

# Audio-Visual Speech Recognition using Red Exclusion and Neural Networks

Trent W. Lewis and 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

*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 promising new technique is presented. This paper examines how useful this extraction technique in combination with several integration architectures and compares it with competing techniques, demonstrates that vision does in fact assist speech recognition when used in a linguistically guided fashion, and gives insight into remaining issues.*

*Keywords: Audio-Visual Speech Recognition, Feature Extraction, Neural Networks, Sensor Fusion*

## 1. INTRODUCTION

Automatic speech recognition (ASR) performs well under restricted conditions with word accuracy rates up to 98–99%. When we step outside the boundaries, however, performance can be severely degraded and the utility of such systems comes under fire (Bregler *et al*, 1993). The question then arises of how are humans able to recognise speech in unfavourable conditions such as a busy office, a train station or a construction site? Is our acoustic apparatus performing an enormous amount of noise filtering and reduction or is it that we are using another source of information? It is in fact the latter which may be an answer to robust speech recognition.

Work from the areas of psychology and linguistics has shed much light on how humans perceive speech, not only how we perceive it acoustically but also visually, such as lip-reading in deaf people. This has evolved into what is now known as *speechreading* (Dodd and Campbell, 1987). The most important finding from this research is that normally hearing people do rely on vision for speech perception and that the set of visually perceivable speech sounds forms a complementary set to that of the acoustically perceivable sounds in the presence of noise. This set of visually perceivable speech sounds have been named *visemes*, that is *visual phonemes* (Summerfield, 1987).

Researchers in the fields of engineering and computer science have taken these ideas and applied them to traditional acoustic speech recognition systems to produce audio-visual speech recognition

---

*Copyright© 2003, Australian Computer Society Inc. General permission to republish, but not for profit, all or part of this material is granted, provided that the JRPIT copyright notice is given and that reference is made to the publication, to its date of issue, and to the fact that reprinting privileges were granted by permission of the Australian Computer Society Inc.*

*Manuscript received: 1 May 2002*  
Communicating Editor: Hugh Williams

(AVSR) systems with very encouraging results (for a comprehensive review see Stork and Hennecke, 1996). Although only minimal improvement is found under optimal conditions, improvements using a degraded acoustic signal have been large (Hennecke, Stork and Prasad, 1996). For example, Meier *et al* (1999) have reported up to a 50% error reduction when vision is incorporated.

AVSR requires expertise in interpreting evidence from psycholinguistics, a solid grounding in the traditional acoustic speech recognition, and a grasp on the computer vision techniques for relevant visual feature extraction. However, a new problem also arises with AVSR and that is how to best combine the acoustic and visual signals without the result being worse than acoustic or visual recognition alone, that is *catastrophic fusion* (Movellan and Mineiro, 1998). This is a lively research area in AVSR and the effectiveness of different techniques, such as early, intermediate and late fusion, are still being decided.

Although our primary interest is in the fusion of acoustic and visual data there are considerable problems to be overcome in other aspects of the system as well, notably in visual feature extraction. Unlike acoustic feature extraction, which is a mature field, visual feature extraction suitable for lip reading is still in its infancy and current techniques for lip feature detection prove to be inadequate or marginal.

In this paper we will initially look at the psycholinguistics in Section 2 and give an overview of AVSR in Section 3. We go into the effect of environmental conditions in Section 4, being particularly interested in AVSR under adverse conditions using cheap off-the-shelf (OTS) hardware. Issues of face and feature extraction are described in Section 5 and the new techniques we have developed for this are explored. The final sections report on AVSR fusion experiments using artificial neural networks (ANN) for the recognition process in both the acoustic and visual signals as well as the fusion process.

## 2. 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.

Probably the most cited article in AVSR literature is “Hearing lips and seeing voices” by McGurk and MacDonald (1976). This paper reported on an effect that definitively demonstrated the influence of vision on speech perception that later became commonly known as the “McGurk effect”. This effect, which is easily replicated, occurs when a person is confronted with, for example, the utterance [ba] and the lip movements [ga], and the person perceives the sound [da]. It is stated that the reason for this is because the listener has combined or fused the two sensory inputs into one, thus altering the perception and demonstrating that visual input is heavily influencing the processing of speech. This effect has been replicated many times and has even been extended to entire sentences (Massaro and Stork, 1998).

Research in this area has also uncovered that the addition of visual cues can enhance normal, human listeners’ accuracy and speed in speech perception. For example, using nine normal-hearing subjects, Grant and Seitz (1998) found that the intelligibility threshold of a sentence with respect to noise was greatest for auditory-visually matched sequence when compared to auditory only and auditory-visually unmatched sequences (surprisingly, visually unmatched sequences did not degrade performance relative to auditory only). Further studies by Grant and Seitz have found that the speed of spoken word recognition is superior for auditory-visual conditions over auditory or visual alone.

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	tongue sides	/l/
RRR	tongue blade	/r/
PAL	palatal veolars	/sh,zh/
WWW	lips/tongue back	/w/

Table 1: Consonant viseme classes

These studies demonstrate that even normal hearing people benefit greatly from the addition of a visual signal.

One reason humans benefit from a visual signal may be that our various speech articulators are visible. Lips, teeth, and tongue have been identified as the primary indicators for visual speech, however, the cheeks, chin and nose are also very useful as secondary indicators (Robert-Ribes *et al*, 1996). To an extent the entire facial expression is used, hence *speechreading*.

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 human perception of consonant phonemes 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 nine distinct, humanly perceivable viseme classes, as well as their common place of articulations (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).

### 3. 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 *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 mention the general aspects of the main techniques.

### 3.1 Face and Mouth Region Extraction

There are some AVSR systems that process 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 is 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 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 (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.

### 3.2 Visual Feature Extraction

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 (Movellan and Mineiro, 1998; Meier *et al*, 1999), whilst others insist that a high-level, model approach is the most efficient way to proceed (Hennecke *et al*, 1996; 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 5 elaborates further on this stage of AVSR.

### 3.3 Acoustic and Visual Integration

A researcher's answers to how to integrate and what learning algorithm to use 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 techniques from traditional ASR can be used, with modifications, for the 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 (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 ANN and the Hidden Markov Model (HMM; Hennecke *et al*, 1996). HMMs 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 (Charniak, 1993). 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. ANNs, on the other hand, are often criticised for their slow trainability and variance due to rate. However, they do have the empowering ability to generalise to unseen data, given large

enough training sets, and, moreover, they do not make any assumptions about the underlying data. Furthermore, they demonstrate graceful degradation in the presence of noise.

