Lip Feature Extraction Using Red Exclusion

Trent W. Lewis **David M.W. Powers**

Computer Science, School of Informatics and Engineering
Flinders University of South Australia
Email: {lewi0146,powers}@infoeng.flinders.edu.au

Abstract

Automatic speech recognition (ASR) performs well under restricted conditions, but performance degrades in noisy environments. Audio-Visual Speech Recognition (AVSR) combats this by incorporating a visual signal into the recognition. This paper briefly reviews the contribution of psycholinguistics to this endeavour and the recent advances in machine AVSR. An important first step in AVSR is that of feature extraction from the mouth region. This paper examines several well-known pixel based techniques - grayscale, horizontal edge, red and hue colour space - and compares how well they work on our naturalistic database. Finally, a novel method of feature extraction, red exclusion, is described that outperforms the others on this data set.

1 Introduction

The major aim of this project is to improve the performance of a standard automatic speech recognition (ASR) system by using information from the traditional, auditory signal as well as a visual signal. In effect, the goal of this research is to enable the computer to "lip-read". The motivation for this endeavour stems from the acknowledgement that although current, commercial ASR systems have been touted with word recognition rates of 98-99% [PC1998, 1998], these rates are usually achieved with one speaker, a close, head-mounted microphone, minimal background noise, and considerable dependence on word prediction models. In a noisy environment, or where wearing a head microphone is not practical the recognition rates of such systems degrade[Bregler et al., 1993]. For a robust recognition solution, additional information is required - here we focus on the provision of this information in the form of a visual image, i.e. audio-visual speech recognition (AVSR). This project is also motivated by the fact that psycholinguistic research has found that visual cues play an important role in speech perception by humans [Dodd and Campbell, 1987]. Therefore, the integration of auditory and visual signals to improve speech recognition is not only of benefit to automatic speech recognition systems but it also has psychological plausibility. Thus, a secondary aim is to better understand the role of visual cues in human speech recognition.

An important first step in AVSR is the the extraction of lip features that (may) contribute to visual speech recognition. These features seem likely to include width, height, and general mouth shape, as well as dynamic features such as velocity and inter-frame motion. In this paper, we present a novel pixel-based approach to lip feature extraction that, using our AV database, outperforms other techniques in terms of correct feature identification. We first outline the context of this work, AVSR - both human and machine, and then the feature extraction technique is described and compared to other well-known algorithms.

2 Context

In AVSR, knowledge from diverse areas needs to be brought together to fully understand the problem at hand. The first two parts of this section give a brief overview of psycholinguistic research in the area and the current progress of machine AVSR[1]. The final part places this current work into the broader aspect of the project that we are undertaking, namely, low-cost AVSR in natural conditions.

2.1 Psycholinguistic Research

The knowledge of both the psychological and linguistic aspects of AVSR by humans are valuable tools for exploration in this rapidly developing field. The way in which humans perceive speech, both acoustically and visually, may not be the best or most efficient in engineering terms, but such work can enlighten how one might start tackling the problem. Thus, instead of blindly attempting to get a machine to recognise speech visually, the work from psycholinguistics can be included to produce a potentially more elegant and refined solution.

One reason why humans may benefit from a visual signal is because our various speech articulators affect the sounds we produce. To produce a sound we must force air from our lungs into the trachea, glottis and larynx before it passes into the vocal tract, formed by the pharynx, nasal and oral cavities[Fromkin et al., 1996]. In terms of visual speech, organs that participate in modulation of the sound wave and are visible are of most importance. Lips, teeth, and tongue have been identified as the primary indicators for visual speech[Robert-Ribes et al., 1996], however, the cheeks, chin and nose are also very useful as secondary indicators. To an extent, the entire facial expression is used and because more than just the lips are used, the term *speechreading* has evolved to take the place of 'lip-reading'.

One of the most important findings in this area is that of the *viseme*. A viseme is the virtual sound attributed to a specific mouth (or face) shape. The viseme is analogous to the phoneme in the auditory domain, however, there does not exist a one-to-one mapping between the two. Phonemes are the distinctive sound segments that contrast or distinguish words, for example, /p/ as in pit and /b/ in bit [Fromkin et al., 1996]. Experiments have found that the ability of humans to distinguish different consonant phonemes in the presence of noise can group phonemes into larger, different

[1] For a comprehensive review the reader is directed to [Stork and Hennecke, 1996] and[Dodd and Campbell, 1987]

