Machine Learning, 32, 85–100 (1998)

Robust Sensor Fusion: Analysis and Application to Audio Visual Speech Recognition

JAVIER R. MOVELLAN AND PAUL MINEIRO {movellan,pmineiro}@cogsci.ucsd.edu
Department of Cognitive Science, University of California San Diego, La Jolla, CA 92093-0515

Editors: Gerhard Widmer and Miroslav Kubat

Abstract. This paper analyzes the issue of catastrophic fusion, a problem that occurs in multimodal recognition systems that integrate the output from several modules while working in non-stationary environments. For concreteness we frame the analysis with regard to the problem of automatic audio visual speech recognition (AVSR), but the issues at hand are very general and arise in multimodal recognition systems which need to work in a wide variety of contexts. Catastrophic fusion is said to have occurred when the performance of a multimodal system is inferior to the performance of some isolated modules, e.g., when the performance of the audio visual speech recognition system is inferior to that of the audio system alone. Catastrophic fusion arises because recognition modules make implicit assumptions and thus operate correctly only within a certain context. Practice shows that when modules are tested in contexts inconsistent with their assumptions, their influence on the fused product tends to increase, with catastrophic results. We propose a principled solution to this problem based upon Bayesian ideas of competitive models and inference robustification. We study the approach analytically on a classic Gaussian discrimination task and then apply it to a realistic problem on audio visual speech recognition (AVSR) with excellent results.

Keywords: catastrophic fusion, Bayesian inference, robust statistics, audio visual speech recognition

1. Introduction

This paper analyzes the issue of "catastrophic fusion" in multimodal systems, and explores a principled solution to this problem inspired by the Bayesian ideas of competitive models (Clark & Yuille, 1990) and inference robustification (Box, 1980; O'Hagan, 1994). The discussion is framed with respect to the task of automatic audio visual speech recognition (AVSR) but the issues at hand are very general and appear whenever multimodal signals need to be integrated in non-stationary environments.

1.1. Audio Visual Speech Recognition

Human speech perception is a multimodal process. Sensitivity to contingencies in audio (A) and video (V) signals has been shown in 4-month-old infants (Kuhl & Meltzoff, 1982). By 6 years of age, humans consistently use visual information to understand speech (Massaro, 1987), and by adulthood, vision has an automatic effect on speech perception. The automatic character of this visual influence is commonly illustrated by the McGurk-MacDonald effect, an auditory illusion created by cross-dubbing of A and V speech signals (McGurk & MacDonald, 1976).

Recent years have seen a dramatic flourishing of the engineering literature on AVSR (Yuhas et al., 1990; Wu et al., 1991; Stork et al., 1992; Bregler et al., 1993b; Cosi et al., 1994; Bregler et al., 1994; Wolff et al., 1994; Hennecke et al., 1994;

de Sa, 1994; Movellan, 1995). Current interest on AVSR is in part due to the popularization of digital multimedia tools, its potential application to automatic speech recognition in noisy environments (e.g., car telephony, airplane cockpits, noisy offices), and its links to fundamental theoretical issues in engineering and in cognitive science (Movellan & Chadderdon, 1996).

One important area of research in AVSR is the development of principled methods to integrate A and V information (Hennecke et al., 1995). One of the problems identified in this area is what we refer to as "catastrophic fusion". For concreteness, let us illustrate the problem with a hypothetical car telephony task which we will simulate in later sections. The task is to recognize spoken phone numbers based on input from a camera and a microphone. The system is supposed to operate intelligently under a wide variety of conditions. For example, at times the V signal may be relatively clean and the A signal may be contaminated by sources like the radio, the engine, and friction with the road. At other times the A signal may be more reliable than the V signal (e.g., the radio is off, but the talker's mouth is partially occluded). We want the recognition system to combine the A and V sources intelligently given the conditions at hand (e.g., give more weight to whichever channel is more reliable at that time). At a minimum we expect that for a wide variety of conditions, the performance after fusion should not be worse to that of independent unimodal systems, a pass/non-pass test for any reasonable fusion algorithm (Bernstein & Benoit, 1996). Catastrophic fusion occurs when component modules before fusion significantly outperform the overall system after fusion (e.g., if the A-system alone does better than the combined AV system).

We will analyze this effect in later sections but for now suffice it to say that because of it AVSR systems turned to a variety of heuristics to adaptively change the weight given to the A and V modules (Bregler et al., 1993b; Adjondani & Benoit, 1995). While these heuristics are intuitively sound it is unclear whether they are optimal in any sense.

The goal of this paper is to formally analyze the problem of catastrophic fusion and to propose a principled solution inspired by the Bayesian ideas of competitive models (Clark & Yuille, 1990) and inference robustification (Box, 1980; O'Hagan, 1994). The paper is organized as follows: Section 2 describes competitive models and robustification. Section 3 illustrates analytically how these ideas apply to a simple problem. Section 4 shows that the approach can be easily applied to current AVSR systems with very good results. Section 5 discusses theoretical implications and limitations of the approach.

2. Competitive Models and Robustification

Clark & Yuille (1990) and Yuille & Bulthoff (1996) analyzed information integration in sensory systems from a Bayesian perspective. Their work focuses mostly on the psychophysics of vision, but the issues at hand are very general and apply to the current work on AVSR.

Modularity is justified in their view by the need to make assumptions that disambiguate the data available to the perceptual system (Clark & Yuille, 1990, p. 5). For example, a shape from shading module may assume a single light source and a Lambertian reflectance function with constant albedo. The problem is that these assumptions restrict the range of proper operation of the module: the module will only work well within a restricted *context*. The solution proposed by Clark & Yuille (1990) is to work with an ensemble of models each of which specializes on restricted contexts and automatically checks whether the context is

correct. The hope is that by working with such an ensemble of models, robustness under a variety of contexts can be achieved (Clark & Yuille, 1990, p. 13).

