Patch-Based Representation of Visual Speech

Patrick Lucey **Sridha Sridharan**

Speech, Audio, Image and Video Research Laboratory
Queensland University of Technology
Brisbane, QLD, 4001, Australia

Abstract

Visual information from a speaker's mouth region is known to improve automatic speech recognition robustness, especially in the presence of acoustic noise. To date, the vast majority of work in this field has viewed these visual features in a holistic manner, which may not take into account the various changes that occur within articulation (process of changing the shape of the vocal tract using the articulators, i.e lips and jaw). Motivated by the work being conducted in fields of audio-visual automatic speech recognition (AVASR) and face recognition using *articulatory features* (AFs) and *patches* respectively, we present a proof of concept paper which represents the mouth region as a ensemble of image patches. Our experiments show that by dealing with the mouth region in this manner, we are able to extract more speech information from the visual domain. For the task of visual-only speaker-independent isolated digit recognition, we were able to improve the relative word error rate by more than 23% on the CUAVE audio-visual corpus.

Keywords: Visual Speech Recognition (VSR), Patches, Articulatory Features (AFs).

1 Introduction

Over the past twenty years, considerable research activity has concentrated on utilizing visual speech extracted from a speaker's face in conjunction with the acoustic signal, in order to improve robustness of automatic speech recognition (ASR) systems (Potamianos, Neti, Gravier, Garg & Senior 2003). Critical to the performance of the resulting audio-visual ASR (AVASR) system is the choice of visual features that contain sufficient information about the uttered speech (Potamianos & Scanlon 2005). Even though the visual features used over this time have shown to improve robustness to the overall AVASR system in extreme noisy conditions, the visual-only speech recognition (VSR) performance in these systems do lag by over a order of magnitude to its acoustic counterpart in clean conditions (Potamianos et al. 2003). This fact, clearly highlights the lack of speech classification power current visual features possess to extract speech information to the level of its acoustic counterpart. It may be the case that the visual modality does not hold as much information as the acoustic modality, however, this has not yet been quantified which motivates this research.

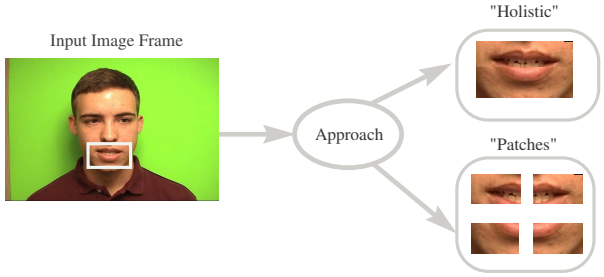

Figure 1: Following the extraction of the ROI, we propose to extract and model the ROI as an ensemble of image "patches" instead of the "holistic" approach which is currently being used in AVASR literature.

In AVASR literature, there have been numerous different methods of extracting visual features from the mouth *region of interest* (ROI) (see Section 2). However, all of these techniques modelled the ROI in a holistic, single stream manner. A potential problem which may arise from this approach is that these features may not take into account all of the various changes that occur within the mouth region during articulation (process of changing the shape of the vocal tract using the articulators, i.e lips and jaw) (Fant 1960). In contrast to the majority of work being conducted in the field of VSR, Saenko et al. has recently proposed the use of multiple streams of hidden *articulatory features* (AFs) to model the visual domain (Saenko, Darrel & Glass 2004). In this work, each sound is described by a unique combination of various articulator states, such as "lip-opened", "lip-rounded", "presence of teeth" etc.

Multi-stream approaches have also been used to good effect in the field of face recognition. Techniques that decompose the face into an ensemble of salient *patches* have reported superior face recognition performance with respect to approaches that treat the face as a whole (Brunelli & Poggio 1993, Moghaddam & Pentland 1997, Martinez 2002, Kanade & Yamada 2003). The idea behind breaking the face into patches is that it is easier to take into account changes in appearance due to the faces complicated three-dimensional shape, in comparison to treating it holistically (Lucey & Chen 2006).

Heavily motivated by the work being conducted with patches in face recognition and AFs in AVASR, we present a novel approach to VSR by breaking the ROI into a series of image patches (see Figure 1).

