

Examples of CDIs obtained are shown in Figure 5.6.

Figure 5.6 Examples of cumulative difference images (*N*=38)

Increasing the number of frames for determining the CDI improves the discrimination of mouth region from the background and other parts of the face, but increases the time for the operation of mouth detection algorithm. Experiments were performed to investigate the effect of the number of frames used on the detection of mouth region, and to determine the number of frames required to extract the motion information needed for accurate mouth region detection with the aim to minimize the processing time. Number of frames, from 13 to 50 with associated delays from half a second to two seconds has been studied and the results for mouth detection shown in Figure 5.7.

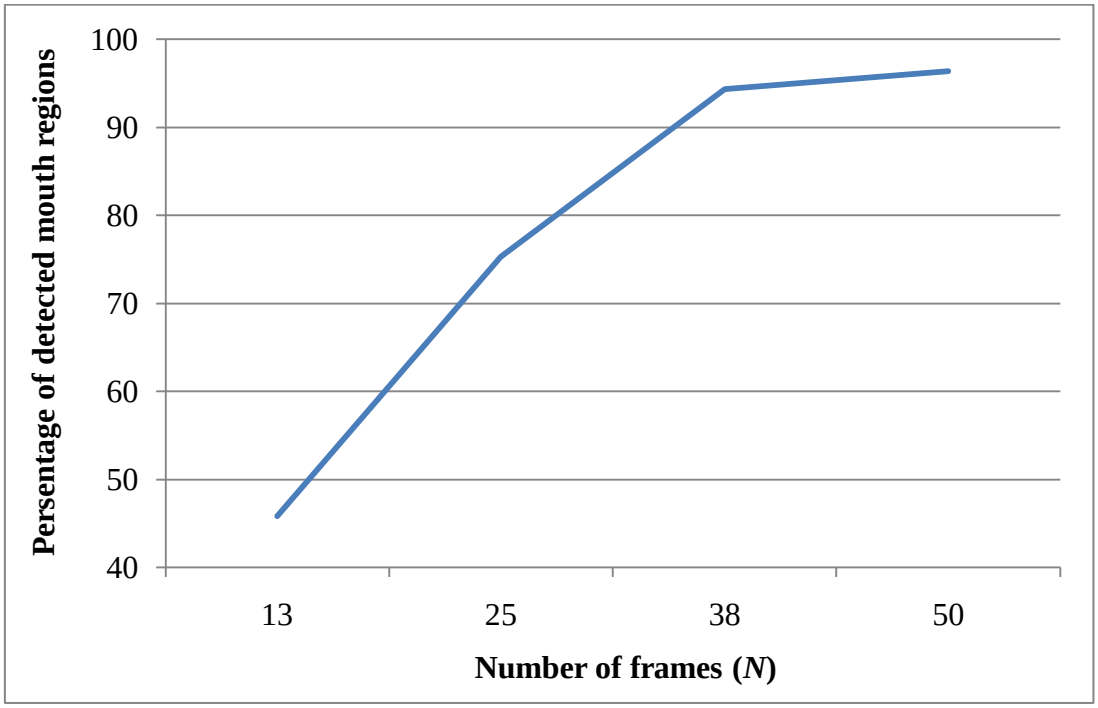
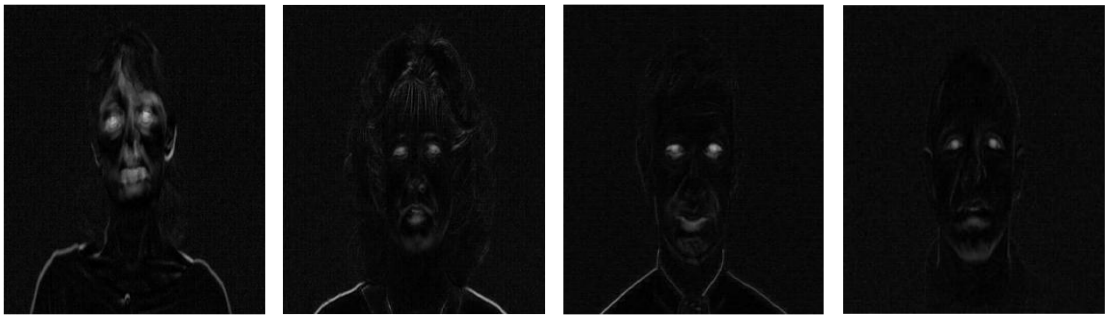
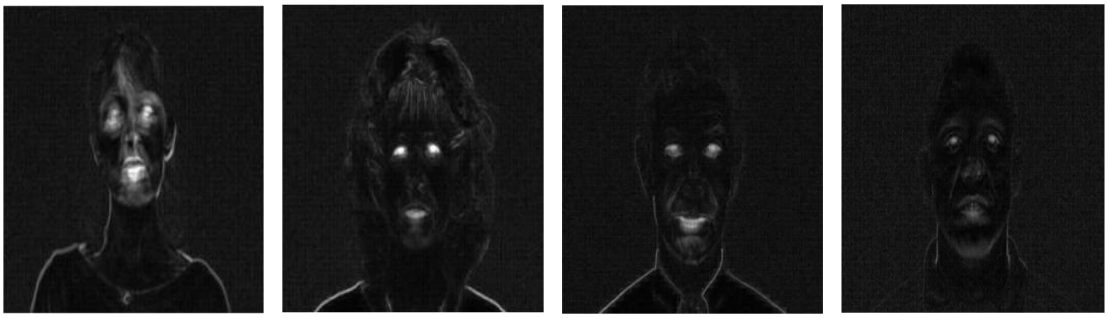


Figure 5.7 Performance of mouth detection with variation in number of frames used for the calculation of CDI

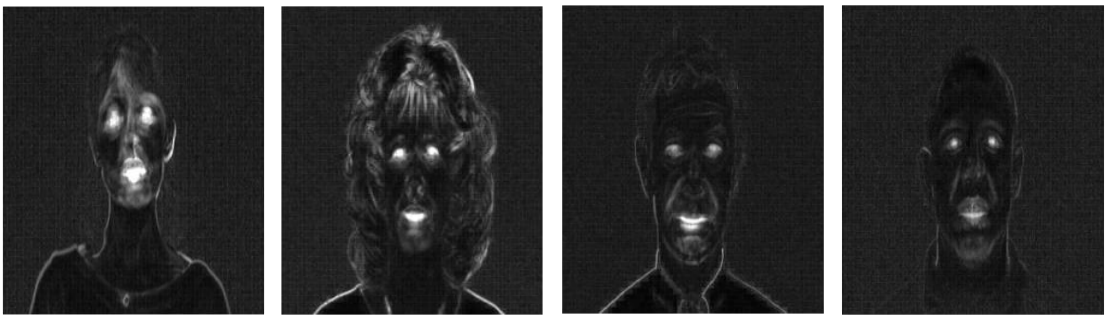
Examples of resulting CDIs for different number of frames are shown in Figure 5.8.



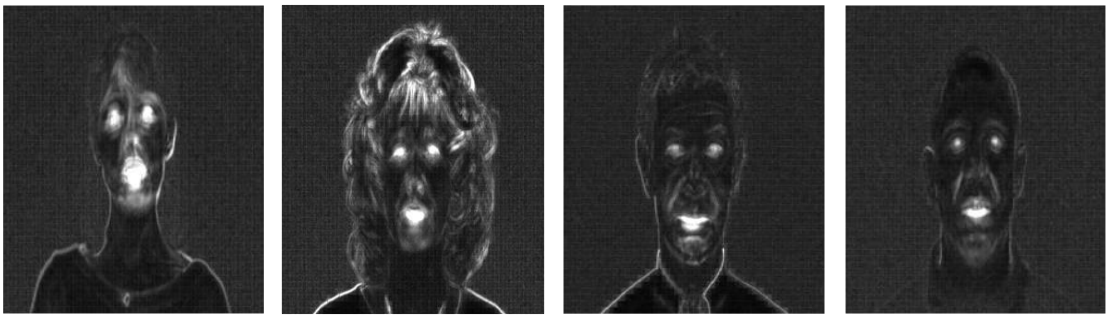
(a) 13 frames



(b) 25 frames



(c) 38 frames



(d) 50 frames

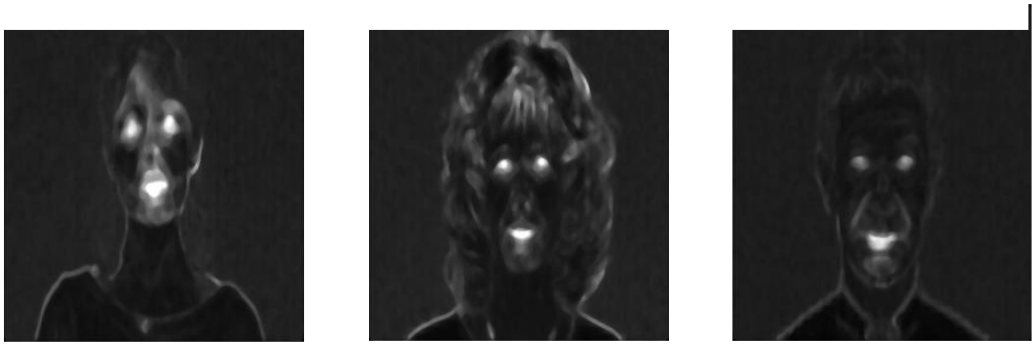
Figure 5.8 Impact of number of frames on CDI

As can be seen from the CDIs in Figure 5.8, the mouth region of the speaker becomes more apparent when frames are used in the process, simplifying its subsequent isolation with respect to other facial parts and the background. The accuracy of mouth detection improved as the number of included frames was increased, but above a value of around 38 frames no further significant gain in performance was apparent. Consequently, 38 frames were used in the subsequent experiments performed in the remainder of this section. Note that in the database used in these experiments, the movement of the speaker's face is limited and normally had little effect on the resulting CDI.

Filtering and thresholding of the CDI

The dominant regions in the CDI represent the mouth and the eyes of the speaker as these features generally exhibit the most motion, but, due to the presence both of edges separating regions of different luminance, particularly at the face boundaries and of 'salt and pepper' noise, filtering of the CDIs is needed to improve the mouth region identification. For the database used in these experiments, a 7x7 median filter was applied to smooth edges that can cause outliers in the later thresholding stage and to remove the 'salt and pepper' noise. The resulting images after the filtering are shown in Figure 5.9.

(a) CDI Before filtering



(b) CDI after filtering

Figure 5.9 filtering of the CDI

The filtering is followed by the conversion to binary of the CDI by applying a suitable threshold level. A number of threshold levels have been used and the impact of changing the threshold level on false positives and true negative mouth and eye region was studied. The results suggested that instead of fixed threshold value, an adaptive thresholding approach dictated by rate of change in foreground, total number of foreground objects and the geometric relation between the foreground objects, were found useful. In the adaptive thresholding approach reported here, the threshold level was initially set to 1 and decremented in small steps of 0.02 and the number of objects in the foreground counted at each step. This process is terminated until three objects being the vertices of a triangle with ratios between the lengths of sides ranging from 1 to 1.4, are obtained. Examples of binary images obtained from the adaptive threshold approach are shown in Figure 5.10.

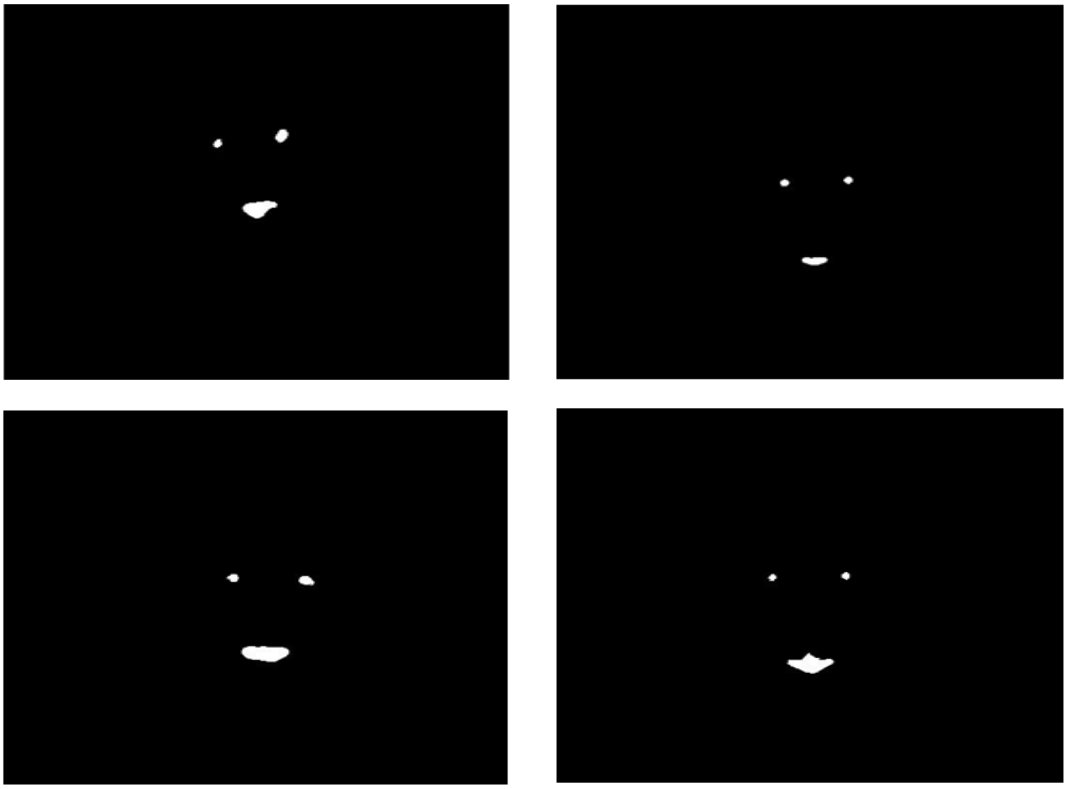


Figure 5.10 Binary images obtained from adaptive thresholding (N = 38)

The adaptive thresholding technique works for the majority of cases to suppress outliers on the face boundary; in a few cases some regions have deceived the triangle rule such as the one's shown in Figure 5.11.

(a) Mouth and ears (b) Face boundary and chin

Figure 5.11 Facial boundaries deceiving the triangle rule

The reason for this failure is that in these cases the face moved significantly in the horizontal, vertical or in both directions, causing motions in the face boundary parts, such as the chin or ears, to become more dominant. These failures could potentially be

avoided by subtracting the global motion so as to minimize the effect of face and camera movements.

Mouth localization

The aim is that the regions obtained following the thresholding will be those that include the two eyes and the mouth. The lower vertex of the triangle is the mouth region of the speaker. The centroid of the mouth region gives an estimate of the centre of the mouth while its height and width represent the opening of the mouth among the frames from which the CDI is calculated. These provide the location of the mouth and an estimate of its size required for the isolation of a ROI for AVASR purposes. For the database used in these experiments, the mouth of the speaker was identified with an accuracy of 94.33 percent.

5.5.2 ROI extraction

The ROI extraction process is depicted in Figure 5.12. The ROI are extracted from the original video frames based on the location and dimensions of the mouth region obtained from the CDI in mouth localisation stage.

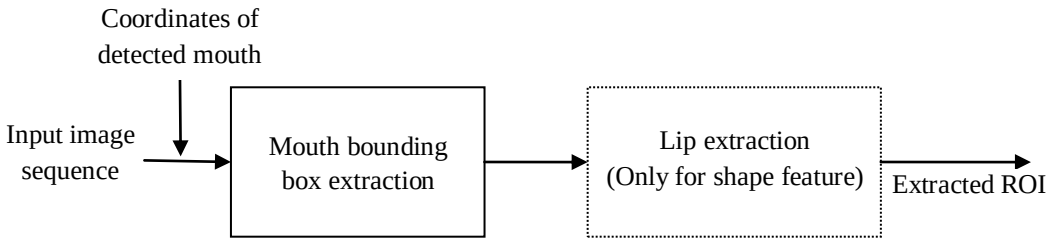
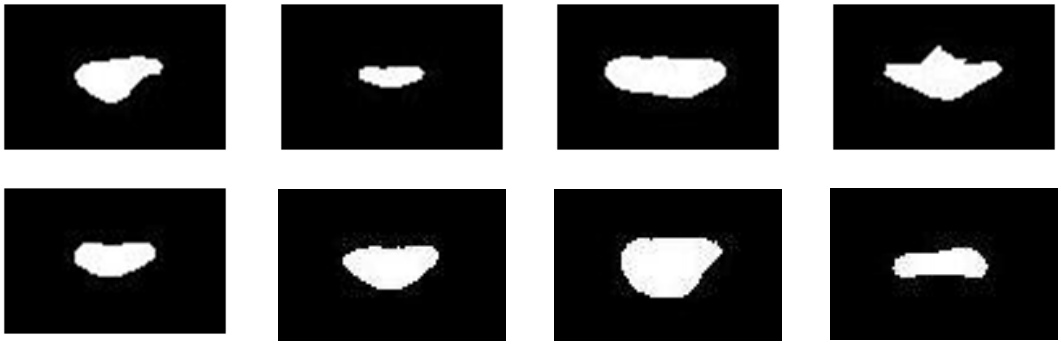


Figure 5.12 ROI extraction

As stated in section 5.1, different feature extraction approaches requires different ROIs. For appearance-based feature approaches, a bounding box of same size needs to be extracted from each frame of the videos. The examples of rectangular bounding box extracted from the CDI containing the detected mouth region and the corresponding ROIs from the first frame of the videos are shown in Figure 5.13. The rectangles in the Figure 5.13 are of size 56x88, larger than the height and width of the detected mouth regions in these experiments.



