An Investigation of HMM Classifier Combination Strategies for Improved Audio-Visual Speech Recognition

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, v.chandran@qut.edu.au

Abstract

The combining of independent audio and visual HMM classifiers (late integration) has been shown to out perform the combination of audio and visual features in a single HMM classifier (early integration) when either or both modalities are presented with distortion for the task of speech recognition. Theoretical foundations for the optimal combination of these audio and video classifiers are still unclear. In this paper a number of strategies for combining these classifiers are investigated. An argument for using a hybrid of the sum and product rules is made based on empirical, theoretical and heuristic evidence.

1. Introduction

The effective combination of an ensemble of classifiers is a topic of particular importance to the pattern recognition community. There is currently empirical evidence [1, 2] that using an ensemble of classifiers for a recognition problem can give superior performance over those classifiers individually under favorable circumstances. Care must be taken, however, as some combinations of classifiers can performs worse than those classifiers individually an effect known as *catastrophic fusion*. There is presently inadequate understanding why some combination schemes are better than others and in what circumstances [1].

This paper investigates methods of combining two classifiers in a system for audio-visual speech recognition. It has been shown [3, 4] that there is much benefit in modeling the audio and visual modalities through separate classifiers via a *late integration* strategy by combining the scores of those classifiers in some manner a depiction of which can be seen in Figure 1. In speech recognition, hidden Markov models (HMMs) are the classifier of choice in both the audio and video modalities [3, 5], due to their ability to effectively model the temporal variations of human speech. However, due to the nature of HMMs and the inevitable fluctuations in recognition performance from changes in train/test conditions (ie. the introduction of noise), the problem of an effective combination scheme that lies above the *catastrophic fusion boundary* is still elusive. The choice of combination strategy cannot be trivialised, as the problem is very complex and there are many issues requiring explanation.

The focus of this paper is on how best to combine the results produced by the audio and video HMM classifiers of our speech recognition system in order to maximise recognition rate. A number of combination techniques have been used for combination in the two classifier case [1, 3, 5, 6], with most techniques being based on either the sum and product rules, each of which will be explained in the due course of this paper. From a Bayesian standpoint [7], the product rule is optimal when one knows that the classifiers are independent and that they give the true *a posteriori* class probabilities due to changes in train/test conditions. In a practical scenario, one can never get the true *a posteriori* class probabilities. The sum rule, using some strict assumptions, has been shown [1] to be more robust to errors if the *a posteriori* probabilities do not deviate greatly from priors. Neither technique works well in the presence of varying noise, with both techniques suffering from severe *catastrophic fusion* at some noise levels. We present results that make a case for using a hybrid of the sum and product rules for late integration HMM based audio-visual speech recognition rather than the conventional sum and product rules.

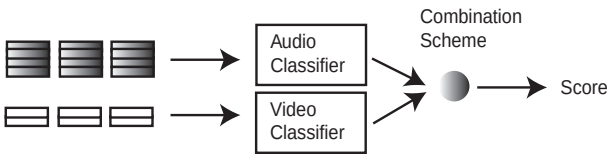

Figure 1: Depiction of late integration strategy for audio-visual speech recognition.

This paper has been organised into the following sections. Section 2 discusses the audio visual database used for our experiments, as well as the topologies of the HMM classifiers for the audio and video modalities. In Section 3, the difference between the ideal *a posteriori* probabilities and likelihood scores given by HMM classifiers is discussed along with the ramifications these differences have on the choice of combination scheme. Section 4 deals with strategies for effective classifier combination, outlining benefits and drawbacks of conventional techniques. A new hybrid approach is then presented in Section 5 based on the sum and product rules.

2. Audio visual data and modelling

The AVLetters database [8] was used for experiments in this paper. The database consists 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 using principal component analysis (PCA) [10] 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 referred to [10] 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 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.