The two most closely followed psychologically derived models of integration are the direct integration (DI) and separate identification (SI) models. In the DI model, feature vectors of the acoustic and visual signals can be simply concatenated together, and then this vector can be used as input into the HMM (Adjoudani and Benoit, 1996) or ANN (Meier *et al*, 1999). When following the DI model sensor integration occurs automatically and it is up to the recognition engine to decide upon the important features. This is the default approach if using ASR already.

Under the more sophisticated SI model integration becomes somewhat trickier. The simplest case is when the outputs of separate ANNs are fed into another ANN 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 combine them in such a way to maximise their cross-product. Late integration (ie. SI) is an evolving area in AVSR and is a difficult issue to contend with because fusing the two signals can lead to *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 ANNs 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).

Which is better – early or late? This is a question still debated within the literature. On theoretical grounds and the necessity of maintaining temporal relationships between the signals, many argue for early integration (Bregler *et al*, 1996; Basu and Ho, 1999). For example, Hennecke *et al* (1996) state that late integration is just a special case of early integration and given the right conditions “... a system that uses early integration should perform at least as well as one that integrates at a later stage. (p. 338, Hennecke *et al*, 1996).” Indeed, if an inadequate set of sensor specific features are used, essential information can be thrown away in late integration. Comparative empirical studies, however, have found that late integration techniques are performing better than early integration even with the loss of synchronisation (Adjoudani and Benoit, 1996; Meier *et al*, 1999). The survey that follows is mainly made up of research involving variants of late integration as this technique has many more issues to overcome. For completeness, work on early integration is also mentioned for comparison.

Potamiaonos and Potamianos (1999) use a multi-stream HMM in which the visual stream is just another parameter to the HMM. The emission probability of the HMM is equal to the product of the sum distributions of each stream. These sum of distributions are augmented by a *stream exponent*  $\lambda$ . This exponent models the *reliability* of each stream and satisfies,

$$0 \leq \lambda_A, \lambda_V \leq 1, \quad \text{and} \quad \lambda_A + \lambda_V = 1 \quad (1)$$

The stream exponents are estimated using a generalised probabilistic descent algorithm. This appears to occur initially during training, but it is unclear as to whether the exponents are dynamically estimated during recognition. Thus in this system the late integration is taking place via a weighted product of the contributions from the acoustic and visual channels. This is probably the most common approach to sensor integration in this field and demonstrates that the AV system is superior to the acoustic or visual alone. Although the word accuracy by this system is high (90.5% for AV) the weights on each stream are determined a priori to test time (i.e. on the training set) and thus if the conditions change enough the weightings might not correctly reflect the reliability of each the signals.

Neti *et al* (2001) and Glotin *et al* (2001) have produced comparative studies of early, late with constant weighting, and late with *dynamic* weighting audio-visual integration schemes. The dynamic technique was based on the degree of voicing present in the audio stream average over the entire utterance such that  $0 \leq \lambda_A = \text{degree of voicing} \leq 1$  and  $\lambda_V = 1 - \lambda_A$ . Overall, the integration system using the dynamic weights outperformed all others on a word recognition task in both clean and noisy acoustic conditions. Interestingly, in clean acoustic conditions some of the late integration techniques were outperformed by the early integration and in some cases even demonstrated catastrophic integration.

Dynamically setting the weights based on the current utterance is a preferred method of integration. This utterance based method, however, is somewhat lacking in its ability to generalise to other situations. For example, if there was a loud, brief sound in the background this might affect the overall average for the utterance and hence distort the weighting considerably. Calculating the median instead of a mean might correct the weights for the majority of the speech segment, but then at extra noisy sections performance would degrade. Dynamically determining the weights needs to occur at a lower level. Moreover, waiting until the end of the utterance to determine weights means that integration can only take place after the entire utterance has been spoken.

Dupont and Leuttin (2000) tackle the problem of *continuous* speech recognition. In continuous speech recognition the system must deal with co-articulation and the fact that the utterance has no predetermined length. They claim that because of these factors waiting until the end of utterance to fuse is too time consuming for late integration architectures and that integration should occur during the utterance. Moreover a list of N-best hypotheses must be kept for each state until integration occurs. Their speech recognition system consists of a multi-stream HMM with NN as HMM state probability estimators. This system uses *anchor points* to denote where individual streams must synchronise (fuse). These anchors may occur on relevant phonological transition points, such as phonemes, syllable or words. Dupont and Leuttin (2000) only test anchor points at the HMM state and word level. Integration is a weighted product of the segment likelihoods. These weights are determined by automatically estimating the acoustic SNR, such that the higher the SNR, the higher the weight to the acoustic information. They make mention that with a clean signal the addition of visual information did not increase accuracy. However, with a clean signal (high SNR) the weight was very high and it might be that the visual system does not have the ability to influence the result given this weighting. Early integration (concatenation) yields inferior results compared to the different late integration techniques. The most successful late integration was with combination at the word level and including phoneme duration models into the HMM further increased the accuracy. They mention that there is considerable temporal asynchrony between the acoustic and visual modalities and that this asynchrony is not stable.

In their work, Adjoudani and Benoit (1996) strive for  $AV > A$  and  $AV > V$  over all testing conditions and explore several progressive model of integration. The first, an early integration method, fails in acoustically noisy conditions because it is dragged down by the inability of the system to capture the contribution of the visual parameters. The first late integration technique is a simple maximisation of the product of the resulting probabilities across each output channel. In high SNR conditions the system is able to take advantage of the complimentary information between the signals with AV outperforming both subsystems. In poor acoustic conditions, however, the system is once again not able to correctly attribute each subsystem.

To overcome the inadequacy of the combination so far, Adjoudani and Benoit (1996) introduced a *certainty factor* to differentially weight each subsystem. This weighting factor differs from previously discussed architectures as it is *not* solely based upon the level of acoustic noise within

the signal. Rather, it is based upon the *dispersion* of the N-best hypotheses in each modality, that is, large differences in probabilities equates to greater certainty, close probabilities to less certainty. The first application of the certainty factor was a binary selection of either the acoustic or visual hypothesis based on which had the greatest certainty. This method satisfy the original criteria set by Adjoudani and Benoit (1996), however it can only ever choose between the votes of the individual subsystems because of its binary nature. A weighted product version of the late integration system based on a normalised dispersion certainty factor combined the acoustic and visual system in a synergistically over all noise levels and can choose a different class from either subsystem.