Table 1: Consonant viseme classes

Label	Place of Articulation	Phoneme(s)
LAB	labial	/p,b,m/
LDF	labiodental fricatives	/f,v/
IDF	interdental fricatives	/th,dh/
LSH	lingual stops and h	/d,t,n,g,k,ng,h/
ALF	alveolar fricatives	/s,z/
LLL	-	/l/
RRR	-	/r/
PAL	palatal veolars	/sh,zh/
WWW	-	/w/

classes [Summerfield, 1987]. Under a signal-to-noise ratio of -6dB, humans are only able to audibly distinguish consonants on the basis of voicing (voiced/voiceless) and nasality. In contrast, visual discrimination doesn't degrade with increasing acoustic noise and hierarchical clustering of human experimental results have found that, from the standpoint of confusion and noise degradation, visemes actually form a *complementary* set to phonemes[Walden et al., 1977]. Table 1 shows the 9 distinct, humanly perceivable viseme classes, as well as their common place of articulations as noted by Cohen, Walker, and Massaro[Cohen et al., 1996]. A further distinction can also be made within the LSH class, which involves a split between the alveolar stops and nasal, /t,d,n/, and the velar/glottal stops and nasal, /g,k,ng,h/[Goldschen et al., 1996].

A key problem both in psycholinguistics and machine AVSR is how the integration of the information gained from the visual and auditory modalities occurs. Several psychological models have been developed to help explain this problem, and they differ mainly in the timing of integration and the representations used[Walden et al., 1977, Robert-Ribes et al., 1996]. The two most widely adhered to models are the Direct Identification (DI) and Separate Identification models. In the DI model, input signals are directly transmitted to a bimodal classifier in which the phoneme is selected. There is no common representation level over the two modalities between the signal and the percept and this model can be considered a type of early integration. On the other hand, in the SI model, the auditory and visual components are separately identified through two parallel identification processes. Thus, fusion takes place after identification has been done in each modality and then a hypothesised phoneme is produced. This is an example of late integration.

2.2 Machine AVSR

Machine AVSR must not only deal with the recognition of the auditory signal, as in ASR, but it must also decide on a number of important design questions concerning visual processing. Some of the questions, pointed out by Hennecke, Stork, and Venkatesh Prasad[Hennecke et al., 1996], are outlined below.

1. How will the face and and mouth region be found?
2. Which visual features to extract from the image?
3. How are auditory and visual channels integrated?
4. What type of learning and/or recognition is used?

Unfortunately, there is still no consensus on the answers to any of these questions. Many different approaches have been developed for each, of which we can only mentioned the general aspects of the main techniques.

1. There are some AVSR systems that processes both the audio and visual channels, and complete recognition in near real-time. These types of systems need to be able to initially locate the face from a cluttered background, a research area in itself, and then extract the mouth region for further analysis. A prime example of this is the Interactive Systems Laboratory complete multi-modal human computer interface, of which part is a movement-invariant AVSR system[Duchnowski et al., 1995]. In this case, as it is with many other systems, the face is found with colour. This simple, but effective, technique works because the colour of human skin (normalised for brightness/white levels) varies little between individuals, and even races[Hunke and Waibel, 1994, Yang and Waibel, 1996]. Once the face is located it is necessary to pinpoint the mouth within the face. This usually achieved using either a triangulation with the eyes (or nose) which are more easily located [Steifelhagen et al., 1997], or by finding an area with high edge-content in the lower half of the face region[Hennecke et al., 1995]. Given the large amount of research already carried out in face locating/recognition[Chelappa et al., 1995], many researches in AVSR opt to skip the stage and start working with pre-cropped mouth images (eg.[Gray et al., 1997], [Movellan, 1995]). This allows for a relatively quicker progression for researchers beginning work in this area.

2. Once the mouth region is found, either automatically or by hand, useful lip features must be extracted that can be used visual or audio-visual speech recognition. It is at this stage where research groups begin to differ greatly in the extraction techniques applied. Some prefer to use low-level, pixel based approaches with minimal alteration to the original image (eg. [Movellan and Mineiro, 1998] or [Meier et al., 1999]), whilst others insist that a high-level, model approach is the most efficient way to proceed (eg. [Hennecke et al., 1996] or [Leuttin and Dupont, 1998]). Section 3 elaborates further on this stage of AVSR and presents a novel extraction technique.