3. Probability estimation from HMMs

Theoretically, when one is dealing with classifiers, one assumes they output *a posteriori* probabilities. Unfortunately, in a practical scenario, one has to deal with likelihood scores which are only representative of these probabilities. In the framework of a hidden Markov model classifier, the output scores are the average of the frame log probability densities for the optimal state sequence $\mathbf{q}$ as dictated by the Viterbi algorithm [9]. We consider the general case when we wish to classify an utterance $\mathbf{x}$ given by,

$$\mathbf{x} = \{x_1, \ldots, x_S\} \tag{1}$$

where S is the number of frames and x_s denotes the feature vector for frame s. Assuming for each frame x_s that the optimal state sequence $\mathbf{q}$ is taken, the likelihood for word w_i given the utterance $\mathbf{x}$ can be expressed as,

$$L(\mathbf{x}|w_i, \mathbf{q}) = \sum_{s=1}^{S} \log a_{q(s),q(s-1)} p_{q(s)}(x_s|w_i) \tag{2}$$

where $a_{i,j}$ is the discrete transition probability from state i to j and $p_j(x_s|w_i)$ is the probability density for state j and frame x_s for word w_i. For our purposes, we can estimate the average probability density estimate for each frame of $\mathbf{x}$ as $p(\mathbf{x}|w_i)$ given by,

$$\log p(\mathbf{x}|w_i) \approx \frac{1}{S} \sum_{s=1}^{S} \log a_{q(s),q(s-1)} p_{q(s)}(x_s|w_i) \tag{3}$$

The average likelihood across frames was used as our class density estimate for two reasons. Firstly, due to differences in synchrony between the audio and video modalities there are 4 times more frames per utterance in the audio than video modalities resulting in likelihoods that are ill scaled due to the multiplication process in Equation 2. Secondly, the average probability density estimate will better facilitate static analysis of classification error with distortion rather than the much more complicated prospect of a formulating a dynamic HMM technique. In reality, $p(\mathbf{x}|w_i)$ will not be the average probability density estimate for each frame as the smaller values of $a_{i,j}$ and $p_j(x_s|w_i)$ in the S length sequence will dominate due to the multiplication process, but the estimation will suffice for this application. Using Bayes [7] rule, we can gain estimates of the *a posteriori* probabilities assuming equal *a priori* class probabilities by taking into account the density estimates of *all* possible classes simultaneously,

$$Pr(w_i|\mathbf{x}) = \frac{p(\mathbf{x}|w_i)}{\sum_{n=1}^{N} p(\mathbf{x}|w_n) + p(\mathbf{x}|\mathbf{S_{tst}} \nsubseteq \mathbf{S_{trn}})} \tag{4}$$

Under similar train/test conditions one can make the assumption,

$$\sum_{n=1}^{N} p(\mathbf{x}|w_n) \gg p(\mathbf{x}|\mathbf{S_{tst}} \nsubseteq \mathbf{S_{trn}}) \tag{5}$$

Which leads to,

$$\hat{Pr}(w_i|\mathbf{x}) = \frac{p(\mathbf{x}|w_i)}{\sum_{n=1}^{N} p(\mathbf{x}|w_n)} \tag{6}$$

used in the experiments to estimate the *a posteriori* probabilities for combination.

Equations 4 and 5 can be understood if we analyse the problem of determining an *a posteriori* probability in terms of subspaces, where the input utterance $\mathbf{x}$ exists in the D-dimensional space $x \in \mathbf{R^D}$. At any given time, we only have at our disposal data existing in the subspace $\mathbf{S_{trn}} \subset \mathbf{R^D}$ or $\mathbf{S_{tst}} \subset \mathbf{R^D}$, representing training and testing data respectively. The estimation of the density functions $p(\mathbf{x}|w_i)$ is based on samples drawn from $\mathbf{S_{trn}}$. When one has to gain an *a posterior* probability estimate $\hat{Pr}(w_i|\mathbf{x})$ of utterance $\mathbf{x} \in \mathbf{S_{tst}}$, one has to make a decision based on HMM models of data lying in $\mathbf{S_{trn}}$ even though $\mathbf{x}$ may not. A depiction of this situation is shown in the Venn diagram in Figure 2(b) where $\mathbf{R^D}$, $\mathbf{S_{trn}}$ and $\mathbf{S_{tst}}$ are visualised as subsets. Within a Bayesian framework, one has to allow for the possibility that $\mathbf{x} \notin \mathbf{S_{trn}}$ even though $\mathbf{x} \in \mathbf{S_{tst}}$. This leads to the situation requiring a density estimate of $p(\mathbf{x}|\mathbf{S_{tst}} \nsubseteq \mathbf{S_{trn}})$ as used in Equation 4 which we call the *reject* class density.

