

IMPROVING VISUAL NOISE INSENSITIVITY IN SMALL VOCABULARY AUDIO VISUAL SPEECH RECOGNITION APPLICATIONS

Simon Lucey, Sridha Sridharan and Vinod Chandran

Speech Research Laboratory, RCSAVT
School of Electrical and Electronic Systems Engineering
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
s.lucey@qut.edu.au, s.sridharan@qut.edu.au and v.chandran@qut.edu.au

ABSTRACT

Visual noise insensitivity is important to audio visual speech recognition (AVSR). Visual noise can take on a number of forms such as varying frame rate, occlusion, lighting or speaker variabilities. In this paper the use of a high dimensional secondary classifier on the word likelihood scores from both the audio and video modalities is investigated for the purposes of adaptive fusion. Preliminary results are presented demonstrating performance above the catastrophic fusion boundary for our confidence measure irrespective of the type of visual noise presented to it. Our experiments were restricted to small vocabulary applications.

1. INTRODUCTION.

The visual modality has proven very difficult to get reliable confidence measures from due to the poor performance of the visual modality in recognition. An accurate measure of visual noise also plagues the use of the visual modality as visual noise can take on a number of forms (eg. varying frame rate, occlusion, lighting or speaker variabilities). AVSR becomes advantageous over conventional audio based speech recognition in noisy conditions due to the complementary nature of the audio and video modalities. Much work has been done on adaptively weighting audio and visual modalities for improved speech recognition performance based on the degradation in either modality. It has been shown that the fusion of likelihood scores from independent audio and video recognisers can improve the recognition rates for speech recognition [6, 7]. This improvement is heavily dependent on how the scores are combined with each score being given a weight based on the confidence of recognition accuracy in that modality. However, most of these weights have been deduced by a measure of noise or dispersion in the audio modality [3, 6] with the visual modality being largely ignored.

If an audio only or a video only modality can out perform an overall system after fusion, clearly there is something wrong with the fusion algorithm, a problem called *catastrophic fusion*. The primary goal of an AVSR scheme is to fuse the audio and video modalities such that their overall performance lies well above this catastrophic fusion boundary. Previous techniques have enjoyed much success by trying to relate the calculation of a confidence measure to the amount of noise in the audio modality [3] and assuming a clean video modality. Unfortunately, such approaches cannot be used when the quality of the video modality fluctuates as unlike the audio modality a video noise measure cannot be easily

found. Other techniques [1, 7] have tried to take advantage of the correlation between the recognition error of a classifier and dispersion measures (ie. entropy, variance, range) of the word likelihood scores from that classifier. These type of approaches can be readily applied to the calculation of a confidence measure in the video modality as they require no real world measure of noise. Unfortunately these techniques have had limited success due to the poor correlation of dispersion measures with classifier recognition error when test conditions differ markedly from those in the training stage (ie. context change) resulting in catastrophic fusion.

In this paper a secondary classifier is employed to gain a confidence measure in the recognition accuracy of the audio and video modalities. The secondary classifier is applied to the word likelihood scores of the audio and visual modalities so as to gain an accurate confidence measure of each modality. Secondary classifiers take advantage of the knowledge that in practical scenarios, due to data deficiencies and mismatches between training and testing data, classifiers will never output the true a posteriori probabilities but rather an estimate [8]. By training up stochastic models on the training data under a variety of degradations a model can be formed that more accurately takes into account the errors associated with these estimated posterior probabilities. In our approach the initial log-likelihoods gained from the word recognisers are concatenated into a N dimensional feature vector ξ for each modality where N is the size of the vocabulary. Stochastic models are then trained for both modalities using a priori knowledge of whether ξ belongs to a correct or incorrect class. This secondary classifier is then used to gauge how confident one is in ξ whereby we can weight the importance of each modality accordingly. This adaptive scheme can be applied to both the audio and video modalities with the performance being less sensitive to the type or amount of noise presented to it in the real world than previous techniques based on finding confidence measures from likelihood scores.

2. FUSION STRATEGIES.

There are basically two topologies for integrating visual and audio modalities with one another [6]: early integration, in which 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. It has been shown in previous work [3, 6] that late integration generally out performs early integration. Late

integration has the following benefits over early integration [6],

- provides robustness to the failure of a modality;
- modalities can have different temporal synchrony;
- easier training and computation as each modality can be processed independently;

The question of how the results of different classifiers should be combined arises when using a late integration topology. Experiments have shown that multiplying the probabilities from each classifier performed better than using either sum, min or max combination schemes.

