COMBINING NOISE COMPENSATION WITH VISUAL INFORMATION IN SPEECH RECOGNITION

Stephen Cox, Iain Matthews, Andrew Bangham
School of Information Systems, University of East Anglia, Norwich NR4 7TJ, U.K.

ABSTRACT

The addition of visual information derived from the speaker's lip movements to a speech recogniser (*speechreading*) can significantly enhance the performance of the recogniser when it is operating under adverse signal-to-noise ratios. However, processing of video signals imposes a large computational demand on the system and there is little point in using speechreading techniques if similar performance gains can be obtained using techniques which operate on only the audio signal and which are less computationally expensive. In this paper, we show that combining visual information with an audio noise compensation technique (spectral subtraction) leads to a performance significantly higher than that obtained using speechreading only or noise compensation only. The optimum method for speech recognition in the presence of noise is to use speech models that are matched to the input speech, and we show that the addition of visual information also gives a performance gain when matched models are used. We also describe a method of "late" integration which uses a measure of confidence derived from information output by the audio recogniser to achieve a performance which is close to optimum.

1. INTRODUCTION

There is currently great interest in increasing the robustness of automatic speech recognition (ASR) to make it more effective in adverse environments e.g. when interfering noise, reverberation, distortion or filtering of the signal is present etc—for a review of work in this area, see [4]. However, these techniques are ultimately limited by the amount of information available in the degraded audio signal and there has recently been interest in augmenting the audio signal with a visual signal derived from an image of the speaker's lips (speechreading) [9]. At present, it is not clear whether using visual information is superior to some of the audio noise compensation techniques which have been developed or whether they can be used successfully in combination. Here, we present results on recognition of isolated words which confirm that the techniques can be combined to give a performance which is superior to using either in isolation. Moreover, our results show that inclusion of visual information is beneficial in the case where the models used for recognition are matched to the input signal in terms of signal-to-noise ratio (SNR). The use of models matched to the input signal is generally agreed to be optimum in cases where the noise is stationary, and so the performance gain obtained from including visual information is significant.

2. INTEGRATION STRATEGIES

There are essentially two approaches to integrating visual information with the audio information: early integration, in which the video and audio information is combined before being processed in a recogniser, and late integration, in which separate recognisers are used for the audio and video channels and their outputs combined in the decision process [6]. Our own experience suggests that late integration is more successful and in these experiments we use an integration scheme similar to the one described in [1]. If we assume that the output of each recogniser is a set of probabilities, one for each of the V vocabulary words, the recognition decision is to choose word w^* where

$$w^* = \arg\max_{i=1,2,\ldots,V} \{\alpha \Pr(w_i|A) + (1-\alpha)\Pr(w_i|V)\} \qquad (1)$$

where $\Pr(w_i|A)$ and $\Pr(w_i|V)$ are the respective probabilities of the i'th word from the audio and video recognisers and α is a weighting factor.

3. DATA, AUDIO AND VISUAL FEATURES AND MODELLING

The audio-visual database used for these experiments has been described in detail elsewhere [8]. Briefly, it consisted of recordings of ten subjects speaking three repetitions of the letters of the alphabet. The video recordings used were of the full face and were made under ordinary studio lighting conditions with no highlighting of the lips and no attempt to restrain the head position. Audio recording was via a high-quality tie-clip microphone. Each utterance movie was hand-segmented so that the mouth position began and ended with the mouth in a neutral position. The database was divided into a training-set which consisted of the first two utterances from each speaker (a total of 520 utterances) and a test-set which consisted of the third utterances from each speaker (260 utterances).

The visual features used were derived from the application of a one-dimensional *sieve* [2] to the image. For each frame, a 20-dimensional vector was derived from an analysis of an 80×60 pixel region centred on the mouth image. For a full description of the datasieve and the derivation of visual features using a datasieve, the reader is referred to [2] and [8].

The audio features were the (unlogged) outputs of a 24 channel filterbank covering the range 100–5000 Hz made using the HTK hidden Markov modelling software [7]. Although filterbank features are known to give lower accuracy than the more commonly used MFCC representation, they enable the use of spectral subtraction which is described in

section 4. It is possible to convert these features to MFCCs after the spectral subtraction process and so improve performance, but this was not done in these experiments. The audio feature vectors were estimated every 20 ms and since the visual vectors were estimated at the PAL frame-rate of 40 ms, an extra vector was interpolated between each pair of visual vectors to give the same audio and video frame-rates.

The signal-to-noise ratio (SNR) of each "clean" audio utterance file was estimated (it was always at least 40 dB) and an appropriate amount of Gaussian noise was added to each file to generate five additional utterance files whose SNRs were 20 dB, 10 dB, 6 dB, 3 dB and 0 dB. Separate hidden Markov models (HMMs) were used for the audio and video utterances. In each case, an utterance was modelled by a 10 state, left-right, HMM with a single Gaussian distribution per state and a diagonal covariance matrix. A set of HMMs was made from the "clean" audio utterances and from the video features.