Similar ideas had actually been advanced in the Bayesian literature with regard to the problem of robust statistical inference. The emphasis in the theory of robust statistics is in the development of estimators that work efficiently over a wide range of distributions (Hoglin et al., 1983). Using a Bayesian approach, Box (1980) noticed that robust statistical inference could be achieved by extending inference models with additional "nuisance" parameters, a process he called Bayesian robustification. The idea is to replace an implicit assumption about the specific value of parameters with a prior distribution that represents uncertainty about those parameters.

The approach presented in this paper combines the ideas of competitive models and Bayesian robustification: each of the channels in the multimodal recognition system is provided with extra parameters that represent non-stationary properties of the environment, what we call a *context model*. By doing so we effectively work with an infinite ensemble of models each of which compete on-line to explain the data. As we will see later, even unsophisticated but variable context models are superior to fixed context models when the environment is non-stationary.

We frame the recognition problem as that of simultaneously choosing the most probable perceptual alternative and context parameters [1]

$$\{\hat{w}, \hat{\sigma}_a, \hat{\sigma}_v\} = \underset{w_i, \sigma_a, \sigma_v}{\operatorname{argmax}}\, p(w_i, \sigma_a, \sigma_v \mid x_a, x_v, \mathcal{M}). \tag{1}$$

where σ_a and σ_v are context parameters for the audio and video, $\mathcal{M}$ symbolizes fixed parameters of the recognition engine, and w_i are the perceptual units being recognized (e.g., words or subwords). The effect of this joint decision approach is that only the most probable context models are allowed to have an influence on the fused percept. In other words, the first component of $(\hat{w}, \hat{\sigma}_a, \hat{\sigma}_v)$ can be obtained as follows

$$\hat{w} = \underset{w_i}{\operatorname{argmax}} \left\{ \max_{\sigma_a, \sigma_v} p(w_i, \sigma_a, \sigma_v \mid x_a, x_v, \mathcal{M}) \right\}, \tag{2}$$

Hereafter, we will refer to this approach as the *robustified approach*.

AVSR systems typically consist of two independent modules, one dedicated to A signals and one to V signals and each capable of independently recognizing the entire lexicon (Bregler et al., 1993a; Wolff et al., 1994; Adjondani & Benoit, 1996; Movellan & Chadderdon, 1996). From a Bayesian perspective this modularity reflects an assumption of conditional independence of A and V signals (i.e., the likelihood function factorizes)

$$p(x_a, x_v \mid w_i, \sigma_a, \sigma_v, \mathcal{M}_a, \mathcal{M}_v) \propto p(x_a \mid w_i, \sigma_a, \mathcal{M}_a)\; p(x_v \mid w_i, \sigma_v, \mathcal{M}_v), \tag{3}$$

where x_a and x_v are the audio and video data, w_i is a perceptual interpretation of the data (i.e., a word in the lexicon) and $\{\mathcal{M}_a, \mathcal{M}_v\}$ symbolize fixed parameters of the audio and video models. In fixed lexicon systems, conditional independence is usually assumed at the word level, with very good results (Stork et al., 1992; Bregler et al., 1993b; Adjondani & Benoit, 1995; Movellan, 1995). While our approach does not require the assumption of conditional independence, it greatly simplifies the computations. In particular,

assuming conditional independence and uninformative priors for (σ_a, σ_v) the robustified approach yields

$$\hat{w} = \underset{w_i}{\operatorname{argmax}} \left\{ \log p(w_i) + \left[\max_{\sigma_a} \log p(x_a \mid w_i, \sigma_a, \mathcal{M}_a) \right] + \left[\max_{\sigma_v} \log p(x_v \mid w_i, \sigma_v, \mathcal{M}_v) \right] \right\}. \tag{4}$$

Thus conditional independence allows a modular implementation of the robustified approach (i.e., the A and V channels do not need to talk to each other until the time to make a joint decision):

1. For each w_i we obtain conditional estimates of the context parameters for the audio and video signals:

$$\hat{\sigma}_{a|w_i} = \underset{\sigma_a}{\operatorname{argmax}} \{ \log p(x_a \mid w_i, \sigma_a, \mathcal{M}_a) \}, \tag{5}$$

and

$$\hat{\sigma}_{v|w_i} = \underset{\sigma_v}{\operatorname{argmax}} \{ \log p(x_v \mid w_i, \sigma_v, \mathcal{M}_v) \}. \tag{6}$$

2. Find the best w_i using the conditional context estimates.

$$\hat{w} = \underset{w_i}{\operatorname{argmax}} \{ \log p(w_i) + \log p(x_a \mid w_i, \hat{\sigma}_{a|w_i}, \mathcal{M}_a) + \log p(x_v \mid w_i, \hat{\sigma}_{v|w_i}, \mathcal{M}_v) \}. \tag{7}$$

In the next sections we analyze this approach on a simple Gaussian discrimination task and then apply it to a realistic AVSR problem.

3. Analytical Example

The purpose of this section is to gain a better understanding of why a naive implementation of Bayesian fusion can have catastrophic results and how Bayesian robustification helps solve this problem. To this goal we chose a Gaussian discrimination, simple enough to obtain analytical results yet closely related to the probabilistic models used in our AVSR system. For ease of presentation we frame this toy problem in terms of audio and video channels. The reader is reminded that in this section no real audio and visual signals are used. Our goal is just to understand mathematically why Bayesian robustification avoids catastrophic interference in this particular problem.