If one assumes that the output of each recogniser is a set of probabilities, one for each of the N vocabulary words, the recognition decision is to choose word w^* where

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

where α is a weighting factor and $Pr(w_i|A)$ and $Pr(w_i|V)$ are the respective probabilities of the i 'th word, estimated from the log-likelihoods taken from the audio and video classifiers. To calculate the α weighting factor to combine the audio and video modalities as used in Equation 1 we used,

$$\alpha = \frac{\zeta_A}{\zeta_A + \zeta_V} \quad (2)$$

where ζ_A and ζ_V are the confidence measures for the audio and video modalities respectively.

3. AUDIO VISUAL DATA AND MODELLING

The AVLetters database [6] was used for experiments. The database consisted of,

- ten subjects (male and female) speaking three repetitions of the letters of the alphabet;
- the visual signal of each utterance being manually cropped into an 80 x 60 pixel region of interest (ROI) containing the mouth image;
- the database being divided into a training set which contained the first two utterances from each speaker (520 utterances) with the test-set containing the third utterance (260 utterances);

For the audio features we used standard HTK [9] mel-frequency cepstral coefficients with mean cepstral subtraction and delta coefficients to create a 26 dimensional feature vector. The visual features were extracted by performing standard principal component analysis (PCA) [2] on the 80 x 60 ROI mouth images obtaining the first 15 *Eigenlip* weights with delta coefficients to obtain a 30 dimensional feature vector. The reader is advised to look to Breglers [2] paper for a full description of the Eigenlip feature extraction technique. Audio features were sampled every 10ms while the video stream was sampled at 40ms intervals.

Separate hidden Markov models (HMMs) were used to model the audio and video utterances using HTK ver 2.2 [9]. For the audio modality, an utterance was modelled using a 4 state, left to right, HMM with 2 mixtures per state and diagonal covariance matrices. A similar topology was used for the visual modality with a 9 state, left to right, HMM with 3 mixtures per state and diagonal covariance matrices.

4. CONFIDENCE MEASURES.

The calculation of a confidence measures for a set of classifiers is important to an effective data fusion scheme. The objective is to determine those classifiers which have the largest additional errors associated with them, and place a lower weighting on these classifiers. The global error of the fused system can thus be minimised, by judicious choice of weights for the system.

Confidence measures within a Bayesian framework are calculated for a N word class problem from a theoretical standpoint by modeling a two class problem of *correct* and *incorrect* classes. A true confidence measure ζ is the *a posteriori* probability that a class selected w_i is the correct class for the input utterance x such that according to Bayes rule,

$$\zeta = Pr(\text{correct}|x) = \frac{P_c p(x|\text{correct})}{P_c p(x|\text{correct}) + P_i p(x|\text{incorrect})} \quad (3)$$

where P_c and P_i are the *a priori* probabilities of being correct and incorrect respectively. Given the true $p(x|w_i)$ and assuming equal a priori word class probabilities,

$$\begin{aligned} j &= \arg \max_{i=1}^N p(x|w_i) \\ p(x|\text{correct}) &= p(x|w_j) \\ p(x|\text{incorrect}) &= \sum_{i=1, i \neq j}^N p(x|w_i) \end{aligned} \quad (4)$$

An assumption is made in Equation 4 that we know the true probability density functions $p(x|w_i)$. In practice we can only get an estimate resulting in estimation errors in the calculation of a confidence measure for our classifier. The log-likelihoods received from our models will be correct within a restricted context. However if the context changes (eg. introduction of noise), the Bayes rule used in Equation 3 no longer guarantees optimal fusion. Estimation errors are also introduced owing to a lack of training data and assumptions made about the type of distribution the training data is modeled by. In practice it was found that as the testing conditions differed from training conditions a blind application of Bayes rule made the wrong classifiers more influential on the final output, just the opposite of what was initially anticipated [1, 7].

To remedy this situation we have taken the approach of gaining our confidence measure indirectly from the log-likelihoods as described in Equation 5. In previous techniques [1, 7] confidence measures ζ have been calculated based on the theoretical probability of error. However, in practical situations the error probability for a given word recognition problem is estimated more accurately by designing a classifier and testing it rather than relying on a theoretical estimation of error [4]. A depiction of this process can be seen in Figure 1.

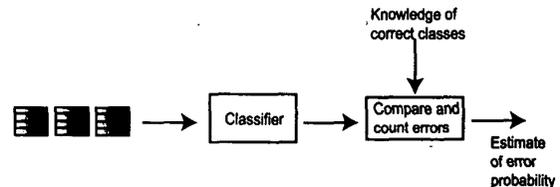


Figure 1: Testing a classifier.