This research was supported by the Australian Research Council Grant No: LP0562101

Copyright ©2006, Australian Computer Society, Inc. This paper appeared at *HCSNet Workshop on the Use of Vision in HCI (VisHCI 2006)*, Canberra, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 56. R. Goecke, A. Robles-Kelly & T. Caelli, Eds. Reproduction for academic, not-for profit purposes permitted provided this text is included.

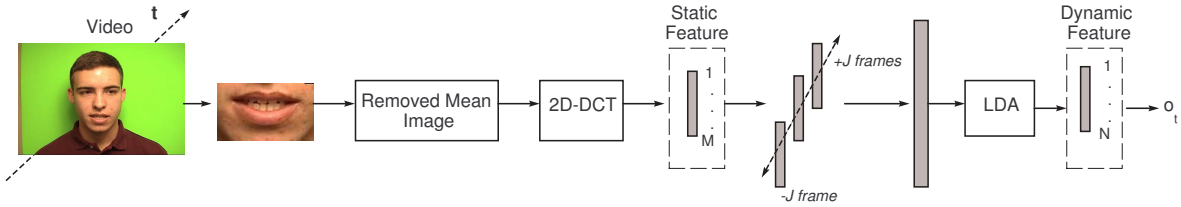

Figure 2: Block diagram of visual feature extraction process.

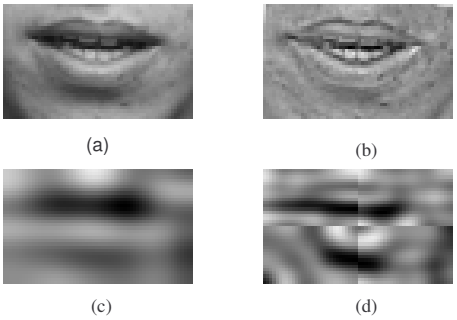

Figure 3: An example ROI from a speaker uttering the phoneme /th/ in the digit "three". (a) original image, (b) mean-removed image, (c) reconstructed "holistic" image showing just the mouth somewhat opened, and (d) reconstructed "patch-based" image, displaying the presence of teeth and lip protrusion.

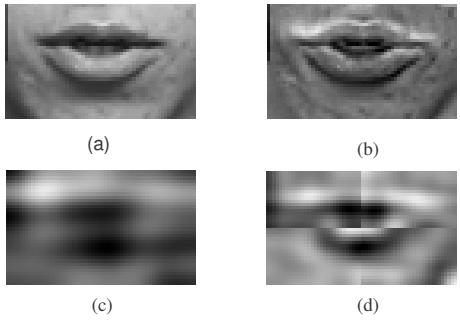

Figure 4: An example ROI from a speaker uttering the phoneme /uh/ in the digit "two". (a) original image, (b) mean-removed image, (c) reconstructed "holistic" image showing just lip openness information, and (d) reconstructed "patch-based" image, displaying the presence of lip roundness and protrusion.

It is hoped by modelling each patch separately, we can take advantage of the local information contained within each patch, and also monitor any dynamic changes that occur during articulation.

By approaching visual speech in this manner, we hope to extract more speech information which will hopefully in turn increase the overall performance of VSR. A benefit of the following approach is that we are able to avoid the *curse of dimensionality* (Chatfield & Collins 1991) by alleviating the restriction of the number of visual features able to be used. This is our main motivation behind this work and is described and discussed in some detail in Section 2.

Following that, Section 3 describes the baseline VSR system, namely ROI detection and tracking, and the holistic visual feature extraction technique and modelling details. Section 4 describes the Patch-based VSR system. Section 5 presents our experimental results, and, finally Section 6 concludes the paper with a summary and a few remarks.

2 Motivation for Patch-Based Approach

Visual speech features can be categorized into two types, namely: area, and contour based representations. Area-based representations are concerned with transforming the whole *region of interest* (ROI) mouth pixel intensity image into a meaningful low-dimensional feature vector. Such transforms used for this approach include *principal component analysis* (PCA) (Bregler & Konig 1994), discrete cosine transform (DCT) (Heckmann, Kroschel, Savariaux & Berthommier 2002), linear discriminant analysis (Belhumeur, Hespanha & Kriegman 1997) or a combination of DCT and LDA (Potamianos et al. 2003). Contour based representations, are concerned with parametrically atomising the mouth, based on a priori knowledge of the components of the mouth (i.e. outer and inner labial contour, tongue, teeth, etc.) (Wark & Sridharan 1998). An *Active Appearance Model* (AAM) (Cootes, Edwards & Taylor 1998), combines both the area and contour parameters together into a single feature vector. None of these above approaches have shown themselves to be clearly superior to each another, but due to its ability to be computed quickly, most researchers have preferred to use the area-based representation, as highlighted by the review conducted by Potamianos et al. (2003).