The *dispersion* idea used by Adjoudani and Benoit (1996) has been implemented by other researchers in various forms (Meier *et al*, 1999; Potamianos and Neti, 2000; Heckmann *et al*, 2001). Using Gaussian mixture model (GMM) to classify phonemes, Potamianos and Neti (2000) use an N-best dispersion method that is framed as the difference between each pair of  $n^{th}$ -best hypotheses, given by,

$$\frac{2}{N(N-1)} \sum_{n=1}^N \sum_{n'=n+1}^N (R_n - R_{n'}), \quad (2)$$

where  $N \geq 2$  and  $R_n$  is equal to the  $n^{th}$  best hypothesis. Interestingly, both Adjoudani and Benoit (1996) and Potamianos and Neti (2000) have found that an N-best of four has been the most successful. Potamianos and Neti (2000) also use a method called N-best likelihood ratio average in which the difference is only calculated against the best hypothesis, that is,

$$\frac{1}{N-1} \sum_{n=2}^N (R_1 - R_n) \quad (3)$$

The best performing system here was the one using dispersion as a confidence measure with a phoneme accuracy of 55.19%. The ratio average achieved an accuracy of 55.05%. Both of these methods were significantly better than the baseline acoustic only system. Another confidence method based on the negative entropy of the stream was unable to achieve accuracy significantly better than the baseline.

Basu and Ho (1999) also used GMMs for recognition but only looked at early integration. In comparison to Potamianos and Neti (2000), the accuracy of the system on the test data was consistently below 50%. Moreover, the combined feature vector provides little increase in accuracy. The value of this research, however, is that they also test the system on a *real-life* data set. That is, one that is not collected in a controlled environment and without specialised equipment. The performance on this data set drops dramatically with 33% for acoustic only and 9% for visual only. This clearly demonstrates that moving out of the experimental environment can severely affect even the “state-of-the-art” systems.

Heckmann *et al* (2001) use a hybrid ANN/HMM AVSR system with the NNs providing the *a posteriori* probabilities for the HMM which provide the phone and word models (language models). Heckmann *et al* (2001) argue for and use a late integration method and use a weighting method they call *Geometric Weighting*. Detecting the most probable phoneme is found by a conditional probability that is augmented by the geometric weights. The value of weight based on another value  $c$  and they want  $c$  to reflect an estimate of the SNR of the acoustic signal. To achieve this they use a similar idea as dispersion by exploiting the distribution of the *a posteriori* probabilities at the output of the MLP, but based on the calculated entropy,

$$H = -\frac{1}{K} \sum_{k=1}^K \sum_{n=1}^N \hat{P}(H_{n,k} | \mathbf{x}_{A,k}) \log_2 \hat{P}(H_{n,k} | \mathbf{x}_{A,k}), \quad (4)$$

where  $N$  is the number of phonemes and  $K$  is the number of frames. They created a mapping between  $c$  and  $H$  through an empirical analysis of the values (optimisation process). Results (WER%) show a synergistic gain using this technique down to -6dB (high noise level) where it starts to perform worse than the visual. The automatic weighting performs similarly to manually setting  $c$ . They have also compared using entropy for setting  $c$  to using a Voicing Index and Dispersion methods, however, the entropy based  $c$  still gave the best results (Heckmann *et al*, 2001).

Using a Multiple State-Time Delayed NN (MS-TDNN), Meier *et al* (1999) use the flexibility of the NN to employ several different integration methods for AVSR. They look at both the traditional early and late integration but also integration on the hidden layer of the NN. The early integration technique included the standard concatenation and also the inclusion of an estimated SNR for the acoustic data. Late integration is explored in two different architectures. The first is a weighted sum of the acoustic and visual systems. The weight was determined either by a piecewise-linear mapping to the SNR of the acoustic signal or by what they called “entropy weights”. The calculation of entropy weights was not fully described in this paper (or previously for that matter, e.g. Meier *et al*, 1996), however, their description of the purpose of the weights, High Entropy = Even Spread = High Ambiguity = Low Accuracy, is reminiscent of the dispersion concept from Adjoudani and Benoit (1996). The entropy weights were further augmented by a bias  $b$  that “... pre-skews the weights to favour one of the modalities (p. 4, Meier *et al*, 1999)”. This  $b$  value was hand set to reflect quality of the acoustic data.

A more interesting and novel technique introduced by Meier *et al* (1999) is the learning of the weights. They used another NN to combine both the acoustic and visual hypotheses with the output being the combined phoneme hypothesis. Theoretically, this technique should be able to at least match the performance of the other late integration techniques as it can not only compute pair-wise comparison but also potentially make comparisons across the phoneme and viseme sets, thus taking advantage of the complementary information contained within the signal better than the simple weighted summation. In fact, best performance was with NN weight learning (except in high noise conditions). As would be expected from the bias  $b$ , entropy and SNR weighting performed similarly throughout. Early and hidden layer integration combinations were, as others have found, poorer in performance.

Movellan and Mineiro (1998) compare standard Bayesian integration technique (sum of log likelihoods) with what they call a robustified approach. They argue that most integration system suffer from catastrophic integration because they make implicit assumptions and degenerate quickly when those assumptions are broken and used outside its original context. The robustified approach makes these assumptions explicit by including extra parameters that represent the non-stationary properties of the environment. These parameters make up what is dubbed the *context model*. This approach works by not only maximising the probability with respect to the word but also to each context model, acoustic and visual. Movellan and Mineiro (1998) prove analytically that their approach is superior to the traditional as when the measurements yield data far from the model the traditional integration system is heavily influenced by this subsystem. In contrast, the robustified approach, limits the influence of signals far from a contextual model. Applied to AVSR using a HMM this technique outperforms the classical in acoustic noise as well as with visual noise, an area not investigated by many researchers. In situations where normal integration exhibits catastrophic integration, the robustified integration is no worse than acoustic or visual subsystems.



Not all of the research conducted follows the rigid late integration architecture of weighted sum/product of hypotheses. For example, Verma *et al* (1999) investigated audio-visual phone recognition using Gaussian mixture models with their second and third late integration techniques being somewhat out of the ordinary. They look at three models of late integration: 1) simple weighted sum, 2) weighted sum but V identifying only viseme and using an associated probability of the phoneme given the viseme, and 3) use both A and V to predict viseme (weighted sum, phase 1), then based on viseme class predict which phoneme class (weighted sum, phase 2). The sum of the weights was equal to 1 and was again adjusted manually. The recognition accuracies of the GMMs were well below that of systems combined with HMM. The third integration technique (multi-phase) performed the best. However, this technique is not the most intuitive and a prime example of a system developed without linguistic knowledge. The very characteristic that is masked by noise in acoustic speech is the one that distinguishes the viseme classes (eg. /b/ from /d/, place of articulation), so that using hypotheses derived from the acoustic data in phase 1 could be more of a hindrance (although this isn't what is found in their experiments). Then in phase 2 they use V to distinguish within viseme classes! This is again very counterintuitive, given the definition of a viseme.