A researcher's answers to questions 3 and 4 are intimately intertwined as the type of recognition algorithm used heavily influences the type, and method of integration used. The recognition problem here is basically a pattern matching problem and many of the recognition techniques from traditional ASR can be used, with modifications, for visual recognition of visemes. Thus, many researchers are biased in the choice of recognition and integration algorithms by what type of ASR system they may have been developing previously and therefore see AVSR as merely an extension to their already powerful ASR system (eg. [Meier et al., 1999]). This is not a problem unless the researcher does not take into account the special characteristics of the visual forms of phonemes, that is, what is practical and what is not.

The two most widely used recognition techniques are the Neural Network (NN) and the Hidden Markov Model (HMM) [Hennecke et al., 1996]. HMMs [Charniak, 1993] have the distinct advantage that they are inherently rate invariant and this is especially important for speaker independent ASR, where different speakers speak at different rates. Another important factor of HMMs concerning recognition, is that there are efficient algorithms for training and recognition, which is hugely beneficial when dealing with the large amounts of visual data that

accumulates, especially if recognition is to be done in real-time. NNs, on the other hand, are often criticised for their slow trainability and variance due to rate. However, they do have the empowering ability of generalisabilty, given large enough training sets, and, moreover, they do not make any assumptions about the underlying data.

As mentioned in the previous section, the two most closely followed models of integration are the DI and SI. In the DI model, feature vectors of the acoustic and visual signals can be, in the simplest form, concatenated together, and then this vector can be used as input into the HMM [Adjoudani and Benoit, 1996] or NN [Meier et al., 1999]. It is obvious that when following the DI model integration occurs automatically, and it is up to the recognition engine to decide upon the important features. However, under a SI model, integration can become somewhat trickier. The simplest case is when the outputs of separate NNs are feed into another NN that effectively performs the integration task. In the case of HMMs the resulting log-likelihoods are combined in some way to produce a final estimate. The most common, and simplest way to integrate the log-likelihoods is to combined them in such a way to maximise their cross-product. Late integration (SI) is an evolving area in AVSR and is a difficult issue to contend with, this is because fusing the two signals can lead to what has become known as *catastrophic fusion* [Movellan and Mineiro, 1998]. This is when the accuracy of the fused outcome is less than the accuracy of both individual systems. Much work is underway, for both HMMs and NNs, in trying to automatically bias one signal, when conditions are adverse for the other [Movellan and Mineiro, 1998, Adjoudani and Benoit, 1996, Meier et al., 1996, Massaro and Stork, 1998].

2.3 The Broader Aspect

Many of the AVSR systems that have been tested are often restricted to operate in well-defined experimental conditions, for example, controlled lighting conditions, and minimal acoustic and visual noise levels. Performance of these systems in adverse conditions is usually tested by artificially increasing the noise levels [Movellan and Mineiro, 1998]. One of the goals of this project is to train and test the AVSR system with naturally degraded input, with an unknown amount of noise, such that the system should perform well in all conditions. This includes the development of a robust visual system for finding lip features, which is the focus of section 3. Figure 1 is a schematic representation of the architecture of the AVSR system that we are developing. Using a low-cost, off-the-shelf (OTS) integrated audio-visual capture device[2], the audio and visual signals are passed through preprocessing stages where feature vectors are built up. Currently this stage is completed off-line, but there is progress being made towards real-time feature extraction. The feature vectors can be further reduced in sized by used a data reduction technique, for example principal components analysis (PCA) or its generalisation, singular valued decomposition (SVD) [Gray et al., 1997, Schifferdecker, 1994] . This is a common trick for overcoming the large amounts of for visual processing, which can improve and speed up training when using NNs. The feature vectors are then passed to a classifier, in this case an NN, where the phoneme (viseme) is identified. This is a stage where this system differs from others, in that we are recognising the sub-word units (phonemes) rather than attempting to identify whole words [Movellan and Mineiro, 1998, Rao and Mersereau, 1994], where gestures and relations are more complex and thus less complexity should be involved. Integration could possible proceed along any of the dotted lines indicated in Figure 1 or at the end, after each subsystem has made its classification.

[2]In this case, a Philips Vesta Pro (PCVC680K).