For area-based features, the current state-of-the-art consists of a hierarchical process. It is based on the hierarchical LDA (or *HiLDA*) process devised by Potamianos et al. (2003) and is shown in Figure 2. Firstly, the mouth ROI is extracted and features extracted using the two-dimensional DCT. The top M energy features are then selected to give a compact representation of the ROI. This resulting vector is called the *static feature*. This static feature vector is then concatenated with $\pm J$ adjacent frames and then LDA is used to project it down to N features giving the resultant *dynamic feature* vector o_t (See Section 3.2 for full description).

In literature, some researchers use only the top 20-30 DCT or PCA (very similar performance to DCT) coefficients for their static feature (Gowdy, Subramanya, Bartels & Bilmes 2004, Heckmann et al. 2002, Liang, Liu, Zhao, Pi & Nefian 2002). Potamianos et al. (2003) use the top 100 features, then use LDA to project it down to 30 features. As dynamic features provide the most information about speech (Goldschen, Garcia & Petajan 1994), it is necessary to keep the number of static features low, as computing the LDA matrix for high input features in computationally prohibitive (hence the reason why 20-30 static features are used). However, it is our contention that limiting the number of static features to around this number limits the amount of available speech stemming from the visual modality. This contention is backed up by the work conducted by Potamianos and Scanlon (Potamianos & Scanlon 2005), as they proposed another way of overcoming the dimensionality problem of the static feature vector. In this work, they made use of the laterally symmetric nature of

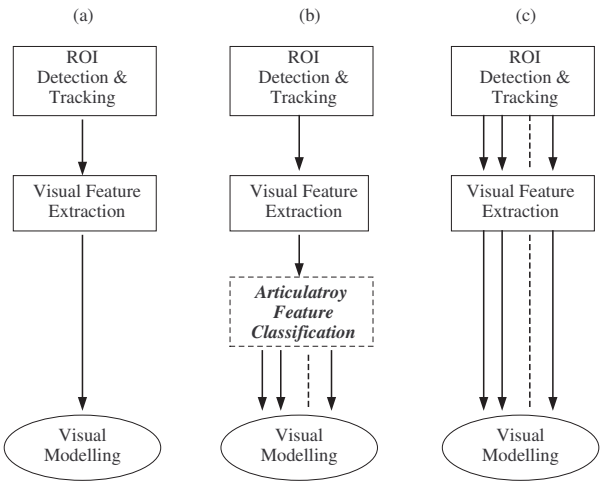

Figure 5: Comparison of the various approaches to visual speech recognition. (a) Shows the holistic approach currently being used in VSR. (b) Shows the multi-stream approach using articulatory features. (c) Shows our patch-based approach, which differs as each patch is treated independently from the initial ROI detection and tracking module.

a speaker's lips by removing the odd frequency discrete cosine transform (DCT) components from the selected visual feature vector. By removing redundancies in the frequency domain, they reported some improvement in visual speech classification.

However, in an effort to get away from conventional holistic techniques and inspired by the work conducted in face recognition with patches, we sought motivation from the following examples shown in Figures 3 and 4. In VSR systems like our baseline one (see Section 3), initially the mean ROI image is subtracted to remove speaker dependencies (Figure 3b and 4b). Due to dimensionality restrictions, only the top 30 DCT coefficients are then extracted from each frame. Upon reconstruction of these images using the 30 top DCT coefficients, it can be seen that not much mouth information is visible (Figure 3c and 4c). Only maybe the mouth being open, and some coarse shape information is retained. However, when you view the original mean-removed images, it can be seen that other important visible articulatory information information such as the presence of teeth (Figure 3b) or lip roundness and protrusion (Figure 4b) is omitted. However, if we break the ROI images into patch quadrants, and use the top 30 DCT coefficients per patch, we are able to gain a closer representation of the original ROI, obviously due to the four-fold increase in features (Figure 3d and 4d). In Figure 3d, teeth information is present, along with lip protrusion and mouth opening information. In Figure 4d not only is it visible that the mouth is open, lip protrusion and roundness information can be seen.

