Audio-Visual Speech Recognition using Red Exclusion and Neural Networks

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

School of Informatics and Engineering
Flinders University of South Australia,
PO Box 2100, Adelaide, South Australia 5001,
Email: [trent.lewis|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 and a technique developed by the authors is breifly presented. This paper examines examine how useful this extraction technique in combination with several integration arhitectures is at the given task, demonstrates that vision does infact assist speech recognition when used in a linguistically guided fashion, and gives insight remaining issues.

Keywords: Audio-Visual Speech Recogition, Neural Networks, Sensor Fusion

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%, 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. This techniques is then used as the visual basis of AVSR experiments. We first outline the context of this work, AVSR - both human and machine, the feature extraction technique is then described, and finally a series of AVSR experiments using nerual networks are discussed.

2 Background

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 are visible. 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].

[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/

Experiments have found that the human perception of consonants systematically group in the presence of noise [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].

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.

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 [Stiefelhagen 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 amountof 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 and this is the approach taken here.

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]). The approach taken here is somewhere in the middle of this continuum; feature points are specifically chosen although no model is constructed. Section 3 elaborates further on this stage of AVSR.

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. Furthermore, they demonstrate gracefull degradation in the presence of noise

The two most closely followed psychologically derived models of integration are the direct integration (DI) and seperate identification (SI) models. 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 theaccuracy 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.

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. In the following sections, this database has been used to test the algorithms explained. Although only a subset of the phonemes have been used for the recognition experiments (see Table 2).

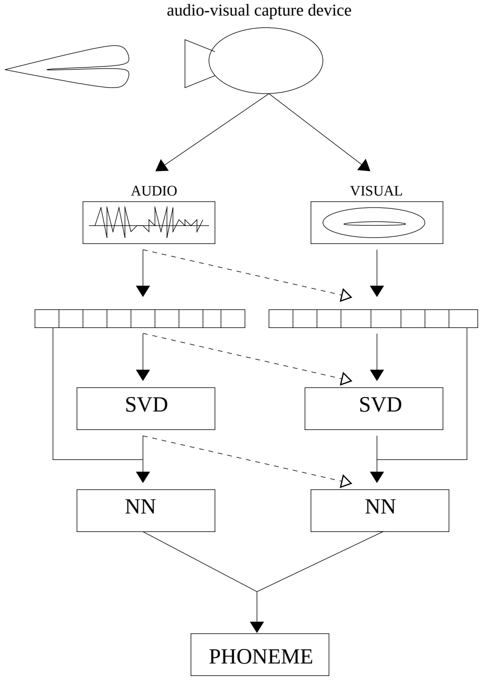

Figure 1: Architecture for AVSR system. A dotted line indicates possible early integration path.

Table 2: Targeted phonemes and words

Targeted	Position		
Phoneme	start	middle	final
/p/	pear/pea	kappa/apple	mop/top
/b/	bear/bag	abba/rabbit	mob/cab
/m/	mare/moon	hammer	tom/ham
/t/	tear/tin	matter/butter	pot/feet
/d/	dare/desk	adder/rudder	pod/bed
/n/	nair/knee	anna/winner	don/bun
/k/	care/kite	hacker/wacky	hock/book
/g/	gair/go	dagger/logging	bog/bag
/ŋ/		banger/singer	bang/song

3 Feature Extraction

3.1 Visual Features

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-

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

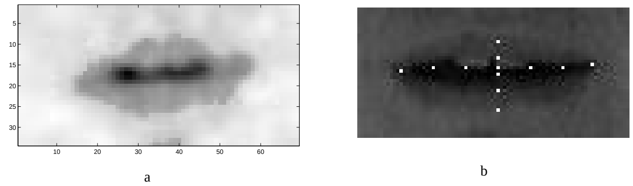

Figure 2: a) Example of red exclusion, and b) Visual features used for recognition.

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].