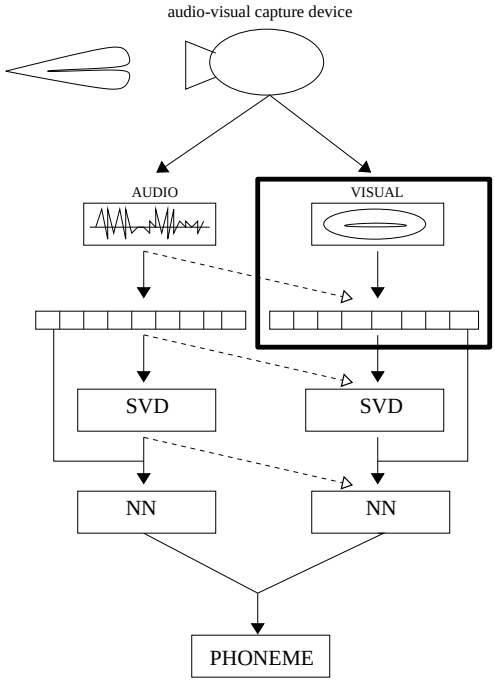

Figure 1: Architecture for AVSR system. A dotted line indicates possible early integration path and the bounded box indicates the focus of this paper

As one of the motivations for this project is AVSR in natural conditions, it was necessary to collect our own data set, that potentially had noise in both acoustic and visual sources. Furthermore, of the datasets that do exist [Web, 2000, Movellan, 1995], they are usually recorded using highly specific recording equipment and another aspect of this project is the use of low-cost, OTS equipment. This data set consisted of words that expressed most of the phonetic contexts of the different phonemes found in (Australian) English, eg. /p/ - pot, apple, cop. These word sets were spoken by three people, 2 male and 1 female, that varied greatly in appearance (see Table 2). In the following sections, this database has been used to test the algorithms explained.

3 Lip Feature Extraction

3.1 Related Work

As mentioned, the accurate extraction of lip features for recognition is very important first step in AVSR. Moreover, the consistency of the extraction is very important if it is to be used in a variety of conditions and people. According to Bregler, Manke, Hild, and Waibel [Bregler et al., 1993], broadly speaking there exist two different schools of thought when it comes to visual processing. At one extreme, there are those who believe that the feature extraction stage should reduce the the visual input to the least amount of hand-crafted features as possible, such as deformable templates [Hennecke et al., 1994]. This type of approach has the advantage that the number of visual inputs are drastically reduced - potential speeding up subsequent processesing and reducing the variability and increasing generalisability. How-

ever, this approach has been heavily criticised as it can be time consuming in fitting a model to each frame [Rao and Mersereau, 1994] and, most importantly, the model may exclude linguistically relevant information [Gray et al., 1997, Bregler et al., 1993]. The opponents of this approach believe that only minimal processing should be applied to the found mouth image, so as to the amount of information lost due to any transformation. For example, Gray et al. [Gray et al., 1997] found that simply using the difference between the current and previous frames produce results that were better than using PCA. However, in this approach the feature vector is equal to the size of the image (40x60 in most cases), which is potentially orders of magnitudes larger than a model based approach. This can potentially become a problem depending on the choice of recognition system and training regime, however, successful systems have been developed using both HMMs and NNs using this approach [Movellan and Mineiro, 1998, Meier et al., 1999].

Of course there are many system that lie in between the two extremes, and the model extrema can also benefit from better feature extraction methods as this is the first step of many models. We will now examine some of the more popular methods for initial feature extraction and how well they work for the subjects in our data set. The first feature set that is usually extracted from the mouth area is the lip corner pair. From this stage many of the algorithms use very similar techniques, such as peak picking [Prasad et al., 1993], and thus the focus will be on how they extract the lip corners.

3.1.1. Horizontal Edges. One of the most common methods for feature extraction of mouth features is the use of the grayscale value and edge detection [Rao and Mersereau, 1994, Steifelhagen et al., 1997]. The initial step, as is with many of these techniques, is the identification of the vertical position of the centre of the mouth. This can be achieved by taking the sum of each row and finding the row with the minimum value, Figure 2a. Then by examining the actuals values of the minimum row, and possibly rows close to it, from the left and right, one can discover the lip corners by setting a threshold. In Figure 2, the threshold was set to the average of the maximum and minimum values for that row. For subject 1 the method works well, however, on subject 2, Figure 3a, the method works poorly due to the slight presence of a beard.

Another common method that makes use of grayscale values, that has been more successful, is the use of horizontal edges [Steifelhagen et al., 1997]. The rationale behind this idea is that the mouth area has a high edge content, especially in the horizontal direction. These horizontal edges can easily be identified by convolving the image with a 3x3, DY prewitt operator, and then the resulting image can be thresholded, at an appropriate edge value, and a similar search method used as before. This algorithm once again works well for subject 1, however, for the bearded subject 2, performance is way below what is acceptable. The beard itself has high edge content in both verticle and horizontal directions and, thus the edge finding technique falls down under this generalisation. Increasing the threshold any further will decrease the amount of beard detected, but, unfortunately, this also results in shrinkage of the detected lip region.