Obviously by using more features, we are able to see more detail in the images. However, this example shows the benefit of using patches, as each patch can be modelled separately, hence overcoming the dimensionality restriction enforced on the static feature vector by the holistic single-stream topology. This approach is similar to Saenko et al. (2004), where they used multiple streams of hidden *articulatory features* (AFs) to model the visual domain. However, this approach requires additional complexity to the overall VSR framework, where each of these articulatory states (such as "lip-opened") require extra classification (via a Support Vector Machine) prior to the

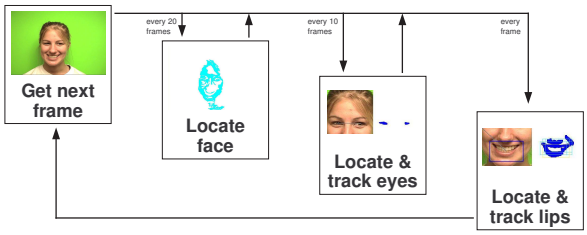

Figure 6: Overview of lip tracking system.

sound classification. The differences between all 3 approaches are shown in Figure 5.

In this paper, we show by representing the ROI as an ensemble of independent patches, we are able to obtain more visible speech information from the static features, in turn improving the overall visual speech recognition performance. This is shown through the improvement in performance for the task of speaker-independent isolated digit recognition on the CUAVE database (Patterson, Gurbuz, Tufekci & Gowdy 2002).

3 Baseline Visual Speech Recognition System

We now proceed to briefly components of our baseline VSR system. There exist three main components, which are over-viewed in the next three subsections: (a) visual front-end; (b) visual feature extraction; and (c) the visual modelling step. This baseline VSR system will be compared our patch-based system in Section 4.

3.1 Visual Front-End

Before the visual speech features can be extracted, the ROI has to be detected and tracked. In an AVASR system, this is performed by the visual front-end. For AVASR to be effective, it is essential that the visual front-end be highly accurate, otherwise these errors will cascade throughout the system and have a large effect on the ability of the final AVASR system to reliably recognize speech. This is known as the *front-end effect*.

In this study, the visual front-end consisted of three stages; face location, eye location and lip location. As shown in Figure 6, each stage was used to help form a search region for the next stage.

3.1.1 Face Location

Before face location was performed on the videos, 10 manually selected skin points for each speaker are used to form thresholds for the red, green and blue (r, g, b) values in colour-space for skin segmentation. The thresholds for each colour-space were calculated from the skin points as

$$\mu_c - \sigma_c \leq p_c \leq \mu_c + \sigma_c, \tag{1}$$

Where $c \in \{r, g, b\}$, μ_c and σ_c are the mean and standard deviation of the 10 points in colour-space c, and p_c is the value of the pixel being thresholded in colour-space c.

Once the thresholds were calculated, they were used for skin segmentation of the video to generate a bounding box of the face region within the frames every 20 frames, and this face location was remembered in the intermediate frames.

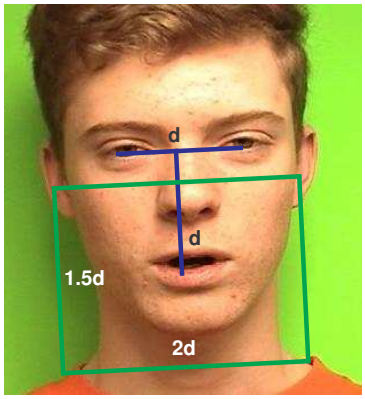

Figure 7: Calculating lip search region from eye locations.

3.1.2 Eye Location and Tracking

When transformed into $YCbCr$ space, the eye region of face images exhibit a high concentration of blue-chrominance, and a low concentration of red-chrominance. Therefore eye detection can be done in the $Cr - Cb$ space with reasonable results. However, eyebrows often appear as false positives and can degrade results. To remove the influence of eyebrows the $Cr-Cb$ image can be shifted vertically and subtracted from the original $Cr - Cb$ image. This will cancel the eyebrow minima by subtracting the eye minima, whereas the eye minima will be subtracted by the high values in the skin region and receive a large negative value suitable for thresholding (Butler, McCool, McKay, Lowther, Chandran & Sridharan 2003).