In a previous paper it was demonstrated that many of the current pixel-based techniques do not adequately identify the lip corners, or even the lip region in some cases [Lewis and Powers, 2000]. 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,

$$log\left(\frac{G}{B}\right) \leq \beta \qquad (1)$$

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. Figure 2a is an example of red exclusion on one of the subjects and 2b is example of visual features used for recognition.

3.2 Acoustic Features

According to Schafer and Rabiner, the choice of the representation of the (acoustic) speech signal is critical [Schafer and Rabiner, 1990]. Many different representations of speech have been developed, including simple waveform codings, time and frequency domain techniques, linear predictive coding, and nonlinear or homomorphic representations. Here, we focus on the homomorphic representations, especially the *mel-cepstrum* representation.

The mel-frequency scale is defined as a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [Davis and Mermelstein, 1990]. This representation is preferred by many in the speech community as it more closely resembles the subjective human perception of sinewave pitch [Brookes, 2000, Rabiner and Juang, 1993]. A compact representation of the phonetically important features of the speech signal can be encoded by a set of mel-cepstrum coefficients, with the cepstral coefficients being the Fourier transform representation of the log magnitude spectrum.

The mel-cepstrum representation of acoustic speech has had great success in all areas of speech processing, including speech recognition. It has been found to be a more robust, reliable feature set for speech recognition than other forms of representation [Davis and Mermelstein, 1990, Rabiner and Juang, 1993]. Thus, it was decided that this was the best representation to be used for the following recognition experiments. Moreover, the cepstrum has been found to be invaluable in identifying the voicing of particular speech segments [Schafer and Rabiner, 1990].

To extract the mel-cepstrum coefficients from the speech signal the Matlab speech processing toolbox *VOICEBOX* was used [Brookes, 2000]. The first 12 cepstral coefficients, 12 delta-cepstral coefficients, 1 log-power and 1 delta log-power [Movellan and Mineiro, 1998]. This is a total of 26 features per acoustic frame, and 130 per data vector (5 frames), which is comparably to the number of visual features.

4 Integration Architectures

This section overviews the three integration architectures tested. The first is a simple early integration technique, whilst the last two are more complicated late integration architetures.

4.1 Early Integration

A very simple approach to early integration has been followed. The acoustic and visual data sets are concatenated together, giving one large input vector from which data transformation and recognition can occur [Hennecke et al., 1996]. This vector is then used as input into a multi-layer perceptron (MLP) with 1 hidden layer. The number of neurons in the hidden layer was equal to the log_2 of the number of input neurons. Supervised training was performed using backpropagation using a mean squared error performance function and a training algorithm known as *resilient* backpropagation. The purpose of resilient backpropagation algorithm is eliminate the potentially harmful effects of the magnitude of the gradient. Basically, it does this by only considering the sign of the derivative to calculate the direction of the weight update. The method is much faster than standard gradient descent and useful for large problems [Demuth and Beale, 1998].

4.2 Late Integration

Many complicated techniques have been developed for integration of acoustic and visual networks, however, an analysis by Meier, Hurst and Duchnowski, found that the best late integration technique was to use a neural network for the integration [Meier et al., 1996, Meier et al., 1999]. A bonus of late integration is that the acoustic and visual data do not have to be in perfect synchrony, because the acoustic and visual subnets effectively act as independent recognisers.

As the subnets are effectively their own recognisers, the training of the late integration network is a little bit more complicated than before and included two phases. The two phases of training and the basic architecture are outlined in figure 3 (ignore part 1b for the momenet).

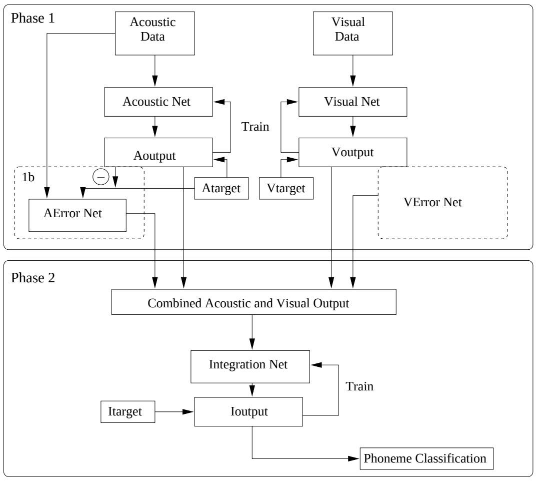

Figure 3: Late Integration with Error Component.