(a) Bounding rectangle of the mouth region obtained from the CDI

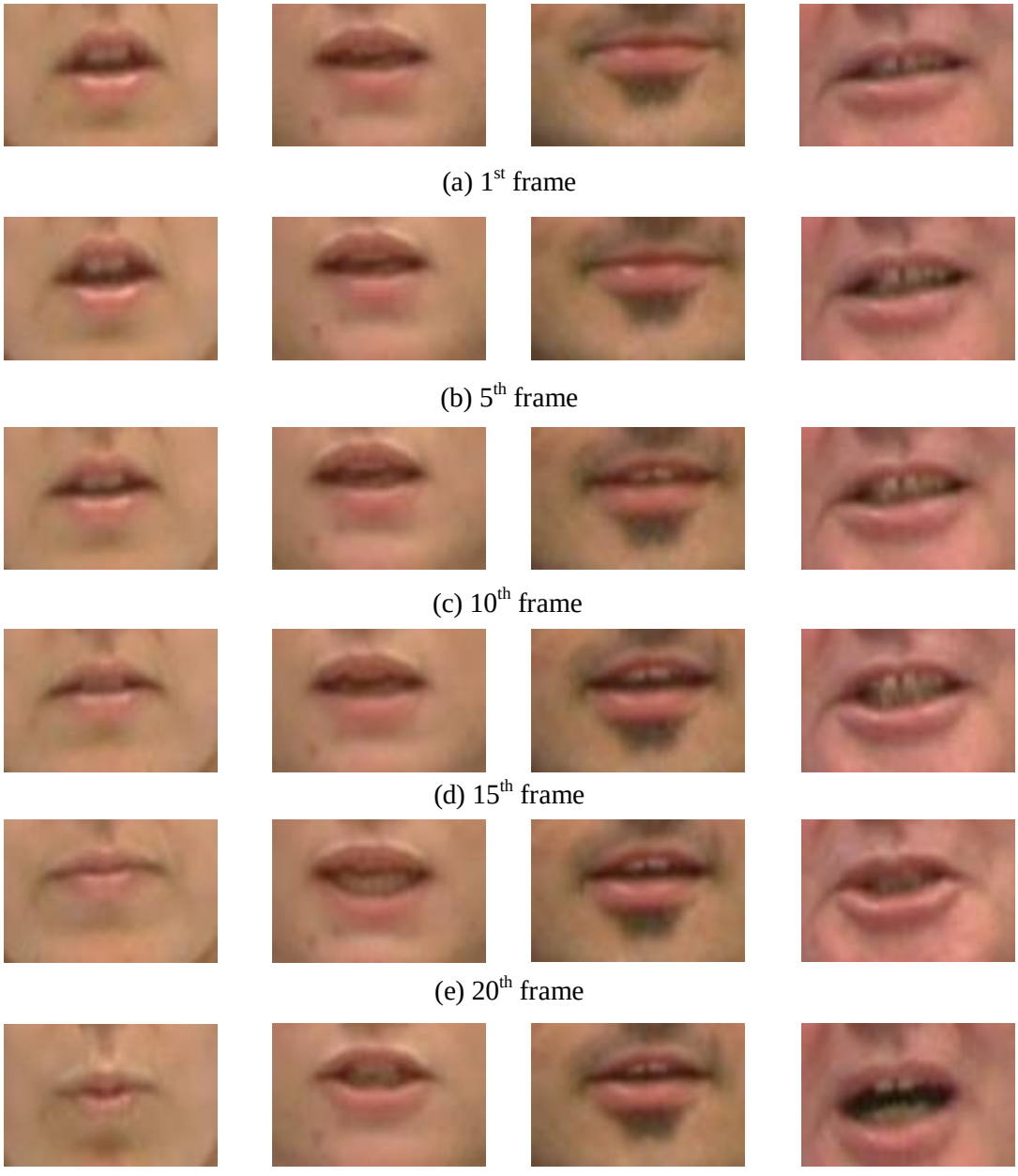
(b) Corresponding mouth region extracted from the first frame of video

Figure 5.13 Examples of the bounding rectangle obtained for the mouth region

The bounding box containing the mouth region of the speaker, such as the one above, is the required ROI for the most commonly-used appearance-based features extraction approaches. However, different sizes of bounding rectangle can be chosen depending on the dimensions of detected mouth region and the purpose of ROI extraction. For example, for mouth-only ROI, a rectangle of size equal to the detected mouth region would normally be appropriate whereas in other appearance-based feature extraction approaches additional facial parts such as jaws and chin may be included by selecting a bigger size of the bounding box.

In Figure 5.13, the bounding rectangles obtained from the original image is the ROI for the first frame of the utterance, however, for visual feature extraction for AVASR purposes, a ROI needs to be extracted from every frame of the video. The ROIs for the remaining frames of videos are obtained by extracting the bounding rectangle from each of the video frames, using the same coordinates used in extracting the ROI

for the first frame. Examples of the ROIs thus obtained from the 1st, 5th, 10th 15th, 20th and 25th frames of the utterances are shown in Figure 5.14.



(a) 1st frame

(b) 5th frame

(c) 10th frame

(d) 15th frame

(e) 20th frame

(f) 25th frame

Figure 5.14 ROI extracted from different frames of video

As for the database used in these experiments, the head movement is limited; the mouth of the speakers remains inside the bounding rectangle and the same coordinates can be used to extract the ROI from all frames of the utterance. In cases where the head of the speaker moves significantly during the speech, the location of the mouth needs to be updated. This can be achieved by implementing the mouth detection

algorithm in recursive mode, where the process is repeated after a certain number of frames, depending on the rate of head movement.

In appearance-based feature approaches, the visual features are obtained from a suitable transformation of the extracted ROI, while in shape-based methods the ROI obtained is processed further to extract the lips of the speaker from which the geometric parameters are then calculated. One such method for lip extraction is to apply an adaptive thresholding approach and a novel implementation of such an approach is described below.

Lip extraction

Suitable processing of the rectangular box around the detected mouth region is needed to extract the lips of the speaker for subsequent determination of the lip geometric parameters for shape-based AVASR. Skin and lip region separation were investigated including RGB (red, green and blue), HSV (hue, saturation and value) and YCbCr (luma and, bue and red chroma) spaces. The examples of the ROIs obtained, represented in these spaces, are shown in Figure 5.15.

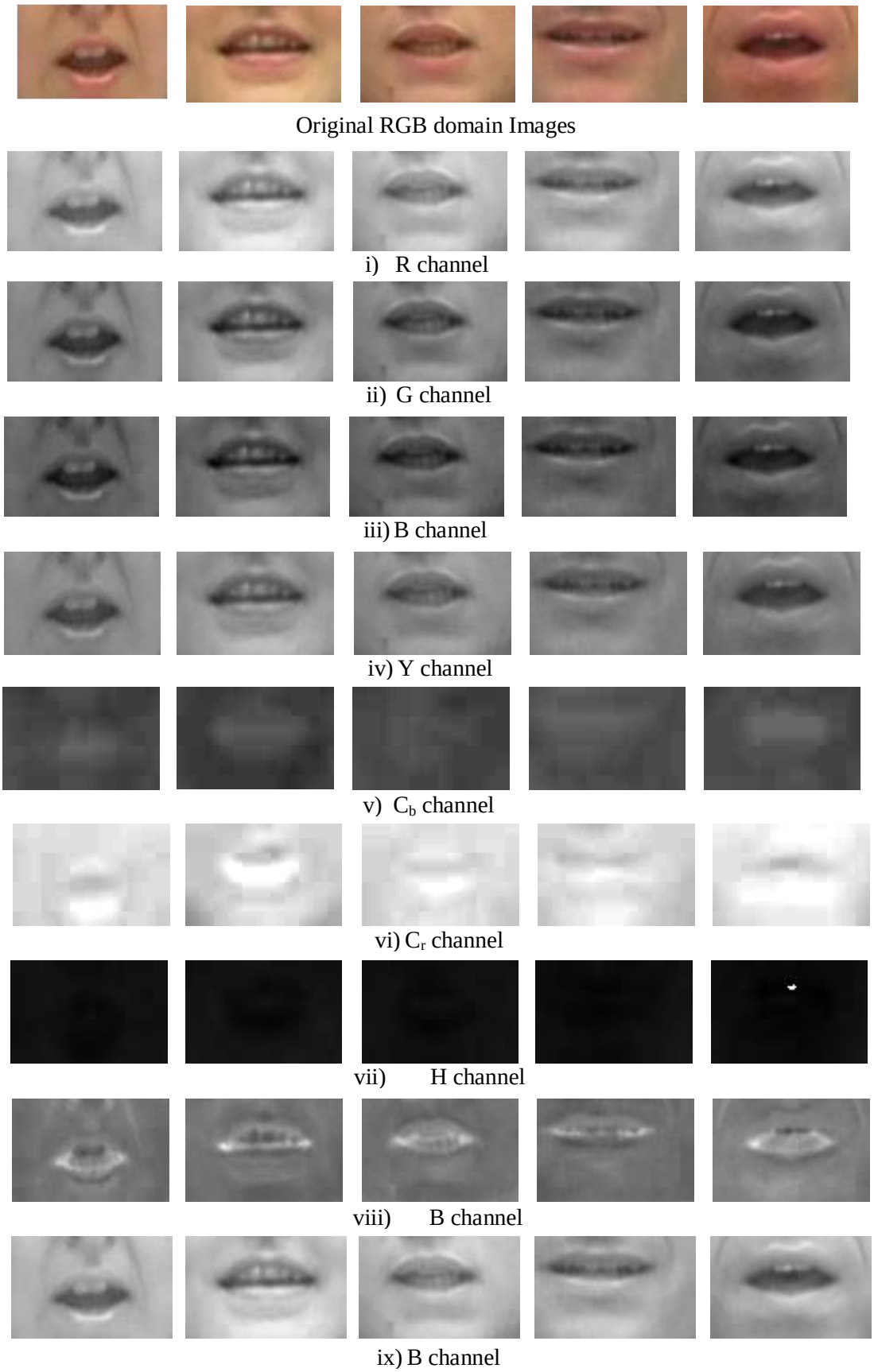
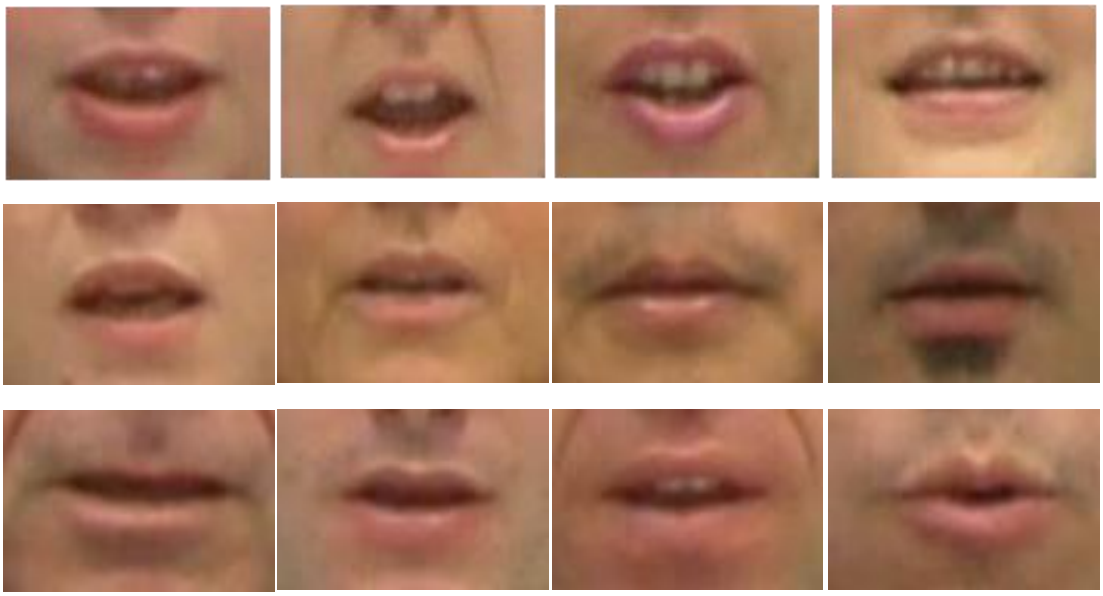
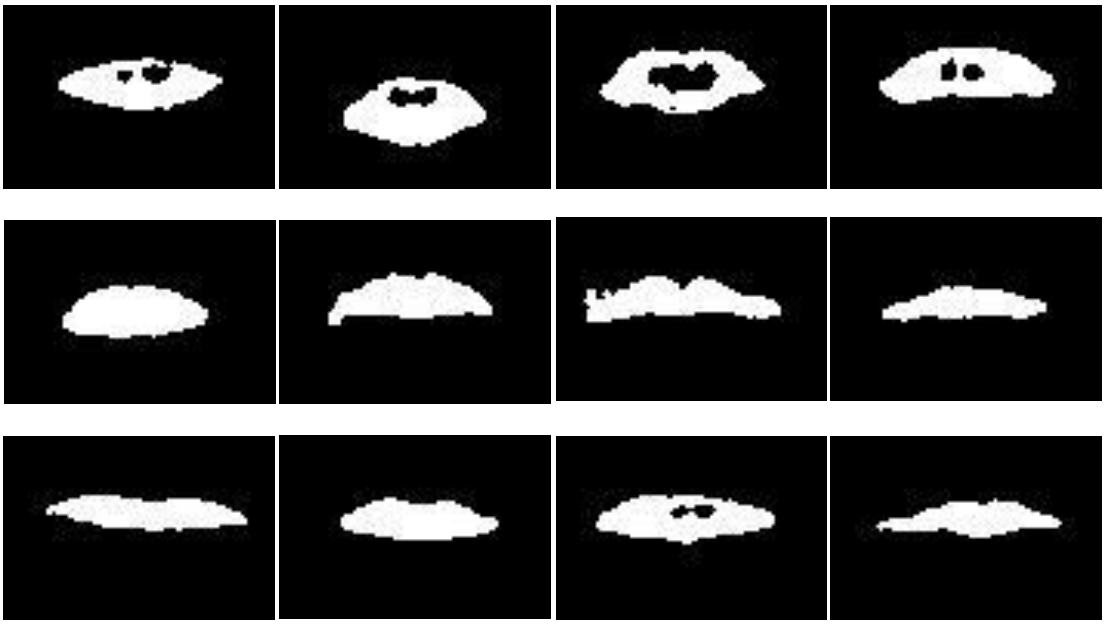


Figure 5.15 ROI representations in different colour spaces

A number of threshold levels were investigated in these spaces in order to separate the lip region from the skin in the mouth region and it was found that no single threshold value was suitable for the segmentation for all the speakers, probably due to the skin colour of the speaker. However an adaptive thresholding approach based on the rate of change of the foreground was able to give the best segmentation on the green component in RGB space for all the speakers. The adaptive thresholding approach was implemented by initially setting the threshold level to 1, decremented in steps of 0.02 and the number of pixels in the foreground counted at each step. As the threshold level decreases the rate of change in the foreground first decreases and then increases. The threshold with minimum rate of change in foreground was found to give the best separation between the skin and lip colours for all the speakers. Typical examples of the extracted lip region obtained from the adaptive thresholding approach are shown in Figure 5.16.