3.1.2. Red, Green, and Blue. To overcome the problem of beards, researchers turned to working with colour images. Taking a leaf out of the face locating research (eg. [Yang and Waibel, 1996]), they have primarily been working with the red colour spectrum for identification of the lip region and features. As an example, Wark, Sridharan, and Chandran [Wark et al., 1998] used equation (1) to identify candidate lip pixels.

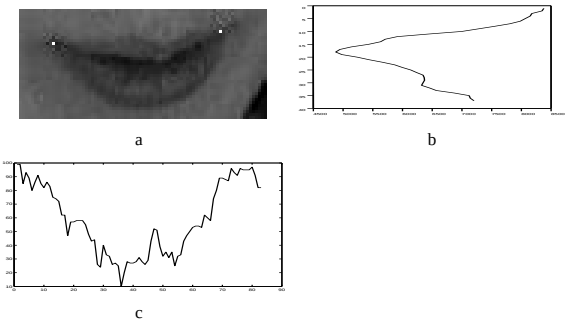

Figure 2: Lip corner extraction using grayscale values for subject 1. a) found lip corners, b) grayscale row sum, and c) grayscale value of minimum row sum.

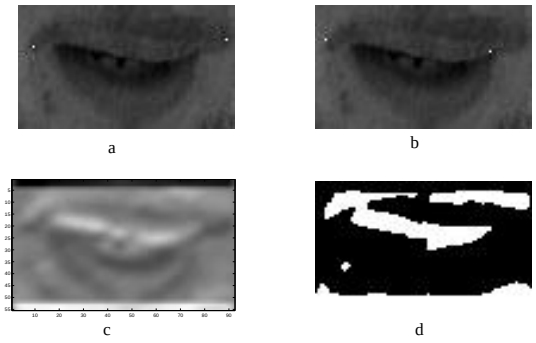

Figure 3: Lip corner extraction using grayscale values and edges for subject 2. a) "found" lip corners - grayscale, b) "found" lip corners - edges, c) horizontal edge magnitude, and d) thresholded edge image (> 10).

$$L_{lim} \leq \frac{R}{G} \leq U_{lim}, \tag{1}$$

where R and G are the red and green colour components, respectively, and L_{lim} and U_{lim} are the lower and upper boundaries that define which values of $\frac{R}{G}$ are considered lip pixels.

After removing some spurious pixels and morphologically opening and closing the image resulting from equation (1), Wark et al. [Wark et al., 1998] were able to accurately define the outer contour of the mouth, a very successful result considering the previous section. When this method was tried on the subjects of our data set, the results were better than previous, however, the was a lack of consistency in identifying the lip corners. Moreover, as can be seen in Figure 4a and c, the lip corners can be identified, but it lacks the ability to further identify other features of the mouth (b and d), eg. the top lip boundary. This would mean that a further processing step would need to be involved to calculate these other features, thus increasing processing time.

3.1.3 Hue, Saturation, and Value. The hue, saturation, and value[3] (HSV) colour space can, and has been exploited for the use of extracting lip information from images [Coianiz et al., 1996, Vogt, 1996]. The main reason a HSV colour space is preferred is that it disentangles illumination from colour, such that variations in lighting should not cause great variation in hue. Thus, Coianiz et al.

[3] also known as intensity

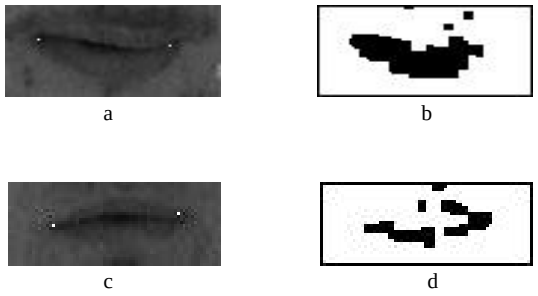

Figure 4: Lip corner extraction using equation (1). a) "found" lip corners - subject 1, b) binary image after application of (refeq:red) - subject 2, c) "found" lip corners - subject 3, and d) binary image after application of (1) - subject 3.