The first phase involves training the acoustic and visual subnets. Once the training of each subnet is completed, the training data is passed through the respective network which produces two outputs - one from each subnet. Phase two of the training uses these outputs by concatenating them together and then this data is used to train the integration network. To test the network, a separate set of acoustic and visual data were passed through the respective subnets. The output from each network was concatenated in the same way as in training and then this data was used to test the integration capabilities of the NN.

Most researchers use the brute force of the algorithm to recognise each phoneme/word, ie each modality attempting to recognise everything. Using late integration, however, one can alter what each subnet is recognising. As would be expected from psycholinguistic research the following were tested: phoneme-phoneme (P-P), phoneme-viseme (P-V), voicing-viseme (Voi-V), where the first is the acoustic subnet and the second is the visual.

4.3 Late Integration with Error

To combat the amount of error that exists in the network, two extra networks have been introduced into the architecture (figure 3 - 1b). The two new networks can be considered as error predicting networks, one for each subnet. The training stage for these NN, part 1b, occurs after the training of the acoustic and visual NNs, but before the integration network. The training data for these networks is the same for which subnet it is predicting the error for. The target pattern is for the error network is,

$$T_E = T_A - O_A, \tag{2}$$

where T_E is the target vector, T_A is the target vector for the acoustic subnet, and O_A is the output of the training acoustic network on the training data. The same is also true for the visual error NN.

The result of equation 2 is in the range [-1,1], thus in order to train the network to produce results in this range a tan sigmoid tranfer function was used on the output layer, rather than the log sigmoid which tranforms data into the range [0,1].

The motivation behind this type of network is to help the integration network decide when an input is useful. Thus, the output of the error NN needs to reflect the usefulness of data. In its present form the output represents a high error as either -1 or 1, and a perfect match with 0. This set up may actually impede the performance of the integration network, thus before the output of the error NN is used for training, it is transformed by,

Table 3: Recognition accuracy (%) of separate acoustic and visual neural networks.

	RAW	NORM	SVD	N/SVD
PHONEME				
Acoustic	11.8	21.2	16.2	20.9
Visual	8.4	11.5	10.9	14.7
VOICING				
Acoustic	54.8	58.4	53.1	53.5
Visual	29.9	29.3	29.5	32.2
VISEME				
Acoustic	42.4	43.3	37.5	42.4
Visual	30.6	54.7	44.1	53.0

$$T_{Etrans} = 1 - |O_E|, \tag{3}$$

which transforms the data such to a perfect classification is ranked as 1 and a high error as 0.

5 Method and Data Preparation

In all the experimental results that follow, a 10-fold procedure, with randomly selected training and testing data for each trial, was adhered to. For each trial the training and testing data were mutually exclusive, however, there was no guarantee of evenly distributed data, even though a uniform random number generator was used.

In addition to the raw data sets, a number of transformations were performed in the hope to improve recognition accuracy. The first transformation was to normalise the data. The normalisation technique chosen here was to scale the data such that it had a zero mean and unity standard deviation [Demuth and Beale, 1998]. SVD was performed on the data and attributes with eignvalues greater than 0.001 were used. We also tested a combination of normalisation and then SVD. Therefore, there were 4 types of data to train each neural network upon - raw, normalised(N), SVD, and N/SVD.

Phoneme, Viseme, or Voicing were the 3 possible classification tasks for a NN to perform. 1) Phoneme classification tasks involved discriminating bewteen the stops /p,b,m,t,d,n,k,g,ŋ/. 2) Viseme classes are defined as labial (/p,b,m/), dental (/t,d,n/), and glottal (/k,g,ŋ/). 3) The voicing task discriminated between unvoiced (/p,t,k/), voiced (/b,d,g/) and nasal stops (/m,n,ŋ/). Thus, the tasks were 9, 3, and 3 item discrimination tasks, respectively.