A more logical approach to integration is presented by Rogozan (1999). This approach is interesting as it uses both early and late integration in the one system. First a HMM based system produces a hypothesis based on a combined bimodal observation (early integration). Then based on the N-best phoneme hypotheses, another system (a HMM or a NN) refines the result using the visual observations. The results of these two systems are then fused (late integration) using a reliability measure based on the dispersion of accuracy of the N-best. In this work the visual processing is used in the late integration to perform visual discrimination to remove any ambiguity of the hypothesis derived from the acoustics signal. This system is much more linguistically sound than the multi-phase system of Verma *et al* (1999).

#### 4. 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 varying amounts 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 5. 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<sup>1</sup>, 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 using 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 data for visual processing and can improve and speed up training when using ANNs. The feature vectors are then passed to a classifier, in this case an ANN, where the phoneme (viseme) is identified. At this point this system differs from

<sup>1</sup> In this case, a Philips VestaPro (PCVC680K) recording at 44.1kHz, 16bit audio and 352x288, 20fps, 24bit video.

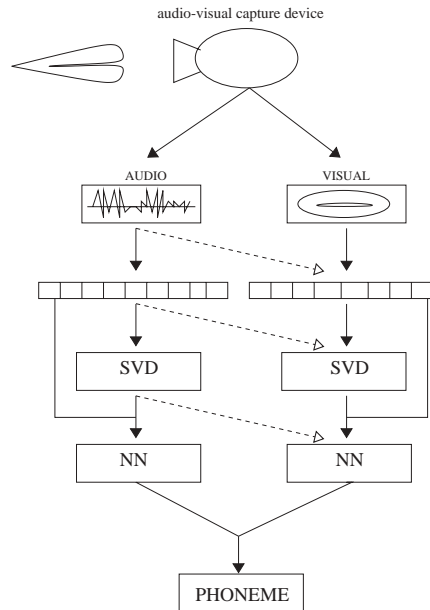


Figure 1: Architecture for AVSR system. A dotted line indicates possible early integration path.

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 possibly 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, datasets that do exist are usually recorded using professional quality equipment and studios whilst our aim is to use low-cost, OTS equipment (M2VTS, 2000; Movellan, 1995). Our data set consisted

Targeted Phoneme	Position		
	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

Table 2: Targeted phonemes and words

of AV recordings of spoken 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, two males and one female, that varied greatly in appearance. In the following sections, this database has been used to test the algorithms explained. However only a subset of the phonemes have been used for the recognition experiments (see Table 2).

## 5. FEATURE EXTRACTION

### 5.1 Visual Features

The accurate extraction of lip features for recognition is a fundamental 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. Broadly speaking there exist two different schools of thought for visual processing (Bregler *et al*, 1993). At one extreme, there are those who believe that the feature extraction stage should reduce 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 processing and reducing the variability and increasing generalisability. However, 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 reduce the amount of information lost due to any transformation. For example, 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 become a problem depending on the choice of recognition system and training regime, however, successful systems have been developed using both HMMs and ANNs using this approach (Movellan and Mineiro, 1998; Meier *et al*, 1999).

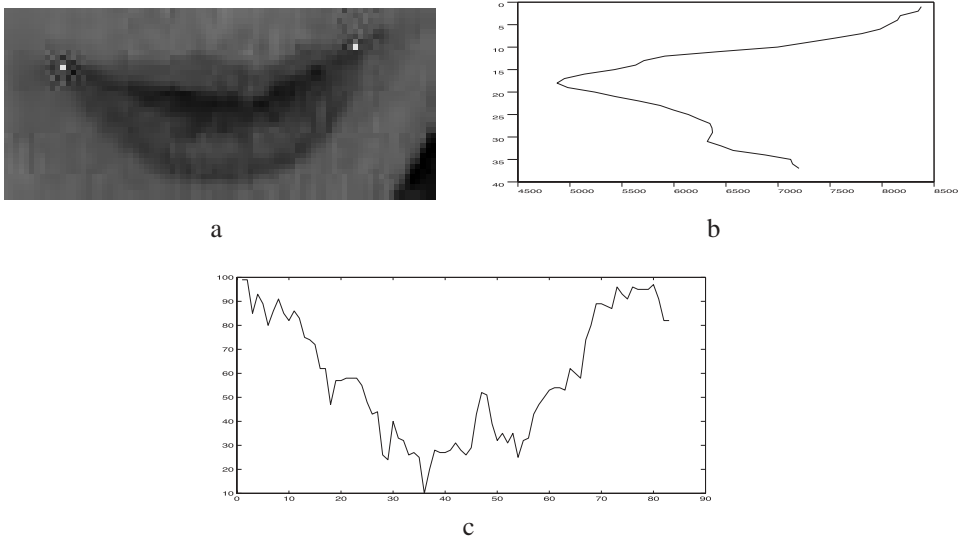
Of course there are many systems that lie between the two extremes, and the model extrema can also benefit from better feature extraction methods as this is the first step of many models. We will now examine some of the more popular methods for initial feature extraction and how well they work for the subjects in our data set. The first feature set that is usually extracted from the mouth area is the lip corner pair. For 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.

#### 5.1.1. Grayscale

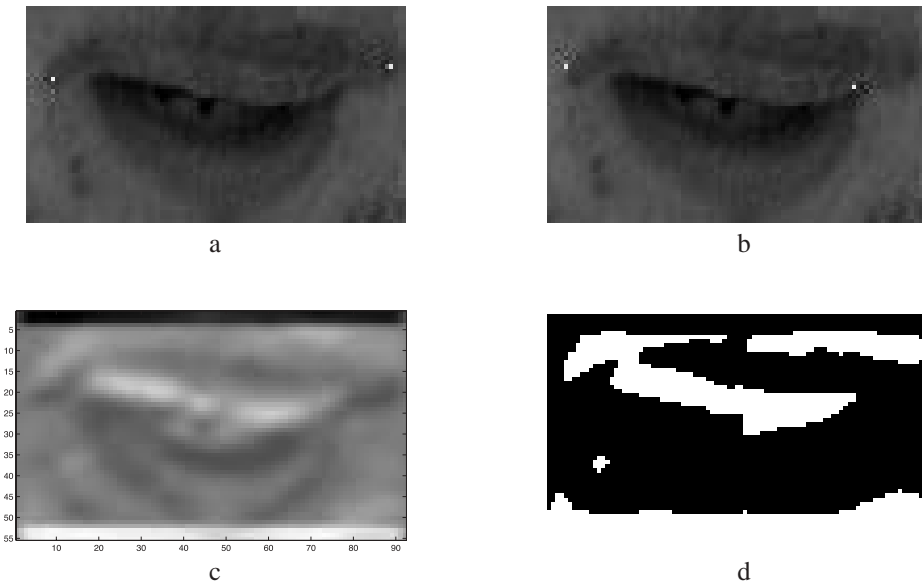
One of the most common methods for feature extraction of mouth features is the use of the gray-scale value and edge detection (Rao and Mersereau, 1994; Stiefelhagen *et al*, 1997). The initial step, as with many of these techniques, is the identification of the vertical position of the centre of the 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 actual 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.