3.1. Baseline results

Results on the training- and test-set using the separate audio and video HMMs are shown in figure 1. It is clear that:

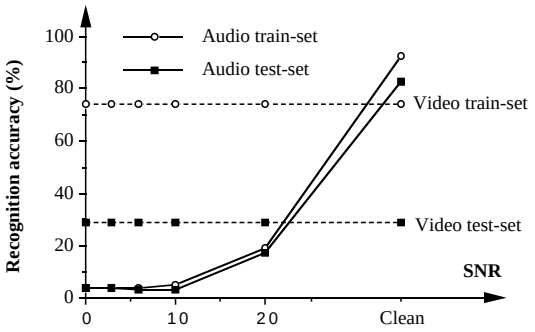

Figure 1: Baseline results for training and test-sets

- audio performance drops rapidly to near chance level as the SNR decreases;
- the large gap in performance between training and testing data for the video features indicates that the video models are undertrained.

The last point shows that, although the audio and video models had identical topologies and received identical amounts of training-data, the video models were undertrained when compared with the audio models. This may indicate that the inherent variability of the video data is higher than that of the audio data or that the video features do not represent the data classes well enough.

The integration technique described by equation 1 was then implemented to make a speechreading system. To establish the best possible performance obtainable from this technique, an exhaustive search was carried out to find values of α that minimized recognition error on the test-set. α was restricted to lie between 0 and 1 and the log-likelihoods from both recognisers were normalised to probabilities before being used in the equation. We would expect the optimum value of α in equation 1 to vary with the signal-to-noise ratio, and it is possible that it would be different for each utterance. Hence two conditions were investigated:

- α optimised for each utterance;
- α optimised for each SNR.

Results are shown in figure 2. The addition of visual information is clearly capable of boosting the accuracy of the recogniser very significantly when a fixed value of α is used for each SNR and the accuracy never drops below that of the visual recogniser. By choosing a different value of α for each utterance, a further significant gain is obtained. However, these results were obtained with prior knowledge of the correct class of the test utterances to establish an upper bound on performance using this integration technique and may not be representative of what can be achieved practically.

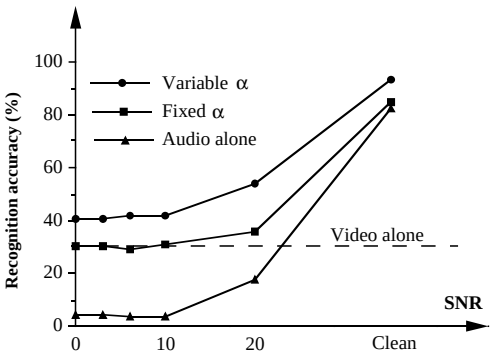

Figure 2: Results for speechreading using uncompensated audio

4. NOISE COMPENSATION TECHNIQUES AND RESULTS

The use of techniques to compensate a speech recognition system when the audio signal is corrupted by additive noise has been extensively studied [5]. Two techniques that have been shown to be successful and which we used here are:

- Matched models, in which the HMMs used are built from speech data corrupted in exactly the same way as the test data;
- Spectral subtraction, in which HMMs built from "clean" speech are used but the input speech is processed to remove some of the noise.

Our spectral subtraction technique was similar to that reported in [3]. An estimate of the power spectrum of the additive noise is made by analysing a section of the signal which is known to consist of silence, and a proportion of this estimate is subtracted from the complete signal before it is input to the speech recognition system. If $X(f)$ is the signal spectrum, $N(f)$ an estimate of the noise spectrum, then the subtracted spectrum $X'(f)$ is:

$$X'(f) = \begin{cases} X(f) - \gamma N(f) & \text{if } X(f) > \gamma N(f) \\ \delta X(f) & \text{otherwise} \end{cases} \quad (2)$$

where the optimum γ and δ are determined experimentally. Results from these experiments are shown in figure 3. Per-

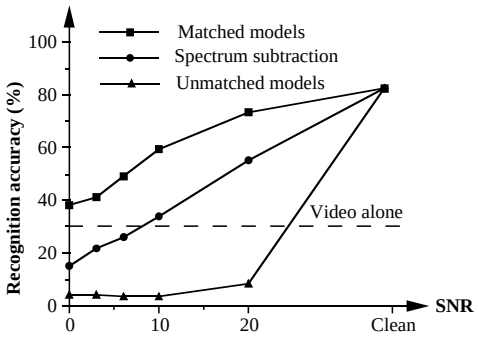

Figure 3: Results for matched models and spectral subtraction (audio only)

formance using matched models is, as expected, superior to that using spectral subtraction, but both techniques give substantial gains over uncompensated audio. Using matched models is superior to the best performance obtained from speechreading without any audio compensation, but the latter is superior to spectral subtraction.