Unfortunately, it is infeasible to gain a model of $p(\mathbf{x}|\mathbf{S_{tst}} \nsubseteq \mathbf{S_{trn}})$, as $\mathbf{x} \in \mathbf{S_{tst}}$ can take on an infinite number of manifestations. However, one can see that if we apply Equation 6 when Equation 5 does not hold (ie. in the case of external noise or an under trained classifier) our *a posteriori* probabilities will become affected two-fold. Firstly, the estimated *a posteriori* probabilities will be ill-scaled, due to the reject class being ignored. Secondly, a classification error will occur, as a class must still be selected again due to the reject class being ignored. This results in a classification error,

$$Pr(w_i|\mathbf{x}) = \hat{Pr}(w_i|\mathbf{x}) + \epsilon(\mathbf{x}) \tag{7}$$

Intuitively, given that we can only work with Equation 6, as the distance between $\mathbf{S_{trn}}$ and $\mathbf{S_{tst}}$ gets bigger one would want to give greater precedence to the classifier whose input is closer to its test conditions, effectively damping the errors.

4. Combination strategies

Assuming that the output of the combined ensemble of classifiers from the audio and visual modalities gives a set of N probabilities, $Pr(w_i)$, where there are N words in the vocabulary, the recognition decision is to choose word w^* such that

$$w^* = \arg \max_{i=1,2,\ldots,N} Pr(w_i) \tag{8}$$

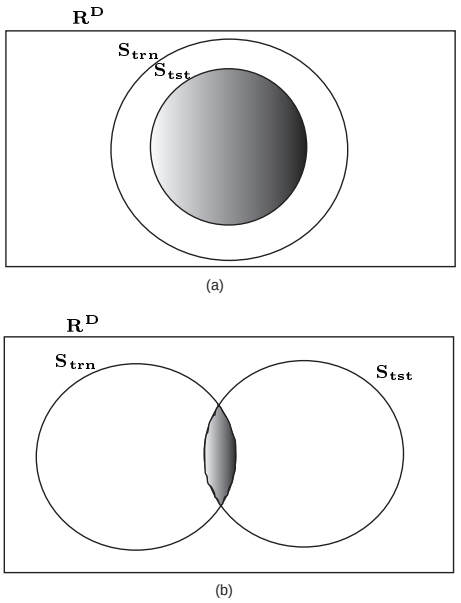

Figure 2: Venn diagram of changes in train/test conditions, (a) $\mathbf{S_{tst}} \subseteq \mathbf{S_{trn}}$ (similar train/test conditions), (b) $\mathbf{S_{tst}} \nsubseteq \mathbf{S_{trn}}$ (different train/test conditions).

The following strategies, as proposed in [1, 6], for the combination of two classifiers were investigated:

Audio only $Pr(w_i) = Pr_A(w_i|\mathbf{x}_A)$

Video only $Pr(w_i) = Pr_V(w_i|\mathbf{x}_V)$

Max $Pr(w_i) = \max[Pr_A(w_i|\mathbf{x}_A), Pr_V(w_i|\mathbf{x}_V)]$, which equates to choosing the classifier with the *highest* probability;

Min $Pr(w_i) = \min[Pr_A(w_i|\mathbf{x}_A), Pr_V(w_i|\mathbf{x}_V)]$, which equates to choosing the classifier with the *lowest* probability;

Product $Pr(w_i) = Pr_A(w_i|\mathbf{x}_A)^{\alpha} \times Pr_V(w_i|\mathbf{x}_V)^{\beta}$

Sum $Pr(w_i) = \alpha Pr_A(w_i|\mathbf{x}_A) + \beta Pr_V(w_i|\mathbf{x}_V)$

where $Pr_A(w_i|\mathbf{x}_A)$ and $Pr_V(w_i|\mathbf{x}_V)$ are the *a posteriori* probabilities for the audio and video utterances respectively. Methods based on voting are not applicable here since only two classifiers are used. An α and β weighting factor can be used on the *sum* and *product* rules to linearly or exponentially dampen the errors occurring in the *a posteriori* probability estimates due to changes in train-test conditions. For our experiments we have decided to set the α and β factors to unity in order to concentrate on how to dampen these errors via selective combination strategies.