#### 5.1.2. Horizontal Edges

Another common method that makes use of gray-scale values, that has been more successful, is the use of horizontal edges (Stiefelhagen *et al*, 1997). The rationale behind this idea is that the mouth



**Figure 2: Lip corner extraction using gray-scale values for Subject 1.**  
 a) found lip corners, b) gray-scale row sum, and c) gray-scale value of minimum row sum.



**Figure 3: Lip corner extraction using gray-scale values and edges for Subject 2.**  
 a) “found” lip corners – gray-scale, b) “found” lip corners – edges, c) horizontal edge magnitude, and d) threshold edge image (> 10).

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 vertical 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, unfortunately, this also results in shrinkage of the detected lip region.

### 5.1.3. 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 (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 *et al* (1998) used Equation (5) to identify candidate lip pixels.

$$L_{lim} \leq \frac{R}{G} \leq U_{lim}, \quad (5)$$

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 (5), 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, there 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.

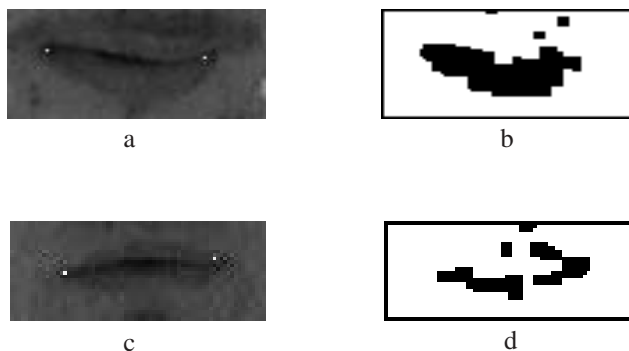


Figure 4: Lip corner extraction using the red-green filter (Equation 5).  
a) “found” lip corners – Subject 1, b) threshold image – Subject 2, c) “found” lip corners – Subject 3,  
and d) threshold image – Subject 3.

### 5.1.4. Hue, Saturation, and Value

The hue, saturation, and value (HSV) colour space can also be 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* (1996) and 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 but not Vogt's<sup>2</sup>. 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,

$$f(h) = \begin{cases} 1 - \frac{(h-h_0)^2}{w^2} & , \quad |h - h_0| \leq w \\ 0 & , \quad otherwise \end{cases} \quad (6)$$

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* (1996) and 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 (Figures 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 (6) to an image. When the hue transform is thresholded at an optimal value, and the result layered over the top of the gray-scale 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.

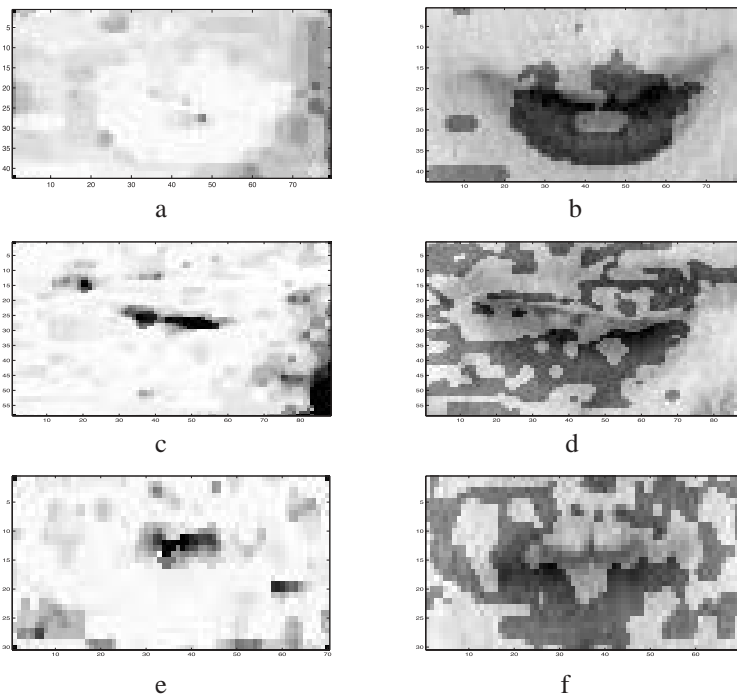


Figure 5: Hue transform (Equation 6) and enhanced gray-scale image. a,b) Subject 1, c,d) Subject 2, and d,e) Subject 3

<sup>2</sup> The major difference between the two algorithms is that Vogt (1996) includes saturation in calculating the likelihood.

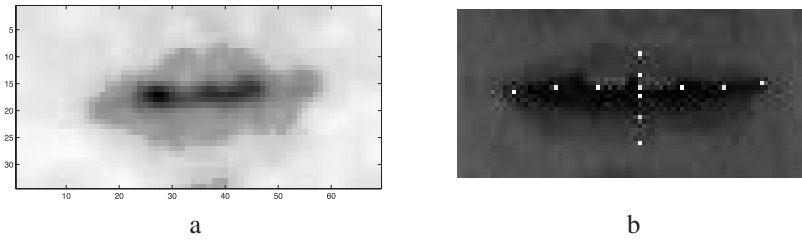


Figure 6: Example of red exclusion (Equation 7) for Subject 3, a) enhance gray-scale and b) visual features used for recognition.

### 5.2 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,

$$\log \left( \frac{G}{B} \right) \leq \beta \tag{7}$$

Using the log scale further enhances the contrast between distinctive areas, and by varying the threshold  $\beta$  the mouth area and the lip features can easily be identified on all three different subjects. Figure 6a is an example of red exclusion on one of the subjects and 6b is an example of the visual features used for recognition.

Using the red exclusion method over a sequence of images to identify the lip corners resulted in near perfect results, as in Figure 7. 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. It is important to note that this method works consistently well over all subjects tested to date, whilst the published algorithms tested did not. The previous analyses are, however, subjective in their interpretation of how they highlight features of the lips. The experiments reported in this paper provide an objective evaluation (see Section 8).