(a) Mouth bounding box (ROI)



(b) Extracted lips region

Figure 5.16 Lip extraction for shape-based AVASR

5.5.3 Comparison of new intensity based ROI detection method with colour based approach

In this section the new motion-based ROI detection method is compared with the baseline colour approach. ROI extraction involves the segmentation of an image into lip and non-lip regions. In practice, the image is often first segmented into face and

non-face candidates with the lip and skin segmentation being the second stage. In the baseline colour method, ROI detection aims to enhance the contrast both between the skin and background and between the skin and the lips. The performance of such ROI extraction is effected by how well separated in colour are the lip and non-lip (skin) regions. A number of colour transformation approaches have been assessed in literature, including RGB, HSV, YC_bC_r and the Pseudo-Hue spaces. In the baseline system described in [18], a mouth map was developed by using a non-linear transformation of a YC_bC_r representation of the mouth region. The mouth map is given by the following equation

$$\text{mouth map} = C_r^2 \,.(C_r^2 - \eta.\frac{C_r}{C_b})^2 \tag{5.8}$$

where C_r, and C_b are the red and blue components of chroma while η is defined as

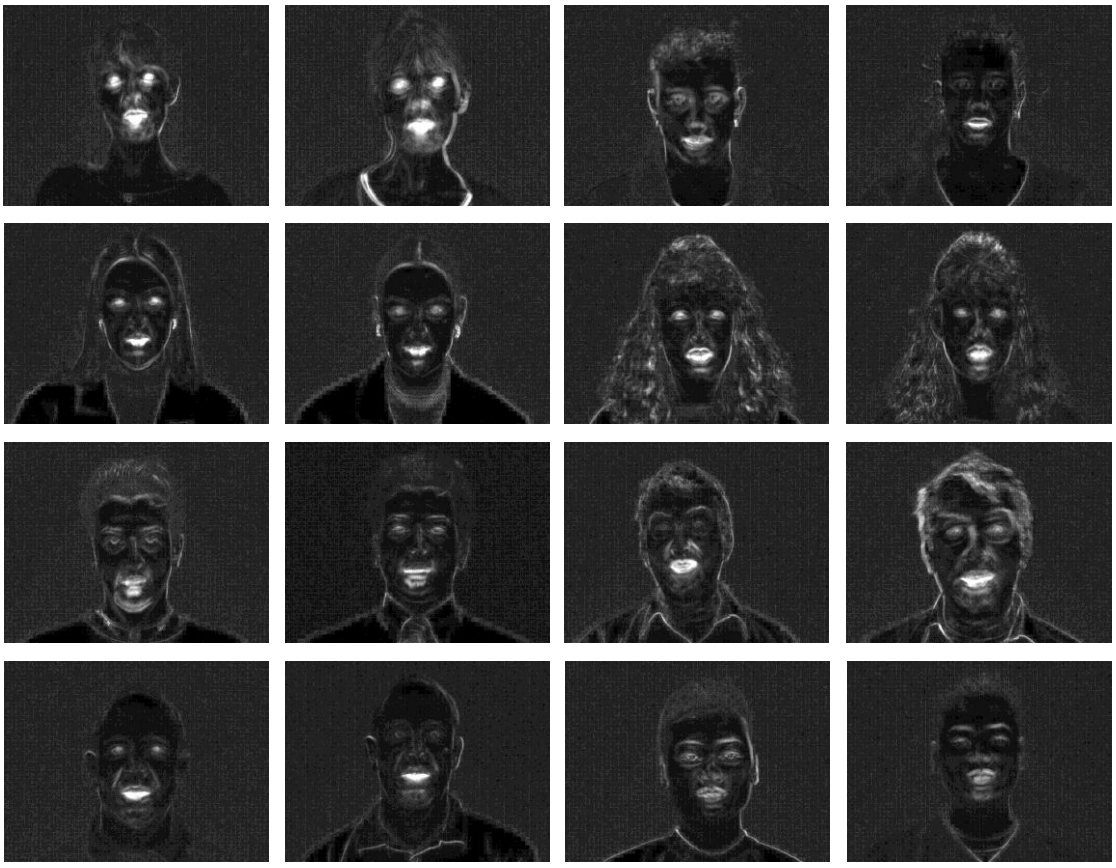
$$\eta = 0.95 \frac{\frac{1}{n}\sum_{(x,y)\in FM} C_r(x,y)^2}{\frac{1}{n}\sum_{(x,y)\in FM} \frac{C_r(x,y)}{C_b(x,y)}} \tag{5.9}$$

where *FM* is the face mask and *n* is the number of pixels in the face mask.

Examples of the resulting images obtained from the first frame of the videos of speakers based on the baseline method and the CDI obtained from the motion based approach are shown in Figure 5.17.

(a) Original images (first frame of video)

(b) Color-based lip segmentation



(c) Motion-based lip segmentation N = 38

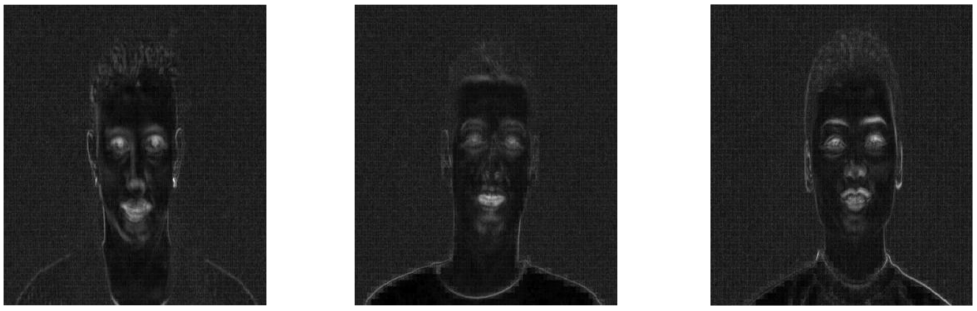
Figure 5.17 Lip segmentation for ROI extraction

While the colour based segmentation approaches have been reported to give acceptable performance for skin detection, our results show that they are not very effective for lip detection purposes. It is because, the lip and skin colours are highly correlated and also the colour based segmentation is affected by the presence of objects having lip-like colours. As in the motion based approach the background is subtracted, these methods are unaffected by these factors and therefore appears to be more robust as compared to the colour-based method.

5.6 FEATURE-BASED ROI EXTRACTION

As discussed in section 5.3, an alternative method for motion estimation in video is the feature-based approach, commonly implemented using the blocks-matching scheme. The motion based mouth detection method described in the previous section was also implemented using the feature-based approach. The block matching approaches reported in [41] were used to calculate the motion vectors, representing the displacement of the macro-blocks in consecutive frames of video. In all the block-matching based motion estimation experiment reported in this section, a block size of 4x4 pixels was used. As in the intensity-based approach described in section 5.6, the motion vectors obtained from a number of frames were accumulated in an attempt to provide better detection. Although motion vectors have widely been used to efficiently capture motion in regions containing edges, the results show that they failed to capture motion in the lip region. The reason for this failure is that lips are non rigid and go through higher deformation during speech and consequently there were few occasions where the shape of the lips persisted sufficiently between frames that the object could be tracked. Examples of CDIs obtained from the motion vector approach and corresponding CDIs from intensity-based approach are shown in Figure 5.18 (a) and (b) respectively.

CDIs obtained from motion vector method



CDIs obtained from intensity-based method

Figure 5.18 Comparison of intensity-based and motion vector approaches

A comparison of the CDIs obtained from the motion vector and intensity-based methods shows that the dominant regions in motion vector CDIs are those representing the regions having rigid outlines such as face and mouth boundaries while the mouth region with non-rigid shape is mostly missed. On the other hand the intensity based approach though giving a weak outline of the face boundary; the dominant regions are mouth and eye regions because of larger intensity changes in these regions, and thus can be easily isolated from the background and other facial parts. The motion vectors in these experiments were calculated using the Three Step Search (TSS) based block matching algorithm. To investigate the effect of search pattern, the approach was implemented using the Adoptive Rood Pattern Search (ARPS) algorithm. Examples of the CDIs obtained by the TSS and ARPS methods are shown in Figure 5.19.

(a) TSS based CDIs

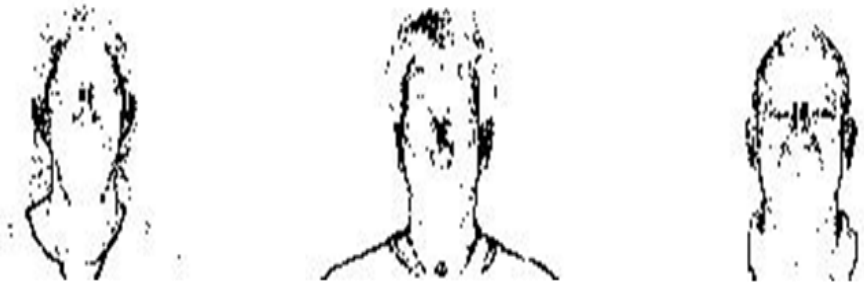
(b) ARPS based CDIs

Figure 5.19 CDIs obtained from two search techniques

The results in Figure 5.19 show that the search pattern has no significant effect on capturing the motion in the mouth region. This is because, although the ARPS has changed the search pattern, the matching criteria remains the same thus having impact on the motion estimation for the regions with sharp edges, thus giving enhanced outline of the face boundary only.

The effects of varying the number of frames for CDI calculation were also studied and the results for the use of 25 and 50 frames are shown in Figure 5.20.

(b) 25 frames



(a) 50 frames

Figure 5.20 CDIs obtained for different numbers of frames

Figure 5.20 shows that, with an increase in the number of frames in calculating the CDI, although the face outline becomes more prominent but has no impact on highlighting the motion present in the mouth region.

5.7 DISCUSSION AND CONCLUSION

This chapter has presented a novel approach for mouth detection and ROI extraction for the purpose of AVASR system design, based on motion information calculated from video sequences of speakers. The ROI extraction stage of the visual front-end provides the input for feature extraction and its accurate estimation is likely to impact on the quality of features that are subsequently obtained and thus on the overall performance of the AVASR system.

The mouth detection performance of the motion-based ROI was compared with the colour-based method and was found to give better performance over the commonly used colour based methods.

For the motion-based ROI reported in this work, both feature-based and intensity-based motion estimation techniques have been investigated. The intensity approach based on the difference in intensity values of the pixels obtained from successive frames was found to isolate the mouth region effectively. In the feature-based approach, due to the high deformation of the lips during speech and the relatively weak edges of lips, the feature-based motion estimation techniques generally do not perform well.

As the motion in the lip region is quite distinct from other parts of face region and background, the intensity-based method was able to achieve mouth region detection for use in AVASR applications. In comparison with a colour-based segmentation method often used in literature, the intensity-based ROI detection approach is able to achieve a more robust extraction of speakers' mouth region and thus potentially improve AVASR performance. The next chapter uses the mouth region identified in the ROI extraction approach described here in order to investigate potential improvement in AVASR performance that may result.

5.8 REFERENCES

[1] Potamianos, G., Neti, C., Luettin, J., and Matthews, I. (2004), "Audiovisual automatic speech recognition: An overview", Bailly, G., Bateson, V. V., and Perrier, P. (Eds.), *Issues in Visual and Audio-Visual Speech Processing*, MIT Press, 2004.

[2] Shiell, D. J., Terry, L. H., Aleksic, P., and Katsaggelos, A. K. (2009), "Audio-Visual and Visual-only Speech and Speaker Recognition - Issues about theory, system design, and implementation", Liew, A., and Wang, S. (Eds.), *Visual Speech Recognition: Lip Segmentation and Mapping*, Hershey, PA, IGI Global.

[3] Potamianos, G., and Neti, C. (2001), "Improved ROI and within frame discriminant features for lipreading", *Proceedings of International. Conference on Image Processing*, Thessaloniki, Greece, pp. 250-253.

[4] Valles, A., Gurban, M., Thiran, J. (2007), "Low-Dimensional Motion Features for Audio-Visual Speech Recognition", *Proceedings of 15th European Signal Processing Conference (EUSIPCO)*, pp. 297-301.

[5] Potamianos, G., Graf, H. P., and Cosatto, E. (1998), "An Image Transform Approach for HMM Based Automatic Lipreading", *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 173-177.

[6] Steifelhagen, R., Yang, J., and Meier, U. (1997), "Real Time Lip Tracking for Lipreading", *Proceeding of Eurospeech, '97*, pp. 142-147.

[7] Petajan, E., Bischoff, B., and Bodoff, D. (1988), "An Improved Automatic Lipreading System to Enhance Speech Recognition", *Proceedings of SIGCHI conference on Human factors in computing systems***,** Washington, D.C., United States, pp.19-25.

[8] Cootes, T. F., Taylor, C. J., and Edward, G. J. (1998), "Active Appearance Models", *Proceedings of European Conference on Computer Vision*, Freiburg, Germany, pp. 484-498.

[9] Kass, M., Witkin, A., and Terzopoulos, D. (1988), "Snakes: Active Contour Models", *International Journal of Computer Vision*, vol. 4, pp. 321-331.

[10] Saenko, K., Glass, J., and Darrell, T. (2005), "Articulatory features for robust visual speech recognition", *Proceedings of ICMI*, State College, PA, pp. 152-158.

[11] Heckmann, M., Berthommier, F., and Kroschel, K. (2001), "A hybrid ANN/HMM audio-visual speech recognition system", *Proceedings of International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 190-195.

[12] Kaynak, M. N., Zhi, Q., Cheok, A. D., Sengupta, K., Jian, Z., and Chung, K. C. (2004), "Analysis of Lip Geometric Features for Audio-Visual Speech Recognition", *IEEE Transaction on System, Man and Cybernetics-Part A: System and Humans*, vol. 34, no. 4, pp. 564-570.

[13] Jian, Z., Kaynak, M. N., Vheok, A. D., and Chung, Ko, C. (2001), "Real-Time Lip Tracking for Virtual Lip Implementation in Virtual Environments and Computer Games", *10th IEEE International conference on Fuzzy systems*, Melbourne, Australia, vol. 3, pp. 1359-1362.

[14] Yanjun, X., Limin, D., Ziqiang, H. (1998), "A novel lip localization method based on shiftable wavelets transform", *Proceedings of fourth International Conference on Signal Processing*, Beijing, China, vol. 2, pp. 1029-1032.

[15] Nilsson, M., Nordberg, J., and Claesson, I. (2007), "Face Detection Using Local SMQT Features and Split up SNoW Classifier", *ICASSP 2007*, pp. 589-592.

[16] Yang, J., and Waibel, A. (1996), "A Real-Time Face Tracker", *Proceedings of 3rd IEEE Workshop on Applications of Computer Vision*, pp.142-147.

[17] Reisfeld, D., and Yeshurun, Y. (1992), "Robust Detection of Facial Features by Generalised Symmetry", *Proceedings of 11th ICPR*, vol. 1, pp. 117-120.

[18] Hsu, R. L., Abdel-Mottaleb, M., and Jain, A. K. (2002), "Face detection in color images", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 696-706.

[19] Zhang, X., Mersereau, R. M. (2000), "Lip Feature Extraction towards an Automatic Speechreading System", *Proceedings of International Conference on Image Processing*, vol. 3, pp. 226-229.

[20] Cristinacce, D., and Cootes, T. F. (2006), "Facial Feature Detection and Tracking with Automatic Template Selection", *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, Southampton, UK, pp. 429-434.

[21] Chandramohan, D., and Silsbee, P. L. (1996), "A Multiple Deformable Template Approach for Visual Speech Recognition", *Proceedings of Fourth International Conference on Spoken Language*, vol. 1, pp. 50-53.

[22] Arsic, I., and Thiran, J. P. (2006), "Mutual Information Eigenlips for Audio-Visual Speech Recognition", *Proceedings of 14th European Signal Processing Conference (EUSIPCO)*, Lecture Notes in Computer Science.

[23] Bevilacqua, V., Filograno, G., and Mastronardi, G. (2008), "Face Detection by Means of Skin Detection" *Proceedings of the 4th international conference on Intelligent Computing: Advanced Intelligent Computing Theories and*

Applications - with Aspects of Artificial Intelligence (ICIC '08), Shanghai, China, pp. 1210-1220.

[24] Yang, M. H., Kriegman, D. J., and Ahuja, N., (2002) "Detecting Faces in Images: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34-58.

[25] Yang, G., Huang, T. S. (1994), "Human Face Detection in Complex Background", *Pattern recognition*, vol. 27, no. 1, pp. 53-63.

[26] Koutlas, A., and Fotiadis, D. I. (2008) "An Automatic Region Based Methodology for Facial Expression Recognition", *IEEE International Conference on Systems, Man and Cybernetics, (SMC 2008)*, pp. 662-666.

[27] Gizatdinova, Y., and Surakka, V. (2007), "Automatic Detection of Facial Landmarks from AU-coded Expressive Facial Images", *Proceedings of 14th IEEE International Conference on Image Analysis and Processing (ICIAP 2007)*, pp. 419-424.

[28] Bevilacqua, V., Ciccimarra, A., Leone, I., and Mastronardi, G. (2008), "Automatic Facial Feature Points Detection", *Proceedings of the 4th international conference on Intelligent Computing: Advanced Intelligent Computing Theories and Applications - with Aspects of Artificial Intelligence*, Shanghai, China, pp. 1142-1149.