Visual information was then added as described in section 3.1. Because of the difficulty of optimising α for each utterance in a real system (to be discussed in section 5), results were obtained using a fixed value of α for each SNR and are shown in figure 4. In both cases (matched models and spectrally subtracted speech) adding visual information improves performance.

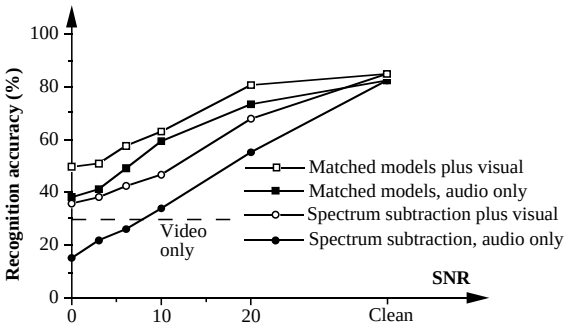

Figure 4: Results for speechreading using compensated audio

5. PRACTICAL INTEGRATION TECHNIQUES

The results obtained for the speechreading experiments in sections 3.1 and 4 were the highest possible accuracies, in that knowledge of the class of each test-set utterance was used to select values of α to maximise recognition accuracy on the test-set. In this section, we concentrate on automatic estimation of the value of α. For the experiments described here, we have used spectral subtraction as our method of noise compensation. Although using matched models gives higher accuracy, they are unlikely to be available in most applications.

Our first approach to estimating α was to estimate confidence-measures c_A from the audio recogniser and c_V from the video recogniser ($0 < c_A, c_V < 1.0$) that the input word was correct. The value of α was then estimated as $\alpha = c_A/(c_A + c_V)$. One advantage of this method is that it does not require an estimate of the SNR to be made. We investigated the following ways of estimating c_A and c_V:

- Using the training-set to estimate the distributions of the likelihoods of correct words and incorrect words. These distributions are then used with the likelihoods produced by an input word to estimate a confidence measure;
- Computing the entropy associated with the likelihoods produced by the recognisers and using this to estimate a confidence-measure;
- Using the ratio of the highest likelihood to the remaining likelihoods.

A problem with all these approaches was that the distribution of likelihoods obtained from the video recogniser when words from the test-set were input was very different from the distribution which resulted when words from the training-set were input. This is predictable from the large difference between the accuracies on the training- and test-set for the video recogniser (figure 1). Because of this effect, it was not attempted to estimate any confidence measures from the output of the video recogniser. In addition, examination of the distributions of likelihoods from the audio recogniser for correct and incorrect words showed that there was little separation between the two distributions and hence using these distributions to estimate a confidence measure for each input utterance was not viable.

We therefore adopted a more robust measure of confidence which was based on the *a priori* uncertainty of the audio recogniser about the identity of the input word at a given SNR. This requires an estimate of the SNR to be made, but an SNR estimate is required anyway to implement the spectral subtraction technique. Denoting the set of legal input words as a random variable X and the set of recognised words as a random variable Y, the probability that word i was input when word j was recognised may be written $\Pr(X = i|Y = j)$ and can be estimated from a confusion-matrix formed by testing the training-set data (after spectral subtraction). At a given SNR, the uncertainty about the identity of the input word given that word j was recognised is given by the conditional entropy $H(X|Y = j)$:

$$H(X|Y=j) = \sum_{i=1}^{26} \Pr(X=i|Y=j)\log_2(\Pr(X=i|Y=j))$$

The average uncertainty after making a recognition decision is given by the average conditional entropy $H(X|Y)$:

$$H(X|Y) = \sum_{j=1}^{26}\sum_{i=1}^{26} \Pr(X=i,Y=j)\log_2(\Pr(X=i|Y=j))$$

The maximum uncertainty about the identity of the input word is $H(X|Y)_{max} = \log_2(26)$ bits and occurs when all elements of the confusion-matrix are equal i.e. all words are equiprobable to have been input when a word is recognised.

Conversely, if the confusion matrix has only a single entry in each column, $H(X|Y) = 0$ and there is no uncertainty about the input utterance. A possible estimate of α is then

$$\alpha = 1 - \frac{H(X|Y)}{H(X|Y)_{max}}. \quad (3)$$

In practice, we found it useful to compress the value of α by using $\alpha' = \alpha^{\gamma}$ in equation 1—a suitable value of γ was 0.5.

Figure 5 compares performance using this entropy-derived confidence measure (EDCM) to estimate α at each SNR with performance using the optimal value of α selected with knowledge of the classification of the test-set utterances (as in figure 4). Performance with no video added is shown for reference. The performance obtained using the EDCM is close to ideal.