To locate the eyes from the face region from the previous stage, the top half of the face region was designated as the eye search-area, which was then searched using the shifted $Cr - Cb$ algorithm for the eye locations. The possible eye candidates were searched for two points that were not too large, too close horizontally, and not too distant vertically. Finally the two candidates which had the largest horizontal distance were chosen to be the eye locations. This process was performed every 10 frames, and the locations were remembered in the intermediate frames.

3.1.3 Lip Location and Tracking

Once the eye locations have been found, they are used to calculate a lip search region, as shown in Figure 7. The lip search region is then rotation-normalised, converted to R/G colour-space, and thresholded. The lip candidates from the thresholding are examined to remove unlikely lip locations (eg. too small, wrong shape). A search-window of 125×75 pixels is then scanned over the lip candidate image to find the windows with the highest concentration of lip candidate regions. The final lip ROI is chosen as the lowest, most central of these windows. Once the ROI was correctly located, the detected ROI was converted to grayscale and downsampled to 60×36 pixels for the experiments.

3.2 Visual Feature Extraction

The visual feature extraction process is given in Figure 2. Following the ROI extraction, the mean ROI over the utterance is removed. For purposes of notation the mouth ROI image matrix $I(x, y)$ is also expressed as the vectorised column vector $y = vec(I)$. So the mean removed mouth sub-image y^* is calculated from a given temporal mouth sub-image sequence $Y = \{y_1, \ldots, y_T\}$ such that,

$$y_t^* = y_t - \bar{y}, \text{ where } \bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \qquad (2)$$

This approach is very similar to cepstral mean subtraction used on acoustic cepstral features to improve recognition performance by providing some invariance to unwanted variations such as speaker dependencies. It is also similar to the *feature mean normalisation* of Potamianos et al. (2003), however in our approach we remove the redundant "DC" component in the image domain, instead of in the feature domain. A two-dimensional, separable, discrete cosine transform (DCT) is then applied to the resulting mean-removed image, with the $M = 30$ top DCT coefficients according to the zig-zag pattern retained, resulting in a "static" visual feature vector. Subsequently, to incorporate dynamic speech information, 21 neighboring such features over $\pm J = 10$ adjacent frames were concatenated, and were projected via an *inter*-frame LDA cascade to $N = 60$ dimensional "dynamic" visual feature vector.

3.3 The Speech Recognition System

In our experiments, we will be comparing two VSR systems: this baseline system, and our patched-based system (see Section 4). Both systems were designed to recognize isolated digits. As we are fusing multiple streams of data together, we saw isolated speech recognition as an ideal way to test our patch-based concept as it is easily implemented by calculating the likelihoods for the visual observations for a given word model. The continuous speech recognition paradigm is a much more complicated task as the number of possible hypothesis of word sequences becomes very large, and the number of best hypothesis obtained for each stream might not necessarily be the same. Our future work will concentrate on the continuous speech scenario, through the implementation of a Dynamic Bayesian Network (DBN) (Gowdy et al. 2004), which provides a framework to combine multiple streams together effectively.

In these experiments, each of the digits were modelled using 9 states and 18 Gaussians per state using HTK (Young, Everman, Hain, Kershaw, Moore, Odell, Ollason, Povey, Valtchev & Woodland 2002). These models were bootstrapped from the timed labelled transcriptions provided with the database. This topology was used as experimental and heuristic evidence showed that this was the optimal configuration.

4 Patch-Based Visual Speech Recognition System

The patch-based VSR system is very similar to that of the holistic baseline system, which was described in the previous section. The overall system is depicted in Figure 8. As it can be seen from the figure, this system is very basic. Essentially it is the baseline system being split into four parallel streams. The intended reason for this simple structure was to show that this configuration could be implemented easily. Also, by only breaking the ROI only into quadrants patches (no overlapping), we wanted to illustrate the benefit of treating parts of the ROI locally instead of as a whole.