[29] Amine, A., Ghouzali, S., and Rziza, M. (2006), "Face detection in still color images using skin color information", *Proceedings of International Symposium on Communications, Control, and Signal Processing (ISCCSP)*, Marrakesh, Morocco.

[30] Viola, P., Jones, M. (2004), "Robust Real-Time Face Detection", *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154.

[31] Xue, Z., Li, S. Z., and Teoh, E. K. (2001), "Facial Feature extraction and image warping using PCA based statistic model", *International Conference on Image Processing*, vol. 2, pp. 689-692.

[32] Moghaddam, B., and Pentland, A. (1995), "Probabilistic visual learning for object detection", *Proceedings of Fifth International Conference on Computer Vision*, Cambridge, MA, USA, pp. 786-793.

[33] Liu, Q., Lu, H., and Ma, S. (2004), "Improving Kernel Fisher Discriminant Analysis for Face Recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 42-49.

[34] Potamianos, G., Neti, C., Gravier, G., Garg, A., and Senior, A. (2003), "Recent advances in the automatic recognition of audiovisual speech", *Proceeding of IEEE*, vol. 91, no. 9, pp. 1306-1326.

[35] Wang, X., Hao, Y., Fu, D., Yuan, C. (2008), "ROI processing for visual features extraction in lip-reading", *Proceeding of International Conference on Neural Networks and Signal Processing*, Nanjing, China, pp. 178-181.

[36] Aggarwal, J. K. and Nandhakumar, N. (1988), "On the computation of motion from sequences of images-a review", *Proceedings of the IEEE*, vol. 76, no. 8, pp. 917-935.

[37] Fleet, D. J., and Weiss, Y. (2006), "Optical flow estimation", Paragios, N., Chen, V., and Faugeras, O. (Eds.), *Handbook of Mathematical Models in Computer Vision*, ch. 15, pp. 239-258, Springer.

[38] Migliore, D., Matteucci, M., Naccari, M., and Bonarini, A. (2006), "A revaluation of frame difference in fast and robust motion detection", *Proceedings of 4th ACM International Workshop on Video Surveillance and Sensor Networks* (*VSSN*), pp. 215-218.

[39] Sanderson, C., and Paliwal, K. K. (2004), "Identity Verification Using Speech and Face Information", *Digital Signal Processing*, vol. 14, no. 5, pp. 449-480.

[40] Black, M. J., and Yacoob, Y. (1997), "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion" *International Journal on Computer Vision*, vol. 25, no. 1, pp. 23-48.

[41] Barjatya, A. (2004), "Block Matching Algorithms for Motion Estimation", *Technical Report*, Utah State University.

CHAPTER 6

MOTION BASED VISUAL FEATURES FOR AVASR

6.1 INTRODUCTION

Videos of speakers can be considered to contain two types of speech information namely, the static speech information of the speaker's mouth region in individual frames in the form of the position of the mouth and other visible articulators and the dynamic information in the form of temporal changes in the video signal [1]. The two commonly used feature extraction approaches in AVASR research, namely the appearance-based and shape-based methods, extract visual speech features from individual frames of video streams and thus these features capture only static speech information. This chapter presents a new motion based approach to visual feature extraction, in which they are obtained from the dynamic speech information in the mouth region of interest (ROI). The work presented in this chapter is based on the motion compensation concepts found in the video compression literature and particularly in MPEG video coding. The visual features obtained are augmented by audio features (here the Mel-frequency cepstral coefficients) to form an audio-visual feature vector. The performance of the motion-based visual features is studied on both visual-only and audio-visual recognizers, both for clean speech and in the presence of a range of different types of audio noise at various signal-to-noise ratios.

The work presented in this chapter is part of the visual front-end design and it relation with the AVASR system of Figure 2.1 is depicted in Figure 6.1.

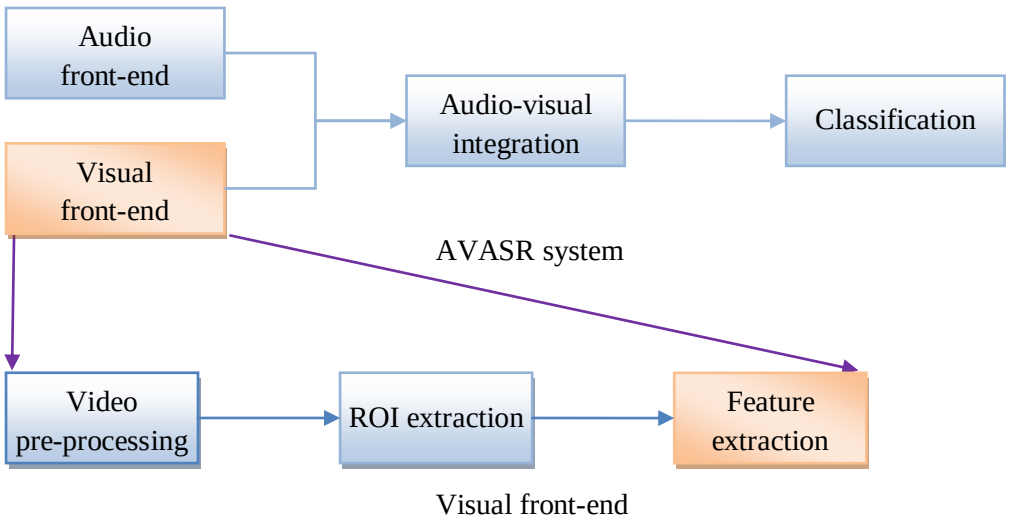


Figure 6.1 Location of the feature extraction process in the general AVASR system

The chapter is organized as follows. Section 6.2 provides the rational for the use of a motion-based approach to AVASR and reviews previous work on the use of motion information for the AVASR task. An overview of the motion compensation in MPEG video compression is provided in sections 6.3. The two popular motion estimation approaches reported in the literature are feature-based motion estimation and intensity-based methods [2]. Both of these approaches have been used in the current work for the extraction of motion-based visual speech features from the mouth region of the speaker. In particular, the feature -based method is implemented using a block matching approach while the intensity-based method is implemented using two alternative approaches; by using an optical-flow field approach and by applying a frame difference approach. Section 6.4 provides a description of the newly proposed motion-based visual feature extraction approaches and the experiments performed to investigate their performance. A description of the experiments to investigate the performance of motion-based features in the presence of noise can be found in section 6.5, while section 6.6 concludes the findings of the work presented in this chapter and discusses the important outcomes of the study.

6.2 MOTION-BASED APPROACH TO AVASR

Although the shape of the mouth and the positions of visible articulators in individual frames of video provide useful information about the utterance, they fail to capture the speech dynamic information necessary for distinguishing certain phonemes [3]. As

speech is inherently a dynamic phenomenon, the motions of the various articulators is likely to add additional information which may not be captured by features extracted from individual frames [4]. For instance, the position of the tongue appears similar when uttering /l/ or /d/, and can only be differentiated by observing the motion of the tongue during the articulation. While the mouth shape provides information for recognizing a set of visemes, the mapping from phoneme to viseme is not one-to-one and several phonemes may correspond to a single viseme. Such phonemes can often be differentiated by utilizing dynamic information obtained from the lips and other visible articulators. Consequently, a suitable representation of the motions of the articulators may potentially improve the overall recognition performance of AVASR systems.

In most appearance-based and geometry-based approaches, the speech dynamics are obtained by taking the temporal derivatives of the extracted frame features. First and second order derivatives are commonly used, while the use of higher orders derivatives have also been reported to yield improved performance [5][5]. However, concatenating the temporal derivatives with static features increases the dimensionality of the feature vector resulting in increased processing time for both training and recognition purposes. In addition, unlike appearance, motion information may well have greater tolerance to changes in lighting conditions and be less influenced by the speaker's skin colour. Furthermore, while extracting motion-based visual features, information unrelated to speech such as static background are filtered out automatically [6].

In [7], Goldschen *et al*. compared the performance of static and dynamic information when applied to speech recognition and found that dynamic features performed better. A joint use of lip texture and motion information was been reported by Cetingul *et al*. [8], for speech and speaker recognition task. In this work, the authors considered lip texture and lip motion as two separate modalities and found that the inclusion of motion features improved system performance. In Pao and Liao [9], motion vectors for specific locations on the lip were used as visual motion features for a digit recognition task. Although some work has recently been reported on the use of motion based features [10], [11], more research is needed to fully explore the potential of speech dynamic information for AVASR tasks.

Feature extraction and redundancy in data

Feature extraction for pattern recognition is the process of isolating discriminating information about the classes present in the data in a compact set of parameters. This is achieved by eliminating irrelevant data and removing the redundancies present in the input data. For the purpose of AVASR this implies that the background and speaker identity information be removed and the ROI be suitably transformed such that the visual speech information is represented in a reasonably small number of dimensions. Video data contains three types of redundancy, namely, spatial, temporal and psychovisual [10][12]. Spatial redundancy means that the pixels in a frame of video are correlated with neighbouring pixels, while temporal redundancy refers to correlation between pixels, in successive frames. Psycovisual redundancy takes advantage of the fact that that the human eye is less sensitive to fine details in the image at objects' edges [13].

In the video compression literature, it has been shown that the number of bytes needed to represent video data can be greatly reduced by reducing redundancy [14]. Intra-frame image transformation techniques are used to reduce spatial redundancies while temporal redundancies are reduced by storing and transmitting the temporal changes in the video signal rather than the separate frames of the raw video sequence. Similar concepts are used in the AVASR literature to represent high dimensional video signals in a more compact form, suitable for use in recognition systems. In appearance-based feature extraction approaches, transformations such as the DCT [15] or DWT [16] have been applied to frames from the speech video to eliminate the spatial redundancy present among the neighbouring pixels. However, to the best of author's knowledge, the use of temporal redundancy has not been fully exploited in the context of AVASR. In this work, a motion-based approach employing DCT transformations, are used for visual feature extraction for AVASR, thus exploiting both the spatial and temporal redundancies present in the video signals of speakers.

6.3 MPEG BASED VIDEO COMPRESSION

MPEG (Moving Picture Expert Group) is the most commonly-used video compression standard in current multimedia applications. MPEG compression exploits spatial, temporal and psychovisual redundancies present in video sequences

to reduce the quantity of input data generated prior to coding [12]. To eliminate the spatial redundancy, inter-frame DCT transformations are applied and DCT coefficients of small value are ignored. These coefficients are generally associated with high-frequency information to which the eye is less sensitive and so its loss has little significant effect on the video quality. For removing the temporal redundancy, inter-frame prediction approaches are adopted, in which all the frames are not transmitted, but rather only a relatively small number of reference frames, with the remaining neighbouring frames being predicted by compensating for the motion between frames. MPEG encodes the input video using a range of data reduction techniques, while the decoder accomplishes the reverse operations at the receiving end for recovery purposes [12]. MPEG compression reduces both the storage and transmission bandwidth requirements. The concepts of intra-frame and inter-frame coding for removing spatial and temporal redundancies are discussed in more detail in the following sub-sections.

6.3.1 Intra frame coding techniques (DCT transformation)

The intra frame coding process of MPEG is depicted in Figure 6.2 and typically involves video filtering, DCT transformation, DCT coefficient quantization and variable length coding (VLC). The DCT transform, when applied to a video frame, represents the image in terms of the spatial frequencies present. In the MPEG standard, the image is subdivided into blocks of size 8x8 pixels and the DCT transform applied to each block [17] resulting in an 8x8 transform matrix, with the first element representing the mean value (or DC component) and the subsequent coefficients ordered in terms of increasing frequency. The DCT transform coefficients are 11 bits in depth, greater than the input spatial domain representation of 8 bits. The high frequency coefficients are usually relatively small in value and can often be removed without discernable degradation of overall quality and can also be justified by psychovisual redundancy which implies that the eye is less sensitive to the fine details in the video contained in the high frequency coefficients. The sequence of zeros can be encoded efficiently using variable length coding.

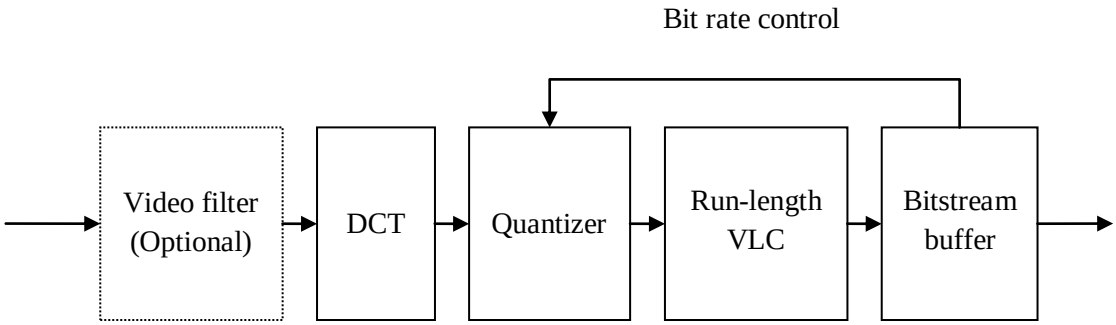


Figure 6.2 Block diagram of MPEG based intra-frame coding [12]

6.3.2 Inter frame motion compensation

Significant video compression is often achieved by reducing time-based redundancies. In MPEG, temporal redundancies are removed using the inter-frame motion compensation approach [18][18]. In MPEG video coding, not every frame is encoded independently rather the frames to be transmitted are predicted from some reference frame among the neighbouring frames. Three types of frames are transmitted: I or intra-frame coding, where each frame is encoded independently using the technique discussed in section 6.3.1; P frames or forward frames that are predicted from the immediately preceding frames; and B frames or bi-directional frames that are predicted from frames before and following the current frame. The motion compensation and encoding process of MPEG is shown in Figure 6.3.

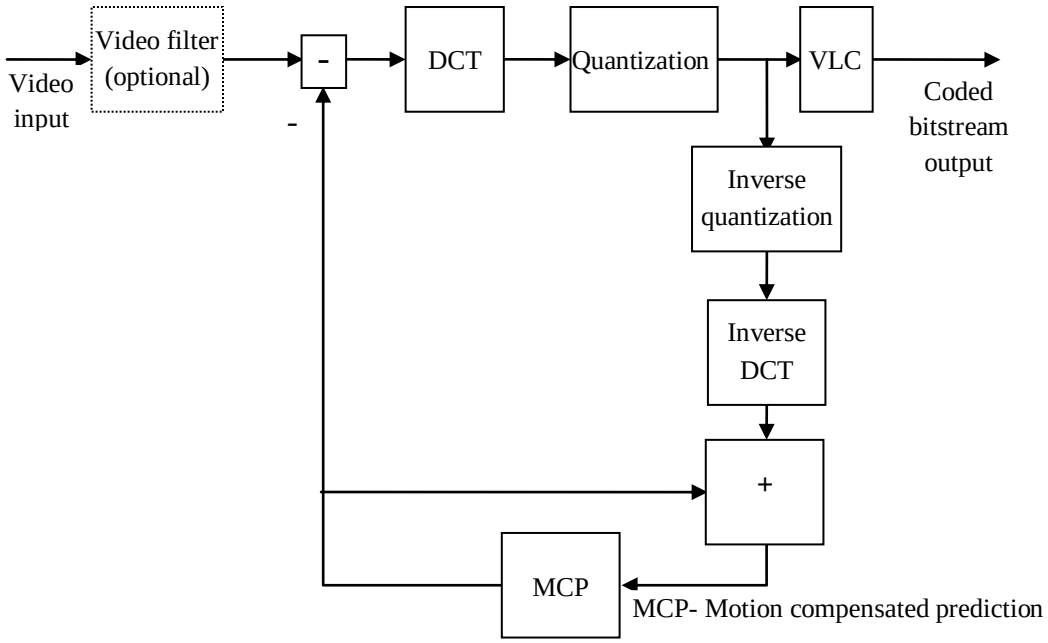


Figure 6.3 Block diagram of MPEG based inter-frame motion compensation

The motion compensated prediction (MCP) process is performed on the image reconstructed after intra-frame coding instead of the original source frame, as shown in Figure 6.3. This is because the bit rate reduction in intra-frame coding introduces distortions, due to which the frames recovered at the decoder are not identical to the original source frame. A local decoder in the transmitter is used to replicate the decoding process at receiver. The motion of objects in neighbouring frames is estimated to predict frames from the previous frame for P frames and both preceding and succeeding frames for B frames. The frame for the current time so predicted is subtracted from the actual current frame and the difference (known as the residue) is encoded and transmitted instead. The more accurate the prediction, the smaller is the residue and the fewer the number of bits required for its encoding. The motion estimation techniques discussed in chapter 5 can be used to find the motion. The choice of motion estimation technique depends on the application and is based on a trade-off between accuracy, processing time and resources. The compliment of this process is accomplished at the decoder to recover the original signal.

