Improved Speech Recognition using Adaptive Audio-Visual Fusion via a Stochastic Secondary Classifier

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

The adaptive fusion of video and audio is one of the fundamental pursuits of audio visual speech recognition (AVSR). 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. Results are presented that lie above or equal to the boundary of catastrophic fusion across a number of audio noise levels.

1. INTRODUCTION.

Verbal communciation uses cues from both the visual and acoustic modalities to convey messages. 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 [2, 6] with the visual modality being largely ignored. 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 and speaker variabilities).

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 [2, 7] have tried to take advantage of the correlation between noise in a modality through dispersion measures (ie. entropy, variance, range) of the word likelihood scores. These techniques have met with limited success due to the low class distinction between correct and incorrect words based on dispersion metrics within high noise environments. Other techniques have calculated a global weighting for certain audio noise types [2] but require a priori knowledge of the degradation in the audio modality.

In this paper the use of a secondary classifier is employed that can be applied to the word likelihood scores of either the audio or visual modalities so as to gain an accurate confidence measure of the recognition accuracy 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 posterior probabilities but rather an estimate [4]. 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_M for each modality M where N is the size of the vocabulary. Stochastic models are then trained for both modalities using a priori knowledge of whether ξ_M belongs to a correct or incorrect class. This secondary classifier is then used to gauge how confident one is in ξ_M for modality M whereby an adaptive weighting is chosen. By taking this approach the same adaptive scheme can be applied to both the audio and video modalities irrespective of the type of noise presented to it in the real world.

2. FUSION STRATEGIES.

It has been shown that there are basically two topologies for integrating visual and audio modalities with

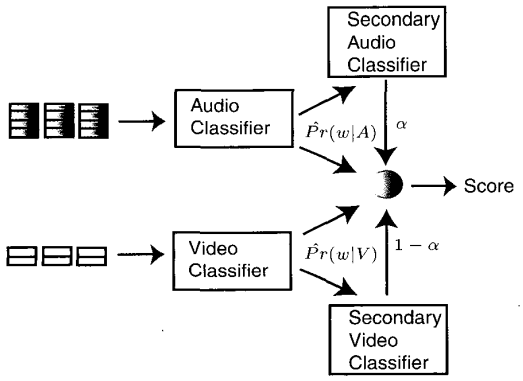

Figure 1: Secondary output classifiers.

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 [2, 6] that late integration generally out performs early integration. Psychological experiments have also suggested that humans combine audio and video sources as if they were conditionally independent [5] similar to a late integration topology. Late integration has the following added benefits over early integration [6],

- provides robustness to the failure of a modality;
- modalities can have different temporal synchrony;
- easier training and computation levels 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 probabilities from each classifier performs better than using either summation, minimum or maximum 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

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

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 normalised likelihoods taken from the audio and video recognisers.

3. AUDIO VISUAL DATA AND MODELLING

The AVLetters database [6] was used for experiments in this paper. 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 [8] 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) [1] 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 [1] 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 [8]. 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. SECONDARY CLASSIFICATION.

In an ideal scenario it would be nice to calculate the adaptive weighting factor α based purely on some confidence measure obtained from the distribution of likelihoods from the N classifers in the vocabulary. However, it has been shown by Cox [2] 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, with a unimodal distribution for each class, it was found this high dimensionality brought great class distinction even though they had very similar expected vectors and dispersion. The class distinction occurs from the difference in covariance matricies between the two classes which can be quantifiably measured by the Bhattacharyya distance as explained in Fukunaga [3]. This explains why confidence measures such as dispersion 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. Class separation between correct and incorrect likelihoods.

The Bhattacharyya distance μ is a convenient measure of the separability of two distributions [3] and gives an approximation of the upper bound of the Bayes error ε_μ between two unimodal distributions,

$$\varepsilon_\mu = \sqrt{P_1 P_2}\exp^{-\mu} \tag{2}$$

where P_1 and P_2 are the a priori probabilites of the two classes which can be assumed to be equal for our purposes. The Bhattacharyya distance μ can be decomposed into the summation of two terms.

$$\begin{aligned} \mu &= \mu_M + \mu_\Sigma \\ \mu_M &= \tfrac{1}{8}(M_2 - M_1)^T \left(\tfrac{\Sigma_1+\Sigma_2}{2}\right)^{-1} \\ \mu_\Sigma &= (M_2 - M_1) + \tfrac{1}{2} ln \frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \end{aligned} \tag{3}$$

where M_1 and M_2 are the means of the two classes and Σ_1 and Σ_2 are the covariance matricies of the two classes. The first term μ_M gives class separability due to mean-difference, while the second term μ_Σ gives the class separability due to the covariance difference. We can apply the Bhattacharyya distance to

Modality	μ	μ_M	μ_Σ
Audio	5.8	0.46	5.34
Video	1.33	0.16	1.17

Table 1: Breakdown of class distinction using the Bhattacharyya distance.

measure the class separability of the audio and video modalities for correct and incorrect words. To ensure both the correct and incorrect distributions 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 also added by reducing the frame rate per second (fps) of the video to rates of 15 fps, 9 fps, 5 fps and 3 fps. The class separation between correct and incorrect models can be seen in Table 1 and show large class distinction for both modalities mainly due to the differences in covariance matricies.