As can be seen in Figure 8, the patch-based system uses the same visual front-end as the baseline system. Once the ROI has been detected and tracked, each grayscale 60×36 ROI image is broken up into $30 \times$

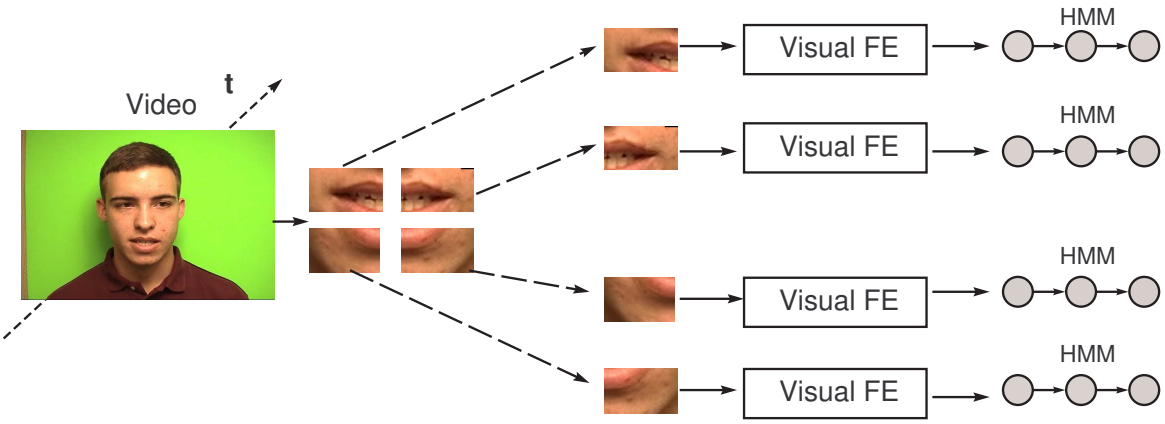

Figure 8: Block diagram of visual feature extraction process using the patch-based representation.

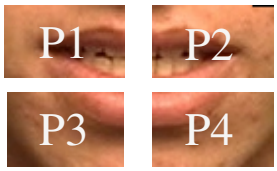

Figure 9: Once the mouth ROI has been detected and tracked, each ROI is broken up into quadrants and labelled.

18 quadrants (labelled as per Figure 9). Each one of these patches are then independently and visual features are extracted and modelled as per the process described in Section 3.2 and 3.3 respectively.

As mentioned previously, as a proof of concept we just conducted these experiments for the task of speaker-independent isolated digit recognition. As this was the case, fusion of the patches was performed via the weighted sum rule. Hence, let each spoken word be represented by a multiple visual speech observations $\mathbf{O}$, defined as

$$\mathbf{O} = \mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_R \tag{3}$$

where $\mathbf{O}_r$ refers to the sequence of visual speech observations with regard to patch r. The isolated digit recognition can then be regarded as that of computing

$$\arg\max_{i=1}^{10}\{\sum_{r=1}^{R}\beta_r P(\omega_i|\mathbf{O}_r)\} \tag{4}$$

where ω_i is the i'th digit and β_r refers to the assigned patch weight. Also it is worth noting that $\sum_r^R \beta_r = 1$, where $0 > \beta_r > 1$.

5 Experimental Results

We now proceed to report a number of experimental results on the performance of the developed patch-based VSR system. The experiments were conducted on the CUAVE database.

5.1 The CUAVE Audio-Visual Corpus

For this work, we compared the speaker-independent visual-only isolated speech recognition performances on our baseline and patch-based systems. Training and evaluation visual speech was taken from the Clemson University, *CUAVE*, audio-visual database (Patterson et al. 2002). The CUAVE database was selected as it is presently the only common audio-visual database which is available for all universities to use. This is important for benchmarking and comparison purposes. The CUAVE database consists of two major sections, one of individual speakers and one of speakers pairs. For this study, only the stationary connected-digit string section of the individual speakers were used. The stationary connected-digit string section of the database consisted of each of the 36 individual speakers uttering the connected digits "zero" to "nine" a total of 5 times each. The 36 individual speakers were divided arbitrarily into a set of 28 training speakers and 8 different test talkers for a completely speaker-independent grouping. As the database is so small, we used 10 different permutations of this configuration to see the effect of having different speakers in the training/testing set.