6.3.3 Motion estimation in MPEG based compression

Motion estimation, the process of finding the motion vector defining the transformation from the reference frame to the current frame, is the central and most time consuming part of MPEG-based video compression. The motion estimation in

MPEG is performed using the block matching approach [17]. The video frames can be divided into macro blocks that are either 8x8 or 16x16 pixels and the position of a macro block in the current frame is determined in the reference frame. The general mechanism of block matching based motion estimation is shown in Figure 6.4. The macro block in the current frame is matched with macro blocks located at a number of candidate positions, known as the search window, in the reference frame. The horizontal and vertical displacement of the macro block are recorded in the form of two dimensional motion vectors. The full search (FS) algorithm checks for a block match at all possible positions in the search window and is the most time consuming method. A number of fast block matching algorithms [19], [20], [21] have been proposed that attempt to reduce the number of search positions tested (and thus the computational time), yet without serious degradation of the accuracy of the motion estimation. A detailed discussion of the alternative block matching algorithms and their search complexities can be found in [22]. The two popular block-matching algorithms are diamond-based search [23] and hexagon-based search [24]. The variable shape search (VSS) algorithm has been used in the current work [25]. It combines the diamond and hexagon search methods and has been found to give performance superior to the individual performances of either of these methods.

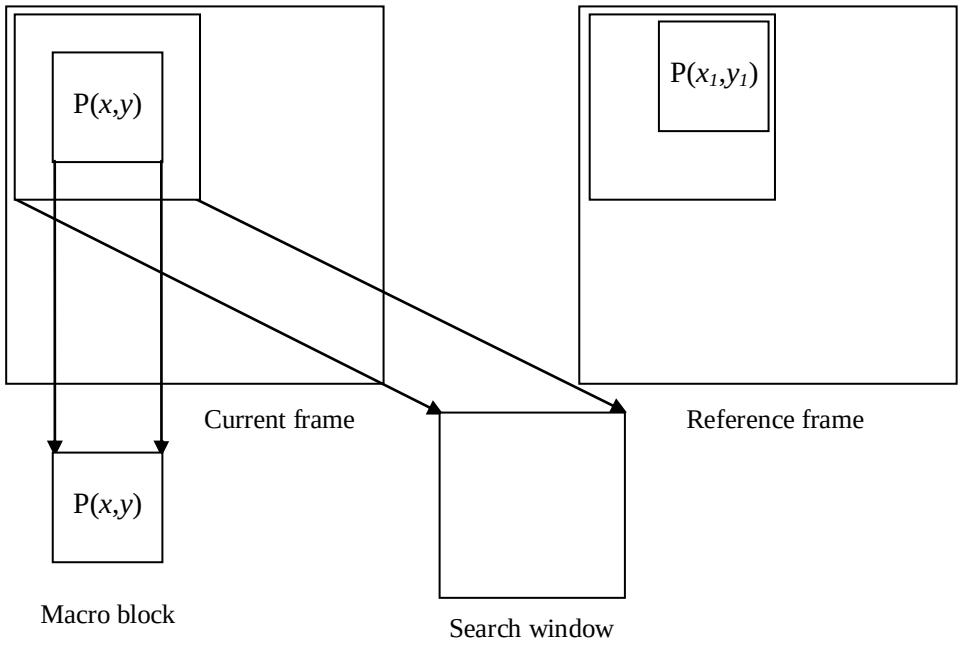


Figure 6.4 Block matching based motion estimation

6.4 MOTION-BASED VISUAL FEATURES FOR AVASR

This section describes the different approaches used in this work for the extraction of motion-based visual features for AVASR, and presents various experiments performed to determine the performance of these approaches. To allow direct comparison of the results obtained from the motion-based approach with those obtained from the appearance-based methods, the experimental setup was the same as that described in chapter 4, in which a visual ROI of dimension 72x96 around the mouth of the speaker was determined from a subset of the VidTIMIT database [26] that contains sentences from 16 speakers. The data thus obtained consists of 256 sentences with a vocabulary of 925 words and was divided into training and test sets such that the vocabulary of the test set is included in training set. Out of the 256 sentences, 216 are used as the training set and the remaining 40 as the test set. The same training and test sets were used in all of the video-only, audio-only and audio-visual experiments presented in this chapter. As in the appearance-based method, a 30 dimensional vector was used as the visual feature set, the same number of features were extracted for each of the motion-based methods used in this work and a five state hidden Markov model (HMM) was trained for each of the 46 phonemes and their context-dependent bi-phonemes and tri-phonemes, using the Cambridge University HTK toolkit [27]. The experiments were performed on the acoustic model employing dictionary search, but without the use of a language model. The results of the new motion based approaches were compared with commonly-used appearance based features on a visual-only ASR task. For the AVASR experiments, MFCC based features were extracted from the speech audio and combined with the extracted visual feature vectors using an early integration strategy. The performance of the AVASR systems was compared with the audio-only ASR for clean speech and in presence of noise at a range of signal-to-noise levels.

6.4.1 Block matching approach

The motion of an object in a sequence of video frames is commonly represented by two dimensional motion vectors. The components of a motion vector represent the horizontal and vertical displacements of object in the plane of a frame. In MPEG-based video compression, the frames are divided into a number of macro blocks and the motion of each macro blocks between successive frames of video is calculated

using the block matching approach. The vectors thus obtained, represent the motion of macro blocks in the sequence of video frames. In this work, the mouth ROI is divided into a number of macro blocks and their motion estimated in each pair of consecutive frames during the utterance. In the experiments described in this section, the variable shape search (VSS) based block-matching algorithm has been used. The horizontal and vertical components of the motion vectors obtained were used as observation vectors for visual speech, from which visual features were extracted by application of either LDA or PCA. Although macro blocks of different dimensions were used, so giving rise to a range of dimensionalities for the observation vectors, these were reduced to the same number of features in each of the experiments.

Experiments and results

The horizontal and vertical components of the motion vectors obtained from the mouth ROI of the speaker represent the horizontal and vertical movements of lips and other articulators during the utterance. As the motion of lips, teeth and tongue during speech is mainly in the vertical direction, it may be expected that the vertical components of the motion vectors would contain more information than the horizontal components. To investigate the performance of the horizontal and vertical components, the two components of motion vectors were used separately to train two separate recognizers. A macro block of size 8x8 was used, giving 108 (9x12) dimensions vector in each of the horizontal and vertical direction and this was reduced to 30 dimensions by applying either LDA or PCA. The performance of the horizontal and vertical components of motion vectors are shown in Table 6.1.

Table 6.1 Word recognition rates for horizontal and vertical components of motion vectors

	Recognition rate (percentage of words correct)	
	Horizontal Component	Vertical Component
PCA	31.18	33.60
LDA	32.26	33.60

As expected, the results showed that the features obtained from the vertical component of the motion vectors gave better recognition performance than the horizontal component. Although the speech produces both horizontal and vertical motion in the mouth region, the dominant direction of motion is vertical and that is why the vertical components of motion vector could better represent the visual speech information compared to the horizontal components. Regarding dimensionality reduction approaches, LDA performed better than to PCA because LDA attempts to separate the classes present in data while PCA preserves the data variance. Methods able to provide better separation of classes is more appropriate in speech recognition applications.

In MPEG compression, macro blocks of sizes 8x8 and 16x16 are commonly used. To assess the effect of the size of macro block, visual features from the vertical components of motion vectors for 4x4, 8x8 and 16x16 macro blocks were investigated. The dimensionality of the motion vector obtained depends on the size of macro block, since the larger the macro block, the smaller the dimensionality of motion vector. To add comparison of results, the motion vectors were reduced to 30 features using LDA or PCA. The recognition results for the different macro block sizes are shown in Figure 6.5.

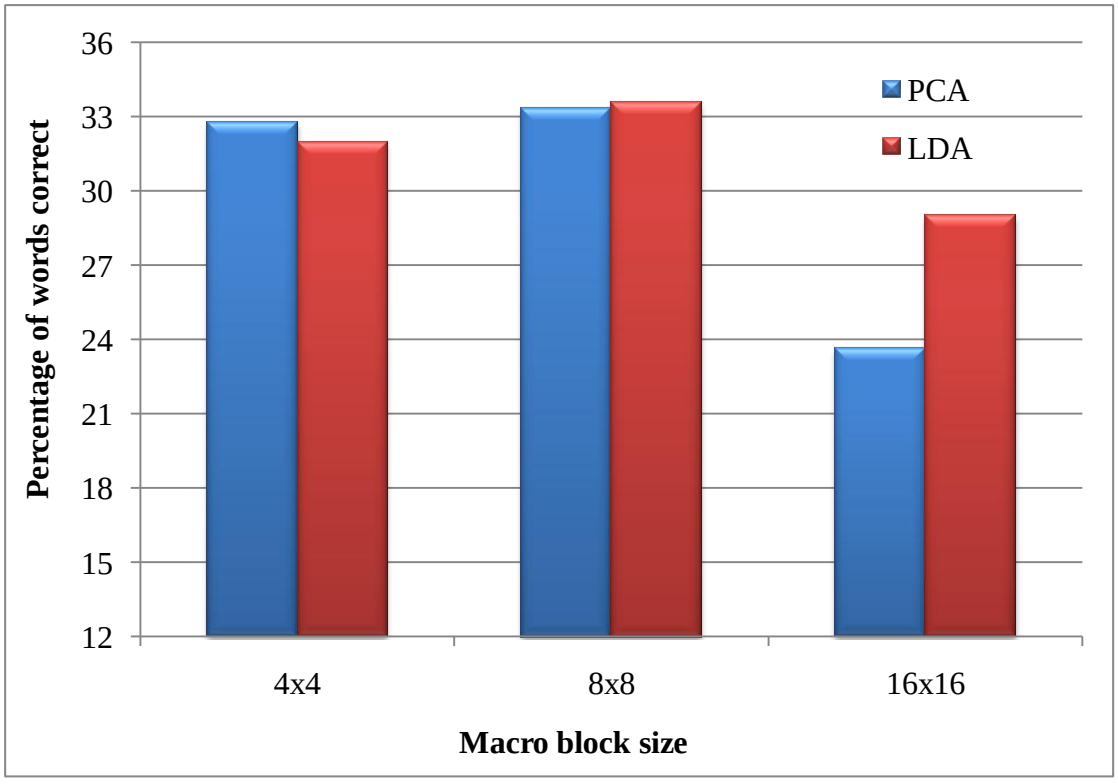
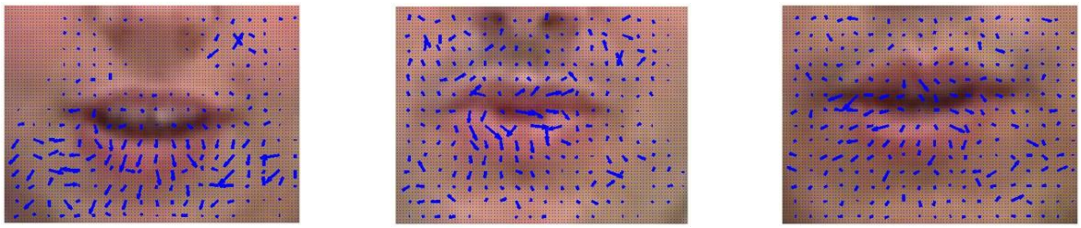


Figure 6.5 Speech recognition performance using different sizes of macro block

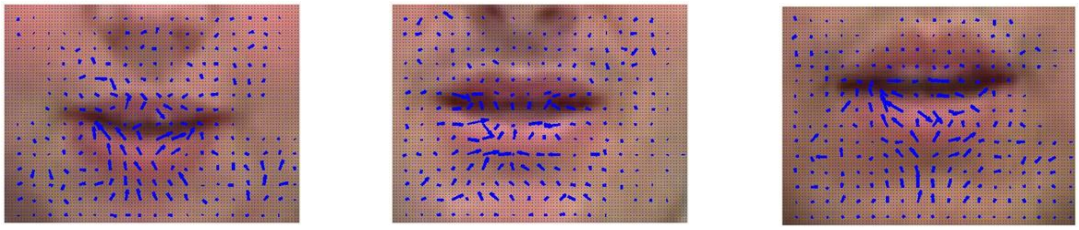
As can be seen from the results the largest macro block (16x16) gave the worst performance. As, such a macro block covers a larger area it may potentially include gross movements that originate from more than one articulator rather than being representative of the movements of a single articulator. Consequently, poor estimation of the motion of the objects in the region that are of interest for speech recognition purposes may well result. Conversely, the smallest size of macro block (4x4) may not capture the salient features of articulators, so resulting in a slightly worse tracking of motion. The best performance is given by the intermediate size macro block of 8x8 pixels. This is probably because the macro block of medium size includes sufficient features to perform suitable motion estimation while still giving an adequate representation of the motion of the individual articulators or their constituent parts.

6.4.2 Optical flow approach

A second commonly used approach for the representation of motion in video is optical flow method. Optical flow is defined as the apparent velocities of the moment of brightness patterns in an image sequence, sometimes described as the velocity of intensity that warps one frame of video to the next. The two most popularly used approaches for optical flow calculation are the Lucas-Kanade [28] and Horn-Schunck [29] methods. In this thesis, the optical flow in the mouth region of the speaker is calculated using the Lucas-Kanade method. The Lucas-Kanade method is a local optical flow calculation method in which the flow field is assumed to be constant in the neighbourhood of a pixel. The method calculates the flow field by solving the basic optical flow equations using a least squares approximation. In contrast to point-wise approaches, the method is less sensitive to noise as the optical flow is calculated based on the assumption of constant flow in the neighbourhood. At the same time, the method gives more accurate estimates of the local motion in mouth regions compared to those operating on the bases of global flow calculations. This effect is similar to that observed in block matching approaches where an intermediate size of macro-block gave better performance compared to those obtained from the smallest and largest sizes studied. Examples of the calculated optical flow field in the mouth region of the speaker are shown in Figure 6.6, where (a) and (b) show the flow fields of the instances of mouth opening and closing respectively.



(a) Mouth opening



(a) Mouth closing

Figure 6.6 Examples of the optical flow field in the mouth region of a speaker

Experiments and results

The optical flow field gives the velocity at each pixel and this is resolved into horizontal and vertical components giving two separate input matrices. To generate a more compact representation, a 2-D DCT was applied to each component of the flow field and each resulting transform matrix divided into four bands of frequencies as performed in section 4.4 in the processing of appearance based features. Each region had a dimensionality of 432 (24x18) which was reduced to 30 using LDA. The performance of the four optical flow based regions for horizontal and vertical component of flow field is shown in Figure 6.7.

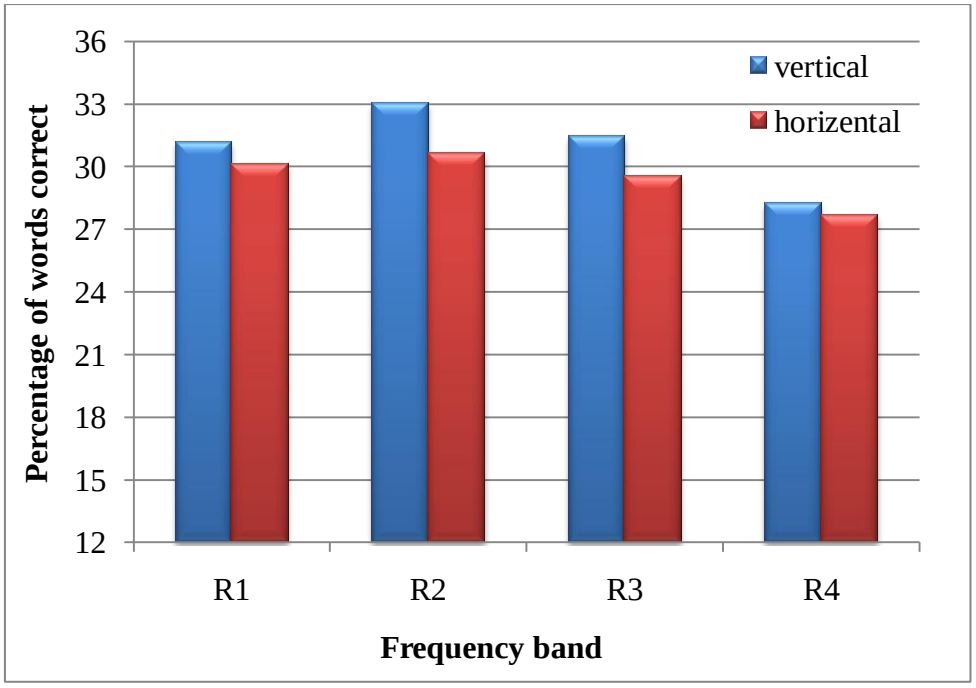


Figure 6.7 Comparison of the speech recognition performance using horizontal and vertical components of the optical flow field

As in the case of the block matching approach, the vertical components of the optical flow field provides better speech recognition than the horizontal components. As was found for the motion vectors, the optical flow in the vertical direction provided a better representation of mouth motion during the utterance of speech than did the horizontal components and this is likely to be because the majority of articulator movements (such as the lips and the tongue) occur in the vertical direction. Moreover, among the four frequency regions used, the medium frequency bands of the DCT transform of the optical flow field gave the best performance, which supports the earlier finding presented in chapter 4.