6 Results

Before presenting the results the reader is reminder that the NN were trained on a very limited set of data; 2 examples of each phoneme/position pair for each of 3 subjects. Furthermore, low-cost OTS equipment was used and each subject was seated 1.5 to 1.8 meters from the recording device. Given this, the results reported below are very promising.

Table 3 shows the overall recognition accuracy of separate acoustic and visual NNs attempting to distinguish between the 9 phonemes, 3 viseme and 3 voicing groupings. It is immeadiately obvious from this table that vision alone is not able to distinguish between the set of 9 phonemes or 3 voicing groups with the accuracies hovering around guessing level (11.1%, and 33.3%, respectively). According to the psycholinguistic work reviewed (e.g. [Dodd and Campbell, 1987]) this is to be expected.

Table 4: Recognition accuracy (%) of early and late integration architectures.

	NORM	N/SVD
PHONEME		
Early	17.0	20.1
Late, P-P	12.1	13.3
Late, P-V	13.9	15.8
Late, Voi-V	**29.0**	24.1
Late/E	19.5	13.2

Significantly, the accuracy of the acoustic network is above this rate. Interestingly, the visual network, as predicted, outperforms the acoustic net on the viseme recognition task. This is very promising for the next stage of integration and indicates that vision alone can differentiate between certain traditional linguistic sound segments.

Another interesting observation from these preliminary investigations is that normalisation of the data greatly increases the accuracy of the network, especially in the case of vision. Thus, in subsequent experiments only normalised or normalised/SVD data was used in testing and training.

Table 4 outlines the results for all of the of integration architectures mentioned. The results gained from the majority of the integration architectures were not quite as good as hoped - and indeed have demonstrated catastrophic fusion. Early, Late P-P, Late P-V, and Late/E all had recall accuracies below the acoustic only NN, which had an accuracy of 21.2% for normalised data. However, the late integration using voicing and viseme subnets an almost 40% increase in accuracy. This clearly demonstrates that the psycholinguistically guided integration architecture can perform better than a stand alone acoustic recogniser when there is a severly degraded signal in both the acoustic and visual modalities.

7 Discussion

This paper, and the research associated with it, has demonstrated the utility of AVSR in an everyday environment using low cost webcams. The following discussion overviews the contributions of this paper and highlights areas of current and possible future research.

7.1 Red Exclusion

Red exclusion, the mouth feature extraction technique described in this paper, was developed because other commonly used techniques did not perform well on the database collected [Lewis and Powers, 2000]. This paper has demonstrated that red exclusion is a viable technique for the extraction of mouth features by its incorporation into this experimental AVSR system with some moderate success.

Investigation into red exclusion has opened up some interesting avenues of research. The spectral reflectance of human skin creates a characteristic “W” shape, with minimums at 546nm and 575nm and the local maximum (middle of the w) at around 560nm [Angelopoulou et al., 2001]. Interestingly, this maximum is also the maximal response of the long wavelength cones of the human retina. Current research is looking at why the relationship might exist and how this can be used to refine the red exclusion technique. It is hypothesized that the red exclusive effect is related to the colour opponent properties of mammalian vision.

7.2 Integration

There could be several factors contributing to the unsatisfactory performance of the early integration network. Firstly, due to the selection procedure the acoustic and visual inputs are not perfectly synchronised. Thus, it makes it difficult for the NN to learn the relative timing between the two concatenated inputs [Hennecke et al., 1996]. This can impede the detection of the voicing of the phoneme, and indeed the acoustic only NN outperformed the early integration network in identifying the voicing, 58.4% versus 52.6%. Furthermore, the NN must also learn the proper weigthing between the acoustic and visual data depending on the noise level. To be effective at this it must be trained at all noise levels likely to occur, thus increasing the required training set size. Therefore, another reason for the poor performance is that because of the small training set, the early integration NN was unable to learn the correct weightings. Another explanation for the failure of integration, and one that is a fundamental problem of NNs, is that NNs are basically linear and produce a kind of weighted average that is inappropriate in the event of competition.