5.2 Isolated Digit Recognition Results

Generally, an accurate measure of how much speech information is contained within the visual features is indicative of how well it performs in the task it is being used for, which in this case is isolated digit VSR. We first performed this on the *static* visual features for the holistic (H), patch-based (P), fused holistic and patch-based features (F), patches concatenated (PC), and patches and holistic concatenated (FC). The first experiment was conducted using the same amount of features as the holistic system (i.e. $M = 30$ for H, P, F, PC and FC). For P, 8 features were used for $P1$ and $P2$ and 7 features for $P3$ and $P4$, and each patch was weighted equally. For F, 6 features where used for each patch quadrant and the holistic patch. For this configuration, the holistic approach was weighted 50% and each patch was weighted 12.5%. The PC and FC experiments were conducted to see the effect of modelling each patch independently instead of in a single stream.

The second experiment was conducted using the same method, however, the same amount of features were used for the patched-based system (i.e. $M = 120$ for H, P, F, PC and FC). For P, 30 features were used for $P1 - P4$. For F, 24 features were used for each patch quadrant and the holistic patch. The experiments were carried out in this way so that we could evaluate how much speech information there is for the same amount of features. The results are given in Table 1.

As can be seen in Table 1, using the same amount of features, the patch-based system outperforms the holistic system using both 30 and 120 features. And

Exp	H	P	F	PC	FC
1	57.10	44.72	***44.27***	66.25	55.16
2	58.69	45.38	44.80	63.73	56.17

Table 1: Isolated WERs of the static features for the: (H) holistic or baseline system, (P) patch-based system, (F) fused holistic and patch-based system, (PC) patches concatenated, (FC) holistic and patches concatenated. For experiment 1, $M = 30$ and for experiment 2, M =120.

Exp	H	P	F
1	30.10	25.95	***22.92***
2	-	28.22	23.68

Table 2: Isolated WERs of the dynamic features concatenating ± 10 frames then using LDA to yield 60 features from the static features given in Table 1.

when the holistic and patch-based system were fused together more improvement was gained. It is somewhat interesting to note that the better performance was gained in experiment 1, and not 2, with the fused holistic and patch-based system achieving the best performance with a word-error-rate (WER) of 44.27% compared to 57.10% for the holistic system. This goes against our initial hypothesis regarding dimensionality, as lower number of features actually obtained around the same or marginally more static speech information. However, it may be the case that the top 30 features contain most of the speech information, whilst the remaining features contain mostly unique speaker information. Another interesting result is that modelling each patch independently seems to achieve better results than concatenating the features and modelling them as one (PC, FC). This may suggest that representation of features is the key to VSR, and not just the sheer number of features used. However, it must be noted that these results may not be significant due to the small size of the database and further investigation is need before any claims can be made about performance.

To gauge the overall performance of the systems using the full system (i.e incorporating the dynamic features); the holistic, patch-based, and fused holistic and patch-based system were compared. The results are given in Table 2. As can be seen from these results, the fused system was again was the best performed following the trend of the previous experiments. For experiment 1, the WER of 22.92% was much better than the holistic one of 30.10%, giving a 23.9% relative improvement. Again these results look very promising, but further investigation really needs to be done before determining whether these results are significant or not. It is also worth noting that no holistic result for experiment 2 could be gain as the dimensionality for the LDA matrix was too large to be computed.

6 Summary and Conclusion

In this paper, we presented a novel patch-based approach to the task of VSR which showed improvement over holistic approaches. Our results show that our concept of breaking up the mouth ROI into patches, instead of just one whole, could extract more speech information from the visual domain. We understand that a major limitation of our experiments was the small size of our training and testing database. However, we believe that the results give an indication that this patch-base approach is worth pursuing on an larger database, as well as on the more complicated task of continuous speech recognition. Our future work will concentrate on the continuous speech recognition scenario, through the implementation of a Dynamic Bayesian Network (DBN), which provides a framework to combine multiple streams together effectively. We believe the DBN framework is a far more prudent way to go rather than using feature fusion as this approach really is not practical as it does not allow us to weight the various patches and may cause catastrophic fusion. Another task we will be undertaking in the future will be investigating which patches in the ROI (or even the face) are most pertinent for visual speech (such as corner of mouths, mouth center, cheeks etc), so as to further enhance VSR.