PCA was used in place of LDA for dimensionality reduction of the vertical component of flow field and the results obtained for speech recognition performance are shown in Figure 6.8.

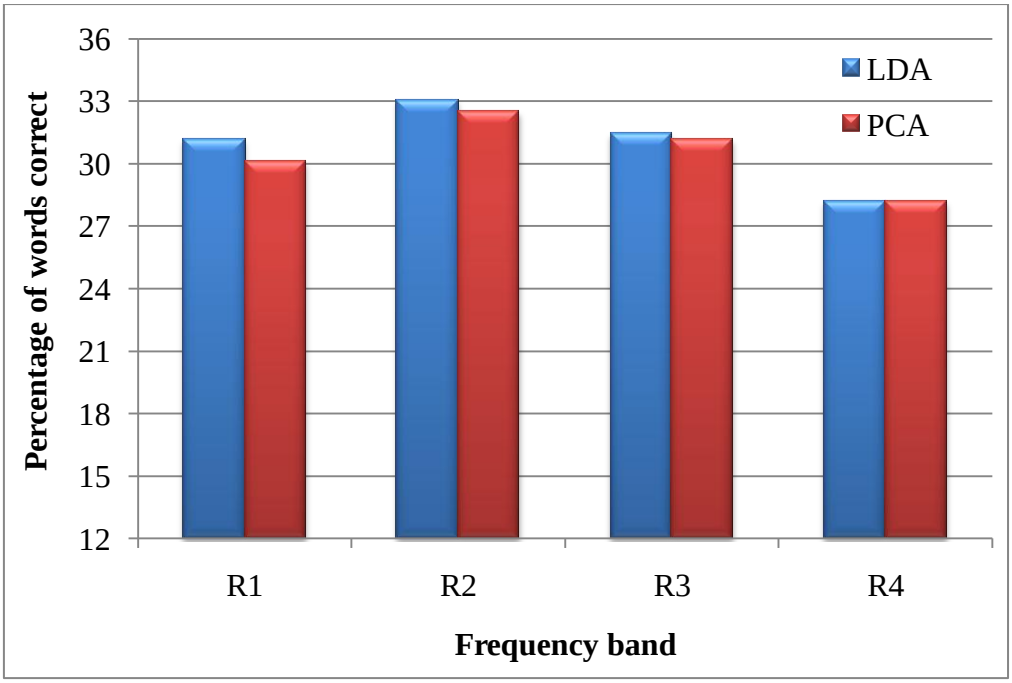


Figure 6.8 Comparison of the use of LDA and PCA for dimensionality reduction of the vertical component of the optical flow method

The results of Figure 6.8 are in agreement with the earlier results obtained using PCA and LDA approaches for dimensionality reduction. In these experiments, LDA has consistently shown better results than the PCA technique. Although both the LDA and PCA approaches are widely used in the AVASR literature to identify underlying patterns in video data, LDA appears to give better separation of speech classes and consequently better recognition performance compared to PCA.

6.4.3 Frame difference approach

The difference between the neighbouring frames of video for motion detection is of importance for compression in video coding [30]. In this thesis the author has investigated the use of the frame difference approach in the extraction of informative motion features for AVASR. The approach can be implemented by subtracting the pixel intensity vales between consecutive frames of video. The frame difference approach naturally filters out undesired data such as the background, yet the difference between the mouth regions (ROI) between consecutive images will be able to provide information about the motion related to speech. The difference image $D_t(x,y)$ is given by,

$$D_t(x,y) = \begin{cases} I_2(x,y) - I_1(x,y) & for\ t = 1 \\ I_t(x,y) - I_{t-1}(x,y) & for\ t = 2,3,\dots,T \end{cases} \quad (6.1)$$

where T is the duration of utterance and $I_t(x,y)$ and $I_{t-1}(x,y)$ are the images at times t and $t-1$. An example of consecutive frames and their corresponding difference image is shown in Figure 6.9.

(a) Previous frame (b) Current frame (c) Difference image

Figure 6.9 Illustration of the frame difference approach

Experiments and results

Similar to optical flow approach, the dimensionality of the difference image is the same as that of the input image and a 2-D DCT is applied and the DCT coefficients divided into four frequency bands and reduced to 30 dimensions using LDA and PCA. The results for the PCA and LDA based features from frame difference approach are shown in Figure 6.10.

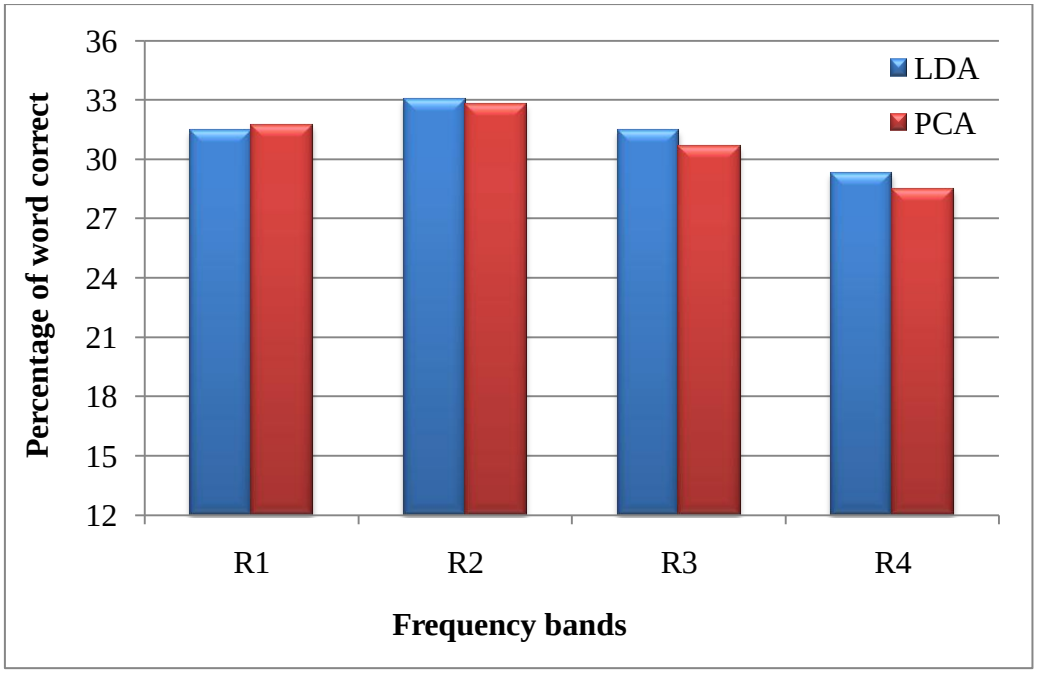


Figure 6.10 Comparison of LDA and PCA in their application to the frame difference approach

The results for the frame difference approach, shows in Figure 6.10 are similar to those obtained by using the optical flow approach, in Figure 6.8. It is probably because both frame difference and optical flow method attempts to extract similar information. However, the frame difference approach is simple to implements and is computationally less expensive as compared to the optical flow approach. These results are also in agreement with earlier results that found that the mid-frequency components obtained from the DCT transform contain information more suitable for AVASR purposes than that obtained from low-frequency components.

6.4.4 Comparison with appearance based features

This section shows a comparison of the results obtained for the motion-based visual feature with those obtained using features extracted from the appearance-based approach. The same number of features was used for all experiments and care was taken to keep all other parameters identical, including training and test set contents and the HMM topology. In the block-matching methods, a 30 dimensional feature vector was obtained by applying PCA or LDA to reduce the dimensionality of the vertical components of the motion vector obtained from 8x8 macro blocks. In each of appearance-based, optical-flow and frame difference method, 30 dimensional feature

vectors were extracted from the four frequency bands of the 2-D DCT transformations of the ROI, optical-flow field and difference image respectively, followed by the application of either PCA or LDA. The best recognition results in terms of word error rate that were obtained for each feature type are shown in Figure 6.11.

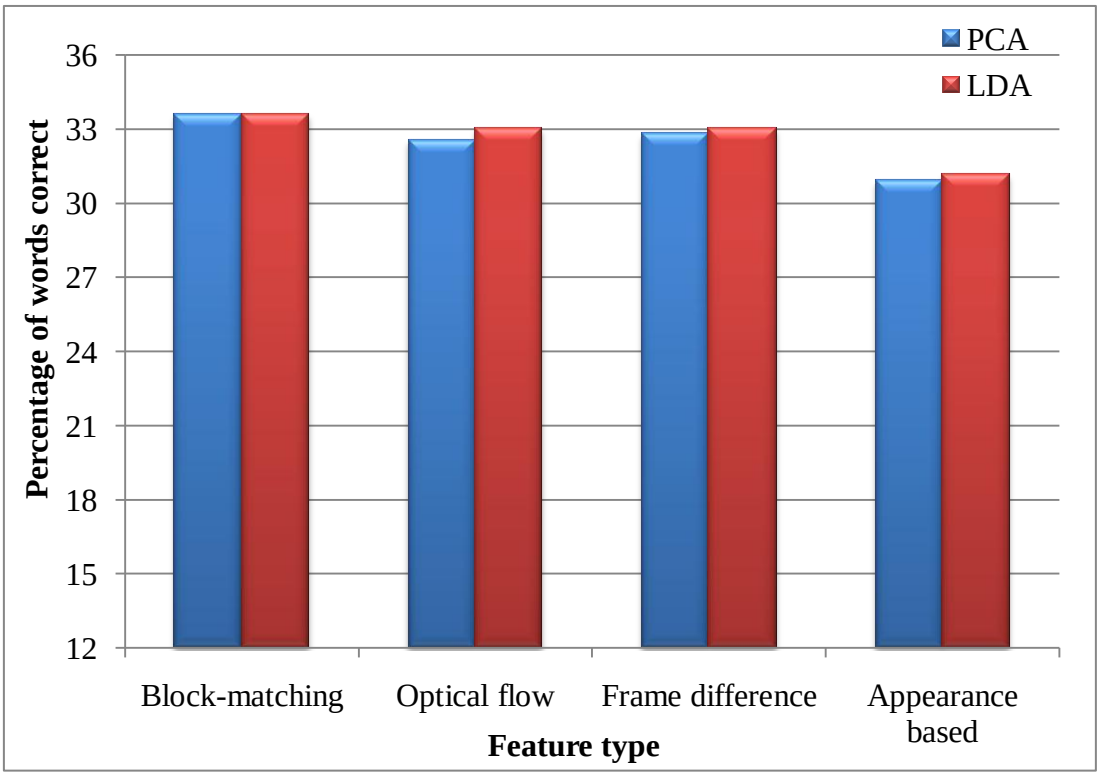


Figure 6.11 Comparison of the performances of the investigated techniques

As can be seen from the results in Figure 6.11, the new motion-based features outperform the commonly-used appearance based features. Moreover, the motion vector approach gives the best performance among all the three motion features described in this chapter. The results obtained from the optical flow approach are similar to those obtained from the frame difference method. Both methods use the intensity variations that occur during speech utterances and therefore have similar information available to them in performing their respective recognition operations.

6.5 NOISE ANALYSIS

The visual modality is inherently inferior to the audio modality in that the videos of speakers convey only a limited portion of the speech information due to the invisibility of the vocal tract and the full or partial occlusion of articulators such as the

teeth and the tongue. This is also evident from the fact that far fewer distinguishable speech units (visemes) exist in the visual modality compared to the number of phonemes found in the audio modality [31][31]. However, the use of the visual modality is important in circumstances where the audio modality is corrupted by acoustic noise. To investigate the noise robustness of the new motion-based features introduced in the current work, their speech recognition performance was studied in the presence of audio noise. AVASR systems were developed for each of the new motion-based approaches and trained using the features obtained by concatenating the MFCC-based audio features with the visual features that gave the best performance in each of the motion-based methods. For comparison purposes, an audio-only ASR system was also developed and trained using the MFCC-based features obtained from the audio training set. MFCC-based audio features from noisy audio signal were then extracted at signal-to-noise ratios (SNR) ranging from 30 dB to -10 dB and combined with each of the new motion-based visual features. Speech noise from the NOISEX database [32][32] was added to the audio speech obtained from the VidTIMIT database to provide noisy speech signals at different SNR levels. The performance of audio-only and audio-visual speech recognition systems in the presence of speech noise is shown in Figure 6.12.

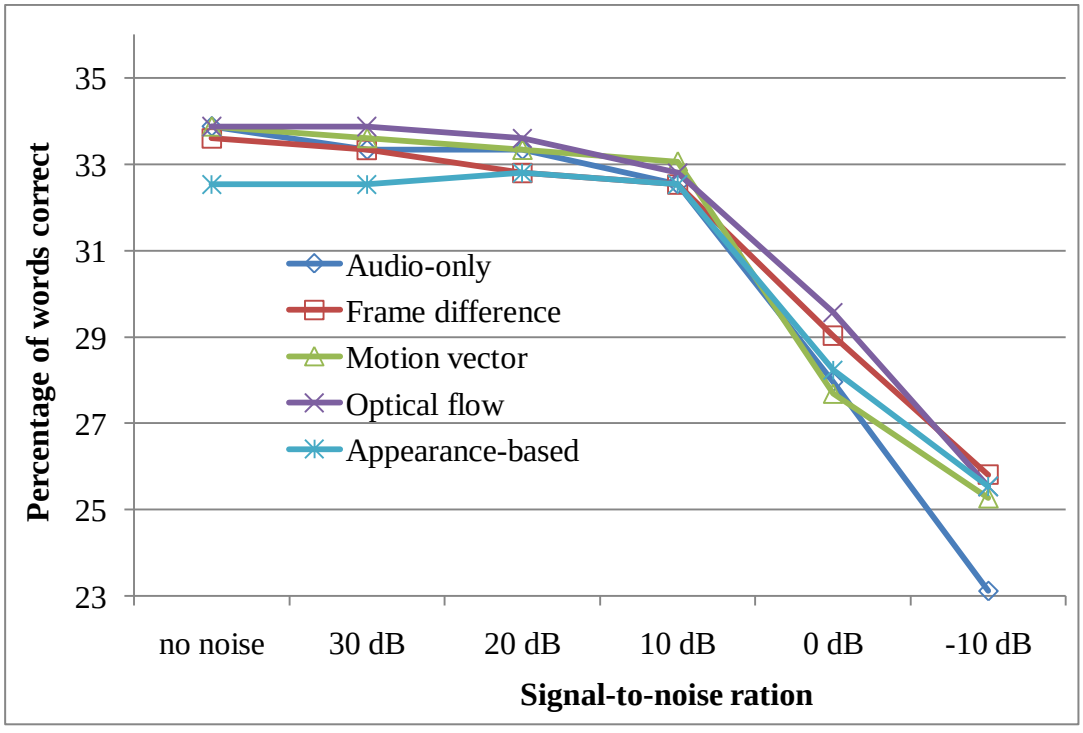


Figure 6.12 Audio-only and audio-visual ASRs performance in presence of speech noise

The performance of the new motion based features was also studied in presence of a range of different types of noise taken from the NOISEX database. The results of audio-only and AVASR in presence of each of these types of noise are given in Table 6.2.

Table 6.2 Audio-only and AVASR performance for different types of noise

Car noise						
	clean	30 dB	20 dB	10 dB	0 dB	-10 dB
Audio-only	33.87	33.87	33.6	33.06	31.99	27.96
Frame difference	33.6	33.6	33.6	33.33	33.33	29.57
Block matching	33.87	33.6	33.6	33.6	33.33	30.38
Optical flow	33.87	33.87	33.87	33.6	33.06	30.11
F16 noise						
	clean	30 dB	20 dB	10 dB	0 dB	-10 dB
Audio-only	33.87	33.87	33.6	31.18	26.08	22.85
Frame difference	33.6	33.33	32.8	32.53	29.57	25.81
Block matching	33.87	33.87	33.33	33.06	29.84	26.34
Optical flow	33.87	33.87	33.6	33.06	28.96	26.08
Factory noise						
	clean	30 dB	20 dB	10 dB	0 dB	-10 dB
Audio-only	33.87	33.87	33.6	33.33	28.13	23.92
Frame difference	33.6	33.33	33.33	33.06	30.15	26.08
Block matching	33.87	33.87	33.87	33.6	30.32	26.61
Optical flow	33.87	33.87	33.6	33.33	29.96	26.34
Operating room noise						
	clean	30 dB	20 dB	10 dB	0 dB	-10 dB
Audio-only	33.87	33.87	33.87	33.6	26.61	23.39
Frame difference	33.6	33.6	33.33	33.06	27.42	26.08

Block matching	33.87	33.87	33.6	33.6	28.15	26.34
Optical flow	33.87	33.6	33.06	32.8	27.49	26.14

The speech recognition results of the audio-only and the AVASR systems for speech with added noise shown in Figure 6.12, demonstrate that AVASR was able to give better performance than audio-only ASR. For the AVASRs methods investigated, the performance of the different motion-based features and that of appearance-based are very similar, but with the optical-flow approach giving marginally better results.