[Coianiz et al., 1996] and Vogt [Vogt, 1996] both use the hue value to calculate candidate lip pixels. Both use a similar algorithm to compute the likelihood of a pixel being part of the lip. We therefore will explain Coianiz and colleagues' algorithm in depth and not Vogt's[4]. The likelihood of a pixel being part of the lips is based on a predefined hue value, h_0 that is representative of lip hue and,

$$f(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & , \quad |h - h_0| \leq w \\ 0 & , \quad otherwise \end{cases} \quad , \qquad (2)$$

where h represents the current hue value and w controls the distance in which the surrounding hue values drop to zero. This function enhances those hue values close to h_0, in this case the lip hue. Thus, as Coianiz et al. [Coianiz et al., 1996] and Vogt [Vogt, 1996] found, this method can be used to identify various lip features, such as width and height. However, once again using this extraction technique did not work to a satisfactory level for all three subjects of our data set (Figure 5a, c, and e). Most noticeably, the mouth region is hardly distinguishable from the surrounding area when viewing the hue transform; that is, that application of equation (2) to an image. When the hue transform is thresholded at an optimal value, and and the result layered over the top of the grayscale image (Figure 5b, d, and f), we can see that this method only partially picks up the mouth area as well as surrounding skin areas. Thus, although this hue transform technique works well under ideal conditions, it has not extended well to our three subjects and conditions. As we are looking for a robust and general feature extraction method, this algorithm is not sufficient to serve out purposes.

3.2 Lip Feature Extraction Using Red Exclusion

The last section showed that many of the current pixel-based techniques do not adequately identify the lip corners, or even the lip region in some cases. This led to us to define our own lip feature extraction technique. This novel technique, rather than looking at the red colour spectrum, focuses on the green and blue colour values. The rationale is that as the face, including lips, are predominantly red, such that any contrast that may develop would be found in the green or blue colour range, red exclusion. Thus, after convolving with a Gaussian filter to remove any noise, the green and blue colours are combined as in,

$$log\left(\frac{G}{B}\right) \leq \beta \qquad (3)$$

[4]the major difference between the two algorithms is that Vogt [Vogt, 1996] includes saturation in calculating the likelihood

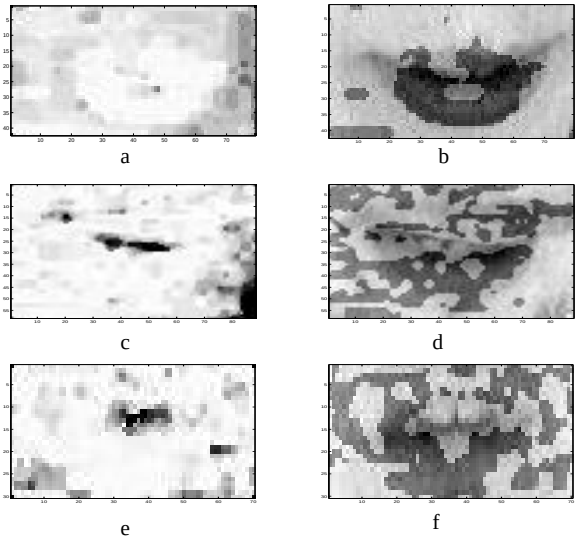

Figure 5: Hue transform (2) and enhanced grayscale image. a,b) subject 1, c,d) subject 2, and d,e) subject 3

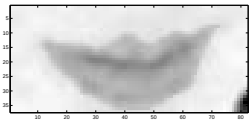

Figure 6: Grayscale enhanced mouth region of subject 1 using red exclusion.

Using the log scale further enhances the contrast between distinctive areas, and by varying the threshold β the mouth area and the lip features can easily be identified on all three different subjects (Figures 6, 7, 8). Currently β is calculated manually for each subject. Furhter analysis of equation (3) and the skin/lip colour variation before machine automation is possible.

Using the red exclusion method over a sequence of images to identify the lip corners resulted in near perfect results, as in Figure 9. Thus, this novel method of mouth identification has successfully extracted the mouth region from three very different subjects, and then this has been extended to tracking the lip corners over a series of images. This simple method has been far more successful at identification across all three subjects (even bearded), than any of methods discussed previously. It is important to note that this method works consistently well over all subjects tested to date, whilst the published algorithms tested did not.

4 Summary

In this paper we have briefly reviewed the area of AVSR and explained how the novel algorithm fits into the overall system that is under development. A com-

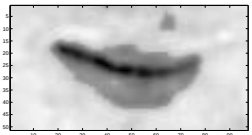

Figure 7: Grayscale enhanced mouth region of subject 2 using red exclusion.

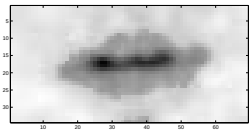

Figure 8: Grayscale enhanced region of subject 3 using red exclusion.