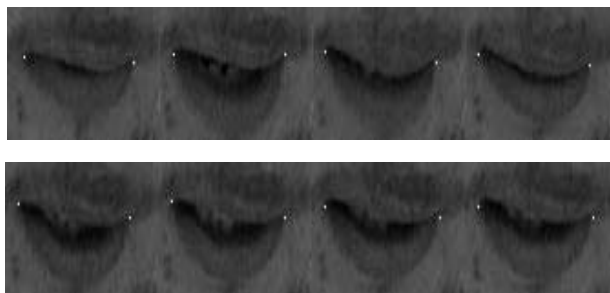


Figure 7: Tracking of lip corners for subject 2 using red exclusion.

### 5.3. Acoustic Features

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).

As per Movellan and Mineiro (1998), the first 12 cepstral coefficients, 12 delta-cepstral coefficients, 1 log-power and 1 delta log-power were extracted from the speech signal. This extraction was performed in Matlab using the speech processing toolbox VOICEBOX (Brookes, 2000) and a final data vector of 130 features (26 features per acoustic frame by 5 frames), which is comparable to the number of visual features, was used for the following experiments.

## 6. 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 architectures.

### 6.1 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 one 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 to 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 converges much faster than standard gradient descent and is useful for large problems (Demuth and Beale, 1998).

### 6.2 Late Integration

Many complicated techniques have been developed for integration of acoustic and visual networks (Section 3.3), however, an analysis by Meier, Hurst and Duchnowski (1996), found that the best late integration technique was to use a neural network for the integration (Meier *et al*, 1996;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.



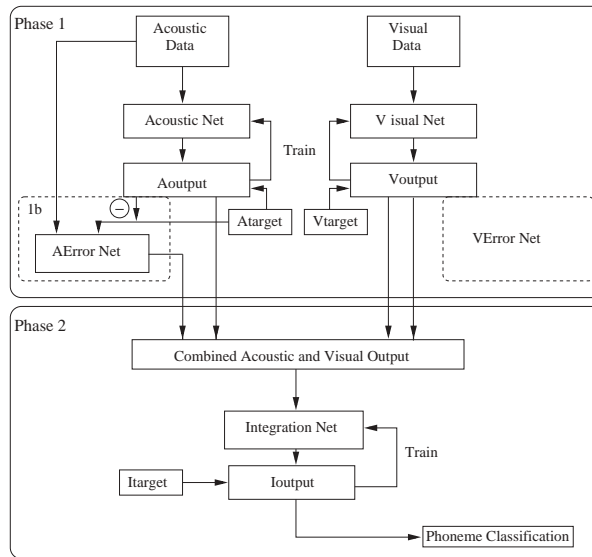


Figure 8: Late Integration with Error Component

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 8 (ignore part 1b for the moment).

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 ANN.

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.

### 6.3 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 8 – 1b). The two new networks can be considered as error predicting networks, one for each subnet. The training stage for these ANN, part 1b, occurs after the training of the acoustic and visual ANNs, 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 for the error network is,

$$T_E = T_A - O_A, \tag{8}$$

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 ANN.

The result of Equation 8 is in the range  $[-1,1]$ , thus in order to train the network to produce results in this range a tan sigmoid transfer function was used on the output layer, rather than the log sigmoid which transforms 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 ANN 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 ANN is used for training, it is transformed by,

$$T_{Etrans} = 1 - |O_E|, \quad (9)$$

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

## 7. METHOD AND DATA PREPARATION

### 7.1 Filter Comparison on AVSR Task

To further assess the utility of red exclusion as a feature extraction technique a comparative analysis of each technique described in Section 5 was carried out. As each filter required to have parameters set *a priori*, images were selected as prototypical examples of each subject and the key features (left corner, right corner, middle top and middle bottom of the mouth) were hand labelled. Using an iterative approach, the parameter spaces were searched to find values that gave the least distance between hand and filter labelled key features. The resulting parameter sets were used for subsequent feature extraction.

A 300 fold bootstrapping procedure, with 100 at a maximum training epoch of 5000 and another 200 with a maximum training epoch of 10,000 were used. Randomly selected training and testing data were used for each trial at a 50:50 ratio without replacement. 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. The data was normalised by scaling the data such that it had a zero mean and unity standard deviation (Demuth and Beale, 1998).

For each image five different filters were used to initially extract the mouth contour. From this contour the same processing applied to each technique to calculate the required features (eg. height and width). Using these features an ANN was trained and tested on its ability to distinguish visemes.

### 7.2 Recognition Experiments

For all the results of the recognition experiments that follow, a 10-fold bootstrapping procedure, with randomly selected training and testing data for each trial, was adhered to.

In addition to the raw and normalised data sets, two other transformations were performed in the hope to improve recognition accuracy. SVD was performed on the data and attributes with eigenvalues greater than 0.001 were used. We also tested a combination of normalisation and then SVD. Therefore, there were four types of data to train each neural network upon – raw, normalised (N), SVD, and N/SVD.

Phoneme, Viseme, or Voicing were the three possible classification tasks for a NN to perform.

1. Phoneme classification tasks involved discriminating between 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. This is more generally useful than the word level preferred by many which are biased by the use of statistical or language models – thus not directly comparable in terms of raw recall accuracy.

**8. RESULTS**

Before presenting the results the reader is reminded that the ANN were trained on a very limited set of low resolution data: two examples of each phoneme/position pair for each of three subjects. Furthermore, low-cost OTS equipment was used and each subject was seated 1.5 to 1.8 metres from the recording device.

**8.1 Red Exclusion**

A one-way ANOVA revealed significant differences between the performance of the filters,  $F(4,495) = 138.235, p < 0.001$ . The means and standard deviations of the ANN recall accuracies for viseme identification are presented in Table 4 and the values for the parameters used for each filter are in Table 3.

The ANN using the red exclusion filter performed better than any of the other filters. Post hoc comparisons using the Fisher LSD test revealed that the red exclusion filter performed significantly better than the edge, red-green, hue and gray filters ( $p < 0.05$ ).