The Bhattacharyya distance is a convenient measure of class separability between models but it does make the assumption that both classes have a unimodal distribution and are accurately modeled by those distributions. This assumption was further tested by gaining classification results over a number of Gaussian mixtures on both practical train and test data. From Table 2 it is clear that:

- audio likelihoods are adequately modeled via a single mixture but optimum results were achieved using a 2 mixture toplogy;

Mixtures	Audio		Video	
	Train	Test	Train	Test
1	13.35	16.23	20.00	39.07
2	10.58	13.85	12.96	39.46
4	10.31	14.00	8.31	38.85

Table 2: Practical class recognition error results (%) for different mixtures.

- the class distinction for the visual modality, as predicted from the initial Bhattacharyya metric in Table 1 is much worse than the audio modality with no clear benefit from employing multimodal models of the distributions;
- the large gap in performance between training and testing data for the visual modality indicates that the HMM models are undertrained as previously stated by [2];

This last point raises problems as the poor performance of the video secondary classifier grossly increases the chance of falsely identifying a incorrect word as correct which requires some type of risk minimisation strategy to prevent catastrophic fusion.

5. PRACTICAL FUSION RESULTS.

To properly evaluate our system we gained an upper AVSR performance boundary via an exhaustive search which found α on a word by word basis. The lower bound (ie. catastrophic fusion boundary) was found by setting $\alpha = 0$ and $\alpha = 1$ for video and audio recognition respectively. Initially the situation of giving each modality equal (ie. $\alpha = 0.5$) a priori weighting was tested with catastrophic fusion occurring at instances of high noise.

Using the secondary classifier outlined in Section 4 an AVSR fusion strategy was developed based on discrete values of α. Fusion based on the audio secondary classifier (*A 2nd-classifier* Figure 2) alone was tested first with words classified as correct setting $\alpha = 1$ and words classified as incorrect setting $\alpha = 0$. In Figure 2 we can see that using just the audio secondary classifier gave results just below those for catastrophic fusion at high noise levels and superior for low audio noise. However, when the audio and video secondary classifiers were combined performance actually dropped below that of the audio secondary classifier. This performance drop can be attributed to the poor class separation of the video classifier which classifies too many false likelihood features positively as correct. Ignoring the video secondary classifier is not an option either as extra class information is still required due to the audio secondary classifier getting results below the catastrophic fusion boundary for high amounts of audio noise. To remedy this situation a risk minimistation strategy was employed to min-

imise the false positives identified by the video and audio secondary classifier via a priori probability threshold.

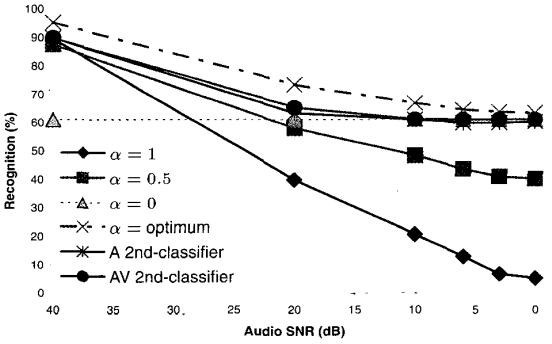

Figure 2: Results for word recognition.

5.1. Risk minimisation.

Receiver operating characteristic (ROC) curves [3] can be used be used to minimise the risk of a false positive classification. By analysing these curves a threshold λ can be found that minimises the chance of a false positive probability to an acceptable level as described by the decision rule in Equation 4.

$$\frac{p(\xi|\omega_i)}{p(\xi|\omega_c)} \begin{matrix} \omega_i \\ > \\ < \\ \omega_c \end{matrix} \lambda \tag{4}$$

where ω_i and ω_c are the incorrect and correct classes, ξ is the likelihood vector and λ is the prior probability threshold ranging from zero to one. For both modalities the false positive probability was chosen emperically to be 0.1. By analysing the ROC curve an acceptable priori probability threshold λ can be found that minimises the number of false positives for each modality. After this minimisation process the weighting factor α is calculated by the following rules,

1. If both the audio and video modalities have been classified as correct set $\alpha = 0.5$.
2. If video is classified as correct set $\alpha = 1$.
3. If audio is classified as correct set $\alpha = 0$.
4. If both the audio and video is classified as incorrect set $\alpha = 0$ or 1 based on priori information.

In our implementation for rule four we classified a number of test words taken under similar conditions to gauge which modality had more correct classifications. The modality with more correct classifications gave an indication of which modality globally had better recognition performance. Our results for this implementation can be seen in Figure 2 (*AV 2nd-classifier*) where performance is better than or equal to catastrophic fusion in all cases and out performing the audio only secondary classifier.

6. DISCUSSIONS.

Our experiments confirm that a high dimensional stochastic secondary classifiers can be used successfully on HMM word likelihood scores to adaptively weight the audio and visual modalities for the purposes of AVSR. Since the technique is based on likelihoods it is insensitive to the type of noise or modality presented to it. The technique is quite useful as it requires no previous knowledge about noise or degradation levels in either modality. Further improvement should be attainable if the performance of the visual modality is improved via improved feature extraction and HMM training techniques. Unfortunately, the technique still requires some priori information if both modalities are classified as incorrect. We are now investigating how effective this system is on different types of audio and visual noise.

7. ACKOWLEDGEMENTS.

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

8. REFERENCES

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

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

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

[4] J. Kittler. Combining Classifiers: A Theoretical Framework. *Pattern Analysis and Applications*, 1(1):18–27, 1998.

[5] D. W. Massaro. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1987.

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