This late integration technique, P-P, could be considered a “no-holds-barred” approach to AVSR, and also a little naive. With enough training the P-P network maybe be able to correctly identify phonemes by being able to correctly weight connections when noise is present. However, even for humans it is very difficult to tell the difference between a /p/ and /b/ when using visual information only. This is because they belong to the same viseme grouping, such that it would be more sensible, and linguistically correct to use the visual data to extract information about visemes, rather than phonemes. This was attempted in the P-V network, yet under these conditions the accuracy was only slightly better and still below that of acoustic only. Thus, following linguistic intuition, the Voi-V late integration network was used was good success.

Even though the Late/E had poor recall accuracy it is still an interesting approach and warrants further investigation with a larger training base. A reason why this network performed badly with respect to the other networks could be to do with the training regime employed. In this case, the error analysis network was trained with the output of the training data. Thus, the subnets were attune to this data and many of the outputs were near perfect. Thus, when unseen data was used the error network may not have acted correctly. A solution to this problem, if enough data is available, is to use a validation set for the error network training. Therefore, the error network will be trained on previously unseen data. This idea could also extend to the integration network of all late integration architectures. So, with a larger training base the gamut of training regime could be explored to find the most efficient and effective method.

7.3 Conclusion

This research has shown that multi-speaker AVSR is useful in a natural office environment where the user is not equiped with specialised eqiupment, eg close head microphone, minimal external noise, etc. Via red exclusion, a visual signal can be integrated into recognition phase to help combat increasing acoustic noise and increasing the accuracy of recognition. Using a knowledge from psycholinguistics, a late integration network was developed that fused the acoutic and visual sources and increased the accuracy by around 40% over an acoustic only NN. AVSR is a flourshing area of research with many avenues still open to investigation, especially in the area of sensor fusion. Current research is aiming to develop a conventional ASR system, using a larger database, that is stable with a distant microphone setup and examine the effect of moving to AVSR with this system.

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.

[Angelopoulou et al., 2001] Angelopoulou, E., Molana, R., and Daniilidis, K. (2001). Multispectral color modeling. Technical Report MS-CIS-01-22, University of Pennsylvania, CIS.

[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.

[Brookes, 2000] Brookes, M. (2000). *VOICEBOX: Speech Processing Toolbox for MATLAB.* World Wide Web, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html.

[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.

[Davis and Mermelstein, 1990] Davis, S. and Mermelstein, P. (1990). Comparision of parametric representations for monosyllabic word recognition in continuously spoken sentences. In Waibel, A. and Lee, K., editors, *Readings in Speech Recognition*, pages 64–74. Morgan Kaufmann Publishers Inc., San Mateo, CA.

[Demuth and Beale, 1998] Demuth, H. and Beale, M. (1998). *Neural Network Toolbox: User's Guide.* The MathWorks, http://www.mathworks.com.

[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.

[Lewis and Powers, 2000] Lewis, T. and Powers, D. (2000). *Lip Feature Extration Using Red Exclusion.* World Wide Web, www.cs.usyd.edu.au/∼vip2000.

[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.

[Rabiner and Juang, 1993] Rabiner, L. and Juang, B. (1993). *Fundamentals of Speech Recognition.* Prentice Hall, Englewood Cliffs, NJ.

[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.

[Schafer and Rabiner, 1990] Schafer, R. and Rabiner, L. (1990). Digital representations of speech signals. In Waibel, A. and Lee, K., editors, *Readings in Speech Recognition*, pages 49–64. Morgan Kaufmann Publishers Inc., San Mateo, CA.

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

[Stiefelhagen et al., 1997] Stiefelhagen, 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].

[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.

[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.