The AVASR results for different noise types in Table 6.2 show that, for all types of noise, AVASRs gives better performance than audio-only ASR, particularly when the signal-to-noise ratio is less than 0dB. Among the different motion-based approaches, the block-matching approach was generally able to achieve the best recognition results.

6.6 DISCUSSION AND CONCLUSION

In this chapter, novel approaches to visual feature extraction for AVASR have been presented, based on motion information taken from the mouth region of the speaker. Speech is a dynamic activity and so the motion of mouth is likely to contain useful information about the contents of the speech. A number of video motion estimation approaches, namely, block matching, optical flow and frame difference methods have been studied. The performance of the novel motion based features extracted from each of these methods was compared with that obtained from features based on an appearance-based approach. The results show that the motion-based features were able to perform the better when experiments were performed using the VidTIMIT database, with the block matching method giving the best performance. In the block matching approach, the block size of 8x8 yielded the best recognition results among the block sizes studied. In their application to dimensionality reduction, LDA gave a better performance than PCA for both visual-only and AVASR. This was as expected since LDA maximizes the variance between classes, which is better suited to the separation of phonemes/visemes, while PCA is designed to retain maximum variance in data rather than attempting to provide a distinction between speech classes. The horizontal and vertical components of motion vectors and optical flow fields were isolated and compared in their speech recognition performance and it was found that the vertical component of both, the motion vectors and optical flow provided better

discrimination. The performances of optical flow and frame difference methods were similar perhaps because they both attempt to represent intensity variation during speech. Lastly, audio-only and audio-visual recognizers were investigated and it was shown that addition of the video modality improved the performance of ASR in the presence of various types of noise. As a consequence, it could be concluded that motion-based features contain useful visual speech information that could be combined with audio features for improved ASR performance in presence of noise.

6.7 REFERENCES

[1] Rosenblum, L. D., and Saldana, H. M. (1998), "Time-varying information for visual speech perception", Campbell, R., Dodd, B., and Burnham, D. (Eds.), *Hearing by Eye II*, Hove, United Kingdom: Psychology Press Ltd. Publishers, pp. 61-81.

[2] Zhang D., and Lu, G. (2000), "An Edge and Color Oriented Optical Flow Estimation Using Block Matching", *proceedings of 5th International Conference on Signal Processing Proceedings, WCCC-ICSP 2000*, Beijing, China. vol. 2, pp. 1026-1032.

[3] Chitu, A. G., Rothkrantz, L. J. M., Wojdel, J. C., and Wiggers, P. (2007), "Comparison Between Different Feature Extraction Techniques for Audio-Visual Speech Recognition", *Journal on Multimodal User Interfaces*, vol. 1, no. 1, pp. 7-20, Springer.

[4] Tamura, S., Iwano, K., and Furui, S. (2004), "Multi-modal speech recognition using optical-flow analysis for lip images", *Journal of VLSI Signal Processing - Systems for Signal, Image, and Video Technology*, vol. 36, no. 2-3, pp. 117-124.

[5] Drugman, T., Gurban, M., and Thiran, J. (2007), "Relevant feature selection for audio–visual speech recognition", *proceedings of the IEEE 9th Workshop on Multimedia Signal Processing (MMSP)*, Chania, Greece, pp. 179-182.

[6] Pao, T. L., and Liao, W. Y. (2005), "A motion feature approach for audio-visual recognition", *proceedings of 48th Midwest Symposium on Circuits and Systems*, vol. 1, pp. 421-424.

[7] Goldschen, A. J., Garcia, O. N., and Petajan. E. D. (1994), "Continuous optical automatic speech recognition by lipreading", *Proceedings of 28th Asilomar Conference on Signals, Systems and Computers*, vol.1, pp. 572-577.

[8] Cetingul, H. E., Erzin, E., Yemez, Y., and Tekalp, A. M. (2006), "Multimodal speaker/speech recognition using lip motion, lip texture and audio", *Journal of Signal Processing*, vol. 86, no. 12, pp. 3549-3558.

[9] Pao, T. L., and Liao, W. Y. (2005), "A motion feature approach for audio-visual recognition", *proceedings of IEEE 48th Midwest Symposium on Circuits and Systems*, vol. 1, pp. 421-424.

[10] Carboneras, A. V., Gurban, M. and Thiran, J. P. (2007), "Low-Dimensional Motion Features for Audio-Visual Speech Recognition", *Proceedings of 15th European Signal Processing Conference*, Poznan, Poland, pp. 297-301.

[11] Pao, T. L., and Liao, W. Y. (2006), "An Audio-Visual Speech Recognition System for Testing New Audio-Visual Databases", *Proceedings of International Conference on Computer Vision Theory and Applications*, Setubal, Portugal, pp. 192-196.

[12] Wiseman, J. (1998), "An Introduction to MPEG Video Compression", http://www.john-wiseman.com/technical/MPEG_tutorial.htm

[13] Chen, H. H. (2004), "Video Compression Tutorial", Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, ROC.

[14] Tudor, P. N. (1995), "MPEG-2 video compression" *Electronics & Communication Engineering Journal*, vol. 7, no. 6, pp. 257-264.

[15] Saenko, K., and Livescu, K. (2006), "An asynchronous DBN for audio-visual speech recognition", *Proceedings of IEEE Workshop on Spoken Language Technology (SLT)*, Palm Beach, Aruba, pp. 154-157.

[16] Potamianos, G., Neti, C., Luettin J., and Matthews, I. (2004), "Audiovisual automatic speech recognition: An overview", Bailly, G., Bateson, V. V. and Perrier, P. (Eds.), *Issues in Visual and Audio-Visual Speech Processing*, MIT Press.

[17] Ebrahimi, T., and Home, C. (2000), "MPEG-4 natural video coding - An overview", *Journal of Signal Processing; Image Communication.* vol. 15, no. 4, pp. 365-385.

[18] Creusere, C. D. (2001), "Motion-compensated video compression with reduced complexity encoding for remote transmission", *Journal of Signal Processing: Image Communication,* vol. 16, no. 7, pp. 627-642.

[19] Thambidurai, P., Ezhilarasan, M., Ramachandran, D. (2007), "Efficient Motion Estimation Algorithm for Advanced Video Coding", *proceedings of the International Conference on Computational Intelligence and Multimedia Applications,* pp. 47-52.

[20] Po, L. M., and Ma, W. C. (1996), "A novel four-step search algorithm for fast block motion estimation", *IEEE Transaction on Circuits Systems and Video Technology,* vol. 6, pp. 313-317.

[21] Zhu, S., and Ma, K. K. (2000), "A new diamond search algorithm for fast block-matching motion estimation", *IEEE Transaction on Image Processing,* vol. 9, pp. 287-290.

[22] Shenolikar, P. C., Narote, S. P. (2009), "Different Approaches for Motion Estimation", *proceedings of International Conference on Control, Automation, Communication and Energy Conservation, INCACEC,* pp.1-4.

[23] Zhu, S., Ma, K. K. (2000), "A new diamond search algorithm for fast block-matching motion estimation", *IEEE Transaction on Image Processing,* vol. 9, no. 2, pp. 287-290.

[24] Zhu, C., Lin, X., Chau, L. P. (2002), "Hexagon-based search pattern for fast block motion estimation", *IEEE Transaction On Circuits & Systems for Video Technology,* vol. 12, no. 5, pp. 349-355.

[25] Hao, L., Wen-Jun, Z., and Jun, C. (2006), "A fast block-matching algorithm based on variable shape search", *Journal of Zhejiang University – Science A,* vol. 7, no. 2, pp. 194-198.

[26] Sanderson, C., Paliwal, K. K. (2003), "Noise compensation in a person verification system using face and multiple speech features", *Journal of Pattern Recognition,* vol. 36, no. 2, pp. 293-302.

[27] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland P. (1999) *The HTK Book*, United Kingdom: Entropic Ltd.

[28] Lucas, B. D., and Kanade, T. (1981), "An iterative image registration technique with an application to stereo vision", *Proceeding of Seventh International Joint Conference on Artificial Intelligence*, p. 674-679.

[29] Horn, B. K., and Schunck, B. G. (1981), "Determining optical flow", *Journal of Artificial Intelligence*, vol. 17, pp. 185-203.

[30] Mignotte, M., and Konrad, I. (2007), "Statistical Background Subtraction Using Spatial Cues", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 12, pp. 1758 -1763.

[31] Lucey, P. (2007), "Lipreading across Multiple Views", *Ph.D. Dissertation*, Queensland University of Technology, Brisbane, Australia.

[32] Varga, A. P., Steenekan, H. J. M., Tomlinson, M., and Jones, D. (1992), "The noisex-92 study on the effect of additive noise on automatic speech recognition," Technical Report, DRA Speech Research Unit.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This chapter draws conclusions from the work presented in this thesis and recommends a number of avenues for further work.

7.1 CONCLUSIONS

Audio is used as principal source of speech information in automatic speech recognition systems, but their performance degrades in presence of noise. To compensate, a number of approaches have been adopted in the ASR literature, of which the use of the visually modality is probably the most suitable candidate being supported by both human speech perception studies and the work reported on AVASR systems.

The main emphasis of this thesis has been to improve the visual front-end of AVASR systems, extracting useful speech information from videos of speakers in order to supplement audio information and resulting in a more robust ASR solution. The work reported in this thesis contains research on two important parts of the visual front-end for use in ASR, namely visual ROI extraction and visual feature extraction. Although AVASR has been an active field of research for the last two decades, the techniques applied mainly extract the speech information only from individual frames of video. Although speech is a dynamic activity and the use of temporal information in audio-only speech recognition is well established, previously explicit use of dynamic information from the video modality has not been exploited. In addition, in this work, a novel approach was taken for the extraction of appearance-based visual features from the DCT and DWT transformations of the speakers' mouth ROI, but with emphasis on the discriminative characteristics of the coefficients in contrast to the traditional data preservation viewpoint. The new visual features extracted by this approach have been tested for a continuous speech recognition task for both visual-only and audio-visual ASR. The popular HMM was used for classification and the HMM based HTK toolkit was used for both training and recognition. The results of the work carried out in this research are reported in three chapters.

In Chapter 4, the new frequency band based approach was used for visual feature extraction from DCT and DWT representations of the mouth ROI. LDA and PCA were used to reduce the dimensionality of the final feature vector for use in the speech recognition system. The features obtained from the frequency bands that gave the best performance on visual-only ASR were used for AVASR in presence of noise. The noise performance of features representing the mid-frequency bands was found to be superior to both audio-only and audio-visual ASRs that used visual features from only the low frequency bands of DCT and DWT transforms. The improvement is due to these bands containing not only useful information about the visual speech but also because they are more robust to variations in illumination.

In chapter 5, a motion-based approach was applied to the detection of the mouth region of the speaker and the extraction of a visual ROI for feature extraction. The motion-based ROI extraction approach offers a reliable solution for the robust detection of the speaker's mouth region in speech videos and the adaptive thresholding method developed in this work was able to successfully remove the outliers caused by cluttered background and sharp edges. The method was shown to outperform the commonly-used colour-based methods. The motion-based approach can be used to extract an appropriate ROI for both appearance-based and shape-based feature extraction approaches. In addition, the lip detection method proposed in this work can also be used to provide reliable initial estimates for the model based AVASR approaches automatically.

In chapter 6, the motion-based approach was used for visual feature extraction in the AVASR task. Three different representations of motion in the mouth region of the speaker were proposed and their results were compared with existing appearance-based approaches. The motion-based approaches were found to give better performance than the appearance-based methods. In particular the motion-vector approach commonly used for motion estimation in MPEG compression was found to give the best motion representation for AVASR purposes.

The specific findings of this research are summarized in the following subsections.

7.1.1 Frequency bands based visual features

The mid-frequency bands in the DCT and DWT transform domains capture important visual speech information. It was found that visual features extracted from such frequency bands resulted in an AVASR system that was able to perform better than one that used only features extracted from low-frequency bands, as commonly used in appearance-based approaches.

The further sub-division of the mid-frequency range into more than four bands was not found to have a significant effect on the AVASR performance, probably due to the mutual correlation of visual speech information contained in the frequency components in the frequency bands.

DCT-based features were found to give better ASR performance than DWT features, while LDA was shown to be a more suitable dimensionality reduction tool than PCA. The DWT transformation may be more appropriate for analysing consonant phonemes, the DCT was found to be more suitable for continuous speech recognition including both the vowels and consonants phonemes. Similarly, PCA, although highly suited for compression applications, was not able to separate the speech classes as well as LDA.

The performance of the AVASR based on the mid-frequency coefficients from both DCT and DWT transform perform better than audio-only ASR in presence of noise. Although the audio features captured more speech information in the absence of noise compared to their video counterparts, they were greatly affected by audible noise and AVASR performed better under these conditions due to the availability of the video modality that remained unaffected.

The use of appropriate stream weights for the two modalities can make better use of the strengths and weaknesses of the two modalities in cases of clean speech and in presence of noise and hence give performance superior to both audio-only ASRs and AVASRs.

7.1.2 Motion based ROI

The motion in the sequence of video frames has shown to be a reliable source of information for the detection of the mouth region of the speaker and a motion-based

approach for the automatic extraction of a region of interest for AVASR purpose has shown to give better performance than the commonly used color-based lip segmentation method.

The approach based on changes in intensity gives better performance compared to feature based approach, probably due to the non-rigid nature of mouth and lips, due to which its features are distorted during speech and can't be captured robustly by the block-matching approaches.

The adaptive thresholding approach used in this work was able to suppress outliers, both for the detection of the mouth region and for lip extraction. For mouth detection, the approach was able to remove false candidates near edges separating regions of contrasting intensities, while for the lip segmentation required in shape-based AVASR, the approach was able to identify the separate lip and skin regions.

7.1.3 Motion-based visual features

The motion information from speech videos contains important information for speech recognition purpose and the motion based features were shown to outperform the static visual features that are usually captured from individual frames of video.

The motion information was represented in three ways, the difference in luminance between successive frames, motion vectors calculated by block-matching method and optical flow. The motion-vector based features showed the best recognition performance among the three representations.

While the performance of audio-only ASR is affected by the presence of audible noise, motion-based AVASR remains relatively robust under such conditions.

In the speech recognition experiments, the AVASR based on motion-based visual features produced better results than those based on appearance based AVASR in the presence of noise.

Although the use of the visual modality in ASRs has been an active research area for the last 25 years, the techniques used are mainly taken from video analysis and data compression research and the video modality has not yet been fully explored from a discriminative point of view. Moreover, the research carried out is mainly restricted to the extraction of spatial information from individual frames, with less attention paid

to the temporal information available in video. In this work presented in this thesis, both of these limitations found in previous research have been addressed by investigating both the discriminative information in individual frames of video and the motion information obtained from temporal changes in the video sequences.

The approaches used in this research have concentrated on delivering improvements to the visual front-end processing and the resulting automated front-end design would be potentially beneficial in the deployment of AVASR system in commercial applications. This work has demonstrated the usefulness of the motion information in the automatic detection of the visual ROI and has implemented a method for feature extraction where there is limited head movement. It is likely that the techniques developed could be extended to more unconstrained environments using techniques that are able to compensate, at least in part, for the speaker and camera motion.

7.2 FUTURE WORK

One of the main results of the current work was that the adoption of visual features extracted from the mid-frequency bands was important in improving the performance of an AVASR system. Consequently, it is clearly important that specific frequency bands should be considered when extracting information from the visual modality and it is probably reasonable to assume that individual phonemes may be best captured from specific frequency bands. Future work could investigate which frequency bands are the most appropriate for extracting features for particular phonemes.

The motion-based approach for AVASR has the potential to further enhance the capability of the video modality to compensate for the degradation of audio-only ASR in the presence of noise by capturing useful information from videos of speech.