Results in Figure 3 are first presented across several audio noise levels. It can be seen that for all combination schemes previously mentioned *catastrophic fusion* occurs at least at one noise level. In the presence of no audio noise, which is approximated as $40dB$ in Figure 3, no combination scheme out performed the audio only scheme. Another interesting result occurs for the product rule at high audio noise, as performance falls well below that of the video only scheme. The sum rule performs well in the presence of high amounts of noise but has poor performance in low noise situations, only achieving a recognition rate of 82% with the product and audio only

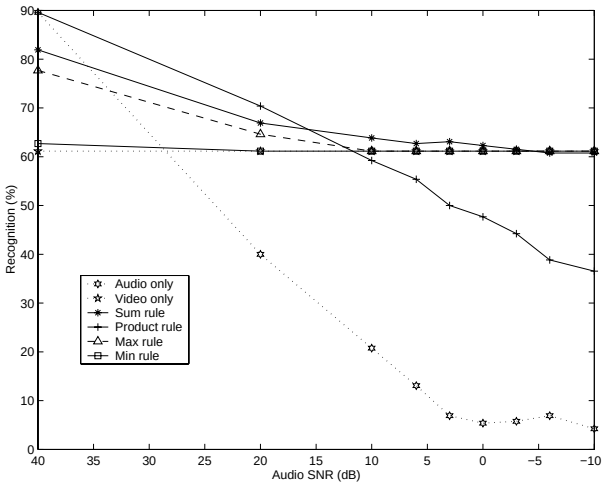

Figure 3: Recognition results for various combination strategies over additive audio noise.

schemes achieving rates of around 90%. The max and min rules gave no results of interest and can be considered as a special case [1] of the sum and product rules respectively.

The results of these experiments can be resolved if one analyses them in terms of the errors that occur within the $\mathbf{S_{trn}}$ and $\mathbf{S_{tst}}$ framework presented in Section 3 and previous work in [1, 11]. The product rule, although optimal in the theoretical case [1], is effectively a severe rule when errors are present, as a single classifier can inhibit a particular class by outputting a probability that is close to zero. The sum rule is a benevolent combination rule, as errors in one classifier have a smaller affect on the final result. The sum rule makes the assumption that the *a posteriori* class probabilities do not deviate greatly from the priors [1]. An ideal scenario would be to use the sum rule when there is little variation between the *a posteriori* class probabilities, as each word class can be assumed to have equal priors, and use the severe product rule in the case of high variation.

5. A hybrid approach.

Kittler [11] hypothesised that a non-linear combination rule may in fact give superior performance over those previously mentioned. In our experimental work, we have devised a hybrid combination scheme using both the sum and product rules based on a theoretical, empirical and heuristic understanding of where they work effectively. The hybrid combination scheme works as follows,

$$Pr(w_i) = \begin{cases} Pr_A(w_i|\mathbf{x}_A) \times Pr_V(w_i|\mathbf{x}_V) & , \sigma_{\boldsymbol{\zeta}(A)} < \lambda \\ Pr_A(w_i|\mathbf{x}_A) + Pr_V(w_i|\mathbf{x}_V) & , \sigma_{\boldsymbol{\zeta}(A)} \geq \lambda \end{cases} \quad (9)$$

The scheme uses the standard deviation $\sigma_{\boldsymbol{\zeta}}(A)$ of the vector $\boldsymbol{\zeta}(A)$ of N estimated audio *a posteriori* probabilities to dictate when the sum or product rule should be used, where

$$\boldsymbol{\zeta}(A) = \{Pr_A(w_1|\mathbf{x}_A), \ldots, Pr_A(w_N|\mathbf{x}_A)\} \quad (10)$$

The decision rule was based purely on the audio probabilities as our experiments were concerned with additive audio noise. The threshold λ used in Equation 9 was determined empirically to optimise performance across all audio noise levels, and in this