7 Acknowledgements

We would also like to thank Clemson University for freely supplying us their CUAVE audio-visual database for our research.

References

Belhumeur, P., Hespanha, J. & Kriegman, D. (1997), 'Eigenfaces vs Fisherfaces: Recognition using class specific linear projection', *IEEE Trans. Pattern Anal. Machine Intell.* **19**(7), 711–720.

Bregler, C. & Konig, Y. (1994), Eigenlips for robust speech recognition, *in* 'International Conference on Acoustics, Speech and Signal Processing', Vol. 2, Adeliade, Australia, pp. 669–672.

Brunelli, R. & Poggio, T. (1993), 'Face recognition: Features versus templates', *IEEE Trans. PAMI* **15**(10), 1042–1052.

Butler, D., McCool, C., McKay, M., Lowther, S., Chandran, V. & Sridharan, S. (2003), Robust face localisation using motion, colour and fusion, *in* C. Sun, H. Talbot, S. Ourselin & T. Adriaansen, eds, 'Seventh International Conference on Digital Image Computing: Techniques and Applications', CSIRO Publishing, Macquarie University, Sydney, Australia.

Chatfield, C. & Collins, A. J. (1991), *Introduction to Mulitvariate Analysis*, London, Unitied Kingdom: Chapman and Hall.

Cootes, T., Edwards, G. & Taylor, C. (1998), Active appearance models, *in* 'Proc. Europ. Conf. Computer Vision', Germany, pp. 484–498.

Fant, G. (1960), Acoustic theory of speech production.

Goldschen, A. J., Garcia, O. N. & Petajan, E. (1994), Continuous optical automatic speech recognition by lipreading, *in* A. Singh, ed., 'Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers', Vol. vol.1, IEEE Comput. Soc. Press, Pacific Grove, CA, USA, pp. 572–577.

Gowdy, J. N., Subramanya, A., Bartels, C. & Bilmes, J. (2004), DBN Based Mulit-Stream Models for Audio-Visual Speech Recognition, *in* 'Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing', Vol. 1, pp. 993–996.

Heckmann, M., Kroschel, K., Savariaux, C. & Berthommier, F. (2002), Dct-based video features for audiovisual speech, *in* 'Proc. Int. Conf. Spoken Language Processing', pp. 1925–1928.

Kanade, T. & Yamada, A. (2003), 'Multi-subregion based probabilistic approach towards pose-invariant face recognition', *IEEE International Symposium on Computational Intelligence in Robotics Automation* **2**, 954–959.

Liang, L., Liu, X., Zhao, Y., Pi, X. & Nefian, A. (2002), Speaker Independent Audio-Visual Continuous Speech Recognition, *in* 'Proc. Int. Conf. on Multimedia and Expo', Vol. 2, pp. 25–28.

Lucey, S. & Chen, T. (2006), Learning patch dependencies for improved pose mismatched face verification, *in* 'IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)'.

Martinez, A. M. (2002), 'Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class', *IEEE Trans. PAMI* **24**(6), 748–763.

Moghaddam, B. & Pentland, A. (1997), 'Probabilistic visual learning for object recognition', *IEEE Trans. PAMI* **19**(7), 696–710.

Patterson, E. K., Gurbuz, S., Tufekci, Z. & Gowdy, J. N. (2002), CUAVE: a new audio-visual database for multimodal human-computer interface research, *in* 'Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing', Orlando.

Potamianos, G., Neti, C., Gravier, G., Garg, A. & Senior, A. W. (2003), 'Recent advances in the automatic recognition of audio-visual speech', *Proc. of the IEEE* **91**(9).

Potamianos, G. & Scanlon, P. (2005), Exploiting lower face symmetry in appearance-based automatic speechreading, *in* 'Proceedings of the Auditory-Visual Speech Processing International Conference 2005', British Columbia, Canada, pp. 79–84.

Saenko, K., Darrel, T. & Glass, J. (2004), Articulatory features for robust visual speech recognition, *in* 'Int. Conf. Mulitmodal Interfaces'.

Wark, T. & Sridharan, S. (1998), An approach to statistical lip modelling for speaker identification via chromatic feature extraction, *in* 'International Conference on Pattern Recognition', pp. 123–125.

Young, S., Everman, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2002), *The HTK Book (for HTK Version 3.2.1)*, Entropic Ltd.