A Novel Lip Geometry Approach for Audio-Visual Speech Recognition

Mohd Zamri bin Ibrahim

B030491

School of Electronic, Electrical and Systems Engineering

Loughborough University

United Kingdom

October 2014

ABSTRACT

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

PUBLICATIONS

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

- M.Z. Ibrahim and D.J. Mulvaney. *Geometry based Lip Reading System using Multi Dimension Dynamic Time Warping*. In IEEE International Conference on Visual Communications and Image Processing (VCIP), San Diego, USA, 27^{th} - 30^{th} November, 2012.

- M.Z. Ibrahim and D.J. Mulvaney. *Robust Geometrical-based Lip-reading using Hidden Markov Models*. In IEEE International Conference on Computer as a Tool (EUROCON), Zagreb, Croatia, 1^{st} - 4^{th} July, 2013.

- M.Z. Ibrahim and D.J. Mulvaney. *Geometrical-Based Lip-Reading using Template Probabilistic Multi-Dimension Dynamic Time Warping*. In Elsevier Journal of Visual Communication and Image Representation (JVCI), Accepted on 20th May 2014 for publication.

- M.Z. Ibrahim and D.J. Mulvaney. *A Lip Geometry Approach for Feature-Fusion based Audio-Visual Speech Recognition*. In IEEE International Symposium on Communications, Control, and Signal Processing (ISCCSP), Athens, Greece, 21^{st} - 23^{rd} May 2014.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

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

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

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

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

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

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

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

1.1 Motivation

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

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

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

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

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

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

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

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

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

1.2 Aim and objectives

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

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

1.3 Original contributions

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

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

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