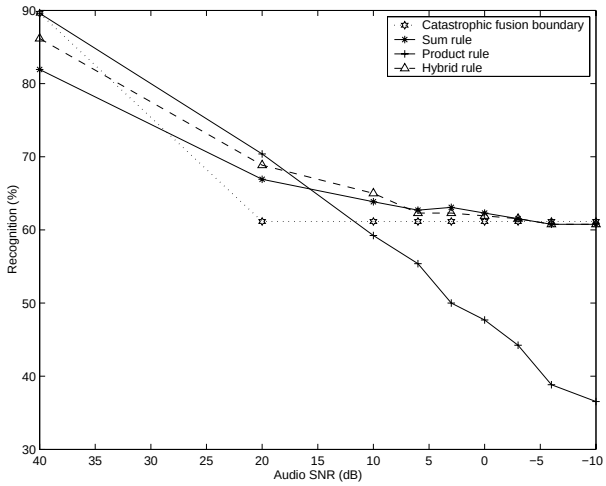

Figure 4: Recognition results for hybrid combination scheme over additive audio noise.

scenario was chosen to be $\lambda = 0.73$. The technique was devised under the assumption that better results would be achieved with the sum rule when there is minimal variation in scores, while the more severe but optimal product rule would be used where there is large variation. Results for this approach are shown in Figure 4

One can see from Figure 4 that for clean conditions the hybrid approach is still below the *catastrophic fusion boundary*, giving a rate of 87% as opposed to the 90% received for the audio only and product rules, however this is well above the 82% received for the sum rule. The catastrophic fusion boundary was calculated in Figure 4 as the best recognition rate for either the audio or video HMM classifier for each noise level. The hybrid rule works well in high noise conditions, actually achieving better results at the $10dB$ audio noise level than the sum or product rules. At high noise levels the hybrid approach maintains performance above or equal to that of the catastrophic fusion boundary and the sum rule. The overall performance of the hybrid combination scheme across various noise levels is superior to the product, sum, audio only or video only schemes.

6. Discussion and Conclusions.

It must be mentioned that superior performance could have been achieved in our hybrid technique if one knew the audio SNR degradation *a priori* as such a measure gives a clear indication of how close the $\mathbf{S_{trn}}$ and $\mathbf{S_{tst}}$ spaces are. However, we are primarily concerned with effective classifier combination based purely on the scores received from the audio and video HMM classifiers without the use of external information. Such an approach, although more difficult, does not rely on the ability of a secondary system to estimate the noise content of an input utterance as this problem is not trivial in the audio modality and is almost impossible in the video modality. Techniques based purely on classifier scores can also be extended to the video modality too where fluctuations in frame rate and picture quality may occur in practical situations.

A hybrid combination scheme has been presented based on the sum and product rules. Results have shown that on the *whole* the technique is superior to the sum, product, max, min, audio only and video only schemes across a wide variety of audio noise levels with no *a priori* knowledge of those noise levels. The technique currently relies purely on a selective combination strategy to dampen errors introduced from changes in train/test conditions. Traditional approaches [1, 4, 5] of applying linear or exponential weighting factors (ie. α and β) for additional damping have not been included into the system. Future work shall try to incorporate such weighting factors into our hybrid system to further improve performance so that recognition performance is above the catastrophic fusion boundary for all noise levels.

7. Acknowledgements.

The authors would like to thank Dr Stephen Cox and Dr Iain Matthews for use of their AVLetters database [8].

8. References

[1] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On Combining Classifiers," *IEEE Transacations on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, March 1998.

[2] D. Pennock, P. Maynard-Reid II, C. Giles, and E. Horvitz, "A Normative Examination of Ensemble Learning Algorithms," in *Proceedings of the Seventeenth International Conference on Machine Learning*, San Francisco, 2000, pp. 735–742.

[3] T. Chen and R. Rao, "Audio-Visual Integration in Multimodal Communication," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

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

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

[6] S. Procter and J. Illingworth, "Combining HMM classifiers in a handwritten text recognition system.," in *IEEE International Conference on Image Processing*, Chicago, Illinois, October 1998.

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

[8] I. Matthews, J. Bangham, and S. Cox, "Audiovisual speech recognition using multiscale nonlinear image decomposition," in *ICSLP '96*, 1996, pp. 38–42.

[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.

[10] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *ICASSP '94*, Adelaide, Australia, 1994, pp. 669–672.

[11] J. Kittler, "Combining Classifiers: A Theoretical Framework," *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, 1998.