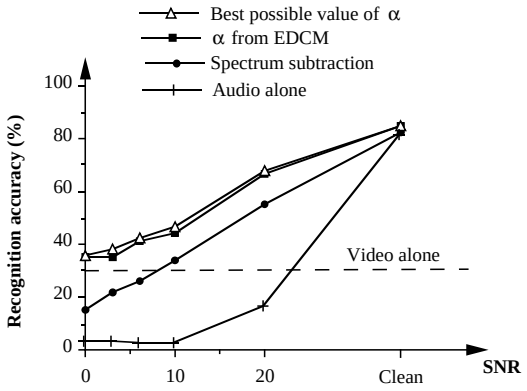

Figure 5: Results for speechreading using EDCM estimate of α

The EDCM technique requires an estimation of the SNR. We classified the test-signal as one of the 6 SNR's used in these experiments by using the training-set to estimate distributions of the audio recogniser likelihoods at each SNR, and then using the likelihoods output by the recogniser to estimate the posterior probability of each SNR when a test utterance was input. This estimate was smoothed by using the previous N probabilities from each SNR i.e. the the SNR of the t'th input utterance is S^* where

$$S^* = \underset{i=1,2,\ldots,6}{\operatorname{argmax}} \prod_{j=t-N+1}^{t} p_i^j \quad (4)$$

and p_i^j is the posterior probability of the j'th utterance being of SNR i. Using this technique, it was found that that the SNR was classified corectly on 85% of occasions. Incorrect SNR classifications were always to an adjacent SNR, which usually did not alter the classification of the utterance, so that the there was virtually no loss of accuracy in automatic estimation of the SNR.

6. DISCUSSION

Adding visual information improves recognition performance over and above the gains obtained using noise compensation techniques. Our experiments confirmed that, when only audio information is available, for additive, stationary noise, the use of matched models leads to very significant gains in performance. The more practical technique of spectral subtraction gives lower but still very significant improvements to accuracy. However, our experiments show a further improvement of 10–15% when visual information is added and we might expect this order of improvement to be observed when other noise compensation techniques not investigated in this paper are used. We used a visually difficult task which resulted in under-trained visual models and a low visual recognition accuracy. This led us to integrate the audio and visual information using an entropy-based measure of confidence derived solely from the audio recogniser. The use of this measure gave results which were as good as could possibly be obtained with prior knowledge of the classification of each utterance. However, if the visual recogniser were more accurate, the approach could be extended by including a confidence measure derived from its output. We are now concentrating on expanding our database and making our visual recogniser more robust.

References

[1] A. Adjoudani and C. Benoît. *On the Integration of Auditory and Visual Parameters in an HMM-based ASR*, pages 461–471. In Stork and Hennecke [9], 1996.

[2] J. Bangham, P. Ling, and R. Harvey. Nonlinear scale-space causality preserving filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:520–528, 1996.

[3] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27:113–120, 1979.

[4] R. Cole et al. The challenge of spoken language systems: research directions for the nineties. *IEEE Transactions on Speech and Audio Processing*, 3(1):1–21, January 1995.

[5] S. Furui. Towards robust speech recognition under adverse conditions. In *Proc. ESCA Workshop in Speech Processing under Adverse Conditions*, pages 31–42, November 1992.

[6] M. E. Hennecke, D. G. Stork, and K. V. Prasad. *Visionary Speech: Looking Ahead to Practical Speechreading Systems*, pages 331–349. In Stork and Hennecke [9], 1996.

[7] J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK book*. Entropic Research Laboratories Inc., 1996.

[8] I. Matthews, J. Bangham, and S. Cox. Audiovisual speech recognition using multiscale nonlinear image decomposition. In *Proc. Int. Conf. on Spoken Language Processing*, pages 38–42, September 1996.

[9] D. G. Stork and M. E. Hennecke, editors. *Speechreading by Humans and Machines: Models, Systems and Applications*. NATO ASI Series F: Computer and Systems Sciences. Springer-Verlag, Berlin, 1996.

```
@inproceedings{Cox1997:Compensation,
    author      = "Cox, Stephen and Matthews, Iain and Bangham, Andrew",
    title       = "Combining noise compensation with visual information in
                    speech recognition",
    pages       = "53--56",
    crossref    = "avsp1997:proceedings"
}

@proceedings{avsp1997:proceedings,
    title       = "Proceedings of the ESCA Workshop on Audio-Visual
                    Speech Processing:~Cognitive and Computational
                    Approaches",
    booktitle   = "Proceedings of the ESCA Workshop on Audio-Visual
                    Speech Processing:~Cognitive and Computational
                    Approaches",
    year        = 1997,
    editor      = "Beno\^{\i}t, Christian and Campbell, Ruth",
    address     = "Rhodes",
    month       = sep,
}
```