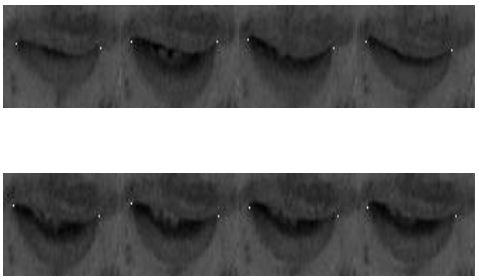

Figure 9: Tracking of lip corners for subject 2 using red exclusion

parison has been made against other lip feature extraction techniques and, on our own data set, this novel feature extraction method outperforms other methods and succeeds extremely well at mouth region finding on all three subjects (table 2). The next step in this project is to use this feature extraction method to extract relevant features from the mouth, like mouth width and height. It is also proposed that instead of using only high-level features or low-level pixels as inputs, this system will take both as inputs into SVD and then the resulting significant features will then be used for training/recognition.

As this method accurately finds the mouth region, it could be used as a precursor to the modelling stage. Many of deformable template models (eg. [Hennecke et al., 1994, Coianiz et al., 1996]) use a pixel-based technique to align the model to image and most use some sort of edge field for this purpose. In this paper we have seen that edges for mouth finding can be variable, especially when the subject has a beard. Thus, a possible extension of the technique outlined would be to incorporate it into a deformable template, which would be beneficial for both techniques - red exclusion would be a better starting point for a deformable template, and the pixel extraction method would be enhanced with long range interactions.

Table 2: Comparison of feature extraction techniques. *does not work for further processing, see text.

Algorithm	Subject 1 female	Subject 2 male, bearded	Subject 3 male, thin lips
Grayscale	$\surd$	$\times$	$\times$
Edge	$\surd$	$\times$	$\surd$
$\frac{R}{G}$*	$\surd$	$\surd$	$\surd$
Hue Transform	$\times$	$\times$	$\times$
Red Exclusion	$\surd$	$\surd$	$\surd$

References

[Adjoudani and Benoit, 1996] Adjoudani, A. and Benoit, C. (1996). On the integration of auditory and visual parameters in an hmm-based asr. In [Stork and Hennecke, 1996], pages 461–471.

[Bregler et al., 1993] Bregler, C., Manke, S., Hild, H., and Waibel, A. (1993). Bimodal sensor integration on the example of "speech-reading". *Proceedings of the IEEE International Conference on Neural Networks*, pages 667–671.

[Charniak, 1993] Charniak, E. (1993). *Statistical language learning*. MIT Press, Cambridge, MA.

[Chelappa et al., 1995] Chelappa, R., Wilson, C., and Sirohey, S. (1995). Human and machine recognition of faces: A survey. In *Proceedings of the IEEE*, volume 83(5), pages 705–739.

[Cohen et al., 1996] Cohen, M., Walker, R., and Massaro, D. (1996). Perception of synthetic visual speech. In [Stork and Hennecke, 1996], pages 153–168.

[Coianiz et al., 1996] Coianiz, T., Torresani, L., and Caprile, B. (1996). 2d deformable models for visual speech analysis. In [Stork and Hennecke, 1996], pages 391–398.

[Dodd and Campbell, 1987] Dodd, B. and Campbell, R., editors (1987). *Hearing by Eye: The pyschology of lip-reading*. Lawrence Erlbaum Associates, Hillsdale NJ.

[Duchnowski et al., 1995] Duchnowski, P., Hunke, P., Busching, M., Meier, U., and Waibel, A. (1995). Toward movement-invariant automatic lip-reading and speech recognition. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing*, Detriot USA.

[Fromkin et al., 1996] Fromkin, V., Rodman, R., Collins, P., and Blair, D. (1996). *An Introduction to Langauge*. Hartcort Brace and Company, Sydney, 3rd edition.

[Goldschen et al., 1996] Goldschen, A., Garcia, O., and Petajan, E. (1996). Rationale for phoneme-viseme mapping and feature selection in visual speech recognition. In [Stork and Hennecke, 1996], pages 505–515.

[Gray et al., 1997] Gray, M., Movellan, J., and Sejnowski, T. (1997). Dynamic features for visual speechreading: A systematic comparision. In Mozer, Jordan, and Persche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge MA.

[Hennecke et al., 1995] Hennecke, M., Prasad, K. V., and Stork, D. (1995). Automatic speech recognition using acoustic and visual signals. Technical Report CRC-TR-95-37, Ricoh Californian Research Centre.