In this work, for the purpose of visual feature extraction, the motion in the ROI was estimated using block-matching, optical-flow and frame-difference approaches, but the exploration of potential motion representation and motion estimation approaches is far from complete. The block-matching approach that gives the best recognition performance was implemented by the variable shape search method, but other search methods could be used. The motion information for feature extraction can also be explored from the perspective of the temporal frequencies of motion patterns in the mouth ROI.

The success of the adaptive threshold method for both motion-based ROI detection and lip region extraction suggests that these methods could be enhanced by incorporating additional constraints. For example, allowing the intensity-based ROI extraction approach to take into account the direction in which changes in intensities are manifest could be beneficial in distinguishing between mouth motion that occurs principally in the vertical direction and the predominantly horizontal motion found at the face boundary. Further, it intuitively would appear to be the case that the motion-based ROI detection and feature extraction approaches are independent both of lighting conditions and of the speaker's skin colour; the latter assuming that there is an adequate colour separation with respect to the background. As the VidTIMIT database used in these experiments is not designed for such investigations, such extensions to the existing motion-based approach will need to be investigated with reference to alternative or specifically established databases. In the database used in this thesis, the video sequences are recorded with limited head movement, but more unconstrained head and camera motions could be permitted should additional compensation for camera and speakers movement be implemented by algorithms able to determine the relative motions of the face and mouth region. The motion-based approach can also provide a computationally efficient way to track the changes in mouth position during speech by its implementation in a recursive mode.

APPENDIX I

Phonemes and their corresponding mouth shapes.

/a/

/i/

/ae/

/ay/

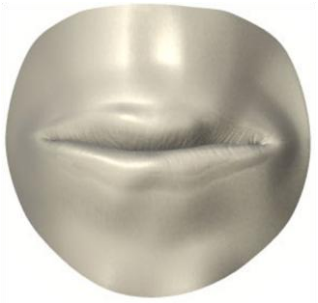
/ah/

/ih/

/f/

/v/

/sil/



/p/

/b/

/m/

/c/

/d/

/j/

/k/

/n/

/r/

/x/

/y/

/iy/

/jh/

/ck/

/l/

/w/

/q/

/aw/

/g/

/h/

/s/

/z/

/ch/

/sh/

/hh/

/zh/

/ng/

/aa/

/uw/

/oy/

/o/

/ao/

/ow/

/uh/

/t/

/it/

/ey/

/eh/

/er/

/u/

/dh/

/th/

/e/

APPENDIX II

The solution of the three problems of HMM namely the evaluation, decoding and the training problems are described in the following.

1. Solution of problem 1 (Evaluation)

The evaluation problem is to find the probability of observation *O* given model ∆. To find out the P(*O*|∆), the probability of getting observation *O* needs to be summed over all possible state sequences *S*.

For a fixed state sequence $S = S_1\ S_2\ \ldots\ S_T$, the probability of getting observation *O* is given by product of P (*O*/ *S*, ∆) (the probability of observation *O*, given state sequence *S* and model ∆) and P (*S* | ∆) (the probability of state sequence S itself). Thus

$$P(O|\ \Delta) = P(O|S,\Delta) * P(S|\ \Delta) \tag{1}$$

where

$$P(O|S,\Delta) = b_{s1}P(O_1) * b_{s2}P(O_2) * b_{s3}P(O_3), \ldots, b_{sT}P(O_T) \tag{2}$$

and

$$P(S|\ \Delta) = \pi_{s1} * a_{s1s2} * a_{s2s3}, \ldots, a_{s(T-1)s(T)} \tag{3}$$

To find the total probability of observation *O* for all state sequences, equation (1) is summed over all possible state sequences, that is

$$P(O|\ \Delta) = \sum_{all\ S} P(O|S,\Delta) * P(S|\ \Delta) \tag{4}$$

This mechanism is pictorially shown in Figure 1, where one such state sequence S is made red.

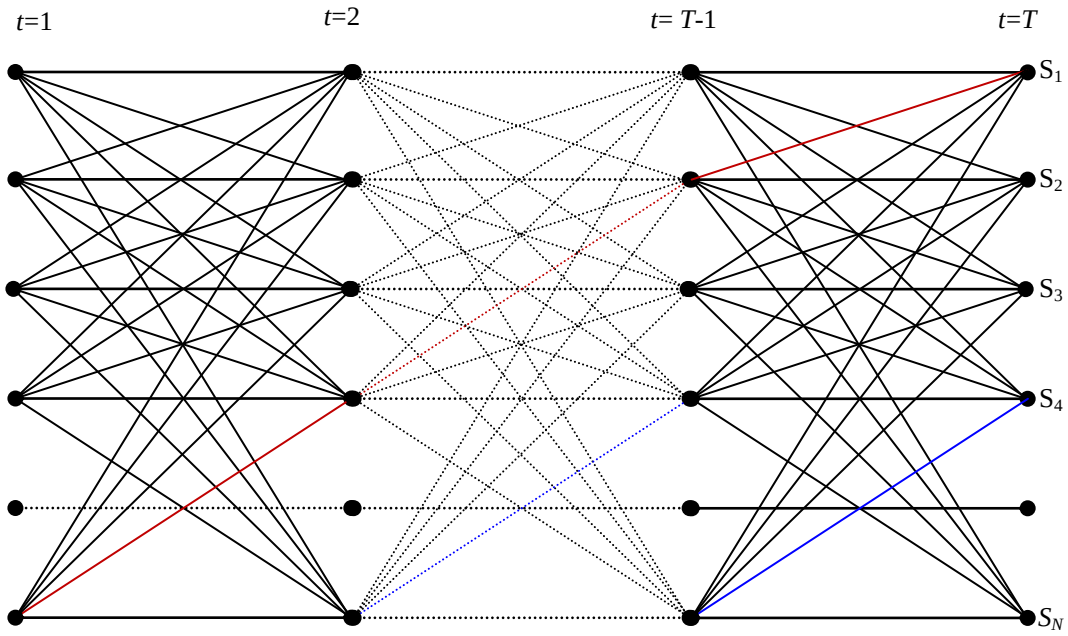


Figure 1 Evaluating probability of observation given model

Although this method is quite straight forward, it is very much computationally expensive, involving $(2T\text{-}1)*N^T$ multiplications and $(N^T\text{-}1)$ addition which is of an order of $2T*\ N^T$. For $N = 6$ and $T = 80$, the number of calculations comes out to be $2*80*6^{100}$, a number that is computationally infeasible.

Fortunately there exists an efficient algorithm known as forward–backward algorithm, which greatly reduces the number of computations. Forward–backward algorithm is an iterative process and is briefly explained here.

Let's define a forward variable, $\alpha_t(j)$ such that

$$\alpha_t(j) = P(O_1 O_2 \dots O_t, S_t = j | \Delta) \quad (5)$$

This gives the probability of partial observation sequence $O = O_1 O_2 \dots O_t$, till time t and being in state j at time t.

$\alpha_t(j)$ can be calculated inductively for successive values of t until $t=T$. Thus $\alpha_T(j)$ gives the probability of getting observation O with final state being j. Summing $\alpha_T(j)$ for all values of j gives the probability of observation O for all states sequences, which is the total probability of observation O. This process can be summarized as

a) Initialization

$$\alpha_1(j) = \pi_j * b_j(O_1) \qquad 1 \leq j \leq N \tag{6}$$

b) Iterations

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j) * a_{ji}\right] * b_i(O_{t+1}) \quad 1 \leq i \leq N \;\&\; 1 \leq t \leq (T-1) \tag{7}$$

c) Final step, summation over all end states

$$P(O|\,\Delta) = \sum_{j=1}^{N} \alpha_T(j) \tag{8}$$

This iterative process greatly reduces the number of computation required compared to the direct method. The number of computations required, using forward variable is N*(N+1)*(T-1). For the system with N=6 and T=80, considered earlier the number of operations will become, 6*7*79=3318, an order of magnitude smaller than the direct method.

The above algorithm can also be implementing by defining a backward variable $\beta_t(j)$ as the probability of being in state j at time t with the partial observation sequence after time t being $O_{t+1}\, O_{t+2} \ldots O_T$, given the model Δ. The backward variable $\beta_t(j)$ can be used to calculate P(O/Δ) in a way similar to that of forward variable $\alpha_t(j)$. Equation (4) can be solved using either forward or backward variable, the backward variable is introduced here as they it is used in the solutions of second and third problem. Solution of equation (4) using backward variable is summarized as follows

a) Initialization

$$\beta_T(j) = 1 \quad 1 \leq j \leq N \tag{9}$$

b) Iterations

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) * a_{ij} * b_j\,(O_{t+1}) \;\; 1 \le i \le N \;\&\; t = (T-1), \ldots, 2, 1 \qquad (10)$$

c) Final step, the total probability of O

$$P(O|\,\Delta) = \sum_{j=1}^{N} \beta_1(j) * b_j(O_1) \qquad (11)$$

2. Second Problem (decoding)

The second problem is, to find the state sequence $S = S_1\, S_2 \ldots S_T$, that maximises the joint probability of observation O and state sequence S, that is P(O,S/Δ).

Contrary to the first problem that has a unique solution, second problem may have different solutions depending on the definition of 'optimality criteria'. Optimality criteria can be defined by optimising the probability of individual states, pair of states, and so on. The optimisation criteria based on optimising the probability of individual states is described first.

Let's define a new variable $\xi_t(j)$ such that

$$\xi_t(j) = P(S_t = j|O,\Delta) \qquad (12)$$

Where $\xi_t(j)$ is the probability of being in state j at time t given the observation O and model Δ. Equation (12) can be re-written in terms of forward and backward variables, $\alpha_t(j)$ and $\beta_t(j)$ as

$$\xi_t(j) = \frac{\alpha_t(j) * \beta_t(j)}{P(O|\,\Delta)} = \frac{\alpha_t(j) * \beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j) * \beta_t(j)} \qquad (13)$$

Equation (13) can be used to find the most likely state S_t at time t

$$S_t = \arg\max_{1 \le j \le N} [\xi_t(j)] \quad 1 \le t \le T \qquad (14)$$

Equation (14) gives the best state at any time instance and entire state sequence can be then found by simply combining the individual states; however this method has a major limitation as it optimizes the probability of individual states with no regards to the resulting state sequence. In practice, some of the state sequences may not be allowed at all. A more practical way is to find out a single best state sequence $S= S_1 S_2 S_3 \ldots S_T$. This is commonly achieved by a method known as Viterbi algorithm, explained bellow.

Let us define a quantity $\sigma_t(j)$ as the 'best path' of all the state sequences, prior to time *t*, and being in state *j* at *t*. Then

$$\sigma_t(j) = \max_{S_1 S_1 \ldots S_{t-1}} P(S_1 S_1 \ldots S_t = j, O_1 O_1 \ldots O_t | \Delta) \tag{15}$$

The best path for the time *t*+1 can then be given by

$$\sigma_{t+1}(j) = \left[\max_i \sigma_t(i) * a_{ij}\right] * b_j(O_{t+1}) \tag{16}$$

The best path (sequence of states) could be found by tracking the state S over all times that maximizes equation (16). This is achieved by defining a new variable $\phi_t(j)$. The mechanism for finding the best state sequence is described as

a) Initialization

$$\sigma_1(j) = \pi_j * b_j(O_1) \quad 1 \le j \le N \tag{17a}$$

$$\phi_1(j) = 0 \tag{17b}$$

b) Iteration

$$\sigma_t(i) = \max_{1 \le j \le N}\left[\sigma_{t-1}(j) * a_{ji}\right] * b_i(O_t) \quad \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{18a}$$

$$\phi_t(i) = \arg\max_{1 \le j \le N}\left[\sigma_{t-1}(j) * a_{ji}\right] \quad \begin{matrix} 2 \le t \le T \\ 1 \le j \le N \end{matrix} \tag{18b}$$

c) End of iteration

$$P = \max_{1 \le j \le N}[\sigma_T(j)] \tag{19a}$$

$$S_T = \arg\max_{1 \le j \le N}[\sigma_T(j)] \tag{19b}$$

d) Tracking best path

$$S_t = \phi_{t+1}(S_{t+1}) \;\; t = T-1, T-2, \ldots, 2,1 \tag{20}$$

3. Third Problem (training)

The third problem is a training problem where the HMM parameters, (A, B, π) are adjusted such that P (*O*/Δ) is maximized. No analytical solution exists to determine the exact values of A, B and π, however a number of iterative method have been proposed in literature, which locally maximise the value of P(*O*/Δ). Out of these, the most popular one is Baun-Welch re-estimation algorithm, explained in the following.

Let's define a variable $\boldsymbol{\varepsilon}_t(i,j)$ such that

$$\varepsilon_t(i,j) = P(S_t = i, S_{t+1} = j | O, \Delta) \tag{21}$$

$\boldsymbol{\varepsilon}_t(i,j)$ is the probability that the system is in state *i* at time *t* and state *j* at time *t*+1 conditional upon the observation O and model Δ, as shown in Figure 2.

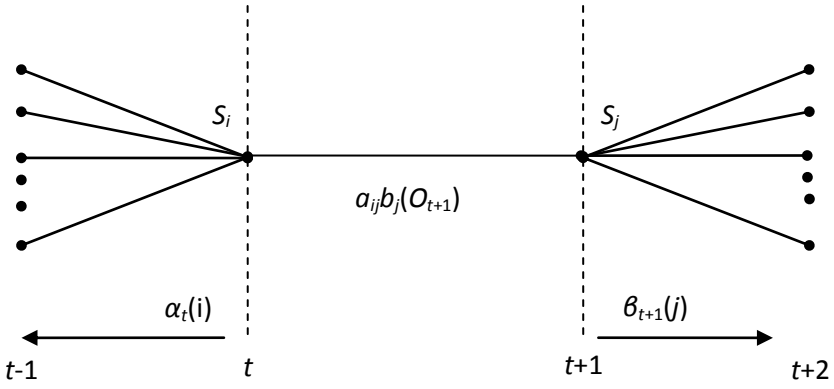


Figure 2 Baun-Welch re-estimation; states of the system at time *t* and *t*+1

$\boldsymbol{\varepsilon}_t(i,j)$ can be expressed in terms of forward and backward variables, $\alpha_t(i)$ and $\beta_t(i)$ as

$$\varepsilon_t(i,j) = \frac{\alpha_t(i) * a_{ij} * b_j(O_{t+1}) * \beta_{t+1}(j)}{P(O|\,\Delta)}$$

$$= \frac{\alpha_t(i) * a_{ij} * b_j(O_{t+1}) * \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) * a_{ij} * b_j(O_{t+1}) * \beta_{t+1}(j)} \tag{22}$$

Here the denominator is a scaling factor that makes it a probability measure. The variable $\boldsymbol{\varepsilon}_t(i,j)$ can be related to the previously defined variable $\xi_t(j)$ (the probability of being in state j at time t given model Δ and observation O), as

$$\xi_t(j) = \sum_{i=1}^{N} \varepsilon_t(i,j) \tag{23}$$

Similarly, summing $\boldsymbol{\varepsilon}_t(i,j)$ over $t = 1$ to $t = T\text{-}1$ gives the expected number of transitions from state i to state j and summing $\xi_t(i,j)$ over the same time interval gives transition from state i, that is

$$\sum_{1}^{T-1} \varepsilon_t(i,j) = expected\ number\ of\ transitions\ from\ state\ i\ to\ j \tag{24a}$$

$$\sum_{1}^{T-1} \xi_t(i) = expected\ number\ of\ transitions\ from\ state\ i \tag{24b}$$

The parameters of model Δ can now be found as follows

$$\pi_i = frequency\ of\ being\ in\ state\ i\ at\ time\ (t=1) = \xi_1(i) \tag{25}$$

The state transition probability

$$a_{ij} = \frac{frequency\ of\ transition\ from\ state\ i\ to\ j}{frequency\ of\ transition\ from\ state\ i} \tag{26a}$$

$$= \frac{\sum_1^{T-1} \varepsilon_t(i,j)}{\sum_1^{T-1} \xi_t(i)} \tag{26b}$$

and

$$a_{ij} = \frac{frequency\ of\ being\ in\ state\ j\ and\ output\ observation\ k}{frequency\ of\ being\ in\ state\ j} \tag{27a}$$

$$= \frac{\sum_{t=1}^{T} \xi_t(j)}{\sum_{t=1}^{T} \xi_t(i)} \tag{27b}$$

The model parameters thus obtained can be reused to find its second estimates. The first set of parameters is denoted as $\Delta_1 = (\pi_i (1), a_{ij}(1), b_j(1)(k))$ and the second as $\Delta_2 = (\pi_i(2), a_{ij}(2), b_j(2)(k))$. When $\Delta_1 = \Delta_2$, this is the optimal set of parameters, otherwise the iterative process is repeated until two consecutive estimates with same value are obtained.