**8.2 Acoustic and Visual Recognition**

Tables 5 and 6 show the overall recognition accuracy of separate acoustic and visual ANNs attempting to distinguish between the nine phonemes, three viseme and three voicing groupings. It is immediately obvious from Table 6 that vision alone is not able to distinguish between the set of nine phonemes or three voicing groups with the accuracies hovering around guessing level (11.1%, and 33.3%, respectively). According to the psycholinguistic work reviewed this is to be expected (Dodd and Campbell, 1987). Significantly, the accuracy of the acoustic network is above this rate. Interestingly, the visual network, as predicted, outperforms the acoustic net on the viseme

Filter	Parameters
edge	threshold = 50
red-green	$L_{lim} = 1.5, U_{lim} = 2.4$
hue	$h_0 = 0.5, w = 0.5$
gray	threshold = 80
red exclusion	$\beta = -0.15$

Table 3: Parameters used for each filter type.

Filter	Mean	Std. Dev.	Min.	Max	Reference
edge	48.46	4.49	36.95	60.79	Stiefelhagen <i>et al</i> (1997)
red-green	43.61	4.70	30.73	60.30	Wark <i>et al</i> (1998)
hue	38.87	4.42	27.18	51.24	Cozniz <i>et al</i> (1996)
gray	50.50	4.60	36.19	63.19	Rao and Mersereau (1994)
red exclusion	51.46	4.69	40.17	66.59	Lewis and Powers (2001)

Table 4: Recall accuracy of each filter for viseme identification.

Class	RAW	NORM	SVD	N/SVD
PHONEME	11.8	21.2	16.2	20.9
VOICING	54.8	58.4	53.1	53.5
VISEME	42.4	43.3	37.5	42.4

Table 5: Recognition accuracy (%) of acoustic neural networks.

Class	RAW	NORM	SVD	N/SVD
PHONEME	8.4	11.5	10.9	14.7
VOICING	29.9	29.3	29.5	32.2
VISEME	30.6	54.7	44.1	53.0

Table 6: Recognition accuracy (%) of visual neural networks.

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.

### 8.3 Integration

Table 7 outlines the results for all of the 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 ANN, which had an accuracy of 21.2% for normalised data. However, the late integration using voicing and viseme subnets gave 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 severely degraded signal in both the acoustic and visual modalities.

## 9. 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.

Architecture	NORM	N/SVD
Acoustic Only	21.2	20.9
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

Table 7: Phoneme Recognition accuracy (%) of early and late integration architectures.

## 9.1 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, 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.

The two best filters when compared to hand-labelled images were the red exclusion and the red-green extraction of Wark *et al* (1998), with the worst being the hue filter of Coianiz *et al* (1996). However, the hand-labelling of images is a very subjective task and there is a systematic variation due to different techniques focussing on different properties of the lip edge. A more sophisticated analysis based on hand-labelling would need to involve a larger data set and the hand-labelling being performed by multiple people for the same image to gain an average for each feature.

Red exclusion performs better than any other method on average for viseme recognition and these results are significant to the .05 level for our current data set. However, with just three Caucasian subjects and a relatively small number of samples of each phoneme taken in the same environment at the same time by the same kind of webcam, significance cannot be claimed for generalisation to different subjects or recordings. The gray filter was the closest to the red exclusion filter even though it did not accurately identify the mouth contours. This may be because it is still accurately and consistently representing the change in the mouth shape. Further analysis on a wider population and standard data sets will test the true effectiveness of red exclusion.

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. Our current research is looking at why the relationship might exist and how this can be used to refine the red exclusion technique. It is hypothesised that red exclusion is related to the colour opponent properties of mammalian vision.

## 9.2 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 ANN 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 ANN outperformed the early integration network in identifying the phonemes, 21.2% versus 20.1%. Furthermore, the ANN must also learn the proper weighting 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 ANN was unable to learn the correct weightings. Another explanation for the failure of integration, and one that is a fundamental problem of ANNs, is that ANNs 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 with 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. One reason why this network performed badly with respect to the other networks has to do with the training regime employed and the lack of additional data for use in training the error networks. The error analysis network was trained using the original training data. Thus the subnets were attuned to this data and most 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, when enough data is available, will be 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 that with a larger training base the gamut of training regime could be explored to find the most efficient and effective method.

## CONCLUSION

This research has shown that multi-speaker AVSR is useful in a natural office environment where the user is not equipped with specialised equipment, 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 acoustic and visual sources and increased the accuracy by around 40% over an acoustic only ANN. AVSR is a flourishing area of research with many avenues still open to investigation, especially in the area of sensor fusion. Our 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

- ADJOUANI, A. and BENOIT, C. (1996): On the integration of auditory and visual parameters in an HMM-based ASR, in STORK and HENNECKE (1996), 461–471.
- ANGELOPOULOU, E., MOLANA, R., and DANILIDIS, K. (2001): Multispectral color modeling. *Technical Report MS-CIS-01-22*, University of Pennsylvania, CIS.
- BASU, M. and HO, T.K. (1999): The learning behavior of single neuron classifiers on linearly separable or nonseparable input. In *Proceedings of the 1999 International Joint Conference on Neural Networks*, Washington, D.C.
- 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*, 667–671.
- BREGLER, C., OMOHUNDRO, S.M., SHI, J., and KONIG, Y. (1996): Towards a robust speechreading dialog system. In STORK and HENNECKE (1996), 410–423.
- 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, E. (1993): *Statistical language learning*. MIT Press, Cambridge, MA.
- CHELAPPA, R., WILSON, C., and SIROHEY, S. (1995): Human and machine recognition of faces: A survey, in *Proceedings of the IEEE*, 83(5): 705–739.
- COHEN, M., WALKER, R., and MASSARO, D. (1996): Perception of synthetic visual speech. In STORK and HENNECKE (1996), 153–168.
- COIANIZ, T., TORRESANI, L. and CAPRILE, B. (1996): 2d deformable models for visual speech analysis. In STORK and HENNECKE (1996), 391–398.
- DAVIS, S. and MERMELSTEIN, P. (1990): Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In WAIBEL, A. and LEE, K., editors, *Readings in Speech Recognition*, 64–74. Morgan Kaufmann Publishers Inc., San Mateo, CA.
- DEMUTH, H. and BEALE, M. (1998): *Neural Network Toolbox: User's Guide*. The MathWorks, <http://www.mathworks.com>.