[Hennecke et al., 1994] Hennecke, M., Prasad, V., and Stork, D. (1994). Using deformable templates to infer visual speech dynamics. In 28^{th} *Annual Asimolar Conference on Signals, Systems, and Computer*, volume 2, pages 576–582, Pacific Grove, CA. IEEE Computer.

[Hennecke et al., 1996] Hennecke, M., Stork, D., and Prasad, K. V. (1996). Visionary speech: Looking ahead to practical speech reading systems. In [Stork and Hennecke, 1996], pages 331–350.

[Hunke and Waibel, 1994] Hunke, M. and Waibel, A. (1994). Face locating and tracking for human-computer interaction. In 28^{th} *Annual Asimolar Conference on Signals, Systems, and Computers*, volume 2, pages 1277–1281. IEEE Computer Society, Pacific Grove CA.

[Leuttin and Dupont, 1998] Leuttin, J. and Dupont, S. (1998). Continuous audio-visual speech recognition. In *Proceedings of the 5^{th} European Conference on Computer Vision*, volume II, pages 657–673.

[Massaro and Stork, 1998] Massaro, D. and Stork, D. (1998). Speech recognition and sensory integration: a 240-year old theorem helps explain how people and machines can integrate auditory adn visual information to understand speech. *American Scientist*, 86(3):236–245.

[Meier et al., 1996] Meier, U., Hurst, W., and Duchnowski, P. (1996). Adaptive bimodal sensor fusion for automatic speechreading. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processing*, volume 2, pages 833–837.

[Meier et al., 1999] Meier, U., Steifelhagen, R., Yang, J., and Waibel, A. (1999). Towards unrestricted lip reading. In *Second International Conference on Multimedia Interfaces*, Hong Kong, http://werner.ir.uks.de/js.

[Movellan, 1995] Movellan, J. (1995). Visual speech recognition with stochastic networks. In Tesauro, G., Toruetzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7, pages 851–858. MIT Press, Cambridge.

[Movellan and Mineiro, 1998] Movellan, J. and Mineiro, P. (1998). Robust sensor fusion: Analysis and application to audio visual speech recognition. *Machine Learning*, 32:85–100.

[PC1998, 1998] PC1998 (1998). Pc magazine.

[Prasad et al., 1993] Prasad, K., Stork, D., and Wolff, G. (1993). Preprocessing video images for neural learning of lipreading. Technical Report CRC-TR-93-26, Ricoh California Research Center.

[Rao and Mersereau, 1994] Rao, R. and Mersereau, R. (1994). Lip modeling for visual speech recognition. In 28^{th} *Annual Asimolar Conference on Signals, Systems, and Computers*, volume 2. IEEE Computer Society, Pacific Grove CA.

[Robert-Ribes et al., 1996] Robert-Ribes, J., Piquemal, M., Schwartz, J., and Escudier, P. (1996). Exploiting sensor fusion and stimuli complementary in av speech recognition. In [Stork and Hennecke, 1996], pages 194–219.

[Schifferdecker, 1994] Schifferdecker, G. (1994). Finding structure in language. Master's thesis, University of Karlsruhe.

[Steifelhagen et al., 1997] Steifelhagen, R., Yang, J., and Meier, U. (1997). Real time lip tracking for lipreading. In *Proceedings of Eurospeech '97.*

[Stork and Hennecke, 1996] Stork, D. and Hennecke, M., editors (1996). *Speechreading by Man and Machine: Models, System, and Applications.* NATO/Springer-Verlag, New York.

[Summerfield, 1987] Summerfield, Q. (1987). *Some preliminaries to a comprehensive account of audio-visual speech perception*, pages 3–52. In [Dodd and Campbell, 1987].

[Vogt, 1996] Vogt, M. (1996). Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In [Stork and Hennecke, 1996], pages 399–407.

[Walden et al., 1977] Walden, B., Prosek, R., Montgomery, A., Scherr, C., and Jones, C. (1977). Effect of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20:130–145.

[Wark et al., 1998] Wark, T., Sridharan, S., and Chandran, V. (1998). An approach to statistical lip modelling for speaker identification via chromatic feature extraction. In *Proceedings of the IEEE International Conference on Pattern Recognition*, pages 123–125.

[Web, 2000] Web, W. W. (2000). *M2VTS Multimodel face database, release 1.0.* World Wide Web, http://www.tele.ucl.ac.be/PROJECTS/M2VTS/.

[Yang and Waibel, 1996] Yang, J. and Waibel, A. (1996). A real-time face tracker. In *Proceedings of WACV'96*, pages 142–147.