Consider an environment which varies randomly between two alternatives w_1 and w_2 (e.g., two different words). Let W be a random variable representing the state of the world, and let $p_W(w_1) = p_W(w_2) = 0.5$. Two sensory channels, arbitrarily labeled A and V, provide conditionally independent information about the state of the world in the form of measurement vectors: $X = (X_a, X_v)$ where X_a is $\mathbb{R}^{N_a}$ valued and X_v is $\mathbb{R}^{N_v}$ valued. The

centroids of these measurements change with the state of the world and are contaminated by an additive Gaussian random vector E:

$$X = E + \sum_i \mu(w_i) 1_{\{W=w_i\}}, \tag{8}$$

where $\mu(w_i) = (\mu_a(w_i), \mu_v(w_i))$ is the centroid when $W = w_i$, and $1_{\{W=w_i\}}$ is an indicator random variable,

$$1_{\{W=w_i\}}(\xi) = \begin{cases} 1 & \text{if } W(\xi) = w_i \\ 0 & \text{if } W(\xi) \neq w_i. \end{cases} \tag{9}$$

The noise vector $E = (E_a, E_v)$ has Gaussian probability density

$$p_E(e \mid w_i, \sigma_a, \sigma_v, \mathcal{M}_a, \mathcal{M}_v) = (\frac{1}{\sigma_a\sqrt{2\pi}})^{N_a} e^{-\|e_a\|^2/2\sigma_a^2} (\frac{1}{\sigma_v\sqrt{2\pi}})^{N_v} e^{-\|e_v\|^2/2\sigma_v^2}, \tag{10}$$

where $e = (e_a, e_v)$, and σ_a and σ_v represent the strengths of the noise in the audio and video channels. The classical Bayesian decision rule to the Gaussian discrimination task simplifies as follows

$$\hat{w} = \underset{w_i}{\operatorname{argmin}} \left\{ \frac{\|x_a - \mu_a(w_i)\|^2}{\sigma_a^2} + \frac{\|x_v - \mu_v(w_i)\|^2}{\sigma_v^2} - \log p(w_i) \right\}. \tag{11}$$

Decisions are made by measuring the Euclidean distance between the data vector (x_a, x_v) and prototypes for each hypothesis $(\mu_a(w_i), \mu_v(w_i))$. The distances between data and prototypes are compared to a bias parameter and the winning prototype is chosen. Those readers familiar with signal detection theory will recognize this solution as the classic signal detection model of observers' sensitivity (Peterson et al., 1954). Note how the σ's weight the relative importance of the two information modules, e.g., if $\sigma_v \gg \sigma_a$ then the audio channel has a higher relative weight in (11).

Let us compare now this standard solution with the robustified approach proposed in this paper. The idea is to use σ_a and σ_v as context parameters and to simultaneously estimate the context parameters and the perceptual hypothesis. In the previous section we saw that this approach leads to the following decision rule

$$\begin{aligned} \hat{w} = \underset{w_i}{\operatorname{argmax}} \big\{ & \log p(w_i) \\ & + \log p(x_a \mid w_i, \hat{\sigma}_{a|w_i}, \mathcal{M}_a) + \log p(x_v \mid w_i, \hat{\sigma}_{v|w_i}, \mathcal{M}_v) \big\}. \end{aligned} \tag{12}$$

where $\hat{\sigma}_{a|w_i}$, $\hat{\sigma}_{v|w_i}$ are the conditional context estimates. In our example these estimates are

$$\hat{\sigma}_{a|w_i} = \frac{\|x_a - \mu_a(w_i)\|}{\sqrt{N_a}}, \tag{13}$$

and

$$\hat{\sigma}_{v|w_i} = \frac{\|x_v - \mu_v(w_i)\|}{\sqrt{N_v}}, \tag{14}$$

substituting them in (12) leads to the robustified decision rule

$$\hat{w} = \underset{w_i}{\operatorname{argmin}} \left\{ N_a \log \|x_a - \mu_a(w_i)\|^2 + N_v \log \|x_v - \mu_v(w_i)\|^2 - \log p(w_i) \right\}. \quad (15)$$

In the standard case, displayed in (11), fusion is based on Euclidean distances between data and prototypes. In contrast, the robustified Bayesian rule calls for a comparison of $\log$ distances, and the σ parameters do not appear in the resulting decision rule.

The difference between these two rules is clear in the gradient of the fused output with respect to the "video data". This gradient reveals the influence of the "video data", x_v, on the fused output. In the standard case the gradient is given by

$$\nabla_{x_v} \left(\frac{\|x_a - \mu_a(w_i)\|^2}{\sigma_a^2} + \frac{\|x_v - \mu_v(w_i)\|^2}{\sigma_v^2} \right) = \frac{2}{\sigma_v^2}(x_v - \mu_v(w_i)). \quad (16)$$

Note that the magnitude of the gradient increases with distance from the prototype. Thus, measurements yielding data far from any prototype become very influential in the final decision. In the robustified rule the gradient is given by

$$\nabla_{x_v} \big(N_a \log \|x_a - \mu_a(w_i)\|^2 + N_v \log \|x_v - \mu_v(w_i)\|^2 \big) = \frac{2N_v}{\|x_v - \mu_v(w_i)\|^2}(x_v - \mu_v(w_i)). \quad (17)$$

It follows that the magnitude of the gradient decreases with the distance from the prototype. This limits the influence of signals far from a prototype, which are treated as indicative of a high noise context.