To obtain this practical estimation of error we need to know the correct classes a priori. Our technique tries to take advantage of this a priori knowledge through the use of a secondary classifier to try and predict whether a word has been correctly or incorrectly classified based purely on the word log-likelihood scores of the primary classifier.

$$\zeta = Pr(\text{correct}|\xi) = \frac{P_{cp}(\xi|\text{correct})}{P_{cp}(\xi|\text{correct}) + P_{ip}(\xi|\text{incorrect})} \quad (5)$$

It has been shown by Cox [3] and reinforced by our own tests that upon examination of the distributions of likelihoods from both audio and video recognisers for correct and incorrect words there is very little class separation between the two distributions. This is very true if one measures class distinction by the expected vectors or dispersion of each class.

However, if one looks at these likelihoods as a feature vector ξ in a N dimensional space, where N is the vocabulary size, it was found this high dimensionality brought great class distinction even though they had very similar expected vectors and dispersion [5]. This explains why adhoc confidence measures based purely on the dispersion of log-likelihood scores do not provide large class distinction between correct and incorrect likelihood scores as they ignore the class distinction provided by the high dimensional space in which the likelihood scores exist.

4.1. Training the secondary classifier.

To ensure both the correct and incorrect distributions for our secondary classifier had enough training data, degradations were introduced to both the audio and video modalities. Additive Gaussian noise was introduced to the audio modality with signal-to-noise-ratios (SNR) of 20 dB, 10 dB, 6 dB, 3 dB and 0 dB. Visual noise was added by reducing the frame rate per second (fps) of the video to rates of 15 fps, 9 fps, 5 fps and 3 fps. We found empirically [5] that effective class distinction was achieved for both the audio and video modalities for a distribution topology of 2 Gaussian mixtures for both classes with full covariance matrices.

5. RESULTS.

We tested our secondary classifier against a number of different confidence measures ζ as listed in Table 1 all being based on the log-likelihoods ξ received from our primary classifiers. The S1 measure gives the confidence measure for ideal classifiers as described previously in Equation 4, with the S2 and S3 being previously used [1, 8] effective adhoc measures based on the dispersion of the log-likelihoods from the primary classifiers. Finally, the S4 measure was based on our own secondary classifier. The a priori class probabilities required to calculate the S4 measure as described in Table 1 were calculated using receiver operating characteristic (ROC) curves in the training stage as described in our previous work [5]. Due to the audio modality giving better results than the video modality for the speech recognition task the audio data used for the AVSR video noise tests was corrupted by 20 dBs of Gaussian noise so that the catastrophic fusion boundary for clean video was not dependent on the audio modality. We tested the dispersion measures defined in Table 1 firstly for video degradations due to varying frame rates in the same fashion as the secondary classifier was trained. The results can be seen in Figure 2. To test the visual noise type invariance properties of our

| Fusion type | Confidence Measure (ζ) |
|-------------|--|
| S1 | $\frac{\exp \xi_{best}}{\sum_{i=1}^N \exp \xi_i}$ |
| S2 | $\frac{\xi_{best} - \xi_{nextbest}}{\sum_{i=1}^N \xi_i}$ |
| S3 | $\xi_{best} - \xi_{nextbest}$ |
| S4 | $\frac{P_{cp}(\xi \text{correct})}{P_{ip}(\xi \text{correct}) + P_{ip}(\xi \text{incorrect})}$ |

Table 1: Legend of confidence measures used (Note S4 is our proposed measure).

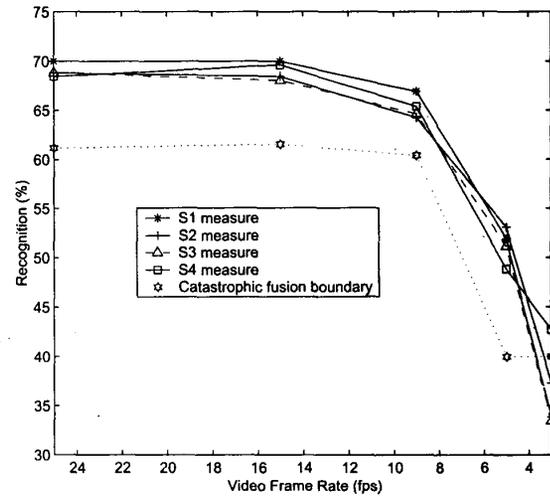


Figure 2: Recognition results for varying frame rate.