- DODD, B. and CAMPBELL, R., editors (1987): *Hearing by Eye: The psychology of lip-reading*. Lawrence Erlbaum Associates, Hillsdale NJ.
- 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*, Detroit USA.
- DUPONT, S. and LEUTTIN, J. (2000): Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151.
- FROMKIN, V., RODMAN, R., COLLINS, P., and BLAIR, D. (1996): *An Introduction to Language*. Hartcourt Brace and Company, Sydney, 3rd edition.
- GLOTIN, H., VERGYRI, D., NETI, C., POTAMIANOS, G., and LUETTIN, J. (2001): Weighting schemes for audio-visual fusion in speech recognition. In *Proc. Int. Conf. Acoust. Speech Signal Process.*
- 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), 505–515.
- GRANT, K. and SEITZ, P. (1998): The use of visible speech cues (speechreading) for directing auditory attention: Reducing temporal and spectral uncertainty in auditory detection of spoken utterances. In *16th International Congress on Acoustics*.
- GRAY, M., MOVELLAN, J., and SEJNOWSKI, T. (1997): Dynamic features for visual speechreading: A systematic comparison. In MOZER, JORDAN, and PERSCHE, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, Cambridge MA.
- HECKMANN, M., BERTHOMMIER, F., and KROSCHEL, K. (2001a): A hybrid ANN/HMM audio-visual speech recognition system. In *Proceedings of AVSP-2001*.
- HECKMANN, M., BERTHOMMIER, F., and KROSCHEL, K. (2001b): Optimal weighting of posteriors for audio-visual speech recognition. In *Proceedings of ICASSP 2001*, Salt Lake City, Utah.
- 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, M., PRASAD, V., and STORK, D. (1994): Using deformable templates to infer visual speech dynamics. In 28th Annual Asimolar Conference on Signals, Systems, and Computer, Pacific Grove, CA. *IEEE Computer*. 2:576–582.
- HENNECKE, M., STORK, D., and PRASAD, K.V. (1996): Visionary speech: Looking ahead to practical speech reading systems. In STORK and HENNECKE (1996), 331–350.
- HUNKE, M. and WAIBEL, A. (1994): Face locating and tracking for human-computer interaction. In 28th Annual Asimolar Conference on Signals, Systems, and Computers, *IEEE Computer Society*, Pacific Grove, CA. 2: 1277–1281.
- LEUTTIN, J. and DUPONT, S. (1998): Continuous audio-visual speech recognition. In *Proceedings of the 5th European Conference on Computer Vision*, 2: 657–673.
- LEWIS, T.W. (2000). Audio visual speech recognition: Extraction, recognition, and intergration.
- LEWIS, T.W. and POWERS, D. (2001): Lip feature extraction using red exclusion. In EADES, P. and JIN, J., editors, *CRPIT: Visualisation*, 2000, 2: 61–70.
- M2VTS (2000): *M2VTS Multimodel face database, release 1.0*. World Wide Web, <http://www.tele.ucl.ac.be/PROJECTS/M2VTS/>.
- 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 and visual information to understand speech. *American Scientist*, 86(3): 236–245.
- MCGURK, H. and MACDONALD, J. (1976): Hearing lips and seeing voices. *Nature*, 264:746–748.
- 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*, 2: 833–837.
- 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, J. (1995): Visual speech recognition with stochastic networks. In Tesauro, G., Toruetzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, 7: 851–858. MIT Press, Cambridge.
- MOVELLAN, J. and MINEIRO, P. (1998): Robust sensor fusion: Analysis and application to audio visual speech recognition. *Machine Learning*, 32: 85–100.
- NETI, C., POTAMIANOS, G., LEUTTIN, J., MATTHEWS, I., GLOTIN, H., and VERGYRI, D. (2001): Large-vocabulary audio-visual speech recognition: A summary of the Johns Hopkins summer 2000 workshop. In *Workshop on Multimedia Signal Processing, Special Session on Joint Audio-Visual Processing*, Cannes.
- POTAMIANOS, G. and NETI, C. (2000): Stream confidence estimation for audio-visual speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, 746–749, Beijing.
- POTAMIAONOS, G. and POTAMIANOS, A. (1999): Speaker adaptation for audio-visual speech recognition. In *Proceedings of EUROSPEECH (3)*, 1291–1294, Budapest.
- 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 Centre.

- RABINER, L. and JUANG, B. (1993): *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ.
- RAO, R. and MERSERAU, R. (1994): Lip modeling for visual speech recognition. In 28th Annual Asimolar Conference on Signals, Systems, and Computers, volume 2. *IEEE Computer Society*, Pacific Grove CA.
- 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), 194–219.
- ROGOZAN, A. (1999): Discriminative learning of visual data for audiovisual speech recognition. *International Journal of Artificial Intelligence Tools*, 8(1):43–52.
- SCHAFER, R. and RABINER, L. (1990): Digital representations of speech signals. In WAIBEL, A. and LEE, K., editors, *Readings in Speech Recognition*, 49–64. Morgan Kaufmann Publishers Inc., San Mateo, CA.
- SCHIFFERDECKER, G. (1994): Finding structure in language. Master's thesis, University of Karlsruhe.
- STIEFELHAGEN, R., YANG, J., and MEIER, U. (1997): Real time lip tracking for lipreading. In *Proceedings of Eurospeech '97*.
- STORK, D. and HENNECKE, M., editors (1996): *Speechreading by Man and Machine: Models, System, and Applications*. NATO/Springer-Verlag, New York.
- SUMMERFIELD, Q. (1987): *Some preliminaries to a comprehensive account of audio-visual speech perception*, 3–52. In DODD and CAMPBELL (1987).
- VERMA, A., FARUQUIE, T., NETI, C., BASU, S., and SENIOR, A. (1999): Late integration in audio-visual continuous speech recognition. In *Automatic Speech Recognition and Understanding*.
- VOGT, M. (1996): Fast matching of a dynamic lip model to color video sequences under regular illumination conditions. In STORK and HENNECKE (1996), 399–407.
- 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, 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*, 123–125.
- YANG, J. and WAIBEL, A. (1996): A real-time face tracker. In *Proceedings of WACV'96*, 142–147.

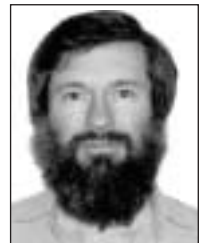
### BIOGRAPHICAL NOTES

*Trent Lewis completed a BSc in Cognitive Science in 1999 and went on to complete Honours in Computer Science in 2000 at the Flinders University of South Australia. He commenced his PhD in July of 2001 also at Flinders University. Trent's main area of study is sensory data fusion with a particular focus on solving this problem using learning and statistical techniques to find the best fusion strategy in different situations. His current application for this area is audio-visual speech recognition.*



Trent W. Lewis

*David Powers is Associate Professor of Computer Science at the Flinders University of South Australia. He has a PhD from the University of NSW in the area of Machine Learning of Natural Language and specialises particularly in unsupervised learning techniques as applied to linguistic and cognitive applications. Current research interests include speech and brain control of an intelligent room, and multimodal sensory-motor processing in the context of a robot baby, with another major application being intelligent web search.*



David M.W. Powers