To gain a better understanding of the performance of the standard and robustified approaches we performed a simulation of the following toy problem.

Simulation

The audio channel was simulated as follows

$$X_a = E_a + \sum_i \mu_a(w_i) 1_{\{W = w_i\}}. \quad (18)$$

where the values of $\mu(w_i)$ are discussed below. The simulated video channel was contaminated by an additional vector kU that simulated non-Gaussian contextual changes caused by occlusion and/or illumination.

$$X_v(k) = kU + E_v + \sum_i \mu_v(w_i) 1_{\{W = w_i\}}. \quad (19)$$

We used non-Gaussian contamination to asses how well standard and robustified fusion works when the system's assumptions are not met. In the simulation U had a uniform probability distribution over the N^v-dimensional sphere of unit radius. The parameter k was systematically varied to control the magnitude of the contamination by non-Gaussian sources.

The independent variables in the experiment were:

1. Decision rule (standard vs. robustified).
2. Dimensionality of the feature vectors (2-D vs. 100-D). The 2-D case represents data-poor problems and the 100-D case data-rich problems.
3. Magnitude of the contamination in the simulated video.

The dependent variable was performance in a two category discrimination task. The simulation was conducted as follows:

1. The dimensionality of the audio and video channels was first fixed at $N_a = N_v = N = 2$, simulating data-poor conditions.
2. The prototype centroids $\{\mu(w_1), \mu(w_2)\}$ for both the audio and video signals were located at the origin for w_1 and at $\{1, \ldots, 1\}$ for w_2. The true context values σ_a and σ_v were fixed at $\sqrt{N}$.
3. For the standard approach, the expected context parameters were fixed at $\sqrt{N}$.
4. The parameter k, which represents the magnitude of the contamination in the video channel, was systematically varied from 0 to 5. For each value of k, estimates of the classification performance of the standard and robustified rules were obtained using Monte Carlo sampling (10000 times).
5. The optimum performance achievable by a system that used only the audio channel was determined analytically.

The entire experiment was then repeated for $N_a = N_v = N = 100$, simulating data-rich conditions.

Figure 1 displays the results of the experiment. The horizontal axis shows the magnitude of the contamination vector kU. The vertical axis shows the performance of the standard and robustified approaches. Note that the standard approach shows catastrophic fusion: when the magnitude of the video contamination is large, the performance after fusion is worse than the performance of the A module only. Bayesian robustification practically eliminates this problem. For the data rich case (100-D), the performance of the robustified approach is never worse than the performance of an optimal audio-only system, thus avoiding catastrophic fusion. In the data-poor case the adaptive system is still clearly more robust than the system with fixed context parameters.

These results suggest that robustification has desirable properties and may help avoid catastrophic fusion. We get an important clue for the success of the robustified approach by examining the difference in evidence assigned to the two word alternatives by the video alone. We represent this difference using the indexed family of random variables $\{D_v(k)\}_{k \in \mathbb{R}}$, where k represents the magnitude of the contamination. Ideally we would like $D_v(k)$, to be unchanged by k. This would mean that the noise has been completely filtered out. In practice this may be effectively impossible and thus a more realistic goal is to have $D_v(k)$ converge to zero as the magnitude of the contamination increases. Note how in the standard Bayesian rule the magnitude of $D_v(k)$ diverges almost surely as k increases.

$$D_v(k) = \frac{1}{\sigma_v^2} \left(\|X_v(k) - \mu_v(w_1)\|^2 - \|X_v(k) - \mu_v(w_2)\|^2 \right) . \tag{20}$$

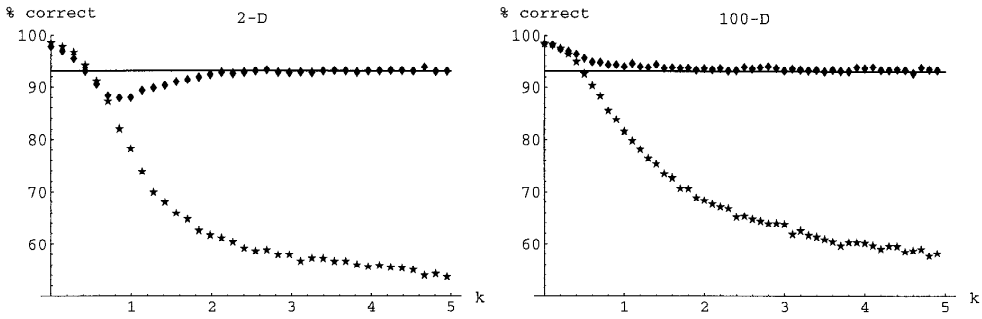

Figure 1. Performance on the two category discrimination task with non-Gaussian noise contamination, when the simulated audio and video data are 2-dimensional and 100-dimensional. Robustified fusion degrades more gracefully than standard fusion. The horizontal line is the theoretical performance limit using audio information only. Note how standard fusion is catastrophic, since when the audio is highly contaminated the performance after fusion is lower than that of the audio system alone.

In other words, as the video becomes more and more contaminated it also becomes more and more influential in the final decision, with catastrophic effects.

However, in the robustified approach $D_v(k)$ converges to zero almost surely[2],

$$D_v(k) = N_v \log \frac{\|X_v(k) - \mu_v(w_1)\|^2}{\|X_v(k) - \mu_v(w_2)\|^2} \tag{21}$$

$$P(\lim_{k\to\infty} D_v(k) = 0) = 1. \tag{22}$$

The robustified approach automatically recognizes the contamination in the video and shuts down its influence by assigning equal weight to the different word hypotheses. This is precisely what we want from a robust procedure, since our assumptions about the video are more incorrect as k increases. In the next section we apply this approach to a realistic AVSR task.

4. Application to AVSR

Bayesian robustification can be easily applied to Hidden Markov Models (HMM), arguably the most successful architecture for AVSR. Hidden Markov models are defined by

- Markovian state dynamics: $p(q_{t+1}|\underline{q}_t) = p(q_{t+1}|q_t)$, where q_t is the state at time t and $\underline{q}_t = (q_1, \cdots q_t)$.
- Conditionally independent sensor models linking observations to states $f(x_t|q_t)$, typically a mixture of multivariate Gaussian densities

$$f(x_t|q_t) = \sum_i p(m_i|q_t)(2\pi)^{-N_a/2} \ |\Sigma|^{-1/2} \ \exp(d(x_t, q_t, \mu_i, \Sigma)), \tag{23}$$

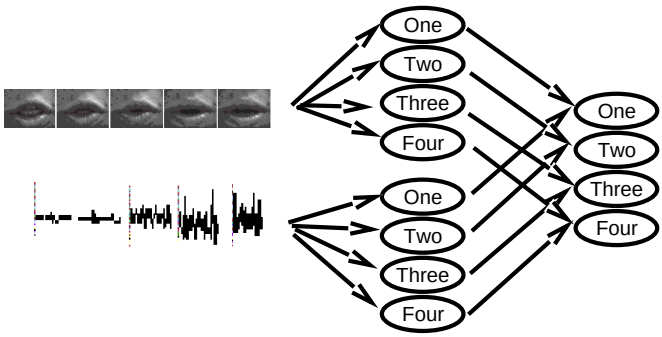

Figure 2. The recognition engine consists of an A module and a V module. Each module is a bank of HMMs one per word hypothesis.

where m_i is the mixture label, $p(m_i|q_t)$ is the mixture distribution given state q_t, Σ is a covariance matrix, and μ_i is the centroid for mixture m_i, and d is the Mahalanobis norm

$$d(x_t, q_t, \mu_i, \Sigma) = (x_t - \mu_i)' \Sigma^{-1} (x_a - \mu_i). \tag{24}$$

Most fixed lexicon systems consist of a bank of HMMs, one per entry in the lexicon. At test time, a new data sequence $\underline{x}_T$ is presented and each HMM calculates the probability of the data given the model. Using the law of total probability in combination with the Markovian property, the probability of the data can be obtained as follows

$$p(\underline{x}_T) = \sum_{\underline{q}_T} p(\underline{q}_T \underline{\mathbf{x}}_T) = \sum_{\underline{q}_T} p(q_o) \prod_{t=1}^{T-1} p(q_{t+1}|q_t) f(x_t|q_t), \tag{25}$$

which is simply a weighted sum of Gaussian densities. Thus the HMM case is formally very similar to the standard Gaussian problem we studied in the previous section.

Current work on AVSR shows that good results are obtained with independent A module and V modules, each of which is a bank of HMMs (see Figure 2). At training time, each model within a module is trained to maximize the likelihood of observation sequences from one of the words in the lexicon. For example, the first model in the A module may be trained to maximize the likelihood of audio-signals from utterances of the word “one”. The first model of the V module may be trained with the corresponding video signals of the same word. This training is usually done using the EM algorithm (Dempster et al., 1977).

At test time the audio part of the test sequence is fed to the A module and the video part to the V module. Each model in the A and V modules computes the probability of the corresponding input sequence. The output of the corresponding A and V models is fused using the standard Bayes rule: the log-likelihoods of the A and V modules are added to the log-priors of each word hypothesis and a decision is made for the hypothesis with largest combined output.

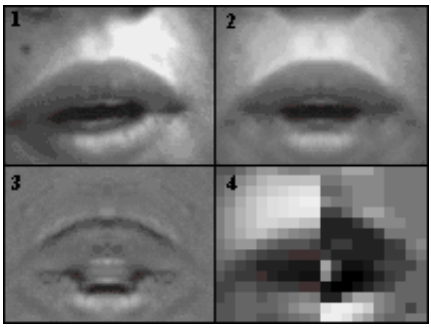

Figure 3. 1) Raw Image. 2) Symmetrized Image. 3) Difference Image. 4) Final Composite.

The robustified approach explored in this paper calls for simultaneous optimization of the variance parameters for all the models at the same time we optimize with respect to the word alternative.

$$\hat{w} = \underset{w_i}{\operatorname{argmax}} \Bigg\{ \log p(w_i) \; + \; \left[\max_{\Sigma_a} \log p(x_a \mid w_i, \mathcal{M}_a) \right] \; + \left[\max_{\Sigma_v} \log p(x_v \mid w_i, \mathcal{M}_v) \right] \Bigg\}. \tag{26}$$

The maximization with respect to the variances can be easily integrated into standard HMM packages by simply applying the EM algorithm on the variance parameters at test time. Thus, in practice, the only difference between the standard approach and the robustified approach is that the latter one retrains the variance parameters of each HMM at test time. In practice this training takes only one or two iterations of the EM algorithm and can be done on-line. We tested this approach on the following AVSR problem.

Training database

We used Tulips1 (Movellan, 1995) a database consisting of 934 images of 9 male and 3 female undergraduate students from the Cognitive Science Department at the University of California, San Diego. For each of these, two samples were taken for each of the digits "one" through "four". Thus, the total database consists of 96 digit utterances. The audio sampling rate is 11.1 kHz, and each sample has an 8-bit representation. Each frame in the video track of a movie is an 8-bit grey-scale, 100x75 pixel image, and each movie is sampled at a visual frame rate of 30 frames per second. The subjects were asked to center and align their lips in the camera during sampling. However significant deviations from ideal conditions occur often (e.g., at times the lips are out of focus or partially occluded by the edge of the rectangular filming window). The database is available at http://cogsci.ucsd.edu.

Signal Processing

We have tried a wide variety of visual processing approaches on this database, including decomposition with local Gaussian templates (Movellan, 1995), PCA-based templates (Gray et al., 1997), and Gabor energy templates (Movellan & Prayaga, 1996). To date, best performance was achieved with the local Gaussian approach. Each frame of the video track is soft-thresholded and symmetrized along the vertical axis, and a temporal difference frame is obtained by subtracting the previous symmetrized frame from the current symmetrized frame. We calculate the inner-products between the symmetrized images and a set of basis images. Our basis images were 10x15 shifted Gaussian kernels with a standard deviation of 3 pixels. The loadings of the symmetrized image and the differential image are combined to form the final observation frame. Each of these composite frames has 300 dimensions (2x10x15). The process is depicted in Figure 3 and explained in more detail in (Movellan, 1995).

LPC/cepstral analysis is used for the auditory front-end. This is a fairly standard preprocessing technique which parameterizes an estimate of the human vocal tract's transfer function. First, the auditory signal is passed through a first-order emphasizer to spectrally flatten it. Then the signal is separated into non-overlapping frames at 30 frames per second. This is done so that there are an equal number of visual and auditory feature vectors for each utterance, which are then synchronized with each other. On each frame we perform the standard LPC/cepstral analysis. Each 30 msec auditory frame is characterized by 26 features: 12 cepstral coefficients, 12 delta-cepstrals, 1 log-power, and 1 delta-log-power. Each of the 26 features is encoded with 8-bit accuracy.

Recognition Engine

In previous work (Chadderdon & Movellan, 1995) a wide variety of HMM architectures were tested on this database including architectures that did not assume conditional independence. Optimal performance was found with independent A and V modules using variance matrices of the form σI, where σ is a scalar and I the identity matrix. The best A models had 5 states and 7 mixtures per state and the best V models had 3 states and 3 mixtures per state. We also determined the optimal weight of A and V modules. Optimal performance is obtained by weighting the output of V times 0.18.

Factorial Contamination Experiment

In this experiment we used the previously optimized architecture and compared its performance under 64 different conditions using the standard and the robustified approaches. We used a $2 \times 8 \times 8$ factorial design. The first factor represents the fusion rule, and the second and third factors represent the context in the audio and video channels. To our knowledge this is the first time an AVSR system is tested with a design of this type. The independent variables were:

1. Fusion rule: Classical, and Robustified.

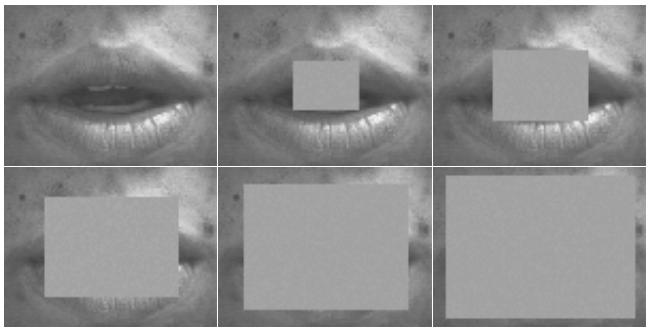

Figure 4. Example of the different occlusions levels. From left to right: 0 %, 10 %, 20 %, 40 %, 60 % 80 %. Percentages are in terms of area.

2. Audio Context: Inexistent, clean, or contaminated at one of the following signal to noise ratios: 12 dB, 6 dB, 0 dB, -6 dB, -12 dB and -100 dB. The contamination was done with audio digitally sampled from the interior of a car while running on a busy highway with the doors open and the radio on a talk-show station.

3. Video Context: Inexistent, clean or occluded by a grey level patch. The percentages of visual area occupied by the patch were 10 %, 20 %, 40 %, 60 %, 80 % and 100 % (see Figure 4).

The dependent variable was performance on the digit recognition task evaluated in terms of generalization to new speakers. In all cases training was done with clean signals and testing was done with one of the 64 contexts under study. Since the training sample is small, generalization performance was estimated using a jackknife procedure (Efron, 1982). Models were trained with 11 subjects, leaving a different subject out for generalization testing. The entire procedure was repeated 12 times, each time leaving a different subject out for testing. Statistics of generalization performance are thus based on 96 generalization trials (4 digits $\times$ 12 subjects $\times$ 2 observations per subject). Standard statistical tests were used to compare the classical and robustified approaches.

The results of this experiment are displayed in Table 1. Note how the experiment replicates the phenomenon of catastrophic fusion. With the classic approach when one of the channels is contaminated, performance after fusion can be significantly worse than performance with the clean channel alone. For example, when the audio is clean, the performance of the audio-only system is 95.83 %. When combined with bad video, this performance drops down to 61.46%, a statistically significant difference[3] F(1,11) = 132.0, $p < 10^{-6}$. In the robustified approach the performance of the joint system is 93.75%, which is not significantly different from the performance of the A system only, F(1,11) = 2.4, p= 0.15. The boxed cells represent regions for which the classical and robustified approaches were significantly different ($\alpha = 0.05$). Contrary to the classic approach, the robustified approach does not exhibit catastrophic fusion.

Table 1. Average generalization performance with classic and robustified fusion. Boxed cells indicate a statistically significant difference $\alpha = 0.05$ between the two fusion approaches.

Performance with Robustified Fusion

	Audio							
Video	None	Clean	12 dB	6 dB	0 dB	-6 dB	-12 dB	-100 dB
None	—	95.83	95.83	90.62	80.20	67.70	42.70	19.80
Clean	84.37	97.92	97.92	94.80	90.62	89.58	81.25	82.20
10 %	73.95	93.75	93.75	94.79	87.50	80.20	71.87	64.58
20 %	62.50	96.87	96.87	94.79	89.58	80.20	62.50	41.66
40 %	37.50	93.75	89.58	87.50	83.30	70.83	43.75	30.20
60 %	34.37	93.75	91.66	88.54	82.29	65.62	42.70	26.04
80 %	27.00	95.83	90.62	86.45	79.16	64.58	46.87	25.00
100 %	25.00	93.75	92.71	84.37	78.12	63.54	44.79	26.04

Performance with Classic Fusion

	Audio							
Video	None	Clean	12 dB	6 dB	0 dB	-6 dB	-12 dB	-100 dB
None	—	95.83	94.79	89.58	79.16	65.62	40.62	20.83
Clean	86.45	98.95	96.87	95.83	93.75	87.50	79.16	70.83
10 %	73.95	93.75	93.75	93.75	89.58	79.16	70.83	52.58
20 %	54.16	89.58	84.41	84.37	84.37	75.00	51.00	43.00
40 %	29.16	81.25	78.12	78.12	67.20	52.08	38.54	34.37
60 %	32.29	77.08	77.08	72.91	62.50	47.91	37.50	29.16
80 %	29.16	70.83	72.91	68.75	54.16	44.79	33.83	28.12
100 %	25.00	61.46	61.45	58.33	51.04	42.70	38.54	29.16

5. Discussion

This paper explored the issue of catastrophic fusion, a problem that occurs when two or more modules need to be fused in non-stationary environments. Catastrophic fusion may occur when modules operate outside their assumed context. The reason for this problem is that in the absence of a context model, deviations from the expected context are interpreted as information about the different perceptual interpretations instead of information about contextual changes.

We proposed a principled solution to this problem inspired by the ideas of competitive models (Clark & Yuille, 1990) and Bayesian robustification (Box, 1980; O'Hagan, 1994). We provided each module with simple white-noise context parameters and jointly estimated the most probable context and perceptual hypothesis. By doing so, context deviations are interpreted as changes in the white noise contamination strength, automatically adjusting the influence of the module.

The approach presented in this paper has two desirable properties: 1) it is very easy to implement; 2) it is framed in Bayesian decision theory and thus we know in which sense it is optimal; 3) even though white noise may be grossly incorrect as a model of context, it automatically shuts down modules with wrong assumptions. However, there is still room for improvement: first of all, the current context models are static. It may be desirable

to use dynamic context models, thus allowing combination of previous context estimates with current context estimates. Second, talkers may articulate different in different noise levels thus affecting both the A and V channels. In such cases it may still be necessary to use a range of audio and visual models explicitly trained with a range of noise levels. Our approach can still be useful in this case by robustifying the different models and providing continuity between the different levels. Finally, we achieve optimality with respect to the joint task of estimating the contexts and deciding for a perceptual hypothesis. While our approach is optimal in this sense, it is not necessarily optimal for each of the marginal tasks (e.g., estimating each context or choosing a perceptual hypothesis). An optimal procedure for the marginal tasks would require integrating over the entire space of context models weighted by their posterior probability. For example, the marginal task of finding the most probable w, (e.g., which word was uttered), can be described as follows

$$\hat{w} = \underset{w_i}{\operatorname{argmax}} \left\{ \int d\sigma_a \; d\sigma_v \; p(x_a x_v | w_i) p(\sigma_a \sigma_v | x_a x_v w_i) p(w_i) \right\}. \tag{27}$$

By redefining the task as joint estimation of w and context $\{\sigma_a, \sigma_v\}$, we avoid the integral in equation 27. From the point of view of the marginal tasks, our approach can be seen as an approximation to the true posterior distribution of the context by a delta function centered at the most probable context (i.e., only the most probable context model is allowed to participate in the fusion process). More sophisticated approximations using the curvature of the posterior probability at the maximum (MacKay, 1996) or Monte Carlo sampling (Neal, 1996) may prove beneficial and help improve the fusion process even more.

Notes

1. For simplicity when possible we identify probability mass and density functions by their arguments. For example, if X is a continuous random variable $p(x)$ will represent $f_X(x)$, the probability density of X evaluated at x.
2. The sole exception being when the vector U is perpendicular to the vector joining the two centroids, an event of probability zero.
3. We use standard notation with $F(a, b)$ standing for Fisher's statistic, with a degrees of freedom in the numerator and b degrees of freedom in the denominator.

References

Adjondani, A. & Benoit, C. (1995). Audio-visual speech recognition compared across two architectures. *Proceedings of the Eurospeech'95 Conference* (pp. 1563–1566). Madrid, Spain.

Adjondani, A. & Benoit, C. (1996). On the Integration of Auditory and Visual Parameters in an HMM-based ASR. In D.G. Stork & M.E. Hennecke (eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications*. New York: NATO/Springer-Verlag.

Bernstein, L. & Benoit, C. (1996). For Speech Perception Three Senses are Better than One. *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, PA.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modeling. *J. Roy. Stat. Soc., A., 143*, 383–430.

Bregler, C., Hild, H., Manke, S., & Waibel, A. (1993). Improving Connected Letter Recognition by Lipreading. *Proc. Int. Conf. on Acoust., Speech, and Signal Processing* (pp. 557–560), Minneapolis. IEEE.

Bregler, C., Manke, S., & Waibel, A. (1993). Bimodal Sensor Integration on the Example of Speech-Reading. *Proceedings of the IEEE International Conference on Neural Networks* (pp. 667–671).

Bregler, C., Omohundro, S.M., & Konig, Y. (1994). A Hybrid Approach to Bimodal Speech Recognition. *28th Annual Asilomar Conference on Signals, Systems, and Computers* (pp. 556–560), Pacific Grove, CA.

Bülthoff, H.H. & Yuille, A.L. (1996). A Bayesian framework for the integration of visual modules. In T. Inui & J.L. McClelland (eds.), *Attention and performance XVI: Information integration in perception and communication*. Cambridge, MA: MIT Press.

Chadderdon, G. & Movellan, J.R. (1995). Testing for Channel Independence in Bimodal Speech Recognition. *Proceedings of 2nd Joint Symposium on Neural Computation* (pp. 84–90). University of California San Diego and California Institute of Technology.

Clark, J.J. & Yuille, A.L. (1990). Data Fusion for Sensory Information Processing Systems. Boston: Kluwer Academic Publishers.

Cosi, P., Magno Caldognetto, E., Vagges, K., Mian, G.A., & Contolini, M. (1994). Bimodal recognition experiments with recurrent neural networks. *Proc. Int. Conf. on Acoust., Speech, and Signal Processing* (pp. 553–556).

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*, 1–38.

de Sa, V. (1994). Combining Uni-Modal Classifiers to Improve Learning. In *Integration of Elementary Functions into Complex Behavior*, volume 2, 19–29.

Efron, A. (1982). *The jacknife, the bootstrap and other resampling plans.* SIAM, Philadelphia, PA.

Gray, M.S., Movellan, J.R., & Sejnowski, T. (1997). Dynamic features for visual speechreading: A systematic comparison. In *Advances in Neural Information Processing Systems, 9*. Cambridge, MA: MIT Press.

Hennecke, M.E., Stork, D.G., & Ventakesh Prasad, K. (1996). Visionary speech: looking ahead to practical speech reading systems. In D.G. Stork & M.E. Hennecke (eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications*. New York: NATO/Springer-Verlag.

Hennecke, M.E., Venkatesh Prasad, K., & Stork, D.G. (1994). Using Deformable Templates to Infer Visual Speech Dynamics. *28th Asilomar Conference on Signals, Systems, and Computers* (pp. 578–582). Pacific Grove, CA: IEEE Computer Society Press.

Hoglin, D.C., Mosteller, F., & Tukey, J.W. (1983). *Understanding robust and exploratory data analysis*. New York: John Wiley.

Kuhl, P.K. & Meltzoff, A.M. (1982). The bimodal perception of speech in infancy. *Science, 218*, 1138–1141.

MacKay, D.J.C. (1996). Hyperparameters: Optimise or interate out?. In G. Heidbreder (ed.), *Maximum entropy and Bayesian methods, Santa Barbara 1993*. Dordrecht: Kluwer.

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

McGurk, H. & MacDonald, J. (1976). Hearing Lips and Seeing Voices. *Nature, 264*, 746–748.

Movellan, J.R. (1995). Visual speech recognition with stochastic neural networks. In G.Tesauro, D.Touretzky, & T. Leen (eds.), *Advances in neural information processing systems*. Cambridge, MA: MIT Press.

Movellan, J.R. & Chadderdon, G. (1996). Channel Separability in the Audio Visual Integration of Speech: A Bayesian Approach. In D.G. Stork & M.E. Hennecke (eds.), *Speechreading by Humans and Machines: Models, Systems, and Applications*. New York: NATO/Springer-Verlag.

Movellan, J.R. & Prayaga, R.S. (1996). Gabor Mosaics: A description of Local Orientation Statistics with Applications to Machine Perception. In *Proceedings of the Eight Annual Conference of the Cognitive Science Society*. Mahwah, NJ: LEA.

Neal, R.M. (1996). *Bayesian learning for neural networks*. New York: Springer.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics: Volume 2B, Bayesian Inference.* Cambridge University Press.

Peterson, W.W., Birdsall, T.G., & Fox, W.C. (1954). The theory of signal detectability. *Transactions IRE Professional Group on Information Theory, 4*, 171–212.

Stork, D.G., Wolff, G.J., & Levine, E.P. (1992). Neural Network Lipreading System for Improved Speech Recognition. *Proceedings International Joint Conference on Neural Networks* (pp. 289–295). IEEE.

Wolff, G.J., Venkatesh Prasad, K., Stork, D.G., & Hennecke, M.E. (1994). Lipreading by Neural Networks: Visual Preprocessing, Learning and Sensory Integration. In J.D. Cowan, G. Tesauro, & J. Alspector (eds.), *Advances in Neural Information Processing Systems, 6*, 1027–1034. San Mateo, CA: Morgan Kaufmann.

Wu, J., Tamura, S., Mitsumoto, H., Kawai, H., Kurosu, K., & Okazaki, K. (1991). Neural network vowel recognition Jointly using voice features and mouth shape image. *Pattern Recognition, 24(10)*, 921–927.

Yuhas, B.P., Goldstein, M.H., Sejnowski, T.J., & Jenkins, R.E. (1990). Neural Network Models of Sensory Integration for Improved Vowel Recognition. *Proc. IEEE, 78(10)*, 1658–1668.

Received August 11, 1997

Accepted December 12, 1997
Final Manuscript March 11, 1998