technique we then ran the same test for video degradations due to an additive Gaussian noise on each Eigenlip feature vector. The results of which can be seen in Figure 3. The performance in Figures 2 and 3 for the confidence measures S1 to S4 is clearly above the catastrophic fusion boundary at low noise levels. The performance of the S1-S4 measures are approximately the same at low noise levels with the S4 being slightly worse in performance. This is expected as the test context has not changed markedly for the primary classifiers making the Bayes theoretical measure of confidence and subsequent dispersion based metrics correct. However, at high noise levels, as shown in previous papers [1, 7, 8] for audio degradations, the effectiveness of these measures drops off markedly due to the change in context. Unlike the other three measures, our S4 measure based on a secondary classifier maintains performance above the catastrophic fusion boundary in all cases. A much more interesting result was that this performance for the S4 measure was maintained for the case of additive Gaussian visual noise even though the secondary classifier was trained up for degradations due to varying frame rate. This clearly demonstrates a visual noise type insensitivity ability for the S4 measure

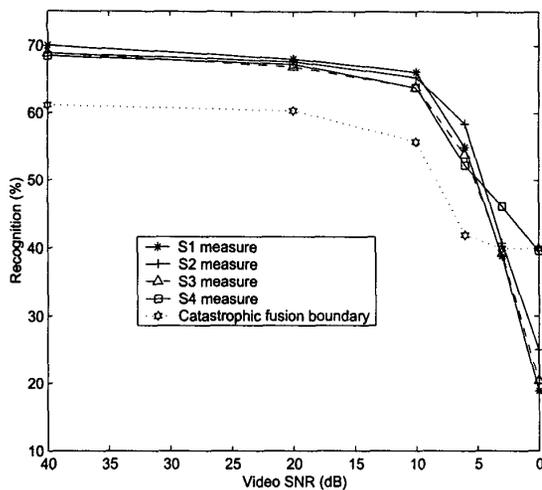


Figure 3: Recognition results for additive Gaussian noise in the visual modality (with secondary classifier being trained on frame rate noise).

for two completely different types of visual noise.

6. DISCUSSION.

We have discussed the use of secondary classifiers as a technique for gaining accurate confidence measures in high noise environments irrespective of the type of noise presented to it. Results above the catastrophic fusion boundary were received for high noise levels even though the secondary classifier was trained up on a different type of visual noise. It is prudent to say that our results are preliminary and we need to test the technique across different types of visual noise instead of the two analysed thus far. However, we think the technique has merit as it calculates a confidence measure based on the log-likelihoods of the preliminary classifier and not on the visual noise directly. Current applications are limited to small vocabulary applications due to the limited amount of data available to train the secondary classifier. Future work will concentrate on testing the system over a number of different visual noise scenarios and larger testing databases.

7. ACKNOWLEDGEMENTS.

The authors would like to thank Dr Stephen Cox and Dr Iain Matthews for providing their AVLetters database [6].

8. REFERENCES

[1] A. Adjoudani and C. Benoit. Audio-Visual Speech Recognition Compared Across Two Architectures. In *EUROSPEECH '95*, pages 1563–1566, Madrid, Spain, September 1995.

[2] C. Bregler and Y. Konig. Eigenlips for robust speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 669–672, Adelaide, Australia, 1994.

[3] S. Cox, I. Matthews, and J. A. Bangham. Combining noise compensation with visual information in speech recognition. In *AVSP*, Rhodes, 1997.

[4] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press Inc., 24-28 Oval Road, London NW1 7DX, 2nd edition, 1990.

[5] S. Lucey, S. Sridharan, and V. Chandran. Improved Speech Recognition Using Adaptive Audio-Visual Fusion via a Stochastic Secondary Classifier. In *International Symposium on Intelligent Multimedia, Video and Speech Processing*, Hong Kong, May 2001. awaiting acceptance.

[6] I. Matthews. *Features for Audio-Visual Speech Recognition*. PhD thesis, School of Information Systems, University of East Anglia, UK, 1998.

[7] J. R. Movellan and P. Mineiro. Modularity and catastrophic fusion: A Bayesian approach with applications to audio visual speech recognition. Technical Report 97.01, Department of Cognitive Science, University of California, San Diego, CA, 1997.

[8] T. Wark. *Multi-modal Speech Processing for Automatic Speaker Recognition*. PhD thesis, Electrical and Electronic Systems Engineering, Queensland University of Technology, October 2000.

[9] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book (for HTK version 2.2)*. Entropic